TECHNISCHE UNIVERSITÄT
CHEMNITZ

Professur für Theoretische Informatik

Spektrale Algorithmen
Mit Eigenwerten schwierige Probleme lösen

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

vorgelegt
der Fakultät für Informatik
der Technischen Universität Chemnitz

von André Lanka

Betreuer:	Prof. Dr. Andreas Goerdt
Gutachter:	PD Dr. Amin Coja-Oghlan, University of Edinburgh
	Prof. Dr. Andreas Goerdt, TU Chemnitz.
	Prof. Dr. Hanno Lefmann, TU Chemnitz.

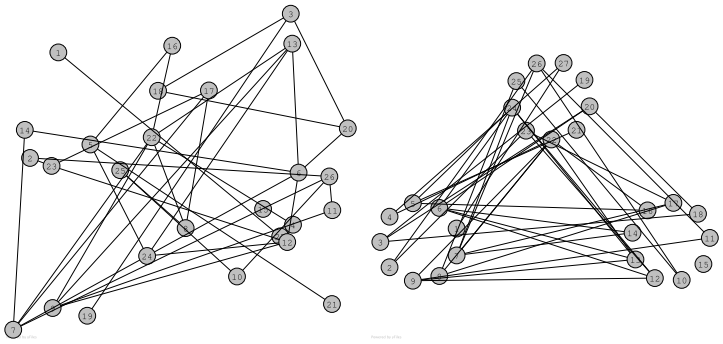
Inhaltsverzeichnis

Kapitel 1

Einleitung

Diese Arbeit beinhaltet neben zahlreichen Beweisen und Analysen zwei effiziente Algorithmen für Probleme, die sich im Allgemeinen nur mit erheblichem Rechenaufwand bewältigen lassen. Der erste Algorithmus (von dem wir drei Varianten untersuchen) dient zur Lösung von Partitionierungsaufgaben in Graphen, der zweite behandelt das Problem Max3Sat.

Ein Graph besteht aus einer Menge $V = \{1, \dots n\}$ von Knoten und einer Menge E von Kanten. Zwei Knoten u und v können durch eine Kante $\{u, v\}$ verbunden werden, um eine Beziehung zwischen beiden Knoten auszudrücken. Beim Partitionieren von Graphen geht es nun darum, gewisse Unregelmäßigkeiten innerhalb des Graphen zu finden und die Knoten anhand gemeinsamer Merkmale zu gruppieren. Der Graph in Abbildung 1.1(a) scheint auf den ersten Blick keinerlei Besonderheiten zu haben. Nach geeigneter Umordnung der Knoten erweist sich dieser Eindruck jedoch als falsch.



(a) ungeordnet (b) partitioniert

Abbildung 1.1: Beispielgraph

Die Knoten können so in drei Gruppen aufgeteilt werden, dass innerhalb der Gruppen keine Kanten verlaufen. Es ist damit möglich, jedem Knoten des Graphen eine Farben zu geben, so dass benachbarte Knoten verschiedene Farben haben. Der Graph ist 3-färbbar.

Algorithmen, die solche Färbungen finden, spielen u. a. im Compilerbau bei der Registeroptimierung eine Rolle. Werden zwei Variablen gleichzeitig gebraucht, können sie nicht im gleichen Register abgelegt werden. Abstrakt gesehen, sind die Variablen die Knoten eines Graphen. Zwei Knoten werden nun durch eine Kante verbunden, wenn die zugehörigen Variablen gleichzeitig gebraucht werden. Ist der konstruierte Graph k-färbbar, dann können die Variablen so auf k Register aufgeteilt werden, dass keine Kollisionen auftreten. Die Aufteilung auf die Register entspricht der Färbung des Graphen: Jede Farbe symbolisiert ein Register.

Zu entscheiden, ob ein gegebener Graph k-färbbar ist, ist für alle $k \geq 3$ $\mathcal{NP}$-vollständig. Eine Konsequenz davon ist, dass kein effizienter Algorithmus bekannt ist, der zu *jedem* gegebenen 3-färbbaren Graphen eine 3-Färbung findet. Solche Aussagen beziehen sich aber nur auf worst-case-Instanzen des Problems, also zunächst nur auf einen kleinen, recht speziellen Teil aller Graphen.

Es stellt sich die Frage, ob derartige Probleme auch für die Graphen schwierig sind, die in der Praxis vorkommen. Eine Möglichkeit, diese Frage zu beantworten, besteht darin, das Verhalten von Algorithmen auf zufälligen Graphen zu untersuchen. Graphen entstehen natürlich nicht komplett zufällig, aber oftmals sind die Abhängigkeiten zwischen den Kanten recht gering, sodass dieses Vorgehen plausibel ist.

Einerseits könnte man den average-case von Algorithmen abschätzen, die im worst-case eine exponentielle Laufzeit haben, andererseits kann man Polynomialzeitalgorithmen betrachten und untersuchen, wie gut diese die optimale Lösung des Problems (z. B. das Finden der Färbung) approximieren. Wir konzentrieren uns auf Letzteres.

Um verwertbare Aussagen treffen zu können, bedarf es eines geeigneten Modells für Zufallsgraphen, das „reale“ Graphen möglichst gut widerspiegelt. Eine häufige Grundlage ist das $G_{n,p}$-Modell. Hier wird jede der $\binom{n}{2} \approx n^2/2$ möglichen Kanten unabhängig mit Wahrscheinlichkeit p in den Graphen eingefügt. Der generierte Graph enthält also erwartungsgemäß $\approx p \cdot n^2/2$ viele Kanten. So erzeugte Graphen haben jedoch im Allgemeinen keine besondere Struktur, da sämtliche Kanten gleichmäßig über den Graphen verteilt werden. Dadurch *müssen* Partitionierungsalgorithmen bei diesen Zufallsgraphen versagen.

Damit der Graph also eine Struktur enthält, die dann gefunden werden kann, pflanzt man eine solche ein. Dafür wird das Modell erweitert und

man lässt verschiedene Kantenwahrscheinlichkeiten zu. Für einen k-färbbaren Graphen werden werden die Knoten zunächst willkürlich in k Gruppen (die Farbklassen) eingeteilt. Anschließend werden Kanten, die Knoten aus verschiedenen Farbklassen verbinden, mit Wahrscheinlichkeit p eingefügt. Innerhalb der Gruppen werden jedoch keine Kanten generiert. Die Kantenwahrscheinlichkeit ist hier 0. Ein so generierter Graph ist in jedem Fall k-färbbar. In der Analyse prüft man dann Algorithmen auf ihre Fähigkeit, die eingepflanzte Färbung zu finden.

Kučera analysierte in [Kuč77] einen einfachen Greedy-Algorithmus zum Färben von Graphen. Er wählte obiges Modell und konnte zeigen, dass der Algorithmus bei derartigen Zufallsgraphen mit hoher Wahrscheinlichkeit optimale Ergebnisse liefert, also eine Färbung mit nur k Farben gefunden wird.

Die Formulierung „mit hoher Wahrscheinlichkeit“ (mhW.) bedeutet in diesem Zusammenhang, dass die Wahrscheinlichkeit gegen 1 geht, wenn n groß wird. Diese Wahrscheinlichkeit bezieht sich jedoch nur auf den Graphen, denn der Algorithmus selbst ist deterministisch.

Ein in diesem Modell erzeugter Graph hat also mhW. die Eigenschaft, durch den Algorithmus optimal gefärbt zu werden. Aufgrund der worst-case-Annahmen über die Schwierigkeit des Problems, kann man nicht erwarten, dass *alle* Graphen diese Eigenschaft haben. Bei Analysen von Algorithmen auf zufälligen Eingaben muss man also immer eine Restwahrscheinlichkeit akzeptieren, dass der Algorithmus auf dem generierten Graph versagt. Man ist jedoch zufrieden, wenn diese Wahrscheinlichkeit mit wachsender Knotenzahl gegen 0 geht. Alle in diesem Kapitel erwähnten Algorithmen haben eine derart kleine Wahrscheinlichkeit zu versagen.

In [Kuč77] war gefordert, dass es sich bei p um eine Konstante handelt, also p unabhängig von n ist. Die typischen Graphen dieses Modells besitzen jedoch sehr viele Kanten. Die meisten der $\binom{n}{2}$ möglichen Kanten werden mit Wahrscheinlichkeit $p = \text{konstant}$ eingefügt. Dadurch liegt die Kantenzahl $|E|$ in der Größenordnung n^2. „Reale Graphen“ dagegen haben meist deutlich weniger Kanten, oft sogar $\leq C \cdot n$ ($C = \text{konstant}$) viele. Um brauchbare Aussagen für praktische Anwendungen zu erhalten, ist es also nötig, die Kantenwahrscheinlichkeit p zu verringern.

An dieser Stelle setzt der l-Pfad-Algorithmus von Blum und Spencer [BS95] an. Dieser funktioniert bereits, wenn der Graph $\geq n^{\varepsilon} \cdot n$ (ε ist eine beliebig kleine positive Konstante) viele Kanten hat. Den Sprung in den linearen Kantenbereich (also $|E| \leq C \cdot n$) schaffte erst der Färbungsalgorithmus von Alon und Kahale [AK97]. Dieser Übergang ist oft besonders schwierig, denn bei diesen dünnen Zufallsgraphen treten Effekte auf, die in dichteren Zufallsgraphen mit $|E| \gg \log n \cdot n$ nicht vorhanden sind. Solche Effekte werfen oft große Probleme bei der Algorithmenanalyse auf, wie auch

wir später sehen werden.

Partitionieren umfasst jedoch mehr als nur das Einfärben von Graphen. Auch beim Problem der minimalen Bisektion müssen die Knoten des Graphen gruppiert werden. Das Ziel dabei ist, zwei gleichgroße Gruppen zu bilden, so dass die Anzahl der Kanten zwischen beiden Gruppen möglichst klein ist. Das gemeinsame Merkmal der Knoten in beiden Mengen ist die verhältnismäßig geringe Anzahl von Nachbarn in der jeweils anderen Menge. Diese Aufgabe bzw. ihre Erweiterung auf $k > 2$ spielt bei der parallelen Programmierung eine wichtige Rolle; eine weitere Anwendung ist die Verteilung von Gattern beim Chipdesign.

Wenn ein Problem parallel gelöst werden soll, zerlegt man es in n kleinere Teilprobleme und verteilt diese auf k Rechner. Mitunter ist zur Lösung der Teilprobleme eine gewisse Kommunikation zwischen ihnen nötig. Da die Kommunikation zwischen Rechnern oft länger dauert als innerhalb eines Rechners, ist man daran interessiert, die Kommunikation zwischen den Rechnern zu minimieren. Stellen wir uns die Teilprobleme als Knoten eines Graphen vor. Wir verbinden zwei Knoten durch eine Kante, wenn die entsprechenden Teilprobleme miteinander kommunizieren müssen. Unser Ziel ist jetzt die Aufteilung der n Knoten in k gleichgroße Gruppen, so dass möglichst wenige Kanten zwischen den Gruppen verlaufen.

Leider sind auch für diesen Spezialfall der Partitionierung keine effizienten Algorithmen bekannt, die auf *allen* Graphen optimale Ergebnisse liefern. Analysen für zufällige Graphen wurden z. B. in [Bop87] und [CO05] gemacht. Während in zuerst genannter Arbeit $|E| \geq \ln n \cdot n$ sein muss, genügt in letzterer $|E| \geq C \cdot n$. Darüber hinaus findet der Algorithmus in [CO05] nicht nur einen gute Näherung an die optimale Lösung, sondern zertifiziert mhW. sogar deren Optimalität.

Zahlreiche weitere Partitionierungsprobleme sind im worst-case ebenfalls nur mit sehr hohem Aufwand lösbar, während es Polynomialzeitalgorithmen gibt, die auf zufälligen Eingaben nahezu optimale Lösungen finden.

Damit nicht für jedes einzelne Partitionierungsproblem ein eigener Algorithmus entworfen werden muss, sucht man nach Verfahren, die die Knoten gruppieren, *ohne* die konkreten gemeinsamen Merkmale zu kennen. Das heißt, dass der Algorithmus im Idealfall ausschließlich den Graphen kennt, aber keine Information darüber hat, ob eine Färbung, eine minimale Bisektion oder Ähnliches gefunden werden soll. Diese adaptiven Algorithmen sind dadurch motiviert, dass man auf der Suche nach Auffälligkeiten nicht unbedingt schon im Vorfeld genau weiß, welche Eigenschaft die Knoten in den Gruppen gemeinsam haben oder wie viele Gruppen es überhaupt gibt.

Für dieses allgemeine Partitionierungsproblem wurde in [CK01] ein Algorithmus vorgestellt, der die gepflanzte Partition rekonstruieren kann, wenn

$|E| \gg n^{1.5}$ ist. Das in [CK01] benutzte Modell war relativ unflexibel und wurde [McS01] von McSherry erweitert: Die Knoten V werden in k Gruppen $V_1, \ldots, V_k$ unterteilt und Kanten zwischen V_i und V_j werden mit Wahrscheinlichkeit p_{ij} eingefügt[1]. McSherry konnte zeigen, dass sein Algorithmus mhW. korrekt partitioniert, solange jedes $p_{ij} \gg \log^6 n/n$ ist, was zu $|E| \gg \log^6 n \cdot n$ führt. 2006 gelang es Coja-Oghlan McSherrys Resultat zu verbessern. Der in [CO06] vorgestellte Algorithmus ist in der Lage, eingepflanzte Partitionen auch dann zu finden, wenn der eingegebene Graph nur $C \cdot n$ viele Kanten hat.

Alle bis hierher vorgestellten Algorithmen haben eine Gemeinsamkeit: Ihre Wirksamkeit – die Rekonstruktion der eingepflanzten Partition $V_1, \ldots, V_k$ – wird in einem Modell für Zufallsgraphen nachgewiesen, das auf $G_{n,p}$ basiert. Alle Kanten zwischen V_i und V_j werden mit der gleichen Wahrscheinlichkeit p_{ij} eingefügt. Eine Konsequenz daraus ist, dass alle Knoten der gleichen Menge V_i die gleiche erwartete Anzahl von Nachbarn haben. Wie sich herausgestellt hat, ist dies ein recht ungewöhnliches Verhalten, das „reale“ Graphen eher selten aufweisen.

Betrachten wir beispielsweise den Graphen des `www`. Dort wird auf bestimmte Domains (z. B. `microsoft.com`, `kernel.org`) sehr häufig verlinkt, während der Großteil der Seiten (u. a. die Mehrzahl der privaten Homepages) Ziel nur weniger Verweise sind. Dies ist jedoch keine Auffälligkeit im Sinne der Partitionierung, sondern eine natürliche Entwicklung, die wenig überrascht. Eine interessante Struktur hingegen wäre folgende: Wir betrachten eine kleine Menge S von Domains. Während kaum Links von außerhalb in S hineinzeigen, sind sehr viele Links innerhalb von S.

Dies kann bedeuten, dass die Seiten ein gemeinsames Thema behandeln, das den Rest der (Netz-)Welt kaum interessiert. Es kann sich aber auch um „Suchmaschinen-Spamming“ handeln.

Suchmaschinen bewerten Webseiten, damit wichtige Seiten bei Suchergebnissen möglichst weit vorn stehen. Dabei spielt es auch eine Rolle, wie oft (und von wem) auf eine Seite verwiesen wird (siehe z. B. Googles PageRank in [BMPW99]). Zur Manipulation der Seitenbewertung erzeugen Seitenbesitzer mitunter künstlich Linkstrukturen, sogenannte Linkfarmen: Es werden zahlreiche Seiten erstellt, die eigens dazu dienen, viele eingehende Links auf eine „Hauptseite“ zu erzeugen, damit diese in den Suchergebnissen möglichst weit vorn einsortiert wird. Da die Besitzer aber keinen Einfluss auf den Rest des `www` haben, erhalten sie nur wenige Links von außerhalb der Linkfarm.

Unabhängig davon, ob man eine Menge von Webseiten inhaltlich strukturieren oder Linkfarmen erkennen will, sucht man in beiden Fällen nach

[1] Falls $i = j$ ist, bedeutet das „innerhalb von V_i“.

Knotengruppen, die sehr viele Verbindungen innerhalb der Gruppe aber nur wenige Nachbarn außerhalb der Gruppe haben.

Aufgrund der starken natürlichen Unterschiede in den Knotengraden[2] werden Graphen wie der des `www` nur schlecht von $G_{n,p}$-basierten Modellen erfasst. Ein Ausweg ist die Erweiterung des Modells, so dass solche Graphen besser modelliert werden. Leider funktionieren die bisher vorgestellten Algorithmen dann nicht mehr [MP02].

Dasgupta, Hopcroft und McSherry präsentierten in [DHM04] einen Algorithmus, der in einem weitaus allgemeineren Modell mhW. erfolgreich partitioniert, solange $|E| \gg \log^6 n \cdot n$ ist. Wir benutzen das gleiche Modell und erläutern es in Abschnitt 2.1 ausführlich. Das Resultat von Dasgupta et al. verbessern wir in zwei Schritten.

Im ersten Schritt heben wir die Restriktion $|E| \gg \log^6 n \cdot n$ auf und zeigen, dass unserem Algorithmus $|E| > C \cdot n$ genügt. Damit beantworten wir die in [DHM04] gestellte offene Frage, nach der Existenz eines solchen Algorithmus, positiv. Sowohl der Algorithmus in [DHM04] als auch der in Abschnitt 2.3 vorgestellte benötigen (abgesehen vom Graphen selbst) noch weitere Eingaben. Bei diesen handelt es sich um Parameter des Zufallsprozesses, die in der Realität nicht zu Verfügung stehen. Beide Verfahren sind also aus *Anwendungssicht* zunächst unbrauchbar.

In Kapitel 3 gelingt es uns aber, unseren Algorithmus weiterzuentwickeln, so dass er außer dem Graphen keinerlei Eingaben benötigt. Trotzdem rekonstruiert er eine – nach unserem Modell eingepflanzte – Partition mhW. fast vollständig.

Im letzten Kapitel wird ein Problem aus der Aussagenlogik behandelt. Die Variablen einer aussagenlogischen Formel können die Werte „wahr“ (=1) und „falsch“ (=0) annehmen. Diese Variablen werden in einer Formel durch beliebige Boole'sche Funktionen miteinander verknüpft, z. B.

$$F = ((x_1 \vee x_2) \wedge x_3) \Rightarrow \neg x_1 \,.$$

Bei einer gegebenen aussagenlogischen Formel zu entscheiden, ob es eine Belegung der Variablen gibt, so dass die Auswertung der Formel unter dieser Belegung „wahr“ ergibt, d. h. ob die Formel erfüllbar ist, zählt ebenfalls zu den $\mathcal{NP}$-vollständigen Problemen.

Aussagenlogische Formeln finden in vielen Bereichen Anwendung, etwa bei der automatisierten Planung in der künstlichen Intelligenz [KS92] oder auch bei der formalen Verifikation von Hard- und Softwaresystemen. EDA-Tools[3] benutzen sie intern, um die Äquivalenz von verschiedenen Chipbe-

[2] Der Grad eines Knotens ist die Anzahl seiner Nachbarn.

[3] Electronic Design Automation

schreibungen (z. B. Register-Transfer-Logik $\leftrightarrow$ Gatterdarstellung) zu überprüfen.

Algorithmisch angenehmer, wenn auch nicht einfacher bezüglich der Komplexität, sind Normalformen zu behandeln. Jede solche aussagenlogische Formel lässt sich in Polynomialzeit in eine konjunktive Normalform, die 3KNF überführen. Eine 3KNF enthält Klauseln, die jeweils UND-verknüpft sind. Innerhalb der Klauseln sind genau 3 Literale (= negierte oder nicht-negierte Variablen) durch ein logisches ODER verknüpft. Formel in 3KNF werden auch oft 3Sat-Formeln genannt. Sat steht dabei für satisfiability, also Erfüllbarkeit.

Die folgende Formel ist eine 3KNF mit $n = 4$ Variablen und $m = 3$ Klauseln

$$F = (x_1 \vee x_2 \vee \neg x_3) \wedge (x_4 \vee \neg x_2 \vee x_1) \wedge (x_1 \vee x_4 \vee x_2)\,.$$

Die Entscheidung, ob eine gegebene 3Sat-Formel erfüllbar ist, ist ebenfalls $\mathcal{NP}$-schwer. Unklar ist aber wieder, ob diese Formeln auch im „typischen“ Fall schwierig zu behandeln sind.

Um zufällige Formeln zu generieren und zu analysieren, wird oft das $\text{Form}_{n,3,p}$-Modell benutzt. Hier wird eine 3Sat-Formel mit n Variablen generiert und jede mögliche Klausel unabhängig von allen anderen mit Wahrscheinlichkeit p eingefügt. Ist eine Formel F nach dem $\text{Form}_{n,3,p}$-Modell generiert, schreiben wir $F \in \text{Form}_{n,3,p}$.

Die Erfüllbarkeit von $F \in \text{Form}_{n,3,p}$ hängt vor allem von p ab. Friedgut zeigte in [Fri99] unter anderem, dass $\text{Form}_{n,3,p}$ einen scharfen Übergang hat: Es gibt $c_3 = c_3(n)$ mit $c_3 \leq$ konstant, so dass $F \in \text{Form}_{n,3,p}$ mhW. erfüllbar ist, wenn die Anzahl der Klauseln $m < (1-\varepsilon)c_3 \cdot n$ ist und mhW. unerfüllbar ist, wenn $m > (1+\varepsilon)c_3 \cdot n$ ist.

Mehrere Arbeiten beschäftigten sich mit diesem Schwellenwert c_3. Die besten bewiesenen Schranken an c_3 sind $3.52 \leq c_3 \leq 4.52$, vgl. [HS03, DBM00, KKL03]. Die Beweise für die obere Schranke an c_3 sind nichtalgorithmisch, so dass sie kein Verfahren liefern, das die Unerfüllbarkeit von $F \in \text{Form}_{n,3,p}$ für $p = O(1/n^2)$ zeigt.

Für größere Werte von p existieren jedoch Verfahren, die mhW. einen Beweis für die Unerfüllbarkeit der gegebenen Formel $F \in \text{Form}_{n,3,p}$ liefern (Tabelle 1.1). Das heißt, diese Verfahren antworten entweder mit „unerfüllbar“ oder mit „weiß nicht“. Erstere Antwort geben sie mit hoher Wahrscheinlichkeit; in *jedem* Fall ist die gegebene Antwort aber korrekt.

Über die reine Erfüllbarkeit hinaus interessiert man sich dafür, mit welcher der 2^n möglichen Belegungen die meisten der Klauseln erfüllt werden. Dieses Problem wird Max3Sat genannt.

Fu	$p = \Omega(1/n)$	[Fu95]
Beame, Karp, Pitassi, Saks	$p = \Omega(1/(n \cdot \log n))$	[BKPS98]
Friedman, Goerdt	$p = \Omega(n^{\varepsilon}/n^{1.5})$	[FG01]
Goerdt, Lanka	$p = \Omega(\ln^6 n/n^{1.5})$	[GL03]
Feige, Ofek	$p = \Omega(1/n^{1.5})$	[FO04]

Tabelle 1.1: $F \in \mathrm{Form}_{n,3,p}$ als unerfüllbar zertifizierbar.

Bei einer Formel, die in jeder Klausel genau drei verschiedene Literale enthält, kann mit einfachen Greedy-Methoden eine Belegung gefunden werden, die mindestens $7/8 \cdot m$ Klauseln erfüllt.

Interessanterweise ist dies auch das beste Ergebnis, das in Polynomialzeit garantiert werden kann. In [Hås01] hat Håstad gezeigt, dass es $\mathcal{NP}$-vollständig ist, zwischen den Formeln in 3KNF zu unterscheiden, die erfüllbar sind und denen, die zu $(7/8 + \varepsilon) \cdot m$ erfüllbar sind, egal wie klein die positive Konstante ε ist. So sind keine effizienten Algorithmen bekannt, die die Güte der gefundenen $7/8 \cdot m$-Lösung bewerten können.

Abbildung 1.2 zeigt exemplarisch den Anteil der erfüllbaren Klauseln für $F \in \mathrm{Form}_{n,3,p}$ bei wachsendem p. Die waagerechte Linie entspricht dem „Garantieanteil“ von 7/8.

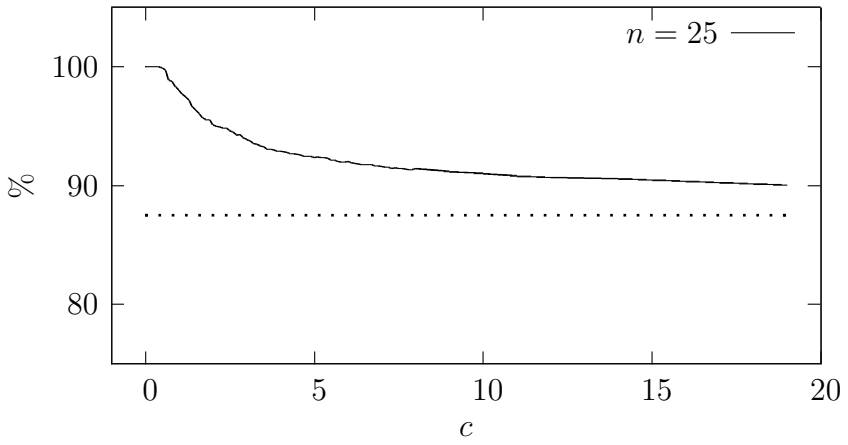


Abbildung 1.2: Erfüllbarer Anteil von $F \in \mathrm{Form}_{n,3,p}$ mit $p = c/n^2$.

Mit einfachen probabilistischen Argumenten kann gezeigt werden, dass die Kurve asymptotisch gegen 7/8 geht. Eine typische Formel aus $\mathrm{Form}_{n,3,p}$ ist also im Fall $m/n \to \infty$ mhW. nur zu ca. $7/8 \cdot m$ erfüllbar.

Nun kann man zwei Wege gehen. Einerseits kann man ähnlich wie bei

den besprochenen Partitionierungsproblemen eine erfüllende Belegung in $F \in \text{Form}_{n,3,p}$ hineinpflanzen. Dieses Szenario untersuchte Flaxman in [Fla03] und kam zu dem Schluss, dass die erfüllende Belegung mhW. gefunden werden kann, wenn $p = \Omega(1/n^2)$ ist.

Andererseits kann man auch versuchen, algorithmisch eine obere Schranke an die Anzahl der erfüllbaren Klauseln zu finden. Dies scheint schwieriger als der erste Weg zu sein, denn während es dort reicht, die erfüllende Belegung anzugeben, muss man bei zweiterem zeigen, dass *keine* Belegung mehr Klauseln erfüllt als vom Algorithmus angegeben. Wir wollen den zweiten Weg gehen.

Es existieren einige „Algorithmen“ für Max3Sat und $F \in \text{Form}_{n,3,p}$, z. B. [VK02, Int04], die solche Schranken angeben. Allerdings gilt: Die Antwort der Algorithmen ist nur mit hoher Wahrscheinlichkeit korrekt. Sicherheit bieten diese Verfahren also nicht.

Solche Algorithmen sind für uns *nicht* von Interesse, denn wir wissen von vornherein, dass eine Formel $F \in \text{Form}_{n,3,p}$ für $m/n \to \infty$ mit hoher Wahrscheinlichkeit nur zu einem Anteil von ca. $7/8$ erfüllbar ist. Damit liefern uns die Algorithmen [VK02] und [Int04] *keine* neuen Informationen über die gegebene, konkrete Formel.

Wir interessieren uns daher für Algorithmen, deren Antwort *stets* korrekt ist. In Kapitel 4.2 zeigen wir einen Algorithmus, der genau das leistet: Bei einer gegebenen Formel $F \in \text{Form}_{n,3,p}$ mit $p \geq \log^4(n)/n^{1.5}$ findet der Algorithmus mhW. einen Beweis, dass keine Belegung mehr als $7/8 \cdot m \cdot (1+4/\ln^{0.25} n)$ Klauseln der Formel erfüllt. Beachte, dass dieser Wert nahezu optimal ist.

Mithilfe dieses Algorithmus können wir (zumindest für die allermeisten Formeln) in Polynomialzeit zeigen, dass die vom Greedy-Algorithmus gefundene Belegung im Wesentlichen optimal ist.

Kapitel 2

Partitionierung

2.1 Das Modell

Das von uns benutzte Modell basiert auf dem Chung-Lu-Modell aus [CL02]. Im Chung-Lu-Modell wird ein Graph folgendermaßen generiert. Jeder Knoten $u \in V$ erhält ein individuelles, positives Gewicht w_u. Das Durchschnittsgewicht bezeichnen wir mit $\overline{w} = \sum_{u \in V} w_u / n$. Die Kante $\{u, v\}$ wird dann unabhängig mit Wahrscheinlichkeit

$$\frac{w_u \cdot w_v}{\sum_{u' \in V} w_{u'}} = \frac{w_u \cdot w_v}{\overline{w} \cdot n}$$

in den Graphen eingefügt. Dabei sollen alle Kantenwahrscheinlichkeiten ≤ 1 sein. Dies ist sichergestellt, wenn alle Gewichte $\leq \sqrt{\overline{w} \cdot n}$ sind.

Es verringert den technischen Aufwand der nachfolgenden Analysen, dass wir auch Schlingen zulassen. Wir zählen eine Schlinge bei Knoten u beim Grad d_u mit 1. Alle Ergebnisse lassen sich jedoch genauso erzielen, wenn Schlingen verboten sind.

Wir sehen, dass Kanten zwischen Knoten mit hohem Gewicht bevorzugt in den Graphen eingefügt werden, da der Zähler in obigem Bruch dann größer ist, während der Nenner gleich bleibt. Die erwartete Anzahl von Nachbarn von Knoten u (bezeichnet durch w'_u) entspricht genau seinem Gewicht:

$$w'_u = \mathbf{E}[d_u] = \sum_{v \in V} \frac{w_u \cdot w_v}{\sum_{u' \in V} w_{u'}} = w_u \cdot \frac{\sum_{v \in V} w_v}{\sum_{u' \in V} w_{u'}} = w_u.$$

Das ist der Grund, warum bei diesem Modell von „Zufallsgraphen mit vorgegebenen erwarteten Graden“ gesprochen wird. Jedem Knoten kann ein eigener erwarteter Grad zugeteilt werden. Mit $\mathbf{w} = (w_1, \ldots, w_n)$ spricht man

auch vom $G(\mathbf{w})$-Modell. Es ist zu beachten, dass das klassische $G_{n,p}$-Modell ein Spezialfall von $G(\mathbf{w})$ mit $w_u = p \cdot n$ für alle $u \in V$ ist.

Wir modifizieren das Modell nun und pflanzen auf natürliche Weise eine Partition in den Graphen. Dazu teilen wir V willkürlich in k disjunkte Teilmengen $V_1, \ldots, V_k$ auf. Außerdem wählen wir uns eine symmetrische $k \times k$-Matrix D mit nichtnegativen Einträgen d_{ij}. Für zwei Knoten $u \in V_i$ und $v \in V_j$ wird die Kante $\{u, v\}$ von nun an mit Wahrscheinlichkeit

$$d_{ij} \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot n}$$

eingefügt. Das heißt, durch große Werte für d_{ij} werden Kanten zwischen V_i und V_j bevorzugt, während Werte < 1 für d_{ij} solche Kanten beim Einfügen benachteiligen.

Wir können auf diese Weise eine Vielzahl von Partitionierungsproblemen modellieren. Sollen beispielsweise k-färbbare Graphen generiert werden, genügt es $d_{ii} = 0$ zu wählen. Dadurch werden keine Kanten innerhalb der V_i eingefügt, sondern nur zwischen den Mengen.

Zur Generierung von Graphen mit einer auffällig dichten Menge, könnte man beispielsweise $k = 2$ sowie

$$D = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

wählen. Innerhalb von V_1 befinden sich dann im Erwartungswert doppelt so viele Kanten als bei dem entsprechenden Zufallsgraphen *ohne* gepflanzte Partition.

Die d_{ij} skalieren die Kantenwahrscheinlichkeiten, wodurch auch der erwartete Grad w'_u von Knoten u ändert. Sei $u \in V_i$, dann ist

$$w'_u = \mathbf{E}[d_u] = \sum_{j=1}^{k} \sum_{v \in V_j} d_{ij} \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot n} = \frac{w_u}{\overline{w} \cdot n} \cdot \sum_{j=1}^{k} \sum_{v \in V_j} d_{ij} \cdot w_v \,.$$

Wir bezeichnen mit $\overline{w}'$ den mittleren erwarteten Grad, also $\sum_{u \in V} w'_u / n$.

Um eine zuverlässige Partitionierung gewährleisten zu können, müssen wir die Freiheiten des Modells leicht einschränken. Falls eine Menge V_i zu wenige Knoten enthält, wird es kaum möglich sein, diese in einem riesigen Graphen wiederzufinden. Daher sollen alle Mengen lineare Größe haben, d. h. $|V_i| \geq \delta n$ für alle i und eine beliebig kleine positive Konstante δ. Da wir unter dieser Bedingung ohnehin nur eine konstante Anzahl ($\leq 1/\delta$) von Mengen V_i bilden können, nehmen wir an, dass k ebenfalls konstant ist.

Weiterhin nehmen wir an, dass die Matrix D konstante (von n unabhängige) Einträge hat, sowie vollen Rang besitzt. Inwieweit diese Forderungen abgeschwächt werden können, diskutieren wir in Abschnitt 3.3.

Bei den Gewichten müssen ebenfalls kleinere Einschränkungen gemacht werden:

1. $\varepsilon \cdot \overline{w} \leq w_u \leq n^{1-\varepsilon}$ für $\varepsilon > 0$ beliebig, aber konstant und alle $u \in V$.
2. $\overline{w} \geq d = d(\varepsilon, D, \delta)$

Punkt 1. sichert, dass kein Knotengewicht deutlich unter $\overline{w}$ liegt. Die obere Schranke an das Gewicht verhindert, dass gleichzeitig Knoten mit kleinem (=konstantem) Gewicht und Knoten mit Gewicht nahe an n vorhanden sind, da dieser Fall die Analyse sehr behindert. Falls alle Gewichte $\gg \log n$ sind, kann auf die obere Schranke verzichtet werden.

$\overline{w} \geq d = d(\varepsilon, D, \delta)$. Der Wert von d hängt also von den vorherigen Konstanten ab. Je kleiner ε, δ und die Einträge in D sind, desto größer ist d (bleibt aber unabhängig von n). Das folgende Lemma, dessen Beweis wir zurückstellen, fasst einige elementare Eigenschaften unseres Modells zusammen.

Lemma 1.

1. *Seien u_1, u_2 zwei Knoten aus der gleichen Menge der gepflanzten Partition. Dann gilt $w_{u_1}/w'_{u_1} = w_{u_2}/w'_{u_2}$.*
2. *Für $C = C(D, \varepsilon, \delta) =$ konstant und groß genug gilt $1/C \leq w_u/w'_u \leq C$ für alle $u \in V$.*
3. *Der mittlere erwartete Grad $\overline{w}' = \sum_{u \in V} w'_u/n$ von G ist $\Theta(\overline{w})$.*

Da w_u/w'_u für alle $u \in V_i$ gleich ist, kürzen wir ab:

$$W_i = w_u/w'_u = \Theta(1) \qquad \text{und} \qquad W = \overline{w}/\overline{w}' = \Theta(1). \tag{2.1}$$

Diese Gleichungen zeigen, dass wir im Wesentlichen noch immer die erwartete Gradsequenz $w'_1, \ldots, w'_n$ durch die Gewichte $w_1, \ldots, w_n$ vorgeben können.

Insbesondere ist es möglich, Gradsequenzen mit sogenanntem heavy-tail zu erzeugen. Dazu zählt unter anderem das power-law: Die Anzahl der Knoten mit Gewicht w_u ist dabei proportional zu $n \cdot {w_u}^{-\beta}$, wobei β konstant ist. Die Gradsequenz vieler sozialer oder auch biologischer Netzwerke folgt einem power-law mit $2 < \beta < 3$. Für weitere Details verweisen wir auf die in [CLV03] zitierten Artikel und [DM03].

Für $\beta > 2$ hat die Gradsequenz einen konstanten Mittelwert $\overline{d}$ und ein großer Anteil von Knoten hat einen Grad $\gg \overline{d}$ bis hin zu $n^{1/\beta}$. Wir weisen darauf hin, dass gerade dieser interessante Fall von unserem Modell trotz der gemachten Einschränkungen erfasst wird.

Beweis von Lemma 1.

Wir zeigen die Punkte 1. und 2. exemplarisch für $u_1, u_2 \in V_1$, also für die erste Menge unserer Partition. Sei $u \in V_1$ beliebig, dann haben wir

$$\mathbf{E}[d_u] = w'_u = \sum_{j=1}^{k} \sum_{v \in V_j} d_{1j} \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot n}.$$

Division durch $w_u > 0$ ergibt

$$\frac{w'_u}{w_u} = \sum_{j=1}^{k} \sum_{v \in V_j} d_{1j} \cdot \frac{w_v}{\overline{w} \cdot n}\,. \tag{2.2}$$

Die rechte Seite ist unabhängig von u, was Punkt 1. beweist.

Wir kommen zu Punkt 2. Da $w_v \geq \varepsilon \cdot \overline{w}$ für alle $v \in V$, gilt mit (2.2)

$$\frac{w'_u}{w_u} \;\geq\; \sum_{j=1}^{k} \sum_{v \in V_j} d_{1j} \cdot \frac{\varepsilon \cdot \overline{w}}{\overline{w} \cdot n} \geq \sum_{j=1}^{k} d_{1j} \cdot |V_j| \cdot \frac{\varepsilon}{n} \geq \delta \cdot \varepsilon \cdot \sum_{j=1}^{k} d_{1j}\,.$$

Da kein d_{ij} negativ ist, gilt $\sum_{j=1}^{k} d_{1j} \geq 0$. Gleichheit können wir ausschließen, da D sonst eine 0-Zeile enthielte und somit D's Rang $< k$ wäre. Also ist $\sum_{j=1}^{k} d_{1j} > 0$ und

$$w'_u / w_u \geq 1/C$$

für eine große positive Konstante $C = C(D, \varepsilon, \delta)$. C ist aber unabhängig von der Wahl der Gewichte $w_1, \ldots, w_n$ und auch von n. Wenn wir (2.2) noch einmal benutzen, erhalten wir

$$\frac{w'_u}{w_u} \leq \max_j\{d_{1j}\} \cdot \sum_{j=1}^{k} \sum_{v \in V_j} \frac{w_v}{\overline{w} \cdot n} = \max_j\{d_{1j}\} \leq C$$

für $C = C(D)$ groß genug. Punkt 3. folgt unmittelbar aus Punkt 2. □

2.2 Notation

An dieser Stelle sammeln wir grundsätzliche Notationen, die wir im Folgenden oft benötigen.

1. $\log \cdot$ bezeichnet den Logarithmus zur Basis 2, falls keine andere Basis angegeben ist.
2. $\|\cdot\|$ ist die l_2-Norm eines Vektors oder einer Matrix.

3. Die Transponierte einer Matrix (oder eines Vektors) M schreiben wir als M^t.
4. Stehen zwei Vektoren v_1, v_2 senkrecht aufeinander, schreiben wir $v_1 \perp v_2$.
5. Für Vektoren $v_1, \ldots, v_k$ bezeichnet $\langle v_1, \ldots, v_k \rangle$ den von diesen Vektoren aufgespannten Unterraum.
6. Für einen Unterraum S ist $S^\perp$ das orthogonale Komplement von S.
7. Wir kürzen $(1, \ldots, 1)^t$ durch $\mathbf{1}$ ab.
8. Die x-te Koordinate eines Vektors v ist mit $v(x)$ bezeichnet.
9. Für $X \subseteq \mathbb{N}$ und einen Vektor v erhält man $v_{|X}$ aus v, indem man $v(x) := 0$ setzt, falls $x \notin X$.
10. Für eine Matrix M und $X, Y \subseteq \mathbb{N}$ ist die durch X und Y induzierte Untermatrix durch $M_{X \times Y}$ bezeichnet. Mit anderen Worten wird $M_{X \times Y}$ aus M konstruiert, indem alle Zeilen x mit $x \notin X$ und alle Spalten y mit $y \notin Y$ aus M *gelöscht* werden. Für einen Vektor v ist v_X analog definiert. Beachte den Unterschied zu $v_{|X}$.
11. Für eine Matrix $M = (m_{uv})$ sei

$$s_M(X, Y) = \sum_{\substack{x \in X \\ y \in Y}} m_{xy}.$$

    In Fällen wie $s_M(\{u\}, Y)$ verzichten wir auf die Klammern und schreiben $s_M(u, Y)$.
12. Grundsätzlich sind alle Konstanten hinter O und Ω positiv, während o-Terme auch negativ sein können. Statt $1 - O(1/n) \leq x \leq 1 + O(1/n)$ schreiben wir $x = 1 \pm O(1/n)$.

2.3 Idee und Algorithmus

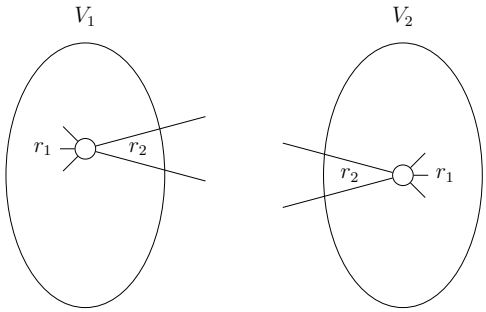
Bei der spektralen Partitionierung werden aus dem gegebenen Graphen geeignete Matrizen konstruiert. Anschließend berechnet man deren Eigenwerte und -vektoren und versucht anhand dieser, Charakteristika des Graphen zu finden. Ein Vektor $e \neq 0$ ist Eigenvektor einer Matrix M zum Eigenwert λ, wenn gilt

$$M \cdot e = \lambda \cdot e.$$

Ist M eine reellwertige, symmetrische Matrix, so sind ihre Eigenwerte und -vektoren ebenfalls reell und können in Polynomialzeit ausreichend genau approximiert werden.

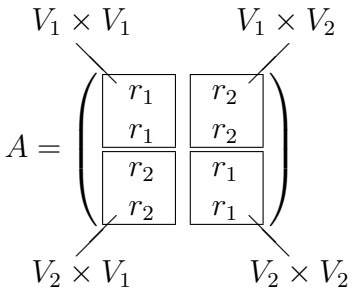
Um das Wesen der spektralen Partitionierung zu verstehen, betrachten wir ein Beispiel. Die Knoten des folgenden Graphen bestehen aus zwei gleich-

großen Gruppen V_1 und V_2, und jeder Knoten hat innerhalb seiner eigenen Gruppe genau r_1 Nachbarn und in der jeweils anderen Gruppe genau r_2 Nachbarn. Nehmen wir nun $r_1 \neq r_2$ an, bildet sich dadurch eine Auffälligkeit im Graphen, die wir finden wollen. Falls $r_1 > r_2$ ist, so sind V_1, V_2 zwei „Communities“, die relativ wenige Verbindungen zueinander haben. Im umgekehrten Fall sind V_1, V_2 zwei relativ dünne Mengen mit relativ vielen Kanten zueinander.



Die zugehörige Adjazenzmatrix A reflektiert die Struktur des obigen Graphen. Zur besseren Übersicht nehmen wir an, dass die Knoten so nummeriert sind, dass zuerst alle Knoten aus V_1 und dann alle Knoten aus V_2 kommen. Natürlich sind die Knoten eines gegebenen Graphen nicht so angenehm angeordnet. Das behindert jedoch nicht den Einsatz von Spektralmethoden: Jede Umsortierung der Knoten kann durch eine entsprechende Umsortierung der Einträge in den Eigenvektoren ausgeglichen werden, so dass wir dieselben Informationen gewinnen können.

Durch die beiden Mengen V_1 und V_2 wird die Adjazenzmatrix A in vier Blöcke zerlegt. Die folgende Abbildung illustriert A's Aufbau. Die angegebenen r_i zeigen an, wie viele Einsen sich in einer Zeile des entsprechenden Blocks befinden.



Multiplizieren wir nun die Matrix mit dem Vektor $\mathbf{1}$, erhalten wir

$$\begin{pmatrix} \boxed{r_1} & \boxed{r_2} \\ \boxed{r_1} & \boxed{r_2} \\ \boxed{r_2} & \boxed{r_1} \\ \boxed{r_2} & \boxed{r_1} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} r_1 + r_2 \\ r_1 + r_2 \\ r_1 + r_2 \\ r_1 + r_2 \end{pmatrix} = (r_1 + r_2) \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Der **1**-Vektor ist also ein Eigenvektor zum Eigenwert $r_1 + r_2$. Interessanter ist jedoch der Vektor, der auf allen zu V_1 gehörenden Koordinaten 1 ist, und auf den zu V_2 gehörenden Koordinaten -1 ist:

$$\begin{pmatrix} \boxed{r_1} & \boxed{r_2} \\ \boxed{r_1} & \boxed{r_2} \\ \boxed{r_2} & \boxed{r_1} \\ \boxed{r_2} & \boxed{r_1} \end{pmatrix} \cdot \begin{pmatrix} \boxed{1} \\ \boxed{1} \\ \boxed{-1} \\ \boxed{-1} \end{pmatrix} = \begin{pmatrix} r_1 - r_2 \\ r_1 - r_2 \\ r_2 - r_1 \\ r_2 - r_1 \end{pmatrix} = (r_1 - r_2) \cdot \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}.$$

Auch dieser Vektor ist ein Eigenvektor, aber zum Eigenwert $r_1 - r_2$. Da $r_1 \neq r_2$ ist, ist der zugehörige Eigenwert ungleich 0. Sind die Kanten jeweils innerhalb der V_i und zwischen V_1 und V_2 gleichmäßig verteilt, so sind alle anderen Eigenwerte ausreichend nahe an 0. So können wir V_1 und V_2 an dem Eigenvektor ablesen, zu dem der betragsmäßig zweitgrößte Eigenwert gehört.

Wichtig ist dabei die Lücke zwischen dem betragsmäßig zweit- und drittgrößten Eigenwert. Den Quotienten zwischen beiden Beträgen bezeichnen wir als „spectral gap“. Je größer diese Lücke ist, desto besser gelingt die Partitionierung. Mitunter bezeichnen wir auch die Lücke zwischen betragsgrößtem und -zweitgrößtem Eigenwert als „spectral gap“. Es wird sich aus dem Kontext erschließen, falls dies der Fall ist.

Der Graph aus obigem Beispiel ist sehr regulär aufgebaut. Typischerweise haben wir beim Partitionieren weniger regelmäßige Graphen zu bearbeiten. Betrachtet man das $G_{n,p}$-Modell mit $p \gg \log n/n$ und einer gepflanzten Partition V_1, V_2, verhält sich die Adjazenzmatrix meistens jedoch ähnlich wie oben.

Die ersten Probleme erhalten wir jedoch, wenn $p = O(1/n)$ ist. Dann nämlich sind die Knotengrade nicht mehr an ihrem Erwartungswert $p \cdot n = O(1)$ konzentriert. Ein so dünner Zufallsgraph hat mhW. für jedes feste l (insbesondere $l = 0$) $\Theta(n)$ Knoten vom Grad l. Ebenso existieren viele Knoten von sehr hohem Grad, etwa $\log \log n$.

Solche Knoten von hohem Grad dominieren das Spektrum von A derart, dass die größten Eigenwerte die Wurzeln der größten Grade sind [KS03] und die zugehörigen Eigenvektoren lediglich diese Knoten anzeigen. Damit wird das Spektrum von *lokalen* Grapheigenschaften dominiert, während die *globalen* Eigenschaften verborgen bleiben.

Um diesem Problem zu begegnen, löscht man die Knoten vom höchsten Grad [AK97] und hinterlässt nur Knoten mit Grad $\leq C \cdot p \cdot n$, wobei $C > 1$ eine Konstante ist. Die Eigenvektoren der Adjazenzmatrix dieses Teilgraphen lassen dann die gepflanzte Partition erkennen.

Leider können wir diesen Trick nicht anwenden, da in unserem Modell die Gradsequenzen irregulär sein *sollen*. Ein Löschen der Knoten mit hohem Grad könnte dazu führen, dass eine Partitionierung des Restgraphen unmöglich ist, da zu viele Kanten (und damit Informationen) aus dem Graphen entfernt würden.

Stattdessen scheint es in unserem Modell nötig zu sein, andere Matrizen als die Adjazenzmatrix zu nutzen. Eine häufig benutzte Alternative ist die normalisierte Laplace-Matrix $\mathcal{L} = (l_{uv})$ mit

$$l_{uv} = \begin{cases} 1 & u = v \text{ und } d_u > 0 \\ -1/\sqrt{d_u \cdot d_v} & u \neq v \text{ und } \{u, v\} \in E \\ 0 & \text{sonst} \end{cases},$$

wobei d_u, d_v die Grade von u bzw. v sind. Der Nutzen von $\mathcal{L}$ in der Graphentheorie ist in [Chu97] ausführlich beschrieben. Für das $G(\mathbf{w})$-Modell wurde $\mathcal{L}$ in [COL06] analysiert.

Eine zu $\mathcal{L}$ ähnliche Normalisierung wurde in [DHM04] benutzt, um gepflanzte Partitionen bei $G(\mathbf{w})$ zu rekonstruieren: Dasgupta, Hopcroft und McSherry teilten jeden Eintrag der Adjazenzmatrix A durch $\sqrt{w_u \cdot w_v}$, wobei w_u und w_v die Gewichte der inzidenten Knoten u und v sind. Es ist zu beachten, dass die Autoren die Gewichte benutzen. Also weder die erwarteten, noch die tatsächlichen Grade. Das macht eine Nutzung des Algorithmus im praktischen Umfeld fraglich, da derartige Informationen für „reale“ Graphen nicht zur Verfügung stehen.

Die gewählte Normalisierung bewirkt, dass die durch $V_i \times V_j$ induzierten Untermatrizen eine Gemeinsamkeit besitzen: Innerhalb eines jeden Blockes sind die Varianzen der Einträge asymptotisch gleich.

Wir wählen eine andere Normalisierung. Der Grund dafür ist zum einen die einfachere Analyse. Zum anderen erzeugt unsere Normalisierung einige interessante Analogien zur Adjazenzmatrix im $G_{n,p}$-Modell mit gepflanzter Partition. Wir teilen jeden Eintrag von $A = (a_{uv})$ durch das Produkt der

erwarteten Grade der inzidenten Knoten. Um die angesprochenen Analogien klarer zu sehen, multiplizieren wir jeden Eintrag anschließend mit $\overline{w}'^2$.

Sei also $M = (m_{uv})$ die normalisierte Matrix, dann ist $m_{uv} = a_{uv} \cdot \overline{w}'^2/(w'_u \cdot w'_v)$. Für $u \in V_i$ und $v \in V_j$ ist

$$\begin{aligned} \mathbf{E}[m_{uv}] &= \frac{\overline{w}'^2}{w'_u \cdot w'_v} \cdot \mathbf{Pr}[\{u,v\} \in E] + 0 \cdot \mathbf{Pr}[\{u,v\} \notin E] \\ &= \frac{\overline{w}'^2}{w'_u \cdot w'_v} \cdot d_{ij} \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot n} \overset{(2.1)}{=} d_{ij} \cdot \frac{W_i \cdot W_j}{W} \cdot \frac{\overline{w}'}{n} \\ &= \Theta\left(d_{ij} \cdot \frac{\overline{w}'}{n}\right). \end{aligned} \tag{2.3}$$

Entscheidend ist, dass der Wert nur von i und j abhängt und nicht von u und v selbst. Alle Einträge innerhalb von $M_{V_i \times V_j}$ haben also *denselben Erwartungswert*. Dies gilt genauso für die Adjazenzmatrix bei $G_{n,p}$ mit eingepflanzter Partition. Dort haben wir einen Erwartungswert von $d_{ij} \cdot p = d_{ij} \cdot \overline{w}/n = \Theta(d_{ij} \cdot \overline{w}'/n)$. Hier sehen wir die erste Analogie zu $G_{n,p}$.

Wie bei der Adjazenzmatrix im $G_{n,p}$-basierten Modell wird das Spektrum unserer Matrix von lokalen Phänomenen dominiert. Wir werden sehen, dass es auch hier genügt, die Zeilen (und Spalten) der Matrix zu löschen, deren Summe weit über dem Erwartungswert liegt.

Aus (2.3) können wir leicht folgern, dass die erwartete Zeilensumme durch

$$\max_{i,j} \left\{ d_{ij} \cdot \frac{W_i \cdot W_j}{W} \right\} \cdot \overline{w}'$$

beschränkt ist.

Wir wollen alle Einträge löschen, deren Zeilensumme größer als $C_1 \cdot \overline{w}'$ ist. Als untere Schranke für C_1 wählen wir

$$C_1 \geq 5 \cdot \max_{i,j} \left\{ d_{ij} \cdot \frac{W_i \cdot W_j}{W} \right\} = \Theta(1). \tag{2.4}$$

Diese sichert uns, dass jeder gelöschte Eintrag eine Summe hat, die mindestens mit dem Faktor 5 über ihrem Erwartungswert liegt. Da eine so große relative Abweichung selten ist, wird mhW. nur eine kleine Anzahl von Einträgen gelöscht.

Um sinnvolle Ergebnisse zu erhalten, darf C_1 aber auch nicht zu groß sein. Wie wir gleich sehen werden, muss C_1 deutlich kleiner als $\overline{w}'$ sein, etwa $C_1 \leq \overline{w}'/\ln \overline{w}'$ da sonst das „spectral gap“ der konstruierten Matrix zu klein und die Partitionierung nicht erfolgreich ist.

Wir bezeichnen die durch diesen Löschschritt aus M erhaltene Matrix mit M^*. M^* hat einige wichtige spektrale Eigenschaften, die uns die Rekonstruktion der eingepflanzten Partition erlauben.

Theorem 2. *Sei G ein in unserem Modell generierter Graph und M^* wie beschrieben konstruiert. Dann gilt mit Wahrscheinlichkeit $1-O(1/n)$ für alle $1 \leq i, j \leq k$ simultan:*

1. $\dfrac{\mathbf{1}^t}{\|\mathbf{1}\|} \cdot {M^*}_{V_i \times V_j} \cdot \dfrac{\mathbf{1}}{\|\mathbf{1}\|} = d_{ij} \cdot \dfrac{W_i \cdot W_j}{W} \cdot \sqrt{|V_i| \cdot |V_j|} \cdot \dfrac{\overline{w}'}{n} \cdot \left(1 \pm O\left(1/\sqrt{\overline{w}'}\right)\right).$

2. *Für alle u, v mit $\|u\| = \|v\| = 1$ sowie $u \perp \mathbf{1}$ oder $v \perp \mathbf{1}$ gilt*

$$\left|u^t \cdot {M^*}_{V_i \times V_j} \cdot v\right| = O\left(\sqrt{C_1 \cdot \overline{w}'}\right).$$

Punkt 1. ist ein Konzentrationsresultat, denn mit (2.3) erhalten wir für $M_{V_i \times V_j}$:

$$\mathbf{E}\left[\frac{\mathbf{1}^t}{\|\mathbf{1}\|} \cdot M_{V_i \times V_j} \cdot \frac{\mathbf{1}}{\|\mathbf{1}\|}\right] = \frac{\mathbf{E}\left[s_M(V_i, V_j)\right]}{\sqrt{|V_i| \cdot |V_j|}} = d_{ij} \cdot \frac{W_i \cdot W_j}{W} \cdot \frac{\overline{w}'}{n} \cdot \sqrt{|V_i| \cdot |V_j|}.$$

Die Untermatrix ${M^*}_{V_i \times V_j}$ verhält sich also in Bezug auf den $\mathbf{1}$-Vektor, wie es $M_{V_i \times V_j}$ im Erwartungswert tut. Das Löschen der Einträge hat folglich nur einen kleinen Effekt.

Theorem 2 zeigt eine weitere Analogie von zum $G_{n,p}$-basierten Modell (vgl. [AK97]): Zwei Einheitsvektoren u und v, die das Produkt $u^t {M^*}_{V_i \times V_j} v$ maximieren, sind fast parallel zu $\mathbf{1}$ und $u^t {M^*}_{V_i \times V_j} v = \Theta(\overline{w}')$. Hingegen gilt für Einheitsvektoren u', v' mit $u' \perp u$ oder $v' \perp v$, dass $\left|u'^t {M^*}_{V_i \times V_j} v'\right|$ deutlich kleiner ist, nämlich $O(\sqrt{C_1 \cdot \overline{w}'})$.

Das große „spectral gap“[1] von ${M^*}_{V_i \times V_j}$ ermöglicht schließlich die Partitionierung mit folgendem Algorithmus. Wir beschränken uns der Einfachheit halber auf den Fall, dass eine Partition mit nur zwei Mengen eingepflanzt wurde. Eine Erweiterung auf mehr Mengen ist unproblematisch, würde die Erklärungen aber unnötig komplizieren. Wir diskutieren diesen Fall in Abschnitt 3.3.

Algorithmus 3.

Eingabe: Die Adjazenzmatrix $A = (a_{uv})$ von $G = (V, E)$ generiert in unserem Modell und die erwarteten Grade $w'_1, \ldots, w'_n$.
Ausgabe: Eine Partition V'_1, V'_2 von V.

1. Berechne den erwarteten Durchschnittsgrad $\overline{w}' = \sum_{u=1}^{n} w'_u / n$.
2. Konstruiere $M = (m_{uv})$ mit $m_{uv} = \overline{w}'^2 \cdot a_{uv} / (w'_u \cdot w'_v)$.
3. Bestimme $U = \{u \in V : \sum_{v=1}^{n} m_{uv} \leq C_1 \cdot \overline{w}'\}$.

[1] Hier beziehen wir uns auf den Quotienten von betragsgrößtem und -zweitgrößtem Eigenwert.

4. Konstruiere $M^* = M_{U \times U}$.
5. Berechne s_1 und s_2: Die Eigenvektoren von M^* zu den beiden betragsgrößten Eigenwerten. Skaliere beide s_i auf Länge $\sqrt{n}$.
6. Falls keiner der beiden s_i die Eigenschaft

   „Es gibt $c_1, c_2 \in \mathbb{R}$ mit $|c_1 - c_2| > 1/4$, so dass jeweils mehr als $n \cdot \sqrt{C_1/\overline{w}'}$ Knoten $v \in U$ $|s_i(v) - c_1| \leq 1/32$ bzw. $|s_i(v) - c_2| \leq 1/32$ erfüllen."

   hat, setze $V_1 = V$ und $V_2 = \emptyset$. Ansonsten sei $s \in \{s_1, s_2\}$ so ein Eigenvektor. Dann sei V_1' die Menge der Knoten, deren Einträge in s näher an c_1 als an c_2 sind. Setze $V_2' := V \setminus V_1'$.

Algorithmus 3 ist nicht in der Lage, den genauen Wert der Schranke in (2.4) zu bestimmen, da wesentliche Informationen aus dem Generierungsprozess (etwa die d_{ij}) fehlen. Um (2.4) zu erfüllen, könnte man $C_1 = \ln \overline{w}'$ (oder auch jede andere mit $\overline{w}'$ wachsende Funktion) wählen. Wegen Punkt 2. auf Seite 13 der Modellbeschreibung und (2.1) gilt $\overline{w}' = \overline{w}/W \geq d/W$. Damit ist C_1 groß genug, wenn d ausreichend groß gewählt wird.

Wir werden später (auf Seite 32, Ungleichung (2.15)) zeigen, dass bei Einhaltung von (2.4) mit Wahrscheinlichkeit $1 - O(1/n)$

$$|V \setminus U| \leq \exp(-\Omega(\overline{w}')) \cdot n \tag{2.5}$$

gilt. Es werden also in Schritt 4. nur wenige Knoten gelöscht.

Es ist zu beachten, dass die Werte von c_1, c_2 in Schritt 6 auch von den Modellparametern abhängen und dem Algorithmus somit unbekannt sind. Durch eine Analyse der s_i ist es jedoch leicht möglich, passende c_i zu bestimmen. Bemerkenswert ist, dass die Eigenschaft in Schritt 6 eine relativ kleine Anzahl von Koordinaten nahe der c_i fordert. Es stellt sich jedoch heraus, dass diese Anzahl bereits genügt, damit fast alle Koordinaten in der Umgebung der c_i liegen.

Dies hängt unmittelbar mir dem „spectral gap" der ${M^*}_{V_i \times V_j}$ zusammen, was laut Theorem 2 die Größenordnung $\Omega(\sqrt{\overline{w}'/C_1})$ hat: Es gibt einen idealen Wert c_1^i für c_1, so dass von den Einträgen, die zu V_1 gehören, nicht mehr als $O(n/(\text{„spectral gap"})^2)$ viele einen Abstand $\geq 1/128$ zu c_1^i haben. In unserem Falle sind das also nicht mehr als $O(n \cdot C_1/\overline{w}')$ viele. Alle anderen Einträge, die zu V_1 gehören, liegen also in einem Intervall der Länge $1/64$.

Algorithmus 3 wählt c_1 nun so, dass deutlich mehr Einträge ($\geq n \cdot \sqrt{C_1/\overline{w}'}$) nahe c_1 sind. Damit hat c_1 von obigem Intervall einen Abstand $\leq 1/32$. Insgesamt liegen also – abgesehen von den $O(n \cdot C_1/\overline{w}')$ vielen – alle Einträge, die zu V_1 gehören in einem Abstand $\leq 1/32 + 1/64 = 3/64$ von c_1. Analoges gilt für c_2 und V_2. Der große Abstand von $1/4 = 16/64$

zwischen c_1 und c_2 gewährleistet eine gute Unterscheidung der Koordinaten, die zu V_1 bzw. V_2 gehören. Exakte Ausführungen dazu sehen wir im Beweis des nachfolgenden Theorems, das die Güte von Algorithmus 3 beschreibt.

Theorem 4. *Sei G ein Graph, der in unserem Modell generiert wurde. Mit Wahrscheinlichkeit $1 - O(1/n)$ (bezogen auf G) konstruiert Algorithmus 3 eine Partition, die sich von der gepflanzten Partition V_1, V_2 nur in $O(C_1 \cdot n/\overline{w}')$ Knoten unterscheidet.*

Der Anteil der falsch klassifizierten Knoten ist also $O(C_1/\overline{w}')$ und verringert sich mit wachsendem $\overline{w}'$, vorausgesetzt $C_1 \ll \overline{w}'$.

Die Erfolgswahrscheinlichkeit in den Theoremen 2 und 4 entspricht der Wahrscheinlichkeit, dass (2.5) gilt. Diese ist – unserer Analyse nach – nur $1 - O(1/n)$. Durch den Einsatz von Martingalen kann man (wie in [COL06]) zeigen, dass (2.5) mit Wahrscheinlichkeit $1 - o(1/n^4)$ gilt. Dann gilt in beiden Theoremen die Schranke $1 - O(1/n^4)$, die dann von anderen Teilen der Beweise bestimmt wird. Da das Martingal-Argument und die in [COL06] benutzte Filtration recht kompliziert sind, begnügen wir uns aber mit der schwächeren Schranke, deren Beweis die Tschebyscheff-Ungleichung benutzt.

2.4 Beweise

2.4.1 Beweis von Theorem 4

Der Beweis von Theorem 4 hat als wichtige Grundlage Theorem 2. Die dort betrachteten Untermatrizen $M^*_{V_i \times V_j}$ können aber unabhängig von Algorithmus 3 analysiert werden. Wir verschieben den Beweis von Theorem 2 daher in Abschnitt 2.4.2.

Wir beginnen den Beweis von Theorem 4 mit einem Lemma über die Eigenwerte von M^*. Dessen Korrektheit basiert hauptsächlich auf Theorem 2 und der Courant-Fischer-Charakterisierung von Eigenwerten. Die folgende Variante des Courant-Fischer-Theorems lässt sich leicht aus Theorem 8.1.2 in [GV01] ermitteln.

Fakt 5. *Sei $A \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix mit Eigenwerten $\lambda_1 \geq \ldots \geq \lambda_n$. Dann gilt für alle $0 \leq j < n$*

$$\begin{aligned} \lambda_{j+1} &= \min_{\substack{S \\ \dim S = j}} \max_{\substack{x \in S^\perp \\ \|x\|=1}} x^t A x \\ \lambda_{n-j} &= \max_{\substack{S \\ \dim S = j}} \min_{\substack{x \in S^\perp \\ \|x\|=1}} x^t A x \,, \end{aligned}$$

wobei $\dim S$ die Dimension des Unterraums S angibt.

Lemma 6. *Mit Wahrscheinlichkeit $1 - O(1/n)$ hat M^* genau zwei Eigenwerte, deren Betrag $\Theta(\overline{w}')$ ist, während alle anderen Eigenwerte nur einen Betrag von $O\left(\sqrt{C_1 \cdot \overline{w}'}\right)$ haben.*

Beweis. Sei U die in Schritt 3 von Algorithmus 3 konstruierte Menge. Seien χ_1 bzw. χ_2 die $|U|$-dimensionalen charakteristischen Vektoren von $V_1 \cap U$ bzw. $V_2 \cap U$ (die u-te Koordinate $\chi_i(u) = 1$ wenn $u \in V_i \cap U$ und ansonsten 0).

Wir betrachten zwei beliebige normierte Vektoren g und h aus dem von χ_1 und χ_2 aufgespannten Raum $\langle \chi_1, \chi_2 \rangle$. Sei dazu

$$\begin{aligned} g &= a_1 \cdot \frac{\chi_1}{\|\chi_1\|} + a_2 \cdot \frac{\chi_2}{\|\chi_2\|} \quad \text{mit} \quad a_1^2 + a_2^2 = 1 \\ \text{und} \quad h &= b_1 \cdot \frac{\chi_1}{\|\chi_1\|} + b_2 \cdot \frac{\chi_2}{\|\chi_2\|} \quad \text{mit} \quad b_1^2 + b_2^2 = 1\,. \end{aligned}$$

Wir erhalten

$$h^t M^* g = \sum_{i,j=1}^{2} b_i \cdot \frac{\chi i}{\|\chi_i\|} \cdot M^* \cdot a_j \cdot \frac{\chi_j}{\|\chi_j\|} = \sum_{i,j=1}^{2} b_i \cdot a_j \cdot \frac{\mathbf{1}^t \cdot {M^*}_{V_i \times V_j} \cdot \mathbf{1}}{\sqrt{|V_i \cap U| \cdot |V_j \cap U|}}\,.$$

Den Zähler schätzen wir mit Punkt 1. von Theorem 2 ab. So gilt mit Wahrscheinlichkeit $1 - O(1/n)$

$$\begin{aligned} h^t M^* g &= \sum_{i,j=1}^{2} b_i \cdot a_j \cdot d_{ij} \cdot \frac{W_i \cdot W_j}{W} \cdot \overline{w}' \cdot \frac{\sqrt{|V_i| \cdot |V_j|}}{n} \cdot \left(1 \pm O\left(\frac{1}{\sqrt{\overline{w}'}}\right)\right) \\ &= \sum_{i,j=1}^{2} \left(b_i \cdot a_j \cdot d_{ij} \cdot \frac{W_i \cdot W_j}{W} \cdot \overline{w}' \cdot \frac{\sqrt{|V_i| \cdot |V_j|}}{n} \right) \pm O\left(\sqrt{\overline{w}'}\right) \\ &= \frac{\overline{w}'}{W} \cdot \begin{pmatrix} b_1 & b_2 \end{pmatrix} \cdot P \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \pm O\left(\sqrt{\overline{w}'}\right)\,, \end{aligned}$$

wobei

$$P = \begin{pmatrix} W_1 \cdot \sqrt{\frac{|V_1|}{n}} & 0 \\ 0 & W_2 \cdot \sqrt{\frac{|V_2|}{n}} \end{pmatrix} \cdot \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix} \cdot \begin{pmatrix} W_1 \cdot \sqrt{\frac{|V_1|}{n}} & 0 \\ 0 & W_2 \cdot \sqrt{\frac{|V_2|}{n}} \end{pmatrix}$$

ist. Wir erinnern uns, dass $0 \neq W_i = \Theta(1)$ ist und $\delta \cdot n \leq |V_i| \leq n$.

Wegen $\delta > 0$ haben die beiden äußeren Faktoren von P in jedem Fall Rang 2 und per Definition gilt das gleiche für D. Somit hat P ebenfalls vollen Rang und alle Eigenwerte von P sind von 0 verschieden. Beachte, dass

weder n noch die konkrete Wahl der Gewichte Einfluss auf diese Eigenschaft haben.

Seien nun $\begin{pmatrix} e_1 & e_2 \end{pmatrix}^t$ und $\begin{pmatrix} f_1 & f_2 \end{pmatrix}^t$ zwei orthonormale Eigenvektoren von P zu den Eigenwerten λ_1 und λ_2. Wir setzen

$$g_1 = e_1 \cdot \frac{\chi_1}{\|\chi_1\|} + e_2 \cdot \frac{\chi_2}{\|\chi_2\|} \qquad \text{und} \qquad g_2 = f_1 \cdot \frac{\chi_1}{\|\chi_1\|} + f_2 \cdot \frac{\chi_2}{\|\chi_2\|} .$$

Es folgt

$$\begin{aligned} \left| g_1^t \cdot M^* \cdot g_1 \right| &= \left| \frac{\overline{w}'}{W} \cdot \begin{pmatrix} e_1 & e_2 \end{pmatrix} \cdot P \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \pm O\left(\sqrt{\overline{w}'}\right) \right| \\ &= \left| \frac{\overline{w}'}{W} \cdot \lambda_1 \pm O\left(\sqrt{\overline{w}'}\right) \right| = \Theta(\overline{w}') \end{aligned}$$

und

$$\begin{aligned} \left| g_1^t \cdot M^* \cdot g_2 \right| &= \left| \frac{\overline{w}'}{W} \cdot \begin{pmatrix} e_1 & e_2 \end{pmatrix} \cdot P \cdot \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \pm O\left(\sqrt{\overline{w}'}\right) \right| \\ &= \left| \frac{\overline{w}'}{W} \cdot 0 \pm O\left(\sqrt{\overline{w}'}\right) \right| = O\left(\sqrt{\overline{w}'}\right) . \end{aligned}$$

Somit haben wir für $1 \leq i, j \leq 2$

$$\left| g_i^t \cdot M^* \cdot g_j \right| = \begin{cases} \Theta(\overline{w}') & \text{für } i = j \\ O\left(\sqrt{\overline{w}'}\right) & \text{für } i \neq j \end{cases} . \tag{2.6}$$

Wegen Fakt 5 haben mindestens zwei Eigenwerte von M^* den Betrag $\Omega(\overline{w}')$: Wir teilen g_1, g_2 anhand des Vorzeichens von $g_i^t M^* g_i$: Seien $g_1, \ldots, g_l$, mit $0 \leq l \leq 2$, die Vektoren, bei denen das Produkt positiv ist und $g_{l+1}, \ldots, g_2$ die, bei denen $g_i^t M^* g_i < 0$ ist.

Sei nun v ein beliebiger Einheitsvektor aus $\langle g_1, \ldots, g_l \rangle$. Dann können wir v darstellen als $v = \sum_{i=1}^{l} \alpha_i \cdot g_i$ mit $\sum_{i=1}^{l} \alpha_i^2 = 1$. Laut (2.6) gibt es nun zwei Konstanten c und C, so dass

$$\begin{aligned} v^t \cdot M^* \cdot v &= \sum_{i,j=1}^{l} \alpha_i \alpha_j \cdot g_i^t M^* g_j = \sum_{i=1}^{l} \alpha_i^2 \cdot g_i^t M^* g_i + \sum_{\substack{i,j=1 \\ i \neq j}}^{l} \alpha_i \alpha_j \cdot g_i^t M^* g_j \\ &\geq \sum_{i=1}^{l} \alpha_i^2 \cdot c \cdot \overline{w}' \; - \; C \cdot \sqrt{\overline{w}'} \cdot \sum_{\substack{i,j=1 \\ i \neq j}}^{l} \alpha_i \alpha_j \\ &\geq c \cdot \overline{w}' \; - \; C \cdot \sqrt{\overline{w}'} \cdot l = \Omega(\overline{w}') . \end{aligned}$$

Die zweite Gleichung von Fakt 5 liefert für die $|U| \times |U|$-Matrix M^* und $j = |U| - l$

$$\lambda_l = \max_{\substack{S \\ \dim S = |U|-l}} \min_{\substack{x \in S^\perp \\ \|x\|=1}} x^t M^* x.$$

Um eine untere Schranke für λ_l zu erhalten, wählen wir $S = \langle g_1, \ldots, g_l \rangle^\perp$, welches Dimension $|U| - l$ hat. Damit ist

$$\lambda_l \geq \min_{\substack{x \in \langle g_1, \ldots, g_l \rangle \\ \|x\|=1}} x^t M^* x \geq c \cdot \overline{w}' \; - \; C \cdot \sqrt{\overline{w}'} \cdot l = \Omega(\overline{w}')\,.$$

Analog zu obiger Vorgehensweise kann bewiesen werden, dass $2-l$ Eigenwerte kleiner als $-\Omega(\overline{w}')$ sind. Dazu benutzt man die erste Gleichung von Fakt 5 mit $j = |U| - (2 - l)$ und $S = \langle g_{l+1}, \ldots, g_2 \rangle^\perp$. Insgesamt sind dann also mindestens 2 Eigenwerte von M^* vom Betrag her $\Omega(\overline{w}')$.

Es ist entscheidend, dass alle anderen Eigenwerte von M^* wesentlich kleiner als $\overline{w}'$ sind. Anderenfalls können wir die Partition kaum an den Eigenvektoren ablesen.

Seien u, v beliebige Einheitsvektoren und sei u senkrecht zu g_1 und g_2. Da $\langle g_1, g_2 \rangle = \langle \chi_1, \chi_2 \rangle$ gilt auch $u \perp \chi_1, \chi_2$. Mit Theorem 2 erhalten wir

$$\left| v^t M^* u \right| = \left| \sum_{i,j=1}^{2} v_{V_i \cap U} \cdot {M^*}_{V_i \times V_j} \cdot u_{V_j \cap U} \right| \leq 4 \cdot O\left(\sqrt{C_1 \cdot \overline{w}'} \right) \tag{2.7}$$

und analog $|u^t M^* v| = O\left(\sqrt{C_1 \cdot \overline{w}'}\right)$. Gleichung 1 von Fakt 5 ergibt

$$\lambda_{l+1} \leq \max_{\substack{x \in \langle g_1, \ldots, g_l \rangle^\perp \\ \|x\|=1}} x^t M^* x\,.$$

Sei x ein Vektor, der die rechte Seite maximiert. Wir schreiben x als $x = \alpha \cdot g + \beta \cdot u$ mit $\alpha^2 + \beta^2 = 1$ sowie $g \in \langle g_{l+1}, \ldots, g_2 \rangle$ und $u \in \langle g_1, g_2 \rangle^\perp$. Wegen der Wahl von l haben wir $g^t M^* g < 0$. Mit (2.7) erhalten wir

$$\lambda_{l+1} \leq x^t M^* x = \alpha^2 \cdot g^t M^* g + 2 \cdot \alpha\beta \cdot g^t M^* u + \beta^2 \cdot u^t M^* u = O(\sqrt{C_1 \cdot \overline{w}'})\,.$$

Benutzen wir die zweite Gleichung von Fakt 5 erhalten wir analog

$$\lambda_{|U|-(2-l)} \geq \min_{\substack{x \in \langle g_{l+1}, \ldots, g_2 \rangle^\perp \\ \|x\|=1}} x^t M^* x \geq -C \cdot \sqrt{C_1 \cdot \overline{w}'}$$

für eine Konstante $C > 0$. Damit sind die restlichen $|U| - 2$ Eigenwerte von M^* im Betrag $O\left(\sqrt{C_1 \cdot \overline{w}'}\right)$. □

Durch den Beweis von Lemma 6 können wir bereits die Intuition entwickeln, dass die Eigenvektoren zu den betragsgrößten Eigenwerten von M^* im Wesentlichen aus dem Raum $\langle \chi_1, \chi_2 \rangle$ stammen und daher das Ablesen der Partition an diesen Vektoren möglich ist. Das folgende Lemma präzisiert dies.

Lemma 7. *Seien s_1, s_2 die Eigenvektoren von M^* zu den betragsgrößten Eigenwerten und χ_1, χ_2 die charakteristischen Vektoren von $V_1 \cap U$ bzw. $V_2 \cap U$. Mit Wahrscheinlichkeit $1 - O(1/n)$ gilt für die eindeutige Zerlegung*

$$s_i = \alpha_i \cdot \chi_1 + \beta_i \cdot \chi_2 + \gamma_i \cdot u_i$$

mit $\|s_i\| = \|u_i\| = \sqrt{n}$ und $u_i \perp \chi_1, \chi_2$

1. *$|\gamma_i| = O\left(\sqrt{C_1/\overline{w}'}\right)$.*
2. *Für ein $i \in \{1, 2\}$ gilt $|\alpha_i - \beta_i| \geq 1/4$.*

Beweis. Wir beginnen mit Punkt 1. Wegen Punkt 2. von Theorem 2 und $u_i \perp \chi_1, \chi_2$ ist

$$\left|s_i^t \cdot M^* \cdot u_i\right| = n \cdot \left|\frac{s_i^t}{\|s_i\|} \cdot M^* \cdot \frac{u_i}{\|u_i\|}\right| = O\left(n \cdot \sqrt{C_1 \cdot \overline{w}'}\right) .$$

Andererseits ist

$$\left|(s_i^t \cdot M^*) \cdot u_i\right| = \left|\Theta(\overline{w}') \cdot s_i^t \cdot u_i\right| = \Theta(\overline{w}') \cdot \left|\gamma_i \cdot u_i^t \cdot u_i\right| = n \cdot \Theta(\overline{w}') \cdot |\gamma_i| ,$$

weswegen $|\gamma_i| = O\left(\sqrt{C_1/\overline{w}'}\right)$ sein muss.

Wir kommen zu Punkt 2. Um einen Widerspruch zu erzeugen, nehmen wir $|\alpha_i - \beta_i| \leq 1/4$ für $i = 1, 2$ an. Aus

$$n = s_i^t \cdot s_i = \alpha_i^2 \cdot |V_1 \cap U| + \beta_i^2 \cdot |V_2 \cap U| + \gamma_i^2 \cdot n$$

erhalten wir nach Division durch n

$$\alpha_i^2 + \beta_i^2 \geq \alpha_i^2 \cdot \frac{|V_1 \cap U|}{n} + \beta_i^2 \cdot \frac{|V_2 \cap U|}{n} = 1 - \gamma_i^2 \geq 1 - O\left(C_1/\overline{w}'\right) . \quad (2.8)$$

Wir sehen, dass $|\alpha_i| > 1/2$ oder $|\beta_i| > 1/2$ gelten muss. Wegen $|\alpha_i - \beta_i| \leq 1/4$ haben α_i und β_i das gleiche Vorzeichen. Also ist

$$|\alpha_1 \cdot \alpha_2 + \beta_1 \cdot \beta_2| = |\alpha_1 \cdot \alpha_2| + |\beta_1 \cdot \beta_2| \geq \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{2} = 1/4 ,$$

was zu

$$\begin{aligned} 0 &= s_1^t \cdot s_2 = \left|\alpha_1 \cdot \alpha_2 \cdot |V_1 \cap U| + \beta_1 \cdot \beta_2 \cdot |V_2 \cap U| + \gamma_1 \cdot \gamma_2 \cdot u_1^t \cdot u_2\right| \\ &\geq (\delta n - |V \setminus U|) \cdot |\alpha_1 \cdot \alpha_2 + \beta_1 \cdot \beta_2| - |\gamma_1 \cdot \gamma_2| \cdot n \\ &\overset{(2.5)}{\geq} n \cdot (\delta \cdot |\alpha_1 \cdot \alpha_2 + \beta_1 \cdot \beta_2| - O(C_1/\overline{w}')) \\ &\geq n \cdot (\delta/4 - O\left(C_1/\overline{w}'\right)) \end{aligned}$$

führt. Da $\delta > 0$ konstant ist und $\overline{w}'$ groß genug ist, erhalten wir den gewünschten Widerspruch. □

Aufgrund des großen „spectral gap“ von $\Omega\left(\sqrt{\overline{w}'/C_1}\right)$, hat γ_i einen kleinen Wert von („spectral gap“)$^{-1} = O\left(\sqrt{C_1/\overline{w}'}\right)$. So ist der Einfluss der Vektors u auf e klein und e und der Anteil von χ_1 und χ_2 entsprechend groß. Der Abstand von $|\alpha_i - \beta_i| \geq 1/4$ in einem der beiden Vektoren s_i ist ausreichend groß, um V_1 von V_2 unterscheiden zu können. Es ist aber nicht klar, welcher der beiden Vektoren von Algorithmus 3 ausgewählt wird.

Sei $s = \alpha \cdot \chi_1 + \beta \cdot \chi_2 + \gamma \cdot u$ der ausgewählte Vektor. Also gibt es c_1, c_2 mit $|c_1 - c_2| > 1/4$ und mehr als $n \cdot \sqrt{C_1/\overline{w}'}$ Einträge v in s erfüllen

$$|s(v) - c_1| \leq \frac{1}{32} \qquad \text{bzw.} \qquad |s(v) - c_2| \leq \frac{1}{32}\,. \tag{2.9}$$

Angenommen, $|\alpha - \beta| \leq 1/16$. Wegen $|c_1 - c_2| \geq 1/4$ hat eines der beiden c_i sowohl von α also auch von β einen Abstand $\geq (1/4 - 1/16)/2 = 3/32$. Für jeden Eintrag v mit $|c_i - s(v)| \leq 1/32$ muss dann $|\gamma_i \cdot u_i(v)| \geq 2/32$ sein. Wegen $|\gamma_i| = O(\sqrt{C_1/\overline{w}'})$ ist $u_i(v) = \Omega(\sqrt{\overline{w}'/C_1})$. Da $n = u_i^t \cdot u_i$ ist, kann es maximal $n \cdot O(C_1/\overline{w}')$ solcher Einträge geben. Laut Algorithmus 3 gibt es aber mindestens $n \cdot \sqrt{C_1/\overline{w}'} \gg O(n \cdot C_1/\overline{w}')$, Widerspruch.

Es muss also $|\alpha - \beta| \geq 1/16$ gelten. Bereits diese kleine Distanz zwischen α und β genügt, um V_1 von V_2 zu separieren. Wie im vorherigen Absatz können wir zeigen: Bis auf $O(n \cdot C_1/\overline{w}')$ Einträge erfüllen alle $v \in U$

$$|s(v) - \alpha| \leq \frac{1}{128} \quad \text{für } v \in V_1 \quad \text{bzw.} \quad |s(v) - \beta| \leq \frac{1}{128} \quad \text{für } v \in V_2. \tag{2.10}$$

Es muss also wenigstens ein $v \in V_1$ geben, das sowohl (2.10) als auch (2.9) erfüllt, vorausgesetzt $\overline{w}'$ ist ausreichend groß.

Angenommen, dieses v erfüllt die erste Ungleichung in (2.9). Wegen der Dreiecksungleichung gilt

$$|\alpha - c_1| \leq |\alpha - s(v)| + |s(v) - c_1| \leq \frac{1}{128} + \frac{1}{32} = \frac{5}{128}$$

und entsprechend $|\beta - c_2| \leq 5/128$. Natürlich ist auch $|\alpha - c_2| \leq 5/128$ und $|\beta - c_1| \leq 5/128$ möglich, was jedoch analog behandelt werden kann. Wegen (2.10) erfüllen – abgesehen von $O(n \cdot C_1/\overline{w}')$ vielen – alle $v \in V_1$

$$|s(v) - c_1| \leq |s(v) - \alpha| + |\alpha - c_1| \leq \frac{1}{128} + \frac{5}{128} = \frac{3}{64}$$

und fast alle $v \in V_2$ erfüllen $|s(v) - c_2| \leq 3/64$. Wegen $|c_1 - c_2| \geq 16/64$ klassifiziert Algorithmus 3 maximal $O(n \cdot C_1/\overline{w}')$ der Knoten $v \in U$ falsch. Das sind dann

$$|U| - O(n \cdot C_1/\overline{w}') \overset{(2.5)}{\geq} n - \exp(-\Omega(\overline{w})) \cdot n - O(n \cdot C_1/\overline{w}') = n \cdot (1 - O(C_1/\overline{w}'))$$

viele. Damit ist der Beweis von Theorem 4 vollständig.

2.4.2 Beweis von Theorem 2

Wie in (2.3) bereits beschrieben, hat jeder Eintrag von $M_{V_i \times V_j}$ den gleichen Erwartungswert. Außerdem hat jeder Eintrag nur zwei mögliche Werte: 0 und eine positive Zahl. Der jeweils mögliche positive Wert variiert von Eintrag zu Eintrag, da wir die Einträge der Adjazenzmatrix mit $\overline{w}'^2/(w'_u \cdot w'_v)$ multiplizieren. Trotzdem können wir uns sicher sein, dass kein Eintrag in M größer als $\overline{w}'^2/(\min_{u \in V} w'_u)^2$ ist. Wir fassen alle diese wichtigen Eigenschaften in folgender Definition zusammen.

Definition 8. Wir nennen eine reelle $n \times m$-Matrix $X = (x_{uv})$ eine *same-mean*-Matrix mit Mittelwert μ und Schranke b, wenn X folgende Eigenschaften hat:

1. Alle x_{uv} sind unabhängige Zufallsvariablen. Falls X symmetrisch ist, sind die trivialen Abhängigkeiten erlaubt.
2. Jedes x_{uv} hat genau zwei mögliche Werte, wovon einer 0 und der andere $\leq b$ ist.
3. $\mathbf{E}[x_{uv}] = \mu > 0$ für alle u, v.

Wir beachten, dass die Ungleichung $x_{uv} \leq b$ stets erfüllt sein muss, unabhängig vom konkreten Ergebnis des Zufallsprozesses.

Wir können leicht nachrechnen, dass $M_{V_i \times V_j}$ eine same-mean-Matrix mit

$$\begin{aligned} \text{Mittelwert} \quad & \mu \overset{(2.3)}{=} d_{ij} \cdot \frac{W_i \cdot W_j}{W} \cdot \frac{\overline{w}'}{n} \overset{(2.1)}{=} \Theta(\overline{w}'/n) \\ \text{und Schranke} \quad & b = \frac{\overline{w}'^2}{(\min_u w'_u)^2} \leq \frac{1}{\varepsilon^2} = \Theta(1) \end{aligned} \tag{2.11}$$

ist. Leider hat ${M^*}_{V_i \times V_j}$ nicht Eigenschaft 1. aus Definition 8. Da wir Knoten mit zu hoher Zeilensumme gelöscht haben, sind die Einträge von ${M^*}_{V_i \times V_j}$ nicht mehr unabhängig. Deshalb konzentrieren wir uns zunächst auf $M_{V_i \times V_j}$ und übertragen dann die Resultate auf ${M^*}_{V_i \times V_j}$.

Das folgende Lemma ist für die Analyse von same-mean-Matrizen äußerst hilfreich. Es ist eine Verallgemeinerung von Lemma 3.4 in [AK97]. Das Lemma in [AK97] beschränkt sich – in unserer Notation gesprochen – auf same-mean-Matrizen mit $x_{uv} \in \{0, 1\}$. Wir beweisen Lemma 9 in Abschnitt 2.4.3

Lemma 9. *Sei X eine same-mean-Matrix mit Mittelwert μ und Schranke b. Sei $\{y_1, \ldots, y_l\}$ eine Menge unabhängiger Einträge aus X und $a_1, \ldots, a_l$ beliebige reelle Zahlen aus dem Intervall $[-a : a]$. Falls S, D und eine Konstante $c > 0$*

$$\sum_{i=1}^{l} a_i^2 \leq D \qquad \textit{und} \qquad S \leq c \cdot \mathrm{e}^c \cdot D \cdot \mu / a,$$

erfüllen, dann gilt für die neue Zufallsvariable $Z = \sum_{i=1}^{l} a_i \cdot y_i$

$$\mathbf{Pr}\left[|Z - \mathbf{E}[Z]| \geq S\right] \leq 2 \cdot \exp\bigl(-S^2/(2\mu \cdot \mathrm{e}^c \cdot D \cdot b)\bigr)\,.$$

Lemma 9 hat eine ähnliche Stärke wie „klassische“ Chernoff-Schranken. Nur wird die *gewichtete* Summe unabhängiger Zufallsvariablen mit gleichem Erwartungswert betrachtet. Der Beweis des Lemmas ist in Abschnitt 2.4.3 zu finden und ähnelt dem der klassischen Chernoff-Schranke.

Punkt 1. von Theorem 2

Der Zähler $\mathbf{1}^t \cdot {M^*}_{V_i \times V_j} \cdot \mathbf{1} = s_M(V_i \cap U, V_j \cap U)$ ergibt sich, indem wir von der Gesamtsumme $s_M(V_i, V_j)$ in $M_{V_i \times V_j}$ die Summe der Einträge abziehen, die wir in Schritt 4. von Algorithmus 3 löschen.

Die Summe aller Einträge in $M_{V_i \times V_j}$ kann mithilfe von Lemma 9 abgeschätzt werden. Im Falle von $i \neq j$ sind alle Einträge in $M_{V_i \times V_j}$ unabhängig. Wir wählen alle a_i's in Lemma 9 zu 1, $D = |V_i| \cdot |V_j|$, $S = \mu \cdot D/\sqrt{\overline{w}'}$ sowie $c = \ln 2$ und erhalten

$$\begin{aligned}
&\mathbf{Pr}\left[\left|s_M(V_i, V_j) - \mu \cdot |V_i| \cdot |V_j|\right| \geq \mu \cdot |V_i| \cdot |V_j| / \sqrt{\overline{w}'}\right] \\
&\qquad \leq \quad 2\exp\left(-\frac{\mu \cdot D}{4 \cdot \overline{w}' \cdot b}\right) = 2\exp\left(-\frac{\Omega(\overline{w}'/n) \cdot \Omega(n^2)}{4 \cdot \overline{w}' \cdot \Theta(1)}\right) \\
&\qquad = \quad \exp(-\Omega(n))\,.
\end{aligned}$$

Aufgrund der Symmetrie sind im Fall $i = j$ nicht alle Einträge unabhängig, weswegen wir Lemma 9 nicht direkt anwenden können. Wir umgehen dieses Problem, indem wir uns auf das obere Dreieck der Matrix beschränken. Bis auf die Konstante hinter dem Ω gilt obige Ungleichung aber genauso für $i = j$.

Demnach gilt mit Wahrscheinlichkeit $1 - \exp(-\Omega(n))$

$$s_M(V_i, V_j) = \mu \cdot |V_i| \cdot |V_j| \cdot (1 \pm O(1/\sqrt{\overline{w}'})). \tag{2.12}$$

Als nächstes beschränken wir die Summe der Einträge, die wir während Schritt 4. löschen. Wir beginnen damit, die *Anzahl* der gelöschten Einträge zu beschränken. Die Idee ist folgende: Wir zeigen, dass $|V \setminus U|$ klein ist und anschließend, dass jede derart kleine Menge nur wenige Nachbarn hat. Da jeder positive Eintrag in M von einer Kante impliziert wird, können deshalb nur wenige positive Einträge gelöscht werden, wobei jeder Eintrag durch $b = \Theta(1)$ beschränkt ist.

Sei also u ein Eintrag, der in Schritt 4. gelöscht wird. Die Zeilensumme $s_M(u, V)$ von u ist demnach $> C_1 \cdot \overline{w}'$. Wie wir uns bereits in Abschnitt 2.3 überlegt haben, bedeutet dies, dass die Zeilensumme mindestens fünfmal so groß wie ihr Erwartungswert ist. Wieder hilft uns Lemma 9, die Wahrscheinlichkeit dafür abzuschätzen.

Wenn $s_M(u, V) \geq 5 \cdot \mathbf{E}[s_M(u, V)]$ dann ist $s_M(u, V_1) \geq 5 \cdot \mathbf{E}[s_M(u, V_1)]$ oder $s_M(u, V_2) \geq 5 \cdot \mathbf{E}[s_M(u, V_2)]$. Beachte, dass wir $s_M(u, V_1)$ und $s_M(u, V_2)$ getrennt voneinander abschätzen müssen, da die Erwartungswerte der Einträge differieren. Wieder setzen wir alle a_i's in Lemma 9 auf 1, $D = |V_i|$, $S = 4 \cdot \mu \cdot |V_i|$ und $c = \ln 4$. Wir erhalten für ein festes u

$$\mathbf{Pr}\left[\, |s_M(u, V_i) - \mathbf{E}[s_M(u, V_i)]| \geq 4 \cdot \mathbf{E}[s_M(u, V_i)] \,\right] \leq 2 \cdot \exp(-2 \cdot \mu \cdot |V_i|/b)\ .$$

Dabei ist $\mu \cdot |V_i|$ die erwartete Zeilensumme in $M_{V_i \times V_j}$. Eine untere Schranke an diesen Wert bildet $c_{\mathrm{m}} \cdot \overline{w}'$ mit

$$c_{\mathrm{m}} = \delta \cdot \min_{\substack{i,j \\ d_{ij} > 0}} \left\{ d_{ij} \cdot \frac{W_i \cdot W_j}{W} \right\} = \Theta(1)\,, \tag{2.13}$$

(vgl. (2.3)). Damit erhalten wir

$$\mathbf{Pr}\left[\, |s_M(u, V_i) - \mathbf{E}[s_M(u, V_i)]| \geq 4 \cdot \mathbf{E}[s_M(u, V_i)] \,\right] \leq 2 \cdot \exp(-c_{\mathrm{m}} \cdot \overline{w}'/b)\ , \tag{2.14}$$

Beachte, dass wegen $b = \Theta(1)$ auch $c_{\mathrm{m}}/b = \Theta(1)$ ist.

Für $u \in V$ sei X_u die Indikatorvariable für das Ereignis $s_M(u,V) \geq 5 \cdot \mathbf{E}[s_M(u,V)]$ und sei $X = \sum_u X_u$. Dann ist $X = |V \setminus U|$ und $\mathbf{E}[|V \setminus U|] = \mathbf{E}[X] = \sum_u \mathbf{E}[X_u]$. Wegen (2.14) ist

$$\mathbf{E}[X_u] = \mathbf{Pr}[X_u = 1] \leq 4 \cdot \exp(-c_{\mathrm{m}} \cdot \overline{w}'/b)$$

und $\mathbf{E}[X] \leq 4 \cdot \exp(-c_{\mathrm{m}} \cdot \overline{w}'/b) \cdot n$. Wir zeigen, dass es ein positives $c < c_{\mathrm{m}}/(2b)$ gibt, so dass $\mathbf{Pr}[X \geq 2 \cdot \exp(-c \cdot \overline{w}') \cdot n] = O(1/n)$ und benutzen dafür die Tschebyscheff-Ungleichung. Dazu müssen wir die Varianz $\mathbf{Var}[X]$ von X abschätzen.

Da die einzelnen X_u fast unabhängig voneinander sind, ergibt sich eine kleine Varianz. Die einzige Abhängigkeit zwischen X_u und X_v besteht im Eintrag m_{uv}. Wir haben für $u \neq v$

$$\begin{aligned}
\mathbf{E}[X_u X_v] &= \mathbf{Pr}[X_u = 1 \text{ und } X_v = 1] \\
&= \mathbf{Pr}[X_u = 1 \text{ und } X_v = 1 \,|\, m_{uv} = 0] \cdot \mathbf{Pr}[m_{uv} = 0] \\
&\quad + \mathbf{Pr}[X_u = 1 \text{ und } X_v = 1 \,|\, m_{uv} > 0] \cdot \mathbf{Pr}[m_{uv} > 0] \,.
\end{aligned}$$

Aus der Definition von X_u ergibt sich sofort

$$\begin{aligned}
\mathbf{Pr}[X_u = 1 \text{ und } X_v = 1 \,|\, m_{uv} = 0] &\leq \mathbf{Pr}[X_u = 1] \cdot \mathbf{Pr}[X_v = 1] \\
&= \mathbf{E}[X_u] \cdot \mathbf{E}[X_v] \,.
\end{aligned}$$

Ähnlich zu (2.14) können wir zeigen

$$\mathbf{Pr}[X_u = 1 \text{ und } X_v = 1 \,|\, m_{uv} > 0] \leq \exp(-3c \cdot \overline{w}')$$

für eine ausreichend kleine Konstante $c > 0$. Insgesamt gilt

$$\mathbf{E}[X_u X_v] \leq \mathbf{E}[X_u] \cdot \mathbf{E}[X_v] + \exp(-3c \cdot \overline{w}') \cdot \mathbf{Pr}[m_{uv} > 0] \,.$$

Um die Varianz $\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}^2[X]$ abzuschätzen, benötigen wir

$$\begin{aligned}
\mathbf{E}\left[X^2\right] &= \mathbf{E}\left[\left(\sum\nolimits_u X_u\right)^2\right] = \mathbf{E}\left[\sum_{u,v} X_u X_v\right] = \mathbf{E}\left[\sum_u X_u^2 + \sum_{\substack{u,v \\ u \neq v}} X_u X_v\right] \\
&= \sum_u \mathbf{E}\left[X_u^2\right] + \sum_{\substack{u,v \\ u \neq v}} \mathbf{E}[X_u X_v] = \sum_u \mathbf{E}[X] + \sum_{\substack{u,v \\ u \neq v}} \mathbf{E}[X_u X_v] \\
&\leq \mathbf{E}[X] + \sum_{\substack{u,v \\ u \neq v}} (\mathbf{E}[X_u] \cdot \mathbf{E}[X_v] + \exp(-3c \cdot \overline{w}') \cdot \mathbf{Pr}[m_{uv} > 0]) \\
&\leq \mathbf{E}[X] + \mathbf{E}^2[X] + \exp(-3c \cdot \overline{w}') \cdot \overline{w}' \cdot n \,.
\end{aligned}$$

So erhalten wir

$$\mathbf{Var}[X] \leq \mathbf{E}[X] + \exp(-2c \cdot \overline{w}') \cdot n \leq 2 \cdot \exp(-2c \cdot \overline{w}') \cdot n$$

für $2c < c_{\mathrm{m}}/b$. Die Tschebyscheff-Ungleichung liefert

$$\mathbf{Pr}\left[|X - \mathbf{E}[X]| > \exp(-c \cdot \overline{w}') \cdot n\right] \leq \frac{\mathbf{Var}[X]}{(\exp(-c \cdot \overline{w}') \cdot n)^2} \leq \frac{2}{n}\,.$$

Also gilt mit Wahrscheinlichkeit $1 - O(1/n)$

$$|V \setminus U| \leq 2 \cdot \exp(-c \cdot \overline{w}') \cdot n\,. \tag{2.15}$$

(Damit ist auch Ungleichung (2.5) auf Seite 21 bewiesen.)

Als nächstes wollen wir die *Summe* der aus $M_{V_i \times V_j}$ gelöschten Einträge beschränken, also

$$s_M(V_i \setminus U, V_j \setminus U) + s_M(V_i \cap U, V_j \setminus U) + s_M(V_i \setminus U, V_j \cap U)\,.$$

Da die Einträge von M nicht-negativ sind, genügt es, eine obere Schranke an $s_M(V_i \setminus U, V_j) + s_M(V_i, V_j \setminus U)$ anzugeben. Wir zeigen die Schranke an den ersten Summanden detailliert, der zweite folgt analog.

Da wir U's Größe nicht genau kennen, betrachten wir nun alle Mengen $X \subset V_i$ mit $|X| = 2{\cdot}\exp(-c \cdot \overline{w}'){\cdot}n$. (Wir verzichten zugunsten der Lesbarkeit auf das Runden.) Natürlich ist $V_i \setminus U$ Teilmenge von mindestens einem solchen X, da $|V_i \setminus U| \leq |V \setminus U| \leq 2{\cdot}\exp(-c \cdot \overline{w}'){\cdot}n$. Es genügt also, $s_M(X, V_j)$ nach oben zu beschränken.

Für ein festes X haben wir

$$\mathbf{E}[s_M(X, V_j)] = \mu \cdot |X| \cdot |V_j| = O\left(\overline{w}' \cdot \exp(-c \cdot \overline{w}') \cdot n\right) \leq \frac{n}{2}\,.$$

für $\overline{w}'$ groß genug.

Im Fall $i = j$ zählen wir in $s_M(X, V_j)$ aus Symmetriegründen einige Einträge doppelt, was bei der Anwendung von Lemma 9 stört. Deshalb verzichten wir jeweils auf den zweiten Eintrag und setzen stattdessen das a_i, das zum ersten Eintrag gehört auf 2, während alle anderen $a_i = 1$ sind. Weiterhin sei $D = 16 \cdot \exp(-c/3 \cdot \overline{w}') \cdot n^2$ und $S = 8 \cdot \exp(-c/3 \cdot \overline{w}') \cdot \mu \cdot n^2$. Die Konstante c *in Lemma 9* setzen wir auf 0.6. So erhalten wir

$$\begin{aligned}
\mathbf{Pr}\left[\,|s_M(X, V_j) - \mathbf{E}[s_M(X, V_j)]| \geq 8 \cdot \exp(-c/3 \cdot \overline{w}') \cdot \mu \cdot n^2\right] & \\
&\leq\ 2 \cdot \exp\left(-\frac{4 \cdot \exp(-c/3 \cdot \overline{w}') \cdot \mu \cdot n^2}{2 \cdot \mathrm{e}^{0.6} \cdot b}\right) \\
&\leq\ 2 \cdot \exp(-2 \cdot \exp(-c/4 \cdot \overline{w}') \cdot \overline{w}' \cdot n)\,.
\end{aligned}$$

Wegen $\binom{n}{k} \leq (\mathrm{e} \cdot n/k)^k$ (vgl. Anhang A, Punkt A.1.) gibt es nur

$$\binom{|V_i|}{|X|} \leq \binom{n}{|X|} \leq \left(\frac{\mathrm{e} \cdot n}{|X|}\right)^{|X|} \leq \exp((1 + c \cdot \overline{w}') \cdot |X|) \leq \exp(\overline{w}' \cdot |X|)$$

verschiedene Mengen $X \subseteq V_i$ der Größe $2 \cdot \exp(-c \cdot \overline{w}') \cdot n$. Der letzte Schritt folgt, da wir von $c < 1$ ausgehen können. Wegen $|X| \leq \exp(-c/4 \cdot \overline{w}') \cdot n$ ist die Gesamtwahrscheinlichkeit für die Existenz *eines* X mit „vielen“ Nachbarn in V_j beschränkt durch

$$2 \cdot \exp(-\overline{w}' \cdot \exp(-c/4 \cdot \overline{w}') \cdot n) = O(1/n)\,.$$

Für die letzte Abschätzung gehen wir davon aus, dass $\overline{w}' \leq 2\ln(n)/c$ gilt. Diese Annahme ist legitim, da ansonsten bereits (2.15) $V \setminus U = \emptyset$ zeigt und sich die weiteren Berechnungen erübrigen.

Wir wissen nun, dass mhW. für alle Mengen X wie oben beschrieben

$$s_M(X, V_j) \leq \mathbf{E}\left[s_M(X, V_j)\right] + 8 \cdot \exp(-c/3 \cdot \overline{w}') \cdot \mu \cdot n^2 \leq \frac{n}{2} + \frac{n}{2} = n$$

gilt, vorausgesetzt $\overline{w}'$ ist groß genug. Auf analogem Weg können wir die gleiche Schranke an $s_M(V_i, Y)$ für alle $Y \subset V_j$ mit $|Y| = 2 \cdot \exp(-c/3 \cdot \overline{w}') \cdot n$ erhalten.

Somit ist mit Wahrscheinlichkeit $1 - O(1/n)$ die Summe aller in Schritt 4. aus $M_{V_i \times V_j}$ gelöschten Einträge durch

$$s_M(V_i \setminus U, V_j) + s_M(V_i, V_j \setminus U) \leq 2 \cdot n$$

beschränkt. Wir erhalten mit (2.12), dass der Term $\mathbf{1}^t \cdot {M^*}_{V_i \times V_j} \cdot \mathbf{1} = s_M(V_i \cap U, V_j \cap U)$ nach unten durch

$$\mu \cdot |V_i| \cdot |V_j| \cdot (1 \pm O(1/\sqrt{\overline{w}'})) - 2n = \mu \cdot |V_i| \cdot |V_j| \cdot (1 \pm O(1/\sqrt{\overline{w}'}))$$

beschränkt ist, da $\mu = \Theta(\overline{w}'/n)$ und $|V_i|, |V_j| \geq \delta n = \Omega(n)$ ist. Letztendlich erhalten wir

$$\begin{aligned}
\frac{\mathbf{1}^t}{\|\mathbf{1}\|} \cdot {M^*}_{V_i \times V_j} \cdot \frac{\mathbf{1}}{\|\mathbf{1}\|} &= \frac{\mathbf{1}^t \cdot {M^*}_{V_i \times V_j} \cdot \mathbf{1}}{\sqrt{|V_i \cap U| \cdot |V_j \cap U|}} \\
&= \frac{\mu \cdot |V_i| \cdot |V_j| \cdot (1 \pm O(1/\sqrt{\overline{w}'}))}{\sqrt{|V_i| \cdot |V_j|} \cdot (1 \pm O(\exp(-c \cdot \overline{w}')))} \\
&= \mu \cdot \sqrt{|V_i| \cdot |V_j|} \cdot (1 \pm O(1/\sqrt{\overline{w}'}))\,.
\end{aligned}$$

Punkt 1 von Theorem 2 folgt unmittelbar mit $\mu = d_{ij} \cdot W_i \cdot W_j / W \cdot \overline{w}'/n$.

Punkt 2. von Theorem 2

Um Punkt 2. zu beweisen, benutzen wir eine Technik, die auf [FKS89] zurückgeht. Dort wurden die Eigenwerte, insbesondere das „spectral gap“ von Adjazenzmatrizen zufälliger regulärer Graphen analysiert. Die Analyse wurde in [AK97] auf das $G_{n,p}$-Modell übertragen. Wir werden sehen, dass das „spectral gap“ unserer Matrix M^* auf ähnlichem Weg wie in [AK97] gezeigt werden kann. Entscheidend dafür ist das folgende Lemma, dessen Beweis in Abschnitt 2.4.4 folgt.

Lemma 10. *Sei X eine $n \times m$–same-mean-Matrix mit Mittelwert μ und Schranke b. Sei $N = n + m$ und $R = \{u : \sum_v x_{uv} \leq d \cdot \mu \cdot N\}$, sowie $C = \{v : \sum_u x_{uv} \leq d \cdot \mu \cdot N\}$ für ein beliebiges $d > 1$.*

Falls $\mu \cdot n \cdot m > b \cdot N$, gilt mit Wahrscheinlichkeit $1 - O(1/N^4)$ für alle *Paare von Vektoren u,v mit $\left\|u_{|R}\right\| = \left\|v_{|C}\right\| = 1$, sowie $u_{|R} \perp \mathbf{1}$ oder $v_{|C} \perp \mathbf{1}$*

$$\left|{u_{|R}}^t \cdot X \cdot v_{|C}\right| = O(\sqrt{b \cdot d \cdot \mu \cdot N}).$$

Wegen ${u_{|R}}^t \cdot X \cdot v_{|C} = u_R^t \cdot X_{R\times C} \cdot v_C =: u' \cdot X' \cdot v'$ (vgl. Abschnitt 2.2) sagt Lemma 10 Folgendes aus: Betrachtet man die Untermatrix X' von X, die nur die Zeilen und Spalten enthält, deren Eintragssumme nicht sehr stark über deren Erwartungswert liegt, so ist typischerweise $|u' \cdot X' \cdot v'|$ klein, solange u' oder v' senkrecht auf $\mathbf{1}$ stehen.

Sei nun u ein $|V_i|$-dimensionaler Vektor und v ein $|V_j|$-dimensionaler Vektor. Dann gilt

$$u_U^t \cdot {M^*}_{V_i \times V_j} \cdot v_U = {u^t}_{|U} \cdot M_{V_i \times V_j} \cdot v_{|U}\,.$$

Um Lemma 10 benutzen zu können, brauchen wir $U \cap V_i \subseteq R$ und $U \cap V_j \subseteq C$.

$M_{V_i \times V_j}$ ist eine same-mean-Matrix mit Schranke $\overline{w}'^2/(\min_u w'_u)^2 = O(1)$, siehe (2.11). U enthält nur die Knoten, deren Zeilensumme (und wegen Symmetrie auch deren Spaltensumme) in M höchstens $C_1 \cdot \overline{w}'$ ist. Wir setzen $d := C_1/(2 \cdot c_{\mathrm{m}})$, mit c_{m} wie in (2.13). Dann gilt für alle $u \in U \cap V_i$

$$\begin{aligned} d \cdot \mu \cdot N &= \frac{C_1}{2 \cdot c_{\mathrm{m}}} \cdot \mu \cdot N = \frac{C_1}{2 \cdot c_{\mathrm{m}}} \cdot \mu \cdot (|V_i| + |V_j|) \geq \frac{C_1}{2 \cdot c_{\mathrm{m}}} \cdot \mu \cdot (2\delta n) \\ &= \frac{C_1}{c_{\mathrm{m}}} \cdot d_{ij} \cdot \frac{W_i \cdot W_j}{W} \cdot \frac{\overline{w}'}{n} \cdot \delta n \geq C_1 \cdot \overline{w}' \geq \sum_{v \in V} x_{uv} \geq \sum_{v \in V_j} x_{uv}\,, \end{aligned}$$

also $u \in R$. Damit ist $U \cap V_i \subseteq R$. Analog erhalten wir $U \cap V_j \subseteq C$.

Mit $\mu \cdot |V_i| \cdot |V_j| = \Theta(\overline{w}' \cdot n)$ und $b \cdot N = b \cdot (|V_i| + |V_j|) = O(n)$ sind (für $\overline{w}'$ ausreichend groß) die Bedingungen von Lemma 10 erfüllt.

Punkt 2. von Theorem 2 ist jetzt leicht zu zeigen. Jeder Vektor u (und jeder Vektor v) mit den Eigenschaften wie im Theorem kann zu einem $|V_i|$-dimensionalen Vektor u' (einem $|V_j|$-dimensionalen Vektor v') erweitert werden, indem er mit 0-en aufgefüllt wird. Natürlich gilt $\|u\| = \|u'\| = 1$ und $\|v\| = \|v'\| = 1$. Wenn $u \perp \mathbf{1}$ (hier ist $\mathbf{1}$ $|V_i \cap U|$-dimensional) gilt, dann gilt wegen $U \cap V_i \subseteq R$ auch $u'_{|R} \perp \mathbf{1}$. Analoges gilt für v und v', falls $v \perp \mathbf{1}$. Mit Lemma 10 erhalten wir

$$\begin{aligned}
\left|u^t \cdot {M^*}_{V_i \times V_j} \cdot v\right| &= \left|u'_{|U} \cdot M_{V_i \times V_j} \cdot v'_{|U}\right| = \left|u'_{|R} \cdot M_{V_i \times V_j} \cdot v'_{|C}\right| \\
&= O\left(\sqrt{b \cdot d \cdot \mu \cdot N}\right) \\
&= O\left(\sqrt{b \cdot C_1/(2 \cdot c_{\mathrm{m}}) \cdot \mu \cdot (|V_i| + |V_j|)}\right) \\
&\overset{(2.11)}{\overset{(2.13)}{=}} O\left(\sqrt{C_1 \cdot \overline{w}'}\right) .
\end{aligned}$$

□

2.4.3 Beweis von Lemma 9

Den Fall $b \neq 1$ können wir auf den Spezialfall $b = 1$ reduzieren, den wir ähnlich zu [AK97] beweisen. Wir beginnen mit $b \neq 1$:

Sei $X = (x_{uv})$ die Matrix aus Lemma 9. Wir konstruieren $X' = (x'_{uv})$, indem wir $x'_{uv} := x_{uv}/b$ setzen. Dann sind alle x'_{uv} unabhängig und durch 1 beschränkt. Der Erwartungswert $\mathbf{E}[x'_{uv}]$ ist $\mathbf{E}[x_{uv}]/b = \mu/b$. Also ist X' eine same-mean-Matrix mit Mittelwert μ/b und Schranke 1.

Seien nun $y_1, \ldots, y_l$ wie in der Behauptung und $y'_1, \ldots, y'_l$ die entsprechenden Einträge in X'. Dann ist $Z' = \sum_{i=1}^{l} a_i \cdot y'_i = Z/b$ und wegen $S \leq c \cdot \mathrm{e}^c \cdot D \cdot \mu/a$ ist

$$S' = \frac{S}{b} \leq \frac{c \cdot \mathrm{e}^c \cdot D \cdot \mu/a}{b} = c \cdot \mathrm{e}^c \cdot D \cdot \mu'/a\,.$$

Wir wenden Lemma 9 mit $b = 1$ auf X' an:

$$\begin{aligned}
\mathbf{Pr}\left[|Z - \mathbf{E}[Z]| \geq S\right] &= \mathbf{Pr}\left[\left|\frac{Z}{b} - \frac{\mathbf{E}[Z]}{b}\right| \geq \frac{S}{b}\right] = \mathbf{Pr}\left[|Z' - \mathbf{E}[Z']| \geq S'\right] \\
&\leq 2 \cdot \exp\left(-S'^2/(2\mu' \cdot \mathrm{e}^c \cdot D)\right) \\
&= 2 \cdot \exp\left(-\left(S/b\right)^2 \;/\; (2 \cdot (\mu/b) \cdot \mathrm{e}^c \cdot D)\right) \\
&= 2 \cdot \exp\left(-S^2/(2 \cdot \mu \cdot \mathrm{e}^c \cdot D \cdot b)\right) .
\end{aligned}$$

Sei nun $b = 1$. In den nachfolgenden Produkten und Summen verzichten wir zugunsten der Lesbarkeit auf die Angabe der Indizes. Der Index i läuft stets von 1 bis l.

Seien $p_1, \ldots, p_l$ die Wahrscheinlichkeiten, dass $y_1, \ldots, y_l$ nicht 0 sind. Das heißt $p_i \cdot y_i = \mu$ falls $y_i \neq 0$ ist. Mit anderen Worten, ist der zweite Wert (außer 0), den y_i annehmen kann, $\mu/p_i \leq b = 1$. Wir erhalten mit Markovs Ungleichung

$$\begin{aligned} \mathbf{Pr}\left[Z - \mathbf{E}[Z] \geq S\right] &= \mathbf{Pr}\left[e^{\lambda(Z-\mathbf{E}[Z])} \geq \mathrm{e}^{\lambda S}\right] = \mathbf{Pr}\left[\mathrm{e}^{\lambda(Z-\mathbf{E}[Z]-S)} \geq 1\right] \\ &\leq \mathbf{E}\left[\mathrm{e}^{\lambda(Z-\mathbf{E}[Z]-S)}\right] = \frac{\mathbf{E}\left[\exp(\lambda Z)\right]}{\exp(\lambda(\mathbf{E}[Z]+S))}\,. \end{aligned} \tag{2.16}$$

Setzen wir $\lambda = S/(\mathrm{e}^c \cdot \mu \cdot D) \leq c/a$, gilt wegen $y_i \leq b = 1$

$$\lambda \cdot a_i \cdot y_i \leq \lambda \cdot a_i \leq \lambda \cdot a \leq c\,.$$

Wir untersuchen zuerst den Nenner von (2.16). Wegen der Unabhängigkeit der y_i gilt

$$\begin{aligned} \mathbf{E}\left[\exp(\lambda Z)\right] &= \mathbf{E}\left[\exp\Big(\lambda \sum a_i \cdot y_i\Big)\right] = \mathbf{E}\left[\prod \exp(\lambda a_i \cdot y_i)\right] \\ &= \prod \mathbf{E}\left[\exp(\lambda a_i \cdot y_i)\right] \\ &= \prod \left(p_i \cdot \exp\left(\lambda \cdot a_i \cdot \frac{\mu}{p_i}\right) + (1-p_i) \cdot \exp(\lambda \cdot a_i \cdot 0)\right) \\ &= \prod \left(1 + p_i \cdot \left(\exp\left(\lambda \cdot a_i \cdot \frac{\mu}{p_i}\right) - 1\right)\right)\,. \end{aligned}$$

Wir schätzen $\exp(\lambda \cdot a_i \cdot \mu/p_i) - 1$ mithilfe von $\mathrm{e}^x - 1 \leq \mathrm{e}^c \cdot x^2/2 + x$ ab (vgl. A.3). Diese Ungleichung gilt für alle $x \leq c$, was für $x := \lambda \cdot a_i \cdot \mu/p_i \leq \lambda \cdot a \cdot 1 \leq c$ zutrifft.

Wir erhalten also

$$\begin{aligned} \mathbf{E}\left[\exp(\lambda Z)\right] &\leq \prod \left(1 + p_i \left(\lambda \cdot a_i \cdot \frac{\mu}{p_i} + \frac{\mathrm{e}^c}{2} \cdot \left(\lambda \cdot a_i \cdot \frac{\mu}{p_i}\right)^2\right)\right) \\ &\leq \prod \left(1 + \lambda \cdot a_i \cdot \mu + \mathrm{e}^c \cdot \lambda^2 \cdot a_i^2 \cdot \frac{\mu^2}{2 \cdot p_i}\right) \\ &\overset{\mu/p_i \leq 1}{\leq} \prod \left(1 + \lambda \cdot a_i \cdot \mu + \mathrm{e}^c \cdot \lambda^2 \cdot a_i^2 \cdot \mu/2\right)\,. \end{aligned}$$

Wegen $1 + x \leq e^x$ für alle $x \in \mathbb{R}$ (vgl. A.2) gilt

$$\begin{aligned}
\mathbf{E}[\exp(\lambda Z)] &\leq \prod \exp\bigl(\lambda \cdot a_i \cdot \mu + \mathrm{e}^c \cdot \lambda^2 \cdot a_i^2 \cdot \mu/2\bigr) \\
&= \exp\Bigl(\sum \lambda \cdot a_i \cdot \mu + \mathrm{e}^c \cdot \lambda^2 \cdot a_i^2 \cdot \mu/2\Bigr) \\
&= \exp\Bigl(\lambda \cdot \mathbf{E}[Z] + \mathrm{e}^c \cdot \lambda^2 \cdot \mu/2 \cdot \sum a_i^2\Bigr) \\
&\leq \exp\bigl(\lambda \cdot \mathbf{E}[Z] + \mathrm{e}^c \cdot \lambda^2 \cdot \mu \cdot D/2\bigr) \\
&= \exp(\lambda \cdot \mathbf{E}[Z] + \lambda \cdot S/2)\ .
\end{aligned}$$

Wir erhalten für (2.16)

$$\begin{aligned}
\mathbf{Pr}[Z - \mathbf{E}[Z] \geq S] &\leq \frac{\mathbf{E}[\exp(\lambda Z)]}{\exp(\lambda(\mathbf{E}[Z] + S))} \leq \frac{\exp(\lambda \cdot \mathbf{E}[Z] + \lambda \cdot S/2)}{\exp(\lambda(\mathbf{E}[Z] + S))} \\
&\leq \exp(-\lambda \cdot S/2) = \exp\left(\frac{-S^2}{2 \cdot e^c \cdot \mu \cdot D}\right).
\end{aligned}$$

Indem wir sämtliche a_i's negieren, erhalten wir auf dem gleichen Weg $\mathbf{Pr}[\mathbf{E}[Z] - Z \geq S] \leq \exp\left(\frac{-S^2}{2 \cdot \mathrm{e}^c \cdot \mu \cdot D}\right)$. □

2.4.4 Beweis von Lemma 10

Genau wie im Beweis von Lemma 9 reduzieren wir den Fall $b \neq 1$ auf den Fall $b = 1$. Dessen Beweis folgt dann den Ideen von [AK97] (Beweis von Lemma 3.3) und [FKS89] (Beweis von Theorem 2.2).

Wie im Beweis von Lemma 9 schreiben wir X als $X = b \cdot X'$. Dann ist X' eine same-mean-Matrix mit Mittelwert μ/b und Schranke 1. Beachte, dass die Mengen R und C für beiden Matrizen X und X' identisch sind und alle Bedingungen erfüllt bleiben. Wir wenden Lemma 10 mit $b = 1$ auf X' an:

$$\begin{aligned}
\bigl|{u_{|R}}^t \cdot X \cdot v_{|C}\bigr| &= \bigl|{u_{|R}}^t \cdot b \cdot X' \cdot v_{|C}\bigr| = b \cdot \bigl|{u_{|R}}^t \cdot X' \cdot v_{|C}\bigr| \\
&= b \cdot O(\sqrt{d \cdot (\mu/b) \cdot N}) = O(\sqrt{b \cdot d \cdot \mu \cdot N}).
\end{aligned}$$

Es bleibt die Korrektheit des Lemmas für $b = 1$ zu zeigen. Für den Rest des Abschnitts basieren sämtliche O- und Ω-Terme auf N und gelten für alle $N > N_0 = \text{konstant}$.

Das folgende Lemma beschreibt, dass in einer same-mean-Matrix typischerweise die Eintragssummen von *allen* Untermatrizen in der Nähe ihres Erwartungswertes liegen.

Lemma 11. *Sei $X = (x_{uv})$ eine $n \times m$-same-mean-Matrix mit Mittelwert μ, Schranke 1 und $\mu \cdot n \cdot m \geq n + m$. Dann gilt für $N = n + m$ mit Wahrscheinlichkeit $1 - O(1/N^4)$ für alle Paare (A, B) von Mengen $A \subseteq \{1, \ldots, n\}$, $B \subseteq \{1, \ldots, m\}$:*

Wenn $K := \max\{|A|, |B|\} \leq N/10$, dann

1. $s_X(A,B) \leq 20 \cdot \mathbf{E}[s_X(A,B)]$ oder
2. $s_X(A,B) \cdot \ln \frac{s_X(A,B)}{\mathbf{E}[s_X(A,B)]} \leq 20 \cdot K \cdot \ln \frac{N}{K}$.

Beweis. Für $K \leq 4$ gilt Ungleichung 2. mit Sicherheit. Es ist $s_X(A,B) \leq |A| \cdot |B|$ und $\mathbf{E}[s_X(A,B)] = |A| \cdot |B| \cdot \mu$. Wir haben also $s_X(A,B)/\mathbf{E}[s_X(A,B)] \leq 1/\mu$.

Wegen $N \leq \mu \cdot n \cdot m \leq \mu \cdot ((n+m)/2)^2 = \mu \cdot N^2/4$ ist $1/\mu \leq N/4$. Wir erhalten

$$\begin{aligned} s_X(A,B) \cdot \ln \frac{s_X(A,B)}{\mathbf{E}[s_X(A,B)]} &\leq |A| \cdot |B| \cdot \ln(1/\mu) \leq K^2 \cdot \ln(N/4) \\ &\leq 20 \cdot K \cdot \ln(N/K)\,. \end{aligned}$$

Wir können im Weiteren also davon ausgehen, dass $K > 4$ gilt. Wir zeigen das Lemma für symmetrische Matrizen. Dieser Fall ist etwas aufwändiger, da die (triviale) Abhängigkeit der Einträge zu beachten ist.

Wir halten zwei Mengen A und B fest und setzen $\eta := \mathbf{E}[s_X(A,B)] = |A| \cdot |B| \cdot \mu$. Dann gibt es eine eindeutige Zahl β, so dass

$$\beta \cdot \ln \beta = 20 \cdot K \cdot \ln(N/K)/\eta\,.$$

Ungleichung 2. im Lemma entspricht dann $s_X(A,B) \leq \beta \cdot \eta$. Wir setzen $\beta' := \max\{20, \beta\}$. Es genügt zu zeigen, dass mhW. kein Paar (A,B) mit $s_X(A,B) > \beta' \cdot \eta$ existiert.

Wir wollen Lemma 9 verwenden. Aufgrund der Symmetrie sind aber möglicherweise nicht alle Einträge unabhängig. Für $u, v \in A \cap B$ mit $u \neq v$ sind die Einträge x_{uv} und x_{vu} gleich. In diesem Fall benutzen wir nur x_{uv} mit $u < v$ und setzen das zugehörige a_i in Lemma 9 auf 2, weil x_{uv} in $s_X(A,B)$ zweimal gezählt wird.

Für die anderen Paare $(u,v) \in A \times B$ setzen wir a_i auf 1. Dadurch ist $a = 2$ und der Wert von $\sum a_i^2$ liegt zwischen $|A| \cdot |B|$ und $2 \cdot |A| \cdot |B|$. Wir setzen $D = 2 \cdot |A| \cdot |B|$ und wählen c, so dass $c \cdot \mathrm{e}^c = \beta' - 1$. Da A und B feste Mengen sind, erhalten wir

$$\begin{aligned} \mathbf{Pr}\left[s_X(A,B) \geq \beta' \cdot \eta\right] &= \mathbf{Pr}\left[s_X(A,B) \geq (c \cdot \mathrm{e}^c + 1) \cdot \eta\right] \\ &\leq \mathbf{Pr}\left[|s_X(A,B) - \eta| \geq c \cdot \mathrm{e}^c \cdot \eta\right]\,. \end{aligned}$$

Wegen $\eta = |A| \cdot |B| \cdot \mu = D \cdot \mu/a$ ergibt sich

$$\begin{aligned}
\mathbf{Pr}\left[s_X(A,B) \geq \beta' \cdot \eta\right] &\leq \mathbf{Pr}\left[|s_X(A,B) - \eta| \geq c \cdot \mathrm{e}^c \cdot D \cdot \mu/a\right] \\
&\leq 2 \cdot \exp\left(-\frac{(c \cdot \mathrm{e}^c \cdot D \cdot \mu/a)^2}{2 \cdot \mu \cdot \mathrm{e}^c \cdot D}\right) \\
&\leq 2 \cdot \exp\left(-\frac{c^2 \cdot \mathrm{e}^c \cdot D \cdot \mu}{2 \cdot a^2}\right) \\
&\leq 2 \cdot \exp\left(-\frac{c^2 \cdot \mathrm{e}^c \cdot \eta}{4}\right).
\end{aligned}$$

Wegen $\beta' \geq 20$ ist $c \geq 2.2$ und $c^2 \cdot \mathrm{e}^c/4 \geq 2/13 \cdot \beta' \cdot \ln \beta'$. Wir erhalten

$$\begin{aligned}
\mathbf{Pr}\left[s_X(A,B) \geq \beta' \cdot \eta\right] &\leq 2\mathrm{exp}(-2/13 \cdot \beta' \cdot \ln \beta' \cdot \eta) \\
&\leq 2\mathrm{exp}(-40/13 \cdot K \cdot \ln(N/K)) < 2\left(\frac{K}{N}\right)^{3K}
\end{aligned}$$

wegen $K < N$. Die Anzahl der möglichen Paare (A, B) ist durch

$$\sum_{i=1}^{K} 2 \cdot \binom{N}{K} \cdot \binom{N}{i} \leq 2K \cdot \binom{N}{K}^2 \leq 2K \cdot \left(\frac{\mathrm{e} \cdot N}{K}\right)^{2K}$$

begrenzt. Also ist die Wahrscheinlichkeit für die Existenz *eines* Paares (A, B) mit $\max\{|A|, |B|\} = K \leq N/10$ und $s_X(A,B) \geq \beta' \cdot \eta$ höchstens

$$2\left(\frac{K}{N}\right)^{3K} \cdot 2K \cdot \left(\frac{\mathrm{e} \cdot N}{K}\right)^{2K} = 4K \cdot \left(\mathrm{e}^2 \cdot \frac{K}{N}\right)^K = O\left(\left(\frac{8K}{N}\right)^K\right).$$

Summieren wir $(8K/N)^K$ über alle möglichen Werte für $K > 4$, erhalten wir eine Schranke von

$$\begin{aligned}
\sum_{K=5}^{N/2} \left(\frac{K}{N}\right)^K &\leq \sum_{K=5}^{\log^2 N} \left(\frac{8K}{N}\right)^K + \sum_{K=\log^2 N}^{N/2} \left(\frac{8K}{N}\right)^K \\
&\leq \sum_{K=5}^{\log^2 N} \left(\frac{8\log^2 N}{N}\right)^K + \sum_{K=\log^2 N}^{\infty} \left(\frac{8}{10}\right)^K \\
&\leq \log^2 N \cdot \frac{1}{N^{4.5}} + 0.8^{\log^2 N} \cdot \sum_{K \geq 0}^{\infty} \left(\frac{4}{5}\right)^K \\
&\leq \log^2 N \cdot \frac{1}{N^{4.5}} + N^{-0.2 \cdot \log N} \cdot O(1) \leq 1/N^4.
\end{aligned}$$

Also ist die Wahrscheinlichkeit, dass X Lemma 11 erfüllt $1 - O(1/N^4)$. □

Wir kommen jetzt zum eigentlichen Beweis von Lemma 10. Wir nehmen an, dass Lemma 11 für X gilt, was es mhW. tut. Wir beweisen den Fall $u_{|R} \perp \mathbf{1}$ detailliert. Der Fall $v_{|C} \perp \mathbf{1}$ folgt auf gleichem Weg.

Offensichtlich gibt es überabzählbar viele Einheitsvektoren u und v über $\mathbb{R}$. Daher ist es schwierig, eine Aussage über *alle* Vektorpaare zu treffen. Um diesem Problem auszuweichen, approximieren wir die reellen Vektoren durch Vektoren über einer diskreten Menge. Sei dazu

$$T_n = \left\{ x \in \left(\frac{\mathbb{Z}}{2 \cdot \sqrt{n}} \right)^n : \|x\| \leq 2 \right\} .$$

Wir verringern dabei die Anzahl der Vektoren erheblich. Die Anzahl der Vektoren in T_n ist $\leq k^n$ für eine Konstante k. Das gewünschte Resultat von Lemma 10 zeigen wir für die Vektoren aus T_n, indem wir klassische Methoden der Wahrscheinlichkeitsrechnung benutzen, die exponentiell scharfe Schranken liefern. Da T_n ausreichend dicht ist, werden die reellen Einheitsvektoren „ausreichend gut“ approximiert, so dass wir die Ergebnisse für Lemma 10 übertragen können.

Zunächst wollen wir die Eigenschaften von T_n untersuchen. Wir beginnen mit $|T_n|$. Stellen wir uns ein n-dimensionales Koordinatensystem vor[2], in das wir n-dimensionale Würfel mit Seitenlänge $1/(2 \cdot \sqrt{n})$ achsparallel platzieren. Die Position der Würfelmittelpunkte ist direkt durch die Vektoren aus T_n gegeben. Der Vektor $(0, \ldots, 0)^t$ zum Beispiel induziert einen Würfel, dessen Mittelpunkt im Ursprung des Koordinatensystems liegt.

Auf diese Weise werden die Würfel überlappungsfrei verteilt. Ihre Mittelpunkte liegen wegen der Bedingung $\|x\| \leq 2$ innerhalb einer n-dimensionalen Kugel mit Radius 2. Die Länge der Würfeldiagonalen ist $\sqrt{n \cdot 1/(2 \cdot \sqrt{n})^2} = 1/2$. Also liegen alle Würfel vollständig innerhalb der n-dimensionalen Kugel mit Mittelpunkt im Ursprung und Radius (1 + halbe Diagonalenlänge) = 2.25.

Die Anzahl der Würfel schätzen wir nach oben ab, indem wir das Kugelvolumen durch das Würfelvolumen teilen. Die Kugel besitzt für $n =$ gerade ein Volumen (vgl. [Zwi03] Abschnitt 4.18.1.4) von

$$\frac{\pi^{n/2}}{(n/2)!} \cdot 2.25^n = \frac{\pi^{n/2}}{\sqrt{2\pi \cdot n/2} \cdot \left(\frac{n}{2\mathrm{e}}\right)^{n/2}} \cdot 2.25^n \cdot (1 + o(1)) .$$

Jeder Würfel hat ein Volumen von $1/(2 \cdot \sqrt{n})^n = 1/(4 \cdot n)^{n/2}$, wodurch sich ergibt, dass es maximal

$$\frac{\pi^{n/2} \cdot 4^{n/2} \cdot (2\mathrm{e})^{n/2} \cdot 2.25^n}{\sqrt{\pi \cdot n}} \cdot (1 + o(1)) \leq k^n$$

[2]Für eine realistische Vorstellung genügt $n = 3$.

Würfel innerhalb der Kugel und damit auch Vektoren in T_n gibt. Wir übergehen den Fall $n =$ ungerade, da offensichtlich $|T^{n+1}| \geq |T^n|$ ist.

Als nächstes wollen wir untersuchen, wie gut die reellen Einheitsvektoren durch unser Gitter approximiert werden. Sei dazu

$$T_R = \left\{x_{|R} : x \in T_n\right\} \qquad \text{und} \qquad T_C = \left\{x_{|C} : x \in T_m\right\},$$

wobei R und C wie in Lemma 10 definiert sind. Beachte $T_R \subseteq T_n$ und $T_C \subseteq T_m$.

Lemma 12. *Seien u, v und X wie in Lemma 10 mit $u_{|R} \perp \mathbf{1}$. Dann gilt*

$$\left|{u_{|R}}^t \cdot X \cdot v_{|C}\right| \leq 4 \cdot \max_{\substack{l \in T_R, l \perp \mathbf{1} \\ r \in T_C}} \left|l^t \cdot X \cdot r\right|.$$

Beweis. Wir setzen $u_0 := u_{|R}$ und konstruieren einen Vektor $l_0 \in T_R$ mit $l_0 \perp \mathbf{1}$ und $\|u_0 - l_0\| \leq 1/2$. Wir gehen koordinatenweise von 1 bis $|R|$ vor. Um $l_0(1)$ zu erhalten, runden wir die erste Koordinate von u_0 auf die Zahl aus $\mathbb{Z}/(2 \cdot \sqrt{n})$, die ihr am nächsten ist[3]. Für $i > 1$ setzen wir $l_0(i) := u_0(i)$, falls $u_0(i) \in \mathbb{Z}/(2n)$ ist. Anderenfalls runden wir $u_0(i)$ ab, wenn $\sum_{j=1}^{i-1} u_0(j) - l_0(j) < 0$ ist. Ist vorherige Summe ≥ 0, runden wir auf. In jeder Koordinate unterscheiden sich l_0 und u_0 um weniger als $1/(2 \cdot \sqrt{n})$, also ist

$$\|u_0 - l_0\|^2 < \sum_{i=1}^{n} \left(\frac{1}{2 \cdot \sqrt{n}}\right)^2 = \frac{1}{4}.$$

Wegen $\|u_0\| \leq 1$ folgt mit der Dreiecksungleichung, dass $\|l_0\| < 1.5$ ist. l_0 gehört also zu T_R. Aufgrund der Konstruktion ist für alle $i = 1, \ldots, n$

$$\left|\sum_{j=1}^{i} u_0(j) - l_0(j)\right| < \frac{1}{2 \cdot \sqrt{n}}.$$

Mit $u_0 \perp \mathbf{1}$ folgt dann $\left|\sum_{j=1}^{n} l_0(j)\right| < 1/(2 \cdot \sqrt{n})$. Da alle Einträge von l_0 ganzzahlige Vielfache von $1/(2 \cdot \sqrt{n})$ sind, folgt $\sum_{j=1}^{n} l_0(j) = 0$, also $l_0 \perp \mathbf{1}$.

Sei nun $u_1 = u_0 - l_0$, dann ist $\|u_1\| \leq 1/2$ und $u_1^t \cdot \mathbf{1} = (u_0^t - l_0^t) \cdot \mathbf{1} = 0$. Wir können obiges Verfahren für $2 \cdot u_1$ wiederholen und $l_1 \in T_R$ finden, so dass $\|2 \cdot u_1 - l_1\| < 1/2$ und $l_1 \perp \mathbf{1}$. Der Vektor $u_2 = u_1 - l_1/2$ erfüllt dann $u_2 \perp \mathbf{1}$ und $\|u_2\| \leq 1/4$. Durch Iteration erhalten wir eine Darstellung von

[3] Gibt es zwei solcher Zahlen, ist die Wahl beliebig.

$u_{|R}$ mithilfe der unendlichen Folge der l_i. Analog gehen wir für $v_{|C}$ vor und finden passende r_i. Damit erhalten wir

$$u_{|R} = \sum_{i\geq 0} \frac{l_i}{2^i} \qquad \text{und} \qquad v_{|C} = \sum_{i\geq 0} \frac{r_i}{2^i}\,.$$

Es gilt

$$\begin{aligned}
\left|{u_{|R}}^t \cdot X \cdot v_{|C}\right| &= \left|\left(\sum_{i\geq 0} \frac{l_i}{2^i}\right)^t \cdot X \cdot \left(\sum_{i\geq 0} \frac{r_i}{2^i}\right)\right| \\
&= \left|\sum_{i\geq 0}\sum_{j\geq 0} \frac{l_i^t \cdot X \cdot r_j}{2^i \cdot 2^j}\right| \\
&\leq \max_{\substack{l\in T_R, l\perp \mathbf{1} \\ r\in T_C}} \left|l^t \cdot X \cdot r\right| \cdot \sum_{i\geq 0} \frac{1}{2^i} \cdot \sum_{j\geq 0} \frac{1}{2^j} \\
&= 4 \cdot \max_{\substack{l\in T_R, l\perp \mathbf{1} \\ r\in T_C}} \left|l^t \cdot X \cdot r\right|\,.
\end{aligned}$$

□

Um also Lemma 10 zu beweisen, genügt es, das Maximum in Lemma 12 durch $O(\sqrt{d\cdot\mu\cdot N})$ zu beschränken. Für den verbleibenden Teil des Abschnitts, arbeiten wir nur noch mit den Vektoren l und r anstelle von $u_{|R}$ und $v_{|C}$. Alle Vorkommen von u (bzw. v) beziehen sich von nun an auf Indizes zwischen 1 und n (bzw. m).

Seien also $l \in T_R$ und $r \in T_C$ mit $l \perp \mathbf{1}$. Um

$$\left|l^t \cdot X \cdot r\right| = \left|\sum_{u,v=1}^{u=n,v=m} l_u x_{uv} r_v\right|$$

abzuschätzen, definieren wir

$$\mathcal{B} = \mathcal{B}(l,r) = \left\{(u,v) : |l_u r_v| \leq \sqrt{d\cdot\mu/N}\right\}\,.$$

Wir werden zeigen, dass mit Wahrscheinlichkeit $1 - O(1/N^4)$

$$\left|\sum_{(u,v)\in\mathcal{B}} l_u x_{uv} r_v\right| = O(\sqrt{d\cdot\mu\cdot N}) \quad \text{und} \quad \left|\sum_{(u,v)\notin\mathcal{B}} l_u x_{uv} r_v\right| = O(\sqrt{d\cdot\mu\cdot N})$$

gilt. Der Beweis der linken Gleichung basiert auf der scharfen Konzentration, die Lemma 9 liefert. Die rechte Gleichung gilt, sobald Lemma 11 auf X zutrifft, was es mit Wahrscheinlichkeit $1 - O(1/N^4)$ tut.

Wir beginnen mit den „kleinen“ Paaren, also denen in $\mathcal{B}$. Seien $l \in T^n$ und $r \in T^m$ mit $l \perp \mathbf{1}$ jetzt zwei *feste* Vektoren. Wegen $\sum_{u=1}^{n} l_u = 0$ gilt

$$\sum_{(u,v)} l_u \cdot r_v = \sum_{v=1}^{m} r_v \sum_{u=1}^{n} l_u = 0\,, \text{also} \qquad \sum_{(u,v)\in B} l_u r_v = -\sum_{(u,v)\notin B} l_u r_v\,.$$

Demnach ist $\left|\sum_{(u,v)\in B} l_u r_v\right| = \left|\sum_{(u,v)\notin B} l_u r_v\right|$. Letzteres beschränken wir mithilfe $|l_u r_v| > \sqrt{d\cdot\mu/N}$ für $(u,v) \notin \mathcal{B}$ durch

$$\begin{aligned}\left|\sum_{(u,v)\notin\mathcal{B}} l_u r_v\right| &= \sqrt{\frac{N}{d\cdot\mu}}\cdot\left|\sum_{(u,v)\notin\mathcal{B}} \sqrt{d\cdot\mu/N}\cdot l_u r_v\right| < \sqrt{\frac{N}{d\cdot\mu}}\cdot\left|\sum_{(u,v)\notin\mathcal{B}} l_u^2 r_v^2\right| \\ &\leq \sqrt{N/(d\cdot\mu)}\cdot\sum_{u=1}^{n} l_u^2\cdot\sum_{v=1}^{m} r_v^2 \leq 16\cdot\sqrt{\frac{N}{d\cdot\mu}}.\end{aligned}$$

Für den Erwartungswert von $\sum_{(u,v)\in\mathcal{B}} l_u x_{uv} r_v$ erhalten wir

$$\begin{aligned}\left|\mathbf{E}\left[\sum_{(u,v)\in\mathcal{B}} l_u x_{uv} r_v\right]\right| &\leq \left|\sum_{(u,v)\in\mathcal{B}} l_u\cdot\mathbf{E}\left[x_{uv}\right] r_v\right| = \left|\mu\cdot\sum_{(u,v)\in\mathcal{B}} l_u r_v\right| \\ &< 16\cdot\sqrt{\frac{\mu\cdot N}{d}}.\end{aligned}$$

Nachdem wir den Erwartungswert beschränkt haben, wollen wir die Wahrscheinlichkeit abschätzen, dass $\sum_{(u,v)\in\mathcal{B}} l_u x_{uv} r_v$ weit von diesem Wert abweicht. Da es sich um eine gewichtete Summe von Einträgen der same-mean-Matrix X handelt, ist Lemma 9 nützlich.

Wir erinnern uns, dass Lemma 9 Unabhängigkeit der gewählten Einträge y_i verlangt. Dies ist nicht der Fall, falls $u \neq v$ und beide Paare (u,v) und (v,u) zu $\mathcal{B}$ gehören, da wir von einer symmetrischen Matrix X ausgehen. Weil $x_{uv} = x_{vu}$ ist, können wir aber $l_u x_{uv} r_v$ und $l_v x_{vu} r_u$ einfach zu $x_{uv}\cdot(l_u r_v + l_v r_u)$ verbinden.

Also sind die y_i für das Lemma die x_{uv} mit $(u,v) \in \mathcal{B}$ (bzw. x_{uv} mit $u < v$, falls $u \neq v$, $(u,v) \in \mathcal{B}$ und $(v,u) \in \mathcal{B}$). Die a_i sind die zugehörigen $l_u \cdot r_v$ (bzw. $l_u r_v + l_v r_u$). Wegen der Definition von $\mathcal{B}$ liegen damit alle a_i im Intervall $[-a : a]$ mit $a = 2\cdot\sqrt{d\cdot\mu/N}$. Wegen $\sum a_i^2 \leq 4\cdot\sum l_u^2\cdot r_v^2 \leq 64$ würde $D = 64$ genügen. Wir sind jedoch an einer sehr großen Abweichung vom Erwartungswert interessiert. Deshalb setzen wir $D = 64{\cdot}d$. Aus Lemma 9

folgt so für feste Vektoren l und r und jede Konstante $c > 0$

$$\mathbf{Pr}\left[\left|\sum_{(u,v)\in\mathcal{B}} l_u x_{uv} r_v - \mathbf{E}\left[\sum_{(u,v)\in\mathcal{B}} l_u x_{uv} r_v\right]\right| \geq 32c \cdot \mathrm{e}^c \cdot \sqrt{d\cdot\mu\cdot N}\right] \leq 2\exp\bigl(-8\cdot c^2\cdot \mathrm{e}^c \cdot N\bigr)\,. \quad (2.17)$$

Die Anzahl der möglichen Vektorpaare (l, r) ist durch

$$|T_R|\cdot|T_C| \leq |T_n|\cdot|T_m| \leq k^{n+m} = k^N = \exp(N\cdot\ln k)$$

beschränkt. Erfüllt die Konstante c die Ungleichung $4\cdot c^2\cdot \mathrm{e}^c > \ln k$, ergibt sich: Mit Wahrscheinlichkeit $1-\exp(-\Omega(N))$ gilt für *alle* $l \in T_R$, $r \in T_C$ mit $l \perp \mathbf{1}$

$$\left|\sum_{(u,v)\in\mathcal{B}} l_u x_{uv} r_v\right| = O(\sqrt{d\cdot\mu\cdot N})\,.$$

Es bleibt, eine ähnliche Schranke für die „großen“ Paare zu zeigen, die nicht in $\mathcal{B}$ sind. Dafür gruppieren wir *alle* Koordinaten von l und r (unabhängig von der Zugehörigkeit zu $\mathcal{B}$). Für $i > 0$ sei

$$A_i = \left\{u : \frac{2^{i-1}}{2\sqrt{N}} \leq l_u < \frac{2^i}{2\sqrt{N}}\right\} \quad \text{und} \quad A_i = \left\{u : \frac{2^{|i|-1}}{2\sqrt{N}} \leq -l_u < \frac{2^{|i|}}{2\sqrt{N}}\right\}$$

für $i < 0$. Sei $a_i = |A_i|$ für alle i.

Beachte: Wegen $\|l\| \leq 2$ gibt es nur $O(\log N)$ nichtleere Mengen A_i. Eine Definition von A_0 erübrigt sich: Jeder Eintrag l_u kleiner als $2^0/(2\sqrt{N})$ und größer als $-2^0/(2\sqrt{N})$ muss wegen $n < N$ und der Definition von T_n genau 0 sein. Solche Koordinaten haben keinerlei Einfluss auf die folgenden Berechnungen, da $l_u x_{uv} r_v$ dann auch 0 ist.

Wir definieren B_j und b_j analog für r. Wir schreiben $i \sim j$, falls $2^{|i|+|j|}/4 > \sqrt{d\cdot\mu\cdot N}$. Für $u \in A_i$, $v \in B_j$ und $(u,v) \notin \mathcal{B}$ haben wir

$$\frac{2^{|i|}}{2\sqrt{N}} > |l_u|\,, \qquad \frac{2^{|j|}}{2\sqrt{N}} > |r_v| \qquad \text{und} \qquad |l_u r_v| > \sqrt{d\cdot\mu/N}\,,$$

so dass $2^{|i|+|j|}/(4N)$ größer als $\sqrt{d\cdot\mu/N}$ sein muss. Daraus folgt $2^{|i|+|j|}/4 > \sqrt{d\cdot\mu\cdot N}$ und damit $i \sim j$. So beschränken wir

$$\left|\sum_{(u,v)\notin\mathcal{B}} l_u x_{uv} r_v\right| \leq \sum_{(u,v)\notin\mathcal{B}} |l_u x_{uv} r_v| \leq \sum_{i\sim j} \sum_{\substack{u\in A_i\\ v\in B_j}} |l_u x_{uv} r_v|\,.$$

Den rechten Teil der Ungleichung können wir in anhand der Vorzeichen von i und j, sowie danach, ob $a_i \geq b_j$ oder $a_i < b_j$ ist, in acht Summen aufteilen. Sei

$$\mathcal{C} = \{(i,j) : i \sim j,\ i,j > 0, a_i < b_j\}.$$

Da die Beweise für alle acht Summen sehr ähnlich sind, beweisen wir lediglich

$$\sum_{(i,j)\in\mathcal{C}} \sum_{\substack{u\in A_i \\ v\in B_j}} |l_u x_{uv} r_v| = O(\sqrt{d\cdot\mu\cdot N}) \tag{2.18}$$

ausführlich. Um die folgenden Berechnungen lesbarer zu machen, führen wir einige Abkürzungen ein:

$$s_{ij} = s_X(A_i, B_j) \qquad \lambda_{ij} = \frac{s_{ij}}{\mu_{ij}} \qquad \alpha_i = \frac{a_i\cdot(2^i)^2}{4N}$$
$$\mu_{ij} = \mathbf{E}[s_{ij}] = a_i\cdot b_j\cdot\mu \qquad \sigma_{ij} = \frac{\lambda_{ij}\cdot\sqrt{d\cdot\mu\cdot N}}{2^{i+j-2}} \qquad \beta_j = \frac{b_j\cdot(2^j)^2}{4N}$$

λ_{ij} beschreibt die relative Abweichung von $s_X(A_i, B_j)$ vom Erwartungswert μ_{ij}. Der Wert von σ_{ij} ist eher technischer Natur. Wegen $i \sim j$ gilt $\sigma_{ij} < \lambda_{ij}$ und σ_{ij}/λ_{ij} wird klein, wenn wir „sehr große“ Paare betrachten ($|l_u r_v| \gg \sqrt{d\cdot\mu/N}$). Der Term α_i beschränkt $\sum_{u\in A_i} l_u^2$:

$$\frac{\alpha_i}{4} \leq \sum_{u\in A_i} l_u^2 < \alpha_i.$$

Summieren wir über alle i, ergibt sich $\sum_i \alpha_i \leq 4\cdot\|l\|^2 \leq 16$ und natürlich genauso $\sum_j \beta_j \leq 16$. Für $i,j > 0$ gilt

$$\begin{aligned}
\sum_{\substack{u\in A_i \\ v\in B_j}} |l_u x_{uv} r_v| &\leq \sum_{\substack{u\in A_i \\ v\in B_j}} \frac{2^{i+j}}{4N}\cdot x_{uv} = \frac{2^{i+j}}{4N}\cdot s_{ij} = \frac{2^{i+j}}{4N}\cdot\lambda_{ij}\cdot\mu_{ij} \\
&= \frac{2^{i+j}}{4N}\cdot\frac{\sigma_{ij}\cdot 2^{i+j-2}}{\sqrt{d\cdot\mu\cdot N}}\cdot\mu_{ij} = \frac{2^{2i}\cdot 2^{2j}}{16N}\cdot\frac{\sigma_{ij}}{\sqrt{d\cdot\mu\cdot N}}\cdot a_i\cdot b_j\cdot\mu \\
&= \alpha_i\cdot\beta_j\cdot\sigma_{ij}\cdot\sqrt{\mu\cdot N/d}.
\end{aligned}$$

Um (2.18) zu zeigen, genügt es demnach $\sum_{(i,j)\in\mathcal{C}} \alpha_i\cdot\beta_j\cdot\sigma_{ij} = O(d)$ zu beweisen.

Wenn wir Lemma 11 auf unsere neue Notation übertragen, erhalten wir: Mit Wahrscheinlichkeit $1 - O(1/N^4)$ gilt für jedes Paar A_i, B_j ein der beiden Ungleichungen

$$\lambda_{ij} \leq 20 \tag{2.19}$$

oder

$$\sigma_{ij} \cdot \alpha_i \cdot \log \lambda_{ij} \leq \frac{20 \cdot 2^{i-j} \cdot \sqrt{d}}{\sqrt{\mu \cdot N}} \cdot (2j - \log \beta_j)\,. \tag{2.20}$$

Beachte, dass Paare (A_i, B_j) mit $b_j > N/10$ nicht direkt vom Lemma erfasst werden. In diesem Fall haben wir aber $\lambda_{ij} < 10d$. Das sehen wir folgendermaßen: l_u ist 0, falls $s_X(u, V) > d \cdot \mu \cdot N$, weil dann $u \notin R$ ist. Also ist u in keinem der A_i enthalten. Demnach gilt, $s_X(u, V) \leq d \cdot \mu \cdot N$ für alle $u \in A_i$. Das führt zu

$$s_{ij} \leq (d \cdot \mu \cdot N) \cdot a_i. \tag{2.21}$$

Ist $b_j > N/10$, dann ist $\mu_{ij} > a_i \cdot N/10 \cdot \mu$ und $\lambda_{ij} = s_{ij}/\mu_{ij} < 10d$.

Wir unterteilen die Paare $(i, j) \in \mathcal{C}$ in fünf Klassen $\mathcal{C}_1, \ldots, \mathcal{C}_5$, so dass $(i, j) \in \mathcal{C}_k$ falls (i, j) die folgende Bedingung k erfüllt, aber keine der Bedingungen $< k$.

1. $\sigma_{ij} \leq 20d$
2. $2^{i-j} > \sqrt{d \cdot \mu \cdot N}$
3. $\log \lambda_{ij} \geq (2j - \log \beta_j)/4$ und $2j > -\log \beta_j$
4. $\log \lambda_{ij} < (2j - \log \beta_j)/4$ und $2j > -\log \beta_j$
5. $2j \leq -\log \beta_j$

Unser Beweis ist beendet, wenn für alle $\mathcal{C}_k$ $\sum_{(i,j)\in\mathcal{C}_k} \alpha_i\beta_j\sigma_{ij} = O(d)$ gezeigt ist.

1. $\sigma_{ij} \leq 20d$

$$\begin{aligned}\sum_{(i,j)\in\mathcal{C}_1} \alpha_i\beta_j\sigma_{ij} &= \sum_{(i,j)\in\mathcal{C}_1} 20d \cdot \alpha_i\beta_j \leq 20d \cdot \sum_{(i,j)} \alpha_i\beta_j \\ &= 20d \cdot \sum_i \alpha_i \cdot \sum_j \beta_j \leq 20d \cdot 16 \cdot 16 = O(d).\end{aligned}$$

Wegen $\sigma_{ij} < \lambda_{ij}$ gilt von nun an $20 < 20d < \sigma_{ij} < \lambda_{ij}$. Das bedeutet insbesondere, dass Ungleichung (2.19) verletzt ist, womit Ungleichung (2.20) gilt.

2. $2^{i-j} > \sqrt{d \cdot \mu \cdot N}$

Aus

$$s_{ij} = \lambda_{ij} \cdot \mu_{ij} = \left(\sigma_{ij} \cdot \frac{2^{i+j-2}}{\sqrt{\mu \cdot d \cdot N}}\right) \cdot (a_i \cdot b_j \cdot \mu)$$

und Ungleichung (2.21) folgt

$$\begin{aligned} \sigma_{ij} \cdot 2^{i+j-2} \cdot b_j \cdot \sqrt{\mu/(d \cdot N)} &\leq d \cdot \mu \cdot N \\ \sigma_{ij} \cdot 2^{i+j-2} \cdot b_j/N &\leq \sqrt{d^3 \cdot \mu \cdot N} \\ \sigma_{ij} \cdot 2^{2j} \cdot b_j/(4N) &\leq \sqrt{d^3 \cdot \mu \cdot N} \cdot 2^{j-i} \\ \sigma_{ij}\beta_j &\leq \sqrt{d^3 \cdot \mu \cdot N} \cdot 2^{j-i} \end{aligned}$$

und so

$$\sum_{(i,j)\in\mathcal{C}_2} \alpha_i\beta_j\sigma_{ij} \leq \sum_{(i,j)\in\mathcal{C}_2} \alpha_i \cdot \sqrt{d^3 \cdot \mu \cdot N} \cdot 2^{j-i}\,.$$

Aus der Voraussetzung $2^{i-j} > \sqrt{d \cdot \mu \cdot N}$ folgt $j < i - \log\sqrt{d \cdot \mu \cdot N}$. Da außerdem $j > 0$ ist, erhalten wir

$$\begin{aligned} \sum_{(i,j)\in\mathcal{C}_2} \alpha_i\beta_j\sigma_{ij} &\leq \sum_i \alpha_i \cdot \sqrt{d^3 \cdot \mu \cdot N} \cdot 2^{-i} \cdot \sum_{j=1}^{i-\log\sqrt{d\cdot\mu\cdot N}-1} 2^j \\ &\leq \sum_i \alpha_i \cdot \sqrt{d^3 \cdot \mu \cdot N} \cdot 2^{-i} \cdot 2^{i-\log\sqrt{d\cdot\mu\cdot N}} \\ &= \sum_i \alpha_i \cdot d = O(d) \end{aligned}$$

3. $\log\lambda_{ij} \geq (2j - \log\beta_j)/4 \qquad$ und $\qquad 2j > -\log\beta_j$

Aus (2.20) und $\log\lambda_{ij} \geq (2j - \log\beta_j)/4$ folgt

$$\sigma_{ij} \cdot \alpha_i \cdot \frac{2j - \log\beta_j}{4} \leq \frac{20 \cdot 2^{i-j} \cdot \sqrt{d}}{\sqrt{\mu \cdot N}} \cdot (2j - \log\beta_j)$$

und mit $2j - \log\beta_j > 0$

$$\sigma_{ij} \cdot \alpha_i \leq \frac{80 \cdot 2^{i-j} \cdot \sqrt{d}}{\sqrt{\mu \cdot N}}\,.$$

So ergibt sich

$$\sum_{(i,j)\in\mathcal{C}_3} \alpha_i\beta_j\sigma_{ij} \leq \sum_{(i,j)\in\mathcal{C}_3} \beta_j \cdot 80 \cdot 2^{i-j} \cdot \sqrt{d/(\mu \cdot N)}$$

Wir wissen, $2^{i-j} \leq \sqrt{d \cdot \mu \cdot N}$ (sonst wären wir in Fall 2.). Also ist

$i \leq j + \log\sqrt{d \cdot \mu \cdot N}$ und

$$\begin{aligned}\sum_{(i,j)\in\mathcal{C}_3} \alpha_i\beta_j\sigma_{ij} &\leq \sum_{j>0} \beta_j \cdot 80 \cdot 2^{-j} \cdot \sqrt{d/(\mu\cdot N)} \cdot \sum_{i=1}^{j+\log\sqrt{d\cdot\mu\cdot N}} 2^i \\ &\leq \sum_{j>0} \beta_j \cdot \frac{80 \cdot 2^{\log\sqrt{d\cdot\mu\cdot N}+1} \cdot \sqrt{d}}{\sqrt{\mu\cdot N}} = \sum_{j>0} \beta_j \cdot 160d \\ &= O(d)\,.\end{aligned}$$

4. $\log\lambda_{ij} < (2j - \log\beta_j)/4 \qquad$ und $\qquad 2j > -\log\beta_j$

   Wegen (2.20) und $2j > -\log\beta_j$ haben wir

$$\sigma_{ij} \cdot \alpha_i \leq \frac{20 \cdot 2^{i-j} \cdot \sqrt{d}}{\sqrt{\mu\cdot N}} \cdot 4j\,.$$

   Beachte: Auf $\log\lambda_{ij} > \log(20) > 1$ wurde auf der linken Seite verzichtet.
   Es gilt

$$\sum_{(i,j)\in\mathcal{C}_4} \alpha_i\beta_j\sigma_{ij} \leq \sum_{(i,j)\in\mathcal{C}_4} \beta_j \cdot 80j \cdot 2^{i-j} \cdot \sqrt{d/(\mu\cdot N)}\,.$$

   Wir haben $\log\lambda_{ij} < (2j - \log\beta_j)/4 \leq j$ bzw. $\lambda_{ij} < 2^j$. Gemeinsam mit der Definition von σ_{ij} erhalten wir daraus $\sigma_{ij} \leq \sqrt{d\cdot\mu\cdot N}/2^{i-2}$. Da der Fall $\sigma_{ij} \leq 20d$ bereits in Bedingung 1. behandelt wurde, gilt jetzt $1 < 20d < \sigma_{ij}$, also $1 < \sqrt{d\cdot\mu\cdot N}/2^{i-2}$ bzw. $2^{i-2} < \sqrt{d\cdot\mu\cdot N}$. Daraus wiederum folgern wir $i < \log\sqrt{d\cdot\mu\cdot N} + 2$ Wir erhalten

$$\begin{aligned}\sum_{(i,j)\in\mathcal{C}_4} \alpha_i\beta_j\sigma_{ij} &\leq \sum_{j>0} \beta_j \cdot 80j \cdot 2^{-j} \cdot \sqrt{d/(\mu\cdot N)} \sum_{i=1}^{\log\sqrt{d\cdot\mu\cdot N}+1} 2^i \\ &\leq \sum_{j>0} \beta_j \cdot 80j \cdot 2^{-j} \cdot 4d = O(d) \cdot \sum_{j>0} \frac{j}{2^j} = O(d)\,.\end{aligned}$$

5. $2j \leq -\log\beta_j$

   Wieder nutzen wir (2.20) und verzichten auf den Term $\log\lambda_{ij} > 1$:

$$\sigma_{ij} \cdot \alpha_i \leq -\frac{40 \cdot 2^{i-j} \cdot \sqrt{d}}{\sqrt{\mu\cdot N}} \cdot \log\beta_j\,.$$

Wegen $2^{i-j} \leq \sqrt{d \cdot \mu \cdot N}$ (vgl. Fall 2.) ist $i \leq j + \log \sqrt{d \cdot \mu \cdot N}$ und somit

$$\begin{aligned}
\sum_{(i,j)\in\mathcal{C}_5} \alpha_i\beta_j\sigma_{ij} &\leq \sum_{(i,j)\in\mathcal{C}_5} -\frac{40 \cdot 2^{i-j} \cdot \sqrt{d}}{\sqrt{\mu \cdot N}} \cdot \beta_j \cdot \log \beta_j \\
&\leq \sum_{j>0} -\beta_j \cdot \log \beta_j \cdot 40 \cdot \sqrt{d/(\mu \cdot N)} \cdot \sum_{i=1}^{j+\log\sqrt{d\cdot\mu\cdot N}} 2^{i-j} \\
&< \sum_{j>0} -\beta_j \cdot \log \beta_j \cdot 80d\,.
\end{aligned}$$

Da $j > 0$ ist, ist $-\log \beta_j$ positiv und $\beta_j < 1$. Für $0 < \beta_j < 1$ können wir $-\log \beta_j$ durch $4/\sqrt{\beta_j}$ nach oben abschätzen (vgl. A.5). Demnach ist

$$-\beta_j \cdot \log \beta_j < \frac{\beta_j \cdot 4}{\sqrt{\beta_j}} = 4 \cdot \sqrt{\beta_j}\,.$$

Wegen der Voraussetzung $2j \leq -\log \beta_j$ ist $\sqrt{\beta}_j \leq 2^{-j}$. Es ergibt sich

$$\sum_{(i,j)\in\mathcal{C}_5} \alpha_i\beta_j\sigma_{ij} < \sum_{j>0} 80d \cdot \sqrt{\beta_j} \leq 80d \cdot \sum_{j>0} 2^{-j} = O(d)\,.$$

Damit ist der Beweis von Lemma 10 abgeschlossen. □

Kapitel 3

Partitionierung Teil II

3.1 Idee und Algorithmus

Algorithmus 3 ging so vor, dass die Adjazenzmatrix „normiert“ wurde, indem jeder Eintrag a_{uv} durch die erwarteten Grade w'_u und w'_v geteilt wurde. Zeilen und Spalten (und damit Knoten), die sich unerwartet verhalten haben, wurden anschließend aus der Matrix entfernt.

Dieser Ansatz ist insofern unpraktikabel, als dass Informationen wie „erwartete Grade“ üblicherweise nicht zur Verfügung stehen. Wir versuchen diese Einschränkung aufzuheben. Dazu teilen wir die Einträge jetzt durch die *tatsächlichen* Grade. Es ist also $m_{uv} = a_{uv}/(d_u \cdot d_v)$.

Das erschwert uns die Analyse jedoch sehr: Die Einträge der Matrix sind jetzt abhängig voneinander. Wir werden versuchen, die Eigenschaften von $M = (m_{uv})$ mithilfe von $\mathbf{M} = (\mathbf{m}_{uv})$ abzuleiten, wobei $\mathbf{m}_{uv} = a_{uv}/(w'_u \cdot w'_v)$ ist. Dies funktioniert jedoch nur für die Einträge, bei denen $d_u \approx w'_u$ und $d_v \approx w'_v$ ist.

Da wir auch den Fall dünner Graphen betrachten, wo $w'_u = O(1)$ ist, sind die tatsächlichen Grade bei vielen Knoten *nicht* nahe ihrem Erwartungswert w'_u. Da wir ein sehr allgemeines Modell zufälliger Graphen betrachten, sind solche Knoten (anders als bei $G_{n,p}$-basierten Modellen) auch algorithmisch nicht zu erkennen.

Wir werden aber in den Analysen zum einen sehen, dass der Einfluss der meisten dieser Knoten auf das Spektrum relativ klein ist und zum anderen, dass die Knoten, deren Einfluss groß ist, durch unseren Algorithmus erkannt werden. Die Analyse geht folgenden Weg. Wir teilen die Knoten des Graphen in „gute“ und „schlechte“ auf. Die „guten“ Knoten sind die, deren Grad nahe an dessen Erwartungswert liegt.

Mithilfe dieser Einteilung zerlegen wir die Matrix M in Untermatrizen

und analysieren die einzelnen Teile getrennt voneinander. Nun kann es aber passieren, dass ein „guter“ Knoten u ungewöhnlich viele „schlechte“ Nachbarn hat. Dann ist zwar $d_u \approx w'_u$, aber in der Untermatrix, die von den „guten“ Knoten induziert wird, sind ungewöhnlich wenige Einträge $m_{u,\cdot}$ bzw. $m_{\cdot,u}$ ungleich 0.

Da solche Effekte die Analyse erschweren, markieren wir so ein u ebenfalls als „schlecht“. Letztendlich erfüllt jeder „gute“ Knoten also

1. $d_u \approx \mathbf{E}[d_u] = w'_u$ und
2. u hat wenige „schlechte“ Nachbarn.

Zur Analyse der Untermatrizen benutzen wir Ideen und Techniken aus [COL06]. Unser Algorithmus kann natürlich nicht alle „schlechten“ Knoten identifizieren, da wir auf die w'_u als Eingabe verzichten. Er kann lediglich die Knoten mit außergewöhnlich geringem Grad finden und deren Einträge unwirksam machen, indem sie auf 0 gesetzt werden. Wir werden sehen, dass dies bereits ausreicht, um eine gute Partitionierung zu ermöglichen. Andererseits ist dies auch notwendig wie wir weiter unten sehen werden.

Wir benutzen nun folgende Variation von Algorithmus 3.

Algorithmus 13.

Eingabe: Die Adjazenzmatrix $A = (a_{uv})$ eines Graphen $G = (V, E)$.
Ausgabe: Eine Partition V'_1, V'_2 von V.

1. Berechne den Durchschnittsgrad $\overline{d} = \sum_{u=1}^{n} d_u/n$ und setze $d_{\mathrm{m}} = \overline{d} \,/\, \ln \overline{d}$.
2. Konstruiere $M = (m_{uv})$ mit $m_{uv} = a_{uv}/(d_u \cdot d_v)$. Falls $d_u \cdot d_v = 0$ ist, setze $m_{uv} = 0$.
3. Bestimme $U = \{u \in V : d_u \geq d_{\mathrm{m}}\}$.
4. Bilde M^* aus M, indem jeder Eintrag m_{uv} durch 0 ersetzt wird, wenn $u \notin U$ oder $v \notin U$.
5. Berechne s_1 und s_2: Die Eigenvektoren von M^* zu den beiden betragsgrößten Eigenwerten. Skaliere beide s_i auf Länge $\sqrt{n}$.
6. Falls keiner der beiden s_i die Eigenschaft

   > „Es gibt $c_1, c_2 \in \mathbb{R}$ mit $|c_1 - c_2| > 1/4$, so dass jeweils mehr als $n/\sqrt{d_{\mathrm{m}}}$ Knoten $v \in U$ entweder $|s_i(v) - c_1| \leq 1/32$ oder $|s_i(v) - c_2| \leq 1/32$ erfüllen.“

   hat, setze $V'_1 = V$ und $V'_2 = \emptyset$. Ansonsten sei $s \in \{s_1, s_2\}$ so ein Eigenvektor. Dann sei V'_1 die Menge der Knoten, deren Einträge in s näher an c_1 als an c_2 sind. Setze $V'_2 := V \setminus V'_1$.

Analog zu Theorem 2 erhalten wir

Theorem 14. *Sei G ein in unserem Modell generierter Graph und M^* konstruiert, wie in Algorithmus 13 beschrieben. Dann gilt mit Wahrscheinlichkeit $1 - O(1/n)$ für alle $1 \leq i, j \leq k$ simultan:*

1. $\dfrac{\mathbf{1}^t}{\|\mathbf{1}^t\|} \cdot {M^*}_{V_i \times V_j} \cdot \dfrac{\mathbf{1}}{\|\mathbf{1}\|} = d_{ij} \cdot W_i \cdot W_j \cdot \dfrac{\sqrt{|V_i| \cdot |V_j|}}{\overline{w} \cdot n} \cdot \left(1 \pm O\left({d_{\mathrm{m}}}^{-0.49}\right)\right).$
2. *Für alle u, v mit $\|u\| = \|v\| = 1$ sowie $u \perp \mathbf{1}$ oder $v \perp \mathbf{1}$ ist*

$$\left|u^t \cdot {M^*}_{V_i \times V_j} \cdot v\right| = O\left({\overline{w}}^{-1.49} + {d_{\mathrm{m}}}^{-1.5}\right).$$

Um zu sehen, dass Schritt 4. notwendig ist, betrachten wir einen Spezialfall unseres Modells. Sei $w_u = \overline{w} = O(1)$ für alle $u \in V$, $d_{11} = n/|V_1|$ und $d_{12} = 0$. Dann entspricht $A_{V_1 \times V_1}$ der Adjazenzmatrix eines zufälligen Graphen G', generiert im $G_{n,p}$-Modell mit $|V_1|$ Knoten und $p = \overline{w}/|V_1|$.

Sei nun $d \in \mathbb{N}$ eine beliebige Konstante und sei T der folgende Baum mit Tiefe 2. In Tiefe 1 befinden sich d Knoten, von denen jeder genau $d-1$ Nachfolger hat. Insgesamt hat T also $1+d+d\cdot(d-1) = d^2+1 = O(1)$ Knoten. Laut Kapitel 3. von [JLR00] kommen in G' mhW. $\Theta(n)$ knotendisjunkte Kopien von T vor. Eine Konsequenz daraus ist, dass sich M (im Gegensatz zu M^*) *nicht* zur Partitionierung eignet.

Während bei ${M^*}_{V_1 \times V_1}$ laut Theorem 14 der größte Eigenvektor annähernd der $\mathbf{1}$-Vektor ist, ist dies bei $M_{V_1 \times V_1}$ mhW. nicht so. Es gilt zwar $\mathbf{1}^t \cdot M_{V_1 \times V_1} \cdot \mathbf{1} = \Theta(1/\overline{w})$, jedoch ist der größte Eigenwert deutlich größer als $O(1/\overline{w})$, nämlich $\geq 1/d^{1.5} = \Omega(1)$, *unabhängig* von $\overline{w}$.

Um dies zu sehen, betrachten wir eine beliebige Kopie von T in G'. Sei v_0 deren Wurzel und $v_1, \ldots, v_d$ die Knoten in Tiefe 1. Sei y ein $|V_1|$-dimensionaler Vektor, wobei die zu v_0 gehörende Koordinate $y(v_0) = \sqrt{d}$, $y(v_i) = 1$ für $i = 1, \ldots, d$ und alle anderen Koordinaten 0 sind. Sei $\xi = M_{V_1 \times V_1} \cdot y$. Dann ist $\xi(v_0) = 1/d$ und $\xi(v_i) = d^{-1.5}$ für $i = 1, \ldots, d$. Die anderen Einträge sind nicht von Belang. Es gilt wegen Fakt 5, der Courant-Fischer-Charakterisierung von Eigenwerten,

$$\begin{aligned} \lambda_1 &\geq \max_{\|x\|=1} x^t \cdot M_{V_1 \times V_1} \cdot x \geq \frac{y^t}{\|y\|} \cdot M_{V_1 \times V_1} \cdot \frac{y}{\|y\|} = \frac{y^t \cdot \xi}{y^t \cdot y} \\ &= \frac{d^{-0.5} + d \cdot d^{-1.5}}{2d} = d^{-1.5}. \end{aligned}$$

Da es $\Theta(n)$ knotendisjunkte Kopien von T gibt, sind sogar die $\Theta(n)$ betragsgrößten Eigenwerte von M^* mindestens $d^{-0.15}$. Die zugehörigen Eigenvektoren reflektieren dabei im Allgemeinen keine *globalen* Eigenschaften des Graphen, sondern Teilgraphen konstanter Größe.

Wir sehen, dass es nötig ist, derartige Strukturen zu entfernen. Wir realisieren dies, indem wir die Einträge der Knoten mit Grad $\leq d_{\mathrm{m}}$ auf 0 setzen. Insbesondere muss $d_{\mathrm{m}} \gg \overline{w}^{2/3}$ gelten, damit $\Theta(1/\overline{w})$ deutlich größer als ${d_{\mathrm{m}}}^{-1.5}$ ist.

Auf der anderen Seite ist es nötig, dass $d_{\mathrm{m}} \ll \min_{u \in V} w'_u$ ist, da sonst U bei vielen Modellparametern mhW. leer ist und eine Partitionierung nicht gelingen kann.

Für den Beweis von Theorem 14 ist es günstig, wenn d_{m} sogar deutlich kleiner als die minimale erwartete Zeilensumme (bzw. Spaltensumme) jeder Matrix $A_{V_i \times V_j}$ mit $d_{ij} > 0$ ist, d. h.

$$d_{\mathrm{m}} \ll \min_{\substack{i,j \\ d_{ij}>0}} \left\{ |V_i| \cdot d_{ij} \cdot \frac{(\min_u w_u)^2}{\overline{w} \cdot n} \right\} . \tag{3.1}$$

Wegen $|V_i| = \Omega(n)$ und $w_u \geq \varepsilon \cdot \overline{w} = \Omega(\overline{w})$ ist die rechte Seite $\Omega(\overline{w})$.

Algorithmus 13 stellt (3.1) mit der Wahl $d_{\mathrm{m}} = \overline{d} \,/ \ln \overline{d}$ „sicher“: Fakt 16 zeigt, dass mit Wahrscheinlichkeit $1 - \exp(-\sqrt{n})$ die Gleichung $\overline{d} = \overline{w}' \cdot (1 + o(1))$ gilt, wobei der erwartete Durchschnittsgrad $\overline{w}' = \Theta(\overline{w})$ erfüllt (Lemma 1). Also ist mit hoher Wahrscheinlichkeit $d_{\mathrm{m}} = O(\overline{w}/\ln \overline{w})$.

Auf dem gleichen Weg wie bei Lemma 6 können wir aus Theorem 14 folgern: Mit Wahrscheinlichkeit $1-O(1/n)$ hat M^* genau zwei Eigenwerte, deren Absolutbetrag $\Theta(1/\overline{w})$ ist, während alle anderen Eigenwerte betragsmäßig

$$O\left(1/\left(\overline{w} \cdot {d_{\mathrm{m}}}^{0.49}\right) + {\overline{w}}^{-1.49} + {d_{\mathrm{m}}}^{-1.5}\right) = O\left(1/\left(\overline{w} \cdot {d_{\mathrm{m}}}^{0.49}\right)\right)$$

sind. Das daraus resultierende „spectral gap“ von ${d_{\mathrm{m}}}^{0.49}$ liefert schließlich das folgende Theorem, dessen Beweis (auf den wir verzichten) analog zu dem von Theorem 4 verläuft.

Theorem 15. *Sei G ein Graph, der in unserem Modell generiert wurde. Mit Wahrscheinlichkeit $1 - O(1/n)$ (bezogen auf G) konstruiert Algorithmus 13 eine Partition, die sich von der gepflanzten Partition V_1, V_2 nur in $O\left(n/{d_{\mathrm{m}}}^{0.98}\right) = O\left(n/{\overline{d}}^{0.97}\right)$ Knoten unterscheidet.*

3.2 Beweise

3.2.1 Beweis von Theorem 14

Während des gesamten Abschnitts benötigen wir für Knotenmengen $U_1, U_2 \in V$ den Term des Volumens:

$$\mathrm{Vol}(U_1, U_2) = \sum_{u \in U_1} \sum_{v \in U_2} d_{\psi(u),\psi(v)} \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot n} ,$$

wobei $\psi(u)$ die Menge der gepflanzten Partition angibt, zu der u gehört. Also $\psi(u) = i$ genau dann, wenn $u \in V_i$.

Beachte: Für feste Mengen U_1, U_2 und einen Graphen aus unserem Modell ist $\mathrm{Vol}(U_1, U_2)$ genau die *erwartete* Anzahl von Kanten zwischen U_1 und U_2, d. h. $\mathrm{Vol}(U_1, U_2) = \mathbf{E}[s_A(U_1, U_2)]$, wenn A die Adjazenzmatrix des Graphen ist. Statt $\mathrm{Vol}(\{u\}, U_2)$ schreiben wir zugunsten der Lesbarkeit $\mathrm{Vol}(u, U_2)$.

Um verschiedene Aussagen über die Adjazenzmatrix machen zu können, benötigen wir die folgende Chernoff-Schranke aus [JLR00] (Theorem 2.1 in Verbindung mit Theorem 2.8).

Fakt 16. *Sei X die Summe unabhängiger 0–1-Zufallsvariablen, dann gilt*

1. $\mathbf{Pr}[X \geq \mathbf{E}[X] + t] \leq \exp\left(-\mathbf{E}[X] \cdot \phi\left(\frac{t}{\mathbf{E}[X]}\right)\right) \leq \exp\left(-\frac{t^2}{2\cdot(\mathbf{E}[X]+t/3)}\right)$
2. $\mathbf{Pr}[X \leq \mathbf{E}[X] - t] \leq \exp\left(-\mathbf{E}[X] \cdot \phi\left(\frac{-t}{\mathbf{E}[X]}\right)\right) \leq \exp\left(-\frac{t^2}{2\cdot\mathbf{E}[X]}\right)$

für alle $t \geq 0$ mit $\phi(x) = (1+x)\cdot\ln(1+x) - x$ für $x > -1$.

Wie bereits in Abschnitt 3.1 beschrieben, teilen wir die Knoten in „gute“ und „schlechte“ auf. Die beiden Mengen der „guten“ Knoten $R_{ij} \subseteq V_i$ und $C_{ij} \subseteq V_j$ sind das Ergebnis des folgenden Prozesses. Den Wert der Konstante D lassen wir vorerst unbestimmt. Er ergibt sich später im Beweis von Lemma 17.

1. $R' = \{u \in V : \forall j' : |s_A(u, V_{j'}) - \mathrm{Vol}(u, V_{j'})| \leq \mathrm{Vol}(u, V_{j'})^{0.51}\}$.
2. $C' = \{v \in V : \forall i' : |s_A(V_{i'}, v) - \mathrm{Vol}(V_{i'}, v)| \leq \mathrm{Vol}(V_{i'}, v)^{0.51}\}$.
3. Setze $R'_{ij} := R' \cap V_i$ und $C'_{ij} := C' \cap V_j$.
4. Solange es ein $u \in R'_{ij}$ mit

   $$s_A(u, V_j \setminus C'_{ij}) \geq \mathrm{Vol}(u, V_j) \cdot D/d_{\mathrm{m}} \qquad \text{gibt, setze} \quad R'_{ij} := R'_{ij} \setminus \{u\}.$$

5. Solange es ein $v \in C'_{ij}$ mit

   $$s_A(V_i \setminus R'_{ij}, v) \geq \mathrm{Vol}(V_i, v) \cdot D/d_{\mathrm{m}} \qquad \text{gibt, setze} \quad C'_{ij} := C'_{ij} \setminus \{v\}.$$

6. Wiederhole die Schritte 4. – 5. bis R'_{ij} *und* C'_{ij} unverändert bleiben.
7. Setze $R_{ij} := R'_{ij}$ und $C_{ij} := C'_{ij}$.

Wir schreiben im Weiteren $\mathcal{R}$ statt R_{ij}, $\mathcal{C}$ statt C_{ij}, $\overline{\mathcal{R}}$ statt $V_i \setminus R_{ij}$ und $\overline{\mathcal{C}}$ statt $V_j \setminus C_{ij}$. Man kann leicht nachrechnen, dass (insbesondere wegen Schritt 1.) für alle $u \in \mathcal{R}$

$$\begin{aligned} |s_A(u, V) - \mathrm{Vol}(u, V)| &\leq 2 \cdot \mathrm{Vol}(u, V)^{0.51} \\ \text{und für alle } v \in \mathcal{C} \qquad |s_A(V, v) - \mathrm{Vol}(V, v)| &\leq 2 \cdot \mathrm{Vol}(V, v)^{0.51} \end{aligned} \tag{3.2}$$

gilt.

Einige Bemerkungen zu obigem Prozess. Sei $u \in V_i$ beliebig. Die typische Abweichung von $s_A(u, V_j)$ vom Erwartungswert $\mathbf{E}[s_A(u, V_j)] = \mathrm{Vol}(u, V_j)$ hat die Größenordnung $O(\mathrm{Vol}(u, V_j)^{0.5})$. Wir haben den Exponenten für Schritt 1. mit 0.51 etwas größer gewählt, weil wir so hoffen können, dass die meisten Knoten aus V_i zu R' gehören. Weiterhin ist die Zugehörigkeit von u zu R' umso wahrscheinlicher, desto größer $\mathrm{Vol}(u, V_j)$ ist.

Unsere Hoffnung ist daher nicht nur, dass viele Knoten zu R' gehören, sondern auch, dass das Volumen der Knoten aus $V_i \setminus R'$ klein ist. Wir werden sehen, dass das mhW. sowohl für $\mathrm{Vol}(V_i \setminus R', V_j,)$ als auch für $\mathrm{Vol}(V_i, V_j \setminus C')$ zutrifft.

Ein Knoten, der in Schritt 4. aus $R'_{ij} \subseteq V_i$ entfernt wird, muss einen relativ großen Anteil seiner Nachbarn innerhalb $V_j \setminus C'_{ij}$ haben. Da aber $\mathrm{Vol}(V_i, V_j \setminus C')$ klein ist, erwarten wir nur wenige Nachbarn innerhalb $V_j \setminus C'$. Wir können daher hoffen, dass nur wenige Knoten in Schritt 4. (und Schritt 5.) gelöscht werden und sowohl R'_{ij} und $\mathrm{Vol}(R'_{ij}, V_j)$ (und C'_{ij} und $\mathrm{Vol}(V_i, C'_{ij})$) groß bleiben.

Wie sich herausstellt, ist dies (zumindest mhW.) tatsächlich der Fall, wie Lemma 17 und dessen Beweis in Abschnitt 3.2.4 zeigen.

Lemma 17. *Sei G ein Graph aus unserem Modell. Dann ist mit Wahrscheinlichkeit $1 - O(1/n)$*

1. $\mathrm{Vol}(\overline{\mathcal{R}}, V_j) \leq n/{d_\mathrm{m}}^4$.
2. $\mathrm{Vol}(V_i, \overline{\mathcal{C}}) \leq n/{d_\mathrm{m}}^4$.
3. $\mathrm{Vol}(\overline{\mathcal{R}}, \overline{\mathcal{C}}) \leq n/{d_\mathrm{m}}^8$.

Eine Konsequenz von Lemma 17 ist, dass weder $\overline{\mathcal{R}}$ noch $\overline{\mathcal{C}}$ viele Knoten enthalten: Da d_m mit Wahrscheinlichkeit $1 - \exp(-\sqrt{n})$ kleiner als die erwartete Zeilensumme in V_i ist (vgl. (3.1)), gilt mit Wahrscheinlichkeit $1 - \exp(-\sqrt{n})$ für alle $u \in V_i$ und alle $v \in V_j$

$$d_\mathrm{m} \leq \mathrm{Vol}(u, V_j) \leq \mathrm{Vol}(u, V) = w'_u \quad \text{und} \quad d_\mathrm{m} \leq \mathrm{Vol}(V_i, v) \leq w'_v\,. \tag{3.3}$$

Somit ist $d_\mathrm{m} \cdot \left|\overline{\mathcal{R}}\right| \leq \mathrm{Vol}(\overline{\mathcal{R}}, V_j) \leq n/{d_\mathrm{m}}^4$, was zu $\left|\overline{\mathcal{R}}\right| \leq n/{d_\mathrm{m}}^5$ führt. Wegen $\delta \cdot n \leq |V_i|$ erhalten wir

$$\left|\overline{\mathcal{R}}\right| \leq \frac{|V_i|}{\delta \cdot {d_\mathrm{m}}^5} \leq \frac{|V_i|}{{d_\mathrm{m}}^4} \quad \text{und} \quad |\mathcal{R}| = |V_i| - \left|\overline{\mathcal{R}}\right| \geq |V_i| \cdot \left(1 - \frac{1}{{d_\mathrm{m}}^4}\right), \tag{3.4}$$

solange $d_\mathrm{m} > 1/\delta$ groß genug ist. Genauso ergibt sich

$$\left|\overline{\mathcal{C}}\right| \leq |V_i|\,/{d_\mathrm{m}}^4 \qquad \text{und} \qquad |\mathcal{C}| \geq |V_i| \cdot \left(1 - 1/{d_\mathrm{m}}^4\right). \tag{3.5}$$

Für die weitere Analyse zerlegen wir ${M^*}_{V_i \times V_j}$ in die vier Teile ${M^*}_{\overline{\mathcal{R}} \times \mathcal{C}}$, ${M^*}_{\mathcal{R} \times \overline{\mathcal{C}}}$, ${M^*}_{\mathcal{R} \times \mathcal{C}}$ und ${M^*}_{\overline{\mathcal{R}} \times \overline{\mathcal{C}}}$ und untersuchen diese getrennt voneinander.

Lemma 18. *Sei G ein in unserem Modell generierter Graph. Dann gilt mit Wahrscheinlichkeit $1 - O(1/n^4)$*

1. $\mathbf{1}^t \cdot {M^*}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} = d_{ij} \cdot W_i \cdot W_j \cdot \frac{|\mathcal{R}| \cdot |\mathcal{C}|}{\overline{w} \cdot n} \cdot \left(1 \pm O(1/{d_\mathrm{m}}^{0.49})\right) = \Theta(n/\overline{w})\,.$
2. *Für alle u, v mit $\|u\| = \|v\| = 1$ und $u \perp \mathbf{1}$ oder $v \perp \mathbf{1}$ ist*
$$\left|u^t \cdot {M^*}_{\mathcal{R}\times\mathcal{C}} \cdot v\right| = O\left(1/\overline{w}^{1.49}\right)\,.$$
3. $\|{M^*}_{\mathcal{R}\times\mathcal{C}}\| = \Theta\left(1/\overline{w}\right).$

Wir zeigen Lemma 18 in Abschnitt 3.2.2. Der Beweis des folgenden Lemmas befindet sich in Abschnitt 3.2.3.

Lemma 19. *Sei G ein Graph, der in unserem Modell generiert wurde.*

1. $\left\|{M^*}_{\mathcal{R}\times\overline{\mathcal{C}}}\right\| = O({d_\mathrm{m}}^{-1.5})$.
2. $\left\|{M^*}_{\overline{\mathcal{R}}\times\mathcal{C}}\right\| = O({d_\mathrm{m}}^{-1.5})$.
3. *Mit Wahrscheinlichkeit $1 - O(1/n^4)$ ist* $\left\|{M^*}_{\overline{\mathcal{R}}\times\overline{\mathcal{C}}}\right\| = O({d_\mathrm{m}}^{-1.5})$.

Zunächst beweisen wir den ersten Punkt von Theorem 14. Es gilt

$$\mathbf{1}^t \cdot {M^*}_{V_i\times V_j} \cdot \mathbf{1} = \mathbf{1}^t \cdot {M^*}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} + \mathbf{1}^t \cdot {M^*}_{\mathcal{R}\times\overline{\mathcal{C}}} \cdot \mathbf{1} + \mathbf{1}^t \cdot {M^*}_{\overline{\mathcal{R}}\times\mathcal{C}} \cdot \mathbf{1} + \mathbf{1}^t \cdot {M^*}_{\overline{\mathcal{R}}\times\overline{\mathcal{C}}} \cdot \mathbf{1}\,. \quad (3.6)$$

Für den ersten Summanden benutzen wir Punkt 1. von Lemma 18.

$$\begin{aligned}\mathbf{1}^t \cdot {M^*}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} &= d_{ij} \cdot W_i \cdot W_j \cdot \frac{|\mathcal{R}| \cdot |\mathcal{C}|}{\overline{w} \cdot n} \cdot (1 \pm O(1/{d_\mathrm{m}}^{0.49})) \\ &\overset{(3.4),(3.5)}{=} d_{ij} \cdot W_i \cdot W_j \cdot \frac{|V_i| \cdot |V_j|}{\overline{w} \cdot n} \cdot (1 \pm O(1/{d_\mathrm{m}}^{0.49}))\,.\end{aligned}$$

Lemma 19 zeigt, dass der zweite Summand in (3.6) durch

$$\begin{aligned}\left|\mathbf{1}^t \cdot {M^*}_{\mathcal{R}\times\overline{\mathcal{C}}} \cdot \mathbf{1}\right| &\leq \sqrt{|\mathcal{R}| \cdot \left|\overline{\mathcal{C}}\right|} \cdot \left\|{M^*}_{\mathcal{R}\times\overline{\mathcal{C}}}\right\| \overset{(3.5)}{\leq} \sqrt{|V_i| \cdot |V_j| \,/ {d_\mathrm{m}}^4} \cdot O({d_\mathrm{m}}^{-1.5}) \\ &= \sqrt{|V_i| \cdot |V_j|} \cdot O\left({d_\mathrm{m}}^{-3.5}\right) = \sqrt{|V_i| \cdot |V_j|} \cdot O(1/(\overline{w} \cdot {d_\mathrm{m}}^2)\end{aligned}$$

beschränkt ist, da $d_\mathrm{m} > \overline{w}^{2/3}$. Die gleiche Schranke gilt für $\left|\mathbf{1}^t \cdot {M^*}_{\overline{\mathcal{R}}\times\mathcal{C}} \cdot \mathbf{1}\right|$ und $\left|\mathbf{1}^t \cdot {M^*}_{\overline{\mathcal{R}}\times\overline{\mathcal{C}}} \cdot \mathbf{1}\right|$. Wir erhalten

$$\begin{aligned}&\frac{\mathbf{1}^t}{\|\mathbf{1}^t\|} \cdot {M^*}_{V_i\times V_j} \cdot \frac{\mathbf{1}}{\|\mathbf{1}\|} \\ &\quad= d_{ij} \cdot W_i \cdot W_j \cdot \frac{\sqrt{|V_i| \cdot |V_j|}}{\overline{w} \cdot n} \cdot (1 \pm O(1/{d_\mathrm{m}}^{0.49})) \pm O\left(\frac{1}{\overline{w} \cdot {d_\mathrm{m}}^2}\right) \\ &\quad= d_{ij} \cdot W_i \cdot W_j \cdot \frac{\sqrt{|V_i| \cdot |V_j|}}{\overline{w} \cdot n} \cdot (1 \pm O(1/{d_\mathrm{m}}^{0.49}))\,.\end{aligned}$$

Um den zweiten Punkt von Theorem 14 zu zeigen, nehmen wir zunächst an, dass $u \perp \mathbf{1}$ ist. Daraus folgt unmittelbar $u^t \cdot (\mathbf{1}_{|\mathcal{R}} + \mathbf{1}_{|\overline{\mathcal{R}}}) = 0$, was zu

$$\left|u^t \cdot \mathbf{1}_{|\mathcal{R}}\right| = \left|u^t \cdot \mathbf{1}_{|\overline{\mathcal{R}}}\right| \leq \|u\| \cdot \left\|\mathbf{1}_{|\overline{\mathcal{R}}}\right\| \leq \sqrt{\left|\overline{\mathcal{R}}\right|} \tag{3.7}$$

führt.

Wir können u als $u = a \cdot \mathbf{1}_{|\mathcal{R}} \,/\, \left\|\mathbf{1}_{|\mathcal{R}}\right\| + b \cdot u_l$ schreiben, wobei $\|u_l\| = 1$ und $u_l \perp \mathbf{1}_{|\mathcal{R}}$. Aus Letzterem folgt sofort $u_{l|\mathcal{R}} \perp \mathbf{1}_{|\mathcal{R}}$. Wir erhalten eine Schranke an a durch

$$|a| = \left| u^t \cdot \frac{\mathbf{1}_{|\mathcal{R}}}{\left\|\mathbf{1}_{|\mathcal{R}}\right\|} \right| \overset{(3.7)}{\leq} \frac{\sqrt{\left|\overline{\mathcal{R}}\right|}}{\|\mathbf{1}\|_{\mathcal{R}}} \overset{(3.4)}{\leq} \sqrt{2/{d_{\mathrm{m}}}^4} < 2/{d_{\mathrm{m}}}^2 \,. \tag{3.8}$$

Für jeden beliebigen Einheitsvektor v folgt dann

$$\begin{aligned}\left|u^t \cdot {M^*}_{V_i \times V_j} \cdot v\right| &= \left|u^t \cdot {M^*}_{V_i \times V_j} \cdot \left(v_{|\mathcal{C}} + v_{|\overline{\mathcal{C}}}\right)\right| \\ &\leq \left|u^t \cdot {M^*}_{V_i \times V_j} \cdot v_{|\mathcal{C}}\right| + \left\|{M^*}_{\mathcal{R} \times \overline{\mathcal{C}}}\right\| + \left\|{M^*}_{\overline{\mathcal{R}} \times \overline{\mathcal{C}}}\right\| \,.\end{aligned}$$

Die beiden letzten Summanden sind $O({d_{\mathrm{m}}}^{-1.5})$. Den ersten Term schätzen wir folgendermaßen ab

$$\begin{aligned}\left|u^t \cdot {M^*}_{V_i \times V_j} \cdot v_{|\mathcal{C}}\right| &= \left|\left(a \cdot \frac{{\mathbf{1}^t}_{|\mathcal{R}}}{\left\|{\mathbf{1}^t}_{|\mathcal{R}}\right\|} + b \cdot u_l\right)^t \cdot {M^*}_{V_i \times \mathcal{C}} \cdot v_{\mathcal{C}}\right| \\ &\leq |a| \cdot \|{M^*}_{\mathcal{R} \times \mathcal{C}}\| + \left|(b \cdot u_l)^t \cdot {M^*}_{V_i \times \mathcal{C}} \cdot v_{\mathcal{C}}\right| \\ &\overset{(3.8)}{\leq} \frac{2}{{d_{\mathrm{m}}}^2} \cdot O(1/\overline{w}) + \left|b \cdot \left(u_{l|\mathcal{R}}^t + u_{l|\overline{\mathcal{R}}}^t\right) \cdot {M^*}_{V_i \times \mathcal{C}} \cdot v_{\mathcal{C}}\right| \\ &\leq O(1/(\overline{w} \cdot {d_{\mathrm{m}}}^2)) + \left|u_{l|\mathcal{R}}^t \cdot {M^*}_{V_i \times \mathcal{C}} \cdot v_{\mathcal{C}}\right| + \left\|{M^*}_{\overline{\mathcal{R}} \times \mathcal{C}}\right\| \,.\end{aligned}$$

Demnach ist

$$\left|u^t \cdot {M^*}_{V_i \times V_j} \cdot v\right| = \left|u_{l|\mathcal{R}}^t \cdot {M^*}_{V_i \times \mathcal{C}} \cdot v_{\mathcal{C}}\right| + O({d_{\mathrm{m}}}^{-1.5}) \,.$$

Es gilt $u_{l|\mathcal{R}}^t \cdot {M^*}_{V_i \times \mathcal{C}} \cdot v_{\mathcal{C}} = u_{l\mathcal{R}}^t \cdot {M^*}_{\mathcal{R} \times \mathcal{C}} \cdot v_{\mathcal{C}}$ und $u_{l|\mathcal{R}} \perp \mathbf{1}_{|\mathcal{R}}$, also $u_{l\mathcal{R}} \perp \mathbf{1}$. Beides zusammen zeigt mit Lemma 18 $\left|u_{l|\mathcal{R}}^t \cdot {M^*}_{V_i \times \mathcal{C}} \cdot v_{\mathcal{C}}\right| = O({\overline{w}}^{-1.49})$. Also ist $\left|u^t \cdot {M^*}_{V_i \times V_j} \cdot v\right| = O({\overline{w}}^{-1.49}) + O\left({d_{\mathrm{m}}}^{-1.5}\right)$.

Der Fall $v \perp \mathbf{1}$ und u beliebig kann analog behandelt werden.

3.2.2 Beweis von Lemma 18

Aufgrund der Konstruktion von $\mathcal{R}$ und $\mathcal{C}$ sowie der Wahl von d_{m} gilt $\mathcal{R}, \mathcal{C} \subseteq U$, und damit ${M^*}_{\mathcal{R} \times \mathcal{C}} = M_{\mathcal{R} \times \mathcal{C}}$.

Die direkte Analyse von $M_{\mathcal{R}\times\mathcal{C}}$'s Spektrum ist sehr kompliziert. Da wir – im Gegensatz zu Kapitel 2 – durch die tatsächlichen Grade der Knoten teilen, entstehen sehr starke Abhängigkeiten zwischen den Einträgen.

Wir verfahren daher in zwei Schritten. Zuerst analysieren wir die Matrix $\mathbf{M}$, die entsteht, indem wir durch die erwarteten Grade teilen. Wir nutzen dabei die Erkenntnisse aus Kapitel 2. Wegen der Konstruktion von $\mathcal{R}$ und $\mathcal{C}$ sind die beiden Matrizen $\mathbf{M}_{\mathcal{R}\times\mathcal{C}}$ und $M_{\mathcal{R}\times\mathcal{C}}$ sehr ähnlich zueinander, so dass wir im zweiten Schritt die Ergebnisse von $\mathbf{M}_{\mathcal{R}\times\mathcal{C}}$ auf $M_{\mathcal{R}\times\mathcal{C}}$ übertragen können.

Sei also $\mathbf{M} = (\mathbf{m}_{uv})$ die $|V_i| \times |V_j|$-Matrix mit $\mathbf{m}_{uv} = a_{uv}/(w'_u \cdot w'_v)$. Wir können leicht überprüfen, dass $\mathbf{M}$ eine same-mean-Matrix im Sinne von Definition 8 auf Seite 28 mit Mittelwert

$$\mu = \mathbf{E}[\mathbf{m}_{uv}] = \frac{1}{w'_u \cdot w'_v} \cdot d_{ij} \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot n} \overset{(2.1)}{=} \frac{d_{ij} \cdot W_i \cdot W_j}{\overline{w} \cdot n} = \Theta\left(\frac{1}{\overline{w} \cdot n}\right) \tag{3.9}$$

und Schranke

$$b = \frac{1}{\left(\min_{u\in V} w'_u\right)^2} =: \frac{1}{w'^2_{\mathrm{m}}} \tag{3.10}$$

ist. w'_{m} steht also im Weiteren für den minimalen erwarteten Grad. Wegen der Wahl von d_{m} ist $d_{\mathrm{m}} \leq w'_{\mathrm{m}}$ (vgl. (3.1)).

Lemma 20. *Mit Wahrscheinlichkeit $1 - O(1/n^4)$ hat $\mathbf{M}_{\mathcal{R}\times\mathcal{C}}$ die folgenden drei Eigenschaften*

1. $\mathbf{1}^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} = \mu \cdot |\mathcal{R}| \cdot |\mathcal{C}| \cdot (1 \pm O(1/d_{\mathrm{m}})) = \Theta(n/\overline{w})$.
2. $|u^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v| = O\left(1/w'^{1.5}_{\mathrm{m}}\right)$, *wenn* $\|u\| = \|v\| = 1$ *sowie* $u \perp \mathbf{1}$ *oder* $v \perp \mathbf{1}$.
3. $\|\mathbf{M}_{\mathcal{R}\times\mathcal{C}}\| = \Theta(\mu \cdot \sqrt{|\mathcal{R}| \cdot |\mathcal{C}|}) = \Theta(1/\overline{w})$.

Beweis. Da die gleichen Techniken in Abschnitt 2.4.2 benutzt werden, ist der Beweis dem von Theorem 2 sehr ähnlich. Wir behandeln den Fall $i \neq j$ im Detail. Der symmetrische Fall $i = j$ folgt auf identischem Weg.

Wir beginnen mit Punkt 1. Natürlich ist $\mathbf{1}^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} = s_{\mathbf{M}}(\mathcal{R}, \mathcal{C})$ nach oben durch $Y = s_{\mathbf{M}}(V_i, V_j)$ beschränkt. Wir wenden Lemma 9 mit $a_i = a = 1$, $c = 1$, $D = |V_i| \cdot |V_j|$ und $S = \mathbf{E}[Y]/d_{\mathrm{m}}$ an. Zusammen mit $b \cdot {d_{\mathrm{m}}}^2 = {d_{\mathrm{m}}}^2/w'^2_{\mathrm{m}} \leq 1$ folgt

$$\begin{aligned} \mathbf{Pr}\left[|Y - \mathbf{E}[Y]| \geq \mathbf{E}[Y]/d_{\mathrm{m}}\right] &\leq 2 \cdot \exp\left(-\mathbf{E}[Y]/(2 \cdot \mathrm{e} \cdot b \cdot {d_{\mathrm{m}}}^2)\right) \\ &\leq 2 \cdot \exp(-\mathbf{E}[Y]/6) \ . \end{aligned}$$

Wegen $\mathbf{E}[Y] = |V_i| \cdot |V_j| \cdot \mu = \Theta(n/\overline{w}') = \omega(n^{\varepsilon})$ gilt mit Wahrscheinlichkeit $1 - O(1/n^4)$

$$s_{\mathbf{M}}(V_i, V_j) = \mu \cdot |V_i| \cdot |V_j| \cdot (1 \pm O(1/d_{\mathrm{m}})) \, .$$

Mit (3.4) und (3.5) ergibt sich

$$s_{\mathbf{M}}(\mathcal{R},\mathcal{C}) \leq \mu \cdot |V_i| \cdot |V_j| \cdot (1 + O(1/d_{\mathrm{m}})) = \mu \cdot |\mathcal{R}| \cdot |\mathcal{C}| \cdot (1 + O(1/d_{\mathrm{m}}))\,.$$

Für die untere Schranke an $s_{\mathbf{M}}(\mathcal{R},\mathcal{C})$ benutzen wir

$$s_{\mathbf{M}}(\mathcal{R},\mathcal{C}) \geq s_{\mathbf{M}}(V_i,V_j) - s_{\mathbf{M}}(\overline{\mathcal{R}}\,,V_j) - s_{\mathbf{M}}(V_i,\overline{\mathcal{C}}\,)$$

und zeigen eine obere Schranke an $s_{\mathbf{M}}(\overline{\mathcal{R}}\,,V_j)$ im Detail. Für $s_{\mathbf{M}}(V_i,\overline{\mathcal{C}}\,)$ folgt der gleiche Wert auf analogem Weg.

Da alle Einträge von $\mathbf{M}$ durch b beschränkt sind, gilt $s_{\mathbf{M}}(\overline{\mathcal{R}}\,,V_j) \leq b \cdot s_A(\overline{\mathcal{R}}\,,V_j)$. Laut Lemma 17 ist $\mathrm{Vol}(\overline{\mathcal{R}}\,,V_j) \leq n/{d_{\mathrm{m}}}^4$ relativ klein und damit – zumindest erwartungsgemäß – $s_A(\overline{\mathcal{R}}\,,V_j)$ ebenfalls. Wir zeigen: MhW. gilt für alle Mengen $T \subseteq V_i$ mit $\mathrm{Vol}(T,V_j) \leq n/{d_{\mathrm{m}}}^4$, dass $s_A(T,V_j) < 2n/\overline{w}$ ist.

Halten wir eine solche Menge T fest. Wegen $\mathrm{Vol}(T,V_j) > d_{\mathrm{m}} \cdot |T|$ ist $|T| \leq n/{d_{\mathrm{m}}}^5$ und die Anzahl solcher Mengen T durch

$$\begin{aligned}\binom{n}{n/{d_{\mathrm{m}}}^5} &\leq \big(\mathrm{e}\cdot {d_{\mathrm{m}}}^5\big)^{n/{d_{\mathrm{m}}}^5} \leq \exp\big(\ln(\mathrm{e}\cdot {d_{\mathrm{m}}}^5)\cdot n/{d_{\mathrm{m}}}^5\big)\\ &< \exp\big(n/{d_{\mathrm{m}}}^4\big) \overset{d_{\mathrm{m}}=\Omega(\overline{w}^{2/3})}{<} \exp\big(n/\overline{w}^2\big)\end{aligned}$$

beschränkt.

Fakt 16 gibt die gewünschte Schranke: Wir setzen $X = s_A(T,V_j)$ und $t = n/\overline{w}$. Wegen $\mathbf{E}[X] = \mathrm{Vol}(T,V_j) \leq n/{d_{\mathrm{m}}}^4 < n/\overline{w}^{7/3}$ ist $\mathbf{E}[X] + t/3 < 2t$. Also gilt

$$\mathbf{Pr}\left[s_A(T,V_j) \geq \mathrm{Vol}(T,V_j) + n/\overline{w}\right] \leq \exp\left(-\frac{n^2/\overline{w}^2}{2\cdot 2n/\overline{w}}\right) \leq \exp(-n/(4\overline{w}))\,.$$

Da es nur $\exp(n/\overline{w}^2)$ mögliche Mengen T gibt, gilt mit Wahrscheinlichkeit $> 1-\exp(-n/(5\overline{w})) = 1-O(1/n^4)$ für *alle* Mengen $T \subseteq V_i$ mit $\mathrm{Vol}(T,V_j) < n/{d_{\mathrm{m}}}^4$

$$s_A(T,V_j) < \mathrm{Vol}(T,V_j) + n/\overline{w} < 2n/\overline{w}\,.$$

Ebenso gilt mit der gleichen Wahrscheinlichkeit $s_A(V_i,T) < 2n/\overline{w}$ für alle $T \subseteq V_j$ mit $\mathrm{Vol}(V_i,T) < n/{d_{\mathrm{m}}}^4$. Es ergibt sich

$$\begin{aligned}s_{\mathbf{M}}(\mathcal{R},\mathcal{C}) &\geq s_{\mathbf{M}}(V_i,V_j) - s_{\mathbf{M}}(\overline{\mathcal{R}}\,,V_j) - s_{\mathbf{M}}(V_i,\overline{\mathcal{C}}\,) \geq s_{\mathbf{M}}(V_i,V_j) - b\cdot 4n/\overline{w}\\ &\geq \mu\cdot|V_i|\cdot|V_j|\cdot(1-O(1/d_{\mathrm{m}})) - O(\mu\cdot n^2/{d_{\mathrm{m}}}^2)\\ &\geq \mu\cdot|V_i|\cdot|V_j|\cdot(1-O(1/d_{\mathrm{m}})) \geq \mu\cdot|\mathcal{R}|\cdot|\mathcal{C}|\cdot(1-O(1/d_{\mathrm{m}}))\,.\end{aligned}$$

Also gilt Punkt 1. mit Wahrscheinlichkeit $1 - O(1/n^4)$.

Wir kommen zu Punkt 2. von Lemma 20. Für $u \in \mathcal{R}$ gilt

$$\begin{aligned} s_{\mathbf{M}}(u, V_j) &= \sum_{v \in V_j} \mathbf{m}_{uv} = \sum_{v \in N(u) \cap V_j} \frac{1}{w'_u \cdot w'_v} \leq \frac{|N(u) \cap V_j|}{w'_{\mathrm{m}} \cdot w'_u} = \frac{s_A(u, V_j)}{w'_{\mathrm{m}} \cdot w'_u} \\ &\overset{(3.2)}{\leq} \frac{2 \cdot \mathrm{Vol}(u, V)}{w'_{\mathrm{m}} \cdot w'_u} = \frac{2}{w'_{\mathrm{m}}} = O\left(\frac{\overline{w}}{w'_{\mathrm{m}}} \cdot \frac{1}{\overline{w}}\right) = O\left(\frac{\overline{w}}{w'_{\mathrm{m}}} \cdot n \cdot \mu\right) \\ &\leq K \cdot \mu \cdot (|V_i| + |V_j|) \end{aligned} \tag{3.11}$$

für $K = O(\overline{w}/w'_{\mathrm{m}})$. Die Konstante hinter dem O hängt lediglich von den Modellparametern ε, δ und D ab. Die gleiche Schranke gilt für die Spaltensumme von $v \in \mathcal{C}$. Da $\mathbf{M}$ die Schranke $b = 1/w'^2_{\mathrm{m}} = O(1/\overline{w}^2)$ und den Mittelwert $\mu = O(1/(\overline{w} \cdot n))$ hat, gilt $\mu \cdot |\mathcal{R}| \cdot |\mathcal{C}| > b \cdot (|\mathcal{R}| + |\mathcal{C}|)$. Damit sind die Voraussetzungen für die Anwendung von Lemma 10 auf $\mathbf{M}$ geschaffen.

Sei u ein beliebiger $|\mathcal{R}|$-dimensionaler Vektor. Wir erweitern u zu einem $|V_i|$-dimensionalen Vektor u', indem wir die neuen Koordinaten auf 0 setzen. Das gleiche können wir für jeden $|\mathcal{C}|$-dimensionalen Vektor v machen, um den $|V_j|$-dimensionalen Vektor v' zu erhalten. Es gilt $u^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v = u'^t \cdot \mathbf{M} \cdot v'$.

Wir wenden Lemma 10 mit $d = K$ auf $\mathbf{M}$ an. Jede von 0 verschiedene Koordinate in u' gehört zu $\mathcal{R}$ und wegen (3.11) ebenfalls zu R (aus Lemma 10). Genauso gehört jede von 0 verschiedene Koordinate in v' zu C. Falls $\|u\| = \|v\| = 1$ und $u \perp \mathbf{1}$ oder $v \perp \mathbf{1}$ ist, erhalten wir mit Lemma 10

$$\begin{aligned} \left|u^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v\right| &= \left|u'^t \cdot \mathbf{M} \cdot v'\right| = \left|(u'_{|R})^t \cdot \mathbf{M} \cdot v'_{|C}\right| \\ &= O\left(\sqrt{K \cdot b \cdot \mu \cdot (|V_i| + |V_j|)}\right) \\ &= O\left(\sqrt{\frac{\overline{w}}{w'_{\mathrm{m}}} \cdot \frac{1}{w'^2_{\mathrm{m}}} \cdot \frac{1}{\overline{w} \cdot n} \cdot 2 \cdot \delta \cdot n}\right) \\ &= O(\sqrt{1/w'^3_{\mathrm{m}}}). \end{aligned}$$

Punkt 3. von Lemma 20 ist eine direkte Folgerung aus den beiden ersten Punkten.

□

Wir kommen nun zum zweiten Schritt, der Übertragung von Lemma 20 auf $M_{\mathcal{R}\times\mathcal{C}}$. In obiger Notation (Gleichung (3.9)) haben wir als erstes

$$\mathbf{1}^t \cdot M^*_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} = \mathbf{1}^t \cdot M_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} = \mu \cdot |\mathcal{R}| \cdot |\mathcal{C}| \cdot (1 + O(1/{d_{\mathrm{m}}}^{0.49}))$$

zu zeigen.

Sei D_l die $|\mathcal{R}| \times |\mathcal{R}|$-dimensionale Diagonalmatrix mit den Einträgen (w'_u/d_u) für $u \in \mathcal{R}$ auf der Diagonale. Analog dazu sei D_r für die Knoten in $\mathcal{C}$ definiert. Dann ist

$$M_{\mathcal{R}\times\mathcal{C}} = D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot D_r \,. \tag{3.12}$$

Wegen (3.2) gilt für alle $u \in \mathcal{R}$, dass $|d_u - w'_u| \leq 2 \cdot w'^{0.51}_u$ ist. Also erfüllen die Diagonaleinträge von D_l

$$\begin{aligned}
\frac{w'_u}{w'_u + 2 \cdot w'^{0.51}_u} \leq \quad & \frac{w'_u}{d_u} \quad \leq \frac{w'_u}{w'_u - 2 \cdot w'^{0.51}_u} \\
1 - \frac{2 \cdot w'^{-0.49}_u}{1 + 2 \cdot w'^{-0.49}_u} \leq \quad & \frac{w'_u}{d_u} \quad \leq 1 + \frac{2 \cdot w'^{-0.49}_u}{1 - 2 \cdot w'^{-0.49}_u} \\
1 - 2 \cdot w'^{-0.49}_u \leq \quad & \frac{w'_u}{d_u} \quad \leq 1 + 3 \cdot w'^{-0.49}_u \,.
\end{aligned}$$

Beachte: $\|D_l\| \leq 1 + 3 \cdot w'^{-0.49}_{\mathrm{m}} \leq 2$. Wir zerlegen $\mathbf{1}^t \cdot D_l$ in

$$\mathbf{1}^t \cdot D_l = a_l \cdot \mathbf{1}^t + u_l^t \qquad \text{mit } u_l \perp \mathbf{1}.$$

Wir erhalten $1 - 2 \cdot w'^{-0.49}_{\mathrm{m}} \leq a_l \leq 1 + 3 \cdot w'^{-0.49}_{\mathrm{m}}$ und $\|u_l\| \leq \sqrt{|\mathcal{R}|} \cdot 5 \cdot w'^{-0.49}_{\mathrm{m}}$.

Analoges gilt für D_r: $\|D_l\| \leq 2$ und für $D_r \cdot \mathbf{1} = a_r \cdot \mathbf{1} + v_r$ mit $v_r \perp \mathbf{1}$ haben wir $1 - 2 \cdot w'^{-0.49}_{\mathrm{m}} \leq a_r \leq 1 + 3 \cdot w'^{-0.49}_{\mathrm{m}}$ und $\|v_r\| \leq \sqrt{|\mathcal{C}|} \cdot 5 \cdot w'^{-0.49}_{\mathrm{m}}$.

Mit diesen Zerlegungen für $\mathbf{1}^t \cdot D_l$ und $D_r \cdot \mathbf{1}$ erhalten wir

$$\begin{aligned}
\mathbf{1}^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot D_r \cdot \mathbf{1} = a_l \cdot a_r \cdot \mathbf{1}^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} + a_l \cdot \mathbf{1}^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v_r + \\
a_r \cdot u_l^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} + u_l^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v_r \,.
\end{aligned} \tag{3.13}$$

Lemma 20 liefert

$$\begin{aligned}
a_l \cdot a_r \cdot \mathbf{1}^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} &= \left(1 \pm O\left(w'^{-0.49}_{\mathrm{m}}\right)\right)^2 \cdot \mu \cdot |\mathcal{R}| \cdot |\mathcal{C}| \cdot (1 + O(1/d_{\mathrm{m}})) \\
&= \mu \cdot |\mathcal{R}| \cdot |\mathcal{C}| \cdot \left(1 \pm O\left(w'^{-0.49}_{\mathrm{m}}\right)\right)
\end{aligned} \tag{3.14}$$

und

$$\begin{aligned}
\left|a_l \cdot \mathbf{1}^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v_r\right| &= a_l \cdot \sqrt{|\mathcal{R}|} \cdot \|v_r\| \cdot \left| \frac{\mathbf{1}^t}{\|\mathbf{1}^t\|} \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \frac{v_r}{\|v_r\|} \right| \\
&\overset{v_r \perp \mathbf{1}}{\leq} 2 \cdot \sqrt{|\mathcal{R}|} \cdot \left(\sqrt{|\mathcal{C}|} \cdot 5 \cdot w'^{-0.49}_{\mathrm{m}}\right) \cdot O\left(1/w'^{1.5}_{\mathrm{m}}\right) \\
&= O\left(\sqrt{|\mathcal{R}| \cdot |\mathcal{C}|} \cdot w'^{-1.99}_{\mathrm{m}}\right) \\
&= O\left(\mu \cdot |\mathcal{R}| \cdot |\mathcal{C}| \cdot w'^{-0.99}_{\mathrm{m}}\right)
\end{aligned} \tag{3.15}$$

wegen $\mu = \Theta(1/(\overline{w} \cdot n)) = \Theta(1/(w'_{\mathrm{m}} \cdot n))$ und $|\mathcal{R}|, |\mathcal{C}| = \Theta(n)$. Dieselbe Schranke (3.15) erhalten wir für $|a_r \cdot u_l^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1}|$ und $|u_l^t \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v_r|$. Also dominiert (3.14) die Gleichung (3.13) und es folgt

$$\begin{aligned}
\mathbf{1}^t \cdot M_{\mathcal{R}\times\mathcal{C}} \cdot \mathbf{1} &\overset{(3.12)}{=} \mathbf{1}^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot D_r \cdot \mathbf{1} \\
&\overset{(3.13)}{=} \mu \cdot |\mathcal{R}| \cdot |\mathcal{C}| \cdot \left(1 \pm O\left(w'^{-0.49}_{\mathrm{m}}\right)\right) \\
&\overset{(3.9)}{=} d_{ij} \cdot W_i \cdot W_j \cdot \frac{|\mathcal{R}| \cdot |\mathcal{C}|}{\overline{w} \cdot n} \cdot (1 \pm O(w'^{-0.49}_{\mathrm{m}})). \\
&= d_{ij} \cdot W_i \cdot W_j \cdot \frac{|\mathcal{R}| \cdot |\mathcal{C}|}{\overline{w} \cdot n} \cdot (1 \pm O(1/{d_{\mathrm{m}}}^{0.49})).
\end{aligned}$$

Nachdem Punkt 1. bewiesen ist, kommen wir zum zweiten Punkt in Lemma 18. Sei $v \perp \mathbf{1}$ ein Einheitsvektor der Dimension $|\mathcal{C}|$. Es gibt eine eindeutige Zerlegung $D_r \cdot v = c \cdot \mathbf{1}/\left\|\mathbf{1}\right\| + v'$ mit $v' \perp \mathbf{1}$, wobei $\|v'\| \leq \|D_r\| \leq 2$ ist.

Mit $\mathbf{1}^t \cdot D_r = (D_r \cdot \mathbf{1})^t = a_r \cdot \mathbf{1}^t + v_r^t$ ergibt sich

$$\begin{aligned}
c = \frac{\mathbf{1}^t}{\sqrt{|\mathcal{C}|}} \cdot D_r \cdot v = \frac{(a_r \cdot \mathbf{1}^t + v_r^t) \cdot v}{\sqrt{|\mathcal{C}|}} &\overset{v\perp\mathbf{1}}{=} \frac{v_r^t \cdot v}{\sqrt{|\mathcal{C}|}} \leq \frac{\|v_r\| \cdot \|v\|}{\sqrt{|\mathcal{C}|}} \\
&\leq 5 \cdot {w'_{\mathrm{m}}}^{-0.49}.
\end{aligned} \tag{3.16}$$

So erhalten wir für jeden Einheitsvektor u

$$\begin{aligned}
\left|u^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot D_r \cdot v\right| &= \left|u^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \left(c \cdot \frac{\mathbf{1}}{\|\mathbf{1}\|} + v'\right)\right| \\
&\leq \|D_l\| \cdot \|\mathbf{M}_{\mathcal{R}\times\mathcal{C}}\| \cdot \left\|c \cdot \frac{\mathbf{1}}{\|\mathbf{1}\|}\right\| + \left|u^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v'\right| \\
&\leq 2 \cdot O(1/\overline{w}') \cdot |c| + \left|u^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v'\right|.
\end{aligned}$$

Sei $u' = u^t \cdot D_l$, dann ist $\|u'\| \leq \|D_l\| \leq 2$. Wir sehen mit $v' \perp \mathbf{1}$ und $\|v'\| \leq 2$, sowie Lemma 20

$$\left|u^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot v'\right| = \|u'\| \cdot \|v'\| \cdot \left|\frac{u'}{\|u'\|} \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot \frac{v'}{\|v'\|}\right| \leq 4 \cdot O\left({w'_{\mathrm{m}}}^{-1.5}\right).$$

Damit ist

$$\begin{aligned}
\left|u^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot D_r \cdot v\right| &\leq 2 \cdot O(1/\overline{w}') \cdot |c| + O\left({w'_{\mathrm{m}}}^{-1.5}\right) \\
&\overset{(3.16)}{\leq} 10 \cdot {w'_{\mathrm{m}}}^{-0.49} \cdot O(1/\overline{w}) + O\left({w'_{\mathrm{m}}}^{-1.5}\right) \\
&= O(1/\overline{w}^{1.49}).
\end{aligned}$$

Mit (3.12) folgt

$$\left|u^t \cdot {M^*}_{\mathcal{R}\times\mathcal{C}} \cdot v\right| = \left|u^t \cdot D_l \cdot \mathbf{M}_{\mathcal{R}\times\mathcal{C}} \cdot D_r \cdot v\right| = O\left(1/\overline{w}^{1.49}\right).$$

Die gleiche Schranke erhalten wir für Einheitsvektoren $u \perp \mathbf{1}$ und v beliebig. Punkt 3. des Lemmas ist eine unmittelbare Folge von 1. und 2. □

3.2.3 Beweis von Lemma 19

Wir verzichten auf den Beweis von $\left\|{M^*}_{\mathcal{R}\times\overline{\mathcal{C}}}\right\| = O({d_\mathrm{m}}^{-1.5})$, der dem von $\left\|{M^*}_{\overline{\mathcal{R}}\times\mathcal{C}}\right\| = O({d_\mathrm{m}}^{-1.5})$ sehr ähnlich ist.

Die Norm von ${M^*}_{\overline{\mathcal{R}}\times\mathcal{C}}$

Sei ξ ein $|\mathcal{C}|$-dimensionaler Vektor mit $\|\xi\| \leq 1$. Wir zeigen, dass $\eta = {M^*}_{\overline{\mathcal{R}}\times\mathcal{C}} \cdot \xi$ eine l_2-Norm von $O({d_\mathrm{m}}^{-1.5})$ hat. Aufgrund der Konstruktion von M^* gilt $\eta_u = 0$ für $u \notin U$ und für $u \in U$ gilt

$$\eta_u = \sum_{v\in\mathcal{C}} \frac{a_{uv}}{d_u \cdot d_v} \cdot \xi_v = \sum_{v\in N(u)\cap\mathcal{C}\cap U} \frac{\xi_v}{d_u \cdot d_v}.$$

Somit ist

$$\|\eta\|^2 = \sum_{u\in\overline{\mathcal{R}}\cap U} \left(\sum_{v\in N(u)\cap\mathcal{C}\cap U} \frac{\xi_v}{d_u \cdot d_v} \right)^2.$$

Wir benutzen jetzt die Cauchy-Schwartz-Ungleichung in der Form[1]

$$\left(\sum a_i \cdot b_i\right)^2 \leq \sum a_i^2 \cdot \sum b_i^2$$

für $a_i \cdot b_i = \frac{\xi_v}{\sqrt{d_u \cdot d_v}} \cdot \frac{1}{\sqrt{d_u \cdot d_v}}$. So erhalten wir

$$\begin{aligned}
\|\eta\|^2 &\leq \sum_{u\in\overline{\mathcal{R}}\cap U} \left(\sum_{v\in N(u)\cap\mathcal{C}\cap U} \frac{\xi_v^2}{d_u \cdot d_v} \right) \cdot \left(\sum_{v\in N(u)\cap\mathcal{C}\cap U} \frac{1}{d_u \cdot d_v} \right) \\
&\overset{u,v\in U}{\leq} \sum_{u\in\overline{\mathcal{R}}\cap U} \left(\sum_{v\in N(u)\cap\mathcal{C}\cap U} \frac{\xi_v^2}{d_\mathrm{m} \cdot d_v} \right) \cdot \left(\sum_{v\in N(u)\cap\mathcal{C}\cap U} \frac{1}{d_u \cdot d_\mathrm{m}} \right) \\
&\leq \frac{1}{{d_\mathrm{m}}^2} \cdot \sum_{u\in\overline{\mathcal{R}}\cap U} \sum_{v\in N(u)\cap\mathcal{C}\cap U} \frac{\xi_v^2}{d_v}.
\end{aligned}$$

[1] Vgl. (1.117b) in [BSMM05].

Weil $v \in \mathcal{C}$ ist, hat v maximal $\mathrm{Vol}(V_i, v) \cdot D/d_\mathrm{m}$ Nachbarn in $\overline{\mathcal{R}}$. Außerdem gilt $\mathrm{Vol}(V_i, v) \leq \mathrm{Vol}(V, v) < 2d_v$, anderenfalls wäre (3.2) falsch. Also wird jeder Term ξ_v/d_v für $v \in \mathcal{C}$ maximal $(2d_v \cdot D/d_\mathrm{m})$-mal in obiger Doppelsumme gezählt. Wir erhalten

$$\|\eta\|^2 \leq \frac{1}{{d_\mathrm{m}}^2} \cdot \sum_{v \in \mathcal{C}} 2d_v \cdot \frac{D}{d_\mathrm{m}} \cdot \frac{\xi_v^2}{d_v} \leq \frac{2D}{{d_\mathrm{m}}^3} \cdot \sum_{v \in \mathcal{C}} \xi_v^2 \leq \frac{2D}{{d_\mathrm{m}}^3}$$

und als unmittelbare Konsequenz $\left\|{M^*}_{\overline{\mathcal{R}} \times \mathcal{C}} \cdot \xi\right\| = \|\eta\| = O({d_\mathrm{m}}^{-1.5})$ für alle ξ mit $\|\xi\| \leq 1$ und damit die Behauptung $\left\|{M^*}_{\overline{\mathcal{R}} \times \mathcal{C}}\right\| = O({d_\mathrm{m}}^{-1.5})$.

Die Norm von ${M^*}_{\overline{\mathcal{R}} \times \overline{\mathcal{C}}}$

Lemma 21. *Sei $G = (V, E)$ ein Graph, der in unserem Modell generiert wurde und U die Menge, die von Algorithmus 13 erzeugt wurde. Dann gilt mit Wahrscheinlichkeit $1 - O(1/n^4)$ für alle Mengen $U_1 \subseteq U \cap V_i$ und $U_2 \subseteq U \cap V_j$ mit $\mathrm{Vol}(U_1, V_j) \leq n/{d_\mathrm{m}}^4$ und $\mathrm{Vol}(V_i, U_2) \leq n/{d_\mathrm{m}}^4$:*

Es gibt Zerlegungen

$$U_1 = U_1^1 \dot\cup U_1^2 \dot\cup \ldots \dot\cup U_1^l \quad \textit{und} \quad U_2 = U_2^1 \dot\cup U_2^2 \dot\cup \ldots \dot\cup U_2^l\,,$$

so dass für alle $m = 1, \ldots, l$, alle $u \in U_1^m$ und alle $v \in U_2^m$ gleichzeitig

$$\sum_{m' \geq m}^{l} s_A(u, U_2^{m'}) \leq d_u \cdot \frac{D}{d_\mathrm{m}} \qquad \textit{und} \qquad \sum_{m' \geq m} s_A(U_1^{m'}, v) \leq d_v \cdot \frac{D}{d_\mathrm{m}}$$

gilt, solange $D = O(1)$ groß genug ist.

Wir beweisen Lemma 21 am Ende dieses Abschnitts.

Wegen Lemma 17 können wir Lemma 21 auf $U_1 = U \cap \overline{\mathcal{R}}$ und $U_2 = U \cap \overline{\mathcal{C}}$ anwenden. Seien ${\overline{\mathcal{R}}}^1 \dot\cup \ldots \dot\cup {\overline{\mathcal{R}}}^l$ $(= U \cap \overline{\mathcal{R}})$ und ${\overline{\mathcal{C}}}^1 \dot\cup \ldots \dot\cup {\overline{\mathcal{C}}}^l$ $(= U \cap \overline{\mathcal{C}})$ die Zerlegungen, die in Lemma 21 beschrieben sind.

Um die Lesbarkeit zu erhöhen, schreiben wir ${\overline{\mathcal{R}}}^{\geq m}$ für $\bigcup_{m' \geq m} {\overline{\mathcal{R}}}^{m'}$, sowie analog ${\overline{\mathcal{R}}}^{<m}$, ${\overline{\mathcal{C}}}^{\geq m}$ und ${\overline{\mathcal{C}}}^{<m}$. In dieser Notation sichert uns Lemma 21 für $u \in {\overline{\mathcal{R}}}^m$ und $v \in {\overline{\mathcal{C}}}^m$

$$\begin{aligned} &\left|\left\{w \in N(u) \cap {\overline{\mathcal{C}}}^{\geq m}\right\}\right| = \sum_{m' \geq m}^{l} s_A(u, {\overline{\mathcal{C}}}^{m'}) < d_u \cdot D/d_\mathrm{m} \\ \text{und} \qquad &\left|\left\{w \in N(v) \cap {\overline{\mathcal{R}}}^{\geq m}\right\}\right| < d_v \cdot D/d_\mathrm{m}\,. \end{aligned} \tag{3.17}$$

Sei ξ ein $\left|\overline{\mathcal{C}}\right|$-dimensionaler Vektor mit $\|\xi\| \leq 1$. Sei weiterhin $\eta = {M^*}_{\overline{\mathcal{R}}\times\overline{\mathcal{C}}} \cdot \xi$ und $u \in \overline{\mathcal{R}}$ beliebig. Falls $u \notin U$ ist, ist wegen M^*'s Konstruktion $\eta_u = 0$. Anderenfalls gibt es ein m, so dass $u \in \overline{\mathcal{R}}^{\,m}$.

Jeder Eintrag m_{uv} von ${M^*}_{\overline{\mathcal{R}}\times\overline{\mathcal{C}}}$ mit $v \in \overline{\mathcal{C}} \setminus U$ ist von Algorithmus 13 auf 0 gesetzt worden. Da also solche Einträge keinen Beitrag zu η_u leisten, werden wir sie im Weiteren ignorieren. Für jeden von 0 verschiedenen Eintrag m_{uv} haben wir $v \in \overline{\mathcal{C}}^{\,<m}$ oder $v \in \overline{\mathcal{C}}^{\,\geq m}$. Es gilt

$$\eta_u = \sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,<m}} \frac{\xi_v}{d_u \cdot d_v} + \sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,\geq m}} \frac{\xi_v}{d_u \cdot d_v}$$

und wegen $(a+b)^2 \leq 2(a^2+b^2)$

$$\begin{aligned}\|\eta\|^2 = \sum_{u \in \overline{\mathcal{R}} \cap U} \eta_u^2 \leq\ & 2 \cdot \sum_{m=1}^{l} \sum_{u \in \overline{\mathcal{R}}^{\,m}} \left(\sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,<m}} \frac{\xi_v}{d_u \cdot d_v} \right)^2 \\ & + 2 \cdot \sum_{m=1}^{l} \sum_{u \in \overline{\mathcal{R}}^{\,m}} \left(\sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,\geq m}} \frac{\xi_v}{d_u \cdot d_v} \right)^2 . \quad (3.18)\end{aligned}$$

Wir beschränken den ersten Summanden ausführlich. Wegen $u \in \overline{\mathcal{R}}^{\,m} \subseteq U$ gilt $d_u \geq d_{\mathrm{m}}$. Mithilfe der Cauchy-Schwartz-Ungleichung folgern wir

$$\begin{aligned} & 2 \cdot \sum_{m=1}^{l} \sum_{u \in \overline{\mathcal{R}}^{\,m}} \left(\sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,<m}} \frac{\xi_v}{d_u \cdot d_v} \right)^2 \\ & \quad \leq\ 2 \cdot \sum_{m=1}^{l} \sum_{u \in \overline{\mathcal{R}}^{\,m}} \left(\sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,<m}} \frac{\xi_v^2}{d_u \cdot d_v} \right) \cdot \left(\sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,<m}} \frac{1}{d_u \cdot d_v} \right) \\ & \quad \leq\ \frac{2}{{d_{\mathrm{m}}}^2} \sum_{m=1}^{l} \sum_{u \in \overline{\mathcal{R}}^{\,m}} \sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,<m}} \frac{\xi_v^2}{d_v} . \end{aligned}$$

Beachte: Die Situation $u \in \overline{\mathcal{R}}^{\,m}$ und $v \in N(u) \cap \overline{\mathcal{C}}^{\,<m}$ gleicht $v \in \overline{\mathcal{C}}^{\,m'}$ und $u \in N(v) \cap \overline{\mathcal{R}}^{\,>m'}$ für ein $m' < m$. Also wird ξ_v^2/d_v für $v \in \overline{\mathcal{C}}^{\,m'}$ maximal $\left|N(v) \cap \overline{\mathcal{R}}^{\,>m'}\right| < d_v \cdot D/d_{\mathrm{m}}$ oft in der obigen Summe gezählt, siehe (3.17). Es folgt

$$\frac{2}{{d_{\mathrm{m}}}^2} \cdot \sum_{m=1}^{l} \sum_{u \in \overline{\mathcal{R}}^{\,m}} \sum_{v \in N(u) \cap \overline{\mathcal{C}}^{\,<m}} \frac{\xi_v^2}{d_v} \leq \frac{2}{{d_{\mathrm{m}}}^2} \cdot \sum_{v \in \overline{\mathcal{C}}} \xi_v^2 \cdot \frac{D}{d_{\mathrm{m}}} \leq \frac{2D}{{d_{\mathrm{m}}}^3} .$$

Die gleiche Schranke erhalten wir analog für den zweiten Summanden in (3.18). Damit ist $\|\eta\|^2 \leq 4D/{d_{\mathrm{m}}}^3$, was $\left\|{M^*}_{\overline{\mathcal{R}}\times\overline{\mathcal{C}}}\right\| = O({d_{\mathrm{m}}}^{-1.5})$ zeigt.

Beweis von Lemma 21

Das folgende Lemma ist das Analogon zu Lemma 11 in Abschnitt 2.4.4, jedoch für die Adjazenzmatrix. Wir halten den Beweis recht knapp, da er sehr ähnlich zu dem von Lemma 11 ist.

Lemma 22. *Sei $G = (V, E)$ ein Graph, der in unserem Modell generiert wurde und A seine Adjazenzmatrix. Mit Wahrscheinlichkeit $1 - O(1/n^4)$ gilt für jedes Paar (U, U') mit $U, U' \subseteq V$*

1. $s_A(U, U') \leq 20 \cdot \mathrm{Vol}(U, U')$ *oder*
2. $s_A(U, U') \cdot \ln(s_A(U, U')/\mathrm{Vol}(U, U')) \leq 20 \cdot u \cdot \ln(n/u)$,

wobei $u = \max\{|U|, |U'|\} \leq n/2$ ist.

Beweis. Seien U, U' fest. Dann ist $\eta = \mathrm{Vol}(U, U')$ der Erwartungswert von $s_A(U, U')$. Es gibt ein eindeutiges β mit $\beta \cdot \ln \beta = 20 \cdot u \cdot \ln(n/u)/\eta$. Wir setzen $\beta' = \max\{20, \beta\}$. Um die Behauptung zu zeigen, schätzen wir die Wahrscheinlichkeit ab, dass $s_A(U, U') \geq \beta' \cdot \eta$ ist.

Mithilfe von Fakt 16 lässt sich zeigen:

$$\begin{aligned}
\mathbf{Pr}\left[s_A(U,U') \geq \beta' \cdot \eta\right] &= \mathbf{Pr}\left[s_A(U,U') \geq \mathbf{E}\left[s_A(U,U')\right] + (\beta'-1)\cdot\eta\right] \\
&\leq \exp(-\eta \cdot \phi\left(\beta'-1\right)) \\
&= \exp(-\eta \cdot (\beta' \cdot \ln\beta' - (\beta'-1))) \\
&\overset{\beta'\geq 20}{<} \exp(-\eta \cdot \beta' \cdot \ln\beta' \cdot 2/3) \\
&= \exp(-40/3 \cdot u \cdot \ln(n/u)) < \left(\frac{n}{u}\right)^{-13u}.
\end{aligned}$$

Da es nur $2 \cdot u \cdot \binom{n}{u}^2 \leq 2u \cdot (\mathrm{e} \cdot n/u)^{2u}$ verschiedene Mengenpaare U, U' mit $\max\{|U|, |U'|\} = u \leq n/2$ gibt, ist die Gesamtwahrscheinlichkeit kleiner als

$$2u \cdot \left(\frac{\mathrm{e}\cdot n}{u}\right)^{2u} \cdot \left(\frac{n}{u}\right)^{-13u} < \left(\frac{n}{u}\right)^{-9u}.$$

Summieren wir den letzten Term über alle $u = 1, \ldots, n/2$ ergibt sich $O(1/n^4)$. □

Im Weiteren nehmen wir an, dass Lemma 22 für den gegebenen Graphen gilt. Der folgende Prozess generiert eine Partition, wie sie von Lemma 21 vorausgesagt wird.

1. Setze $U_1' := U_1,\ \ U_2' := U_2$.
2. Setze $T_1 := T_2 := \emptyset$.
3. Solange es $u \in U_1'$ mit $s_A(u, U_2') \leq d_u \cdot D/d_{\mathrm{m}}$ gibt, füge u zu T_1 hinzu.
4. Solange es $v \in U_2'$ mit $s_A(U_1', v) \leq d_v \cdot D/d_{\mathrm{m}}$ gibt, füge v zu T_2 hinzu.
5. Setze $U_1' := U_1' \setminus T_1,\ \ U_2' := U_2' \setminus T_2$.
6. Wiederhole die Schritte 2. – 5. bis $|U_1'| = |U_2'| = 0$.

Wir beweisen Lemma 21, indem wir zeigen, dass obige Prozedur mit Wahrscheinlichkeit $1 - O(1/n^4)$ terminiert. Um einen Widerspruch zu erzeugen, nehmen wir das Gegenteil an: Es gibt ein Zeitpunkt, so dass in Schritt 6. $T_1 = T_2 = \emptyset$, wohingegen weder U_1' noch U_2' leer sind.

Zu diesem Zeitpunkt haben wir für jedes $u \in U_1'$

$$s_A(u, U_2') > d_u \cdot D/d_{\mathrm{m}} = s_A(u, V) \cdot D/d_{\mathrm{m}}$$

und $s_A(U_1', v) > s_A(V, v) \cdot D/d_{\mathrm{m}}$ für alle $v \in U_2'$. Wir erhalten

$$s_A(U_1', U_2') \geq \max\{s_A(U_1', V), s_A(V, U_2')\} \cdot \frac{D}{d_{\mathrm{m}}} \tag{3.19}$$

$$\overset{U_1', U_2' \subseteq U}{\geq} D \cdot \max\{|U_1'|, |U_2'|\}. \tag{3.20}$$

Wegen $n/{d_{\mathrm{m}}}^4 \geq \mathrm{Vol}(U_1, V_j) \geq \mathrm{Vol}(U_1', V_j) \geq |U_1'| \cdot d_{\mathrm{m}}$ gilt $|U_1'| \leq n/{d_{\mathrm{m}}}^5$ und analog $|U_2'| \leq n/{d_{\mathrm{m}}}^5$.

Sei nun $r = \max\{|U_1'|, |U_2'|\}$ und $s = s_A(U_1', U_2')$. Dann gilt also $r \leq n/{d_{\mathrm{m}}}^5$ und wegen (3.20) $s \geq D \cdot r$. Wir unterscheiden einige Fälle, um die Existenz von U_1', U_2' mit obigen Eigenschaften zu widerlegen.

1. $\mathrm{Vol}(U_1', U_2') \leq r^{1.5}/\sqrt{n}$.
   (a) Die erste Ungleichung von Lemma 22 gilt. Dann ist wegen $r < n$

   $$20r > 20 \cdot r^{1.5}/\sqrt{n} \geq 20 \cdot \mathrm{Vol}(U_1', U_2') \geq s \overset{(3.20)}{\geq} D \cdot r,$$

   was für $D \geq 20$ falsch ist.
   (b) Die zweite Ungleichung von Lemma 22 gilt. Dann ist

   $$20 \cdot r \cdot \ln(n/r) \geq s \cdot \ln(s/\mathrm{Vol}(U_1', U_2')).$$

   Die rechte Seite ist für $s > \mathrm{Vol}(U_1', U_2')$ monoton wachsend in s, was wegen $s \geq D \cdot r > D \cdot r^{1.5}/n^{1.5} \geq D \cdot \mathrm{Vol}(U_1', U_2')$ gilt. Wir ersetzen s durch $D \cdot r$, um die rechte Seite nach unten zu beschränken.

Für $D \geq 40$ entsteht dabei ebenfalls ein Widerspruch:

$$\begin{aligned} 20 \cdot r \cdot \ln(n/r) &\geq D \cdot r \cdot \ln\left(\frac{D \cdot r}{\mathrm{Vol}(U_1', U_2')}\right) \\ &\geq D \cdot r \cdot \ln \frac{D \cdot r}{r^{1.5}/\sqrt{n}} = \frac{D \cdot r}{2} \cdot \ln(D^2 \cdot n/r)\,. \end{aligned}$$

2. $\mathrm{Vol}(U_1', U_2') > r^{1.5}/\sqrt{n}$ und $s \geq \sqrt{\mathrm{Vol}(U_1', U_2') \cdot n}/d_{\mathrm{m}}$.
   (a) Ungleichung 1. von Lemma 22 gilt. Dann ist
   $$20 \cdot \mathrm{Vol}(U_1', U_2') \geq s \geq \sqrt{\mathrm{Vol}(U_1', U_2') \cdot n}/d_{\mathrm{m}}$$
   und
   $$\mathrm{Vol}(U_1', U_2') \geq n/(400 \cdot {d_{\mathrm{m}}}^2)\,.$$
   Da wir $d_{\mathrm{m}} > 20$ annehmen können, ist dies wegen $n/{d_{\mathrm{m}}}^4 \geq \mathrm{Vol}(U_1', U_2')$ falsch.
   (b) Ungleichung 2. von Lemma 22 gilt
   $$20 \cdot r \cdot \ln(n/r) \geq s \cdot \ln\left(s/\mathrm{Vol}(U_1', U_2')\right)\,.$$
   Die rechte Seite schätzen wir wie in Fall 1. (b) nach unten mithilfe von $s \geq \sqrt{\mathrm{Vol}(U_1', U_2') \cdot n}/d_{\mathrm{m}} \geq \mathrm{Vol}(U_1', U_2')$ ab.
   Also ist
   $$20 \cdot r \cdot \ln(n/r) \geq \frac{\sqrt{\mathrm{Vol}(U_1', U_2') \cdot n}}{d_{\mathrm{m}}} \cdot \ln \frac{\sqrt{n}}{d_{\mathrm{m}} \cdot \sqrt{\mathrm{Vol}(U_1', U_2')}}\,.$$
   Die rechte Seite ist monoton wachsend in $\mathrm{Vol}(U_1', U_2')$, solange $\mathrm{Vol}(U_1', U_2') \leq n/(\mathrm{e} \cdot d_{\mathrm{m}})^2$ ist (vgl. A.4). Wegen $\mathrm{Vol}(U_1', U_2') \leq n/{d_{\mathrm{m}}}^4$ ist dies erfüllt und wir können $\mathrm{Vol}(U_1', U_2')$ durch die untere Schranke $r^{1.5}/\sqrt{n}$ ersetzen und erhalten
   $$20 \cdot r \cdot \ln(n/r) \geq \frac{r^{0.75} \cdot n^{0.25}}{d_{\mathrm{m}}} \cdot \ln\left(\frac{n^{0.75}}{d_{\mathrm{m}} \cdot r^{0.75}}\right)\,.$$
   Wegen $r \leq n/{d_{\mathrm{m}}}^5$ ist das Argument des „ln“ auf der rechten Seite > 1 und die gesamte rechte Seite monoton wachsend in n. Wir verkleinern diesen Term, indem wir die untere Schranke $r \cdot {d_{\mathrm{m}}}^5$ an n einsetzen:
   $$\begin{aligned} 20 \cdot r \cdot \ln(n/r) &\geq \frac{r^{0.75} \cdot (r \cdot {d_{\mathrm{m}}}^5)^{0.25}}{d_{\mathrm{m}}} \cdot \ln\left(\sqrt{\frac{n}{r}} \cdot \frac{(r \cdot {d_{\mathrm{m}}}^5)^{0.25}}{d_{\mathrm{m}} \cdot r^{0.25}}\right) \\ &= r \cdot {d_{\mathrm{m}}}^{0.25} \cdot \ln\left(\sqrt{\frac{n}{r}} \cdot {d_{\mathrm{m}}}^{0.25}\right) \\ &= \frac{{d_{\mathrm{m}}}^{0.25}}{2} \cdot r \cdot \ln\left(\frac{n}{r} \cdot \sqrt{d_{\mathrm{m}}}\right)\,. \end{aligned}$$

Wieder erhalten wir einen Widerspruch, wenn d_{m} ausreichend groß ist.

3. $\mathrm{Vol}(U_1', U_2') > r^{1.5}/\sqrt{n}$ und $s < \sqrt{\mathrm{Vol}(U_1', U_2') \cdot n}/d_{\mathrm{m}}$.
Man kann leicht nachrechnen, dass

$$\mathrm{Vol}(U_1', U_2') = \frac{\mathrm{Vol}(U_1', V_j) \cdot \mathrm{Vol}(V_i, U_2')}{\mathrm{Vol}(V_i, V_j)}$$

gilt. Angenommen $\mathrm{Vol}(U_1', V_j)$ ist der größere Faktor im Zähler, also

$$\mathrm{Vol}(U_1', V_j) \geq \sqrt{\mathrm{Vol}(U_1', U_2') \cdot \mathrm{Vol}(V_i, V_j)}\,.$$

(Im Fall $\mathrm{Vol}(U_1', V_j) \leq \mathrm{Vol}(V_i, U_2')$ ist der Beweis analog.) Wegen der Annahme $\mathrm{Vol}(U_1', U_2') > r^{1.5}/\sqrt{n}$ haben wir

$$\begin{aligned} \mathrm{Vol}(U_1', V_j) &\geq \sqrt{\mathrm{Vol}(U_1', U_2') \cdot \mathrm{Vol}(V_i, V_j)} \qquad (3.21) \\ &> \sqrt{\frac{r^{1.5}}{\sqrt{n}} \cdot (d_{\mathrm{m}} \cdot \delta n)} > r^{0.75} \cdot n^{0.25} \geq n^{0.25} \cdot |U_1'|^{0.75}\,. \end{aligned}$$

Da das Volumen zwischen U_1' und V_j relativ groß ist, sind mhW. viele Kanten zwischen U_1' und V_j: Sei $W \subseteq V_i$ eine feste Menge mit $w = |W| \leq n/{d_{\mathrm{m}}}^5$ und $\mathrm{Vol}(W, V_j) \geq n^{0.25} \cdot |w|^{0.75}$. Fakt 16 zeigt

$$\begin{aligned} \mathbf{Pr}\left[s_A(W, V_j) \leq \mathrm{Vol}(W, V_j)/2\right] &\leq \exp(-\mathrm{Vol}(W, V_j)/8) \\ &\leq \exp\bigl(-w^{0.75} \cdot n^{0.25}/8\bigr)\,. \end{aligned}$$

Es gibt nur $\binom{n}{w} \leq (\mathrm{e}{\cdot}n/w)^w$ mögliche Mengen W wie oben mit $|W| = w$. Die Gesamtwahrscheinlichkeit für die Existenz einer solchen Menge W mit „wenigen“ Kanten zwischen W und V_j ist dann

$$\exp\bigl(-w^{0.75} \cdot n^{0.25}/8 + w \cdot \ln(\mathrm{e} \cdot n/w)\bigr)\,.$$

Für $0 < w \leq n/{d_{\mathrm{m}}}^5$ ist der Exponent konvex in w, vorausgesetzt d_{m} ist groß genug. Indem wir die Intervallgrenzen $w = 1$ und $w = n/{d_{\mathrm{m}}}^5$ überprüfen, sehen wir, dass der Exponent höchstens $-n^{0.25}/10$ ist. (Beachte, dass $d_{\mathrm{m}} \leq \sqrt[5]{n}$ gilt.)
Summieren wir über alle w, erhalten wir die Gesamtwahrscheinlichkeit von $1 - w \cdot \exp(-n^{0.25}/10) = 1 - O(1/n^4)$, dass für alle betrachteten W (inklusive $W = U_1'$) $s_A(W, V_j) \geq \mathrm{Vol}(W, V_j)/2$ gilt. Wir haben also

$$s_A(U_1', V) \geq s_A(U_1', V_j) \geq \mathrm{Vol}(U_1', V_j)/2\,. \qquad (3.22)$$

Wegen der Annahme $s_A(U_1', U_2') = s \leq \sqrt{\mathrm{Vol}(U_1', U_2') \cdot n}/d_{\mathrm{m}}$ haben wir

$$\frac{\sqrt{\mathrm{Vol}(U_1', U_2') \cdot n}}{d_{\mathrm{m}}} \geq s_A(U_1', U_2') \overset{(3.19)}{\geq} s_A(U_1', V) \cdot \frac{D}{d_{\mathrm{m}}} .$$

Mit (3.22) ergibt sich daraus

$$\begin{aligned} \sqrt{\mathrm{Vol}(U_1', U_2') \cdot n} &\geq \mathrm{Vol}(U_1', V_j) \cdot \frac{D}{2} \\ &\overset{(3.21)}{>} \sqrt{\mathrm{Vol}(U_1', U_2') \cdot \mathrm{Vol}(V_i, V_j)} \cdot \frac{D}{2} \end{aligned}$$

und damit $4n/D^2 > \mathrm{Vol}(V_i, V_j)$. Das wiederum ist für $D > 2$ falsch, da wir von $\mathrm{Vol}(V_i, V_j) \geq \delta n \cdot d_{\mathrm{m}} > n$ ausgehen. □

3.2.4 Beweis von Lemma 17

Zunächst zeigen wir, dass mhW. sowohl $\mathrm{Vol}(V \setminus R', V)$ als auch $\mathrm{Vol}(V, V \setminus C')$ maximal $n/(2{d_{\mathrm{m}}}^4)$ sind. Dafür teilen wir die Knoten $u \in V$ anhand von $\mathrm{Vol}(u, V_j)$ ein. Sei

$$I^{t,j} = \{u \in V : 2^t \cdot d_{\mathrm{m}} \leq \mathrm{Vol}(u, V_j) < 2^{t+1} \cdot d_{\mathrm{m}}\}.$$

Dadurch wird V für jedes $j \in \{1, 2\}$ durch die Mengen $I^{0,j}, I^{1,j}, I^{2,j}, \ldots$ in maximal $\log n$ Mengen aufgeteilt.

Wir halten jetzt j und t fest. Sei $u \in I^{t,j}$ und X_u die Indikatorvariable für das Ereignis: „$u \notin R'$, weil die Anzahl von u's Nachbarn in V_j nicht wie erwartet ist“. Also ist $\mathbf{Pr}[X_u = 1]$ wegen Fakt 16 maximal

$$\begin{aligned} \mathbf{Pr}\left[|s_A(u, V_j) - \mathrm{Vol}(u, V_j)| \geq \mathrm{Vol}(u, V_j)^{0.51}\right] &\leq 2 \cdot \exp\left(-\mathrm{Vol}(u, V_j)^{0.02}/4\right) \\ &\leq 2 \cdot \exp\left(-(2^t \cdot d_{\mathrm{m}})^{0.02}/4\right) . \end{aligned}$$

Die erwartete Anzahl von Elementen $u \in I^{t,j} \setminus R'$ ist $\mathbf{E}\left[\sum_{u \in I^{t,j}} X_u\right]$ und ist damit maximal $2 \cdot \exp(-(2^t \cdot d_{\mathrm{m}})^{0.02}/4) \cdot |I^{t,j}|$. Wir benutzen analog zur Rechnung in Abschnitt 2.4.2 auf den Seiten 30 – 32 die Tschebyscheff-Ungleichung. Wir erhalten

$$\sum_{u \neq v} \mathbf{E}[X_u \cdot X_v] \leq \mathbf{E}^2[X] + \exp\left(-(2^t \cdot d_{\mathrm{m}})^{0.02}/4\right) \cdot \sum_{u \neq v} \mathbf{Pr}[m_{uv} > 0] .$$

Letztere Summe ist durch

$$\sum_u \mathrm{Vol}(u, V) = \sum_u O(\mathrm{Vol}(u, V_j)) = \sum_u O(2^{t+1} \cdot d_{\mathrm{m}}) = O(2^t \cdot d_{\mathrm{m}} \cdot \left|I^{t,j}\right|)$$

beschränkt. Wir erhalten für d_{m} groß genug

$$\begin{aligned}\mathbf{Var}[X] &= \mathbf{E}[X] + \exp\bigl(-(2^t \cdot d_{\mathrm{m}})^{0.02}/4\bigr) \cdot O(2^t \cdot d_{\mathrm{m}} \cdot \bigl|I^{t,j}\bigr|) \\ &\leq \mathbf{E}[X] + \exp\bigl(-(2^t \cdot d_{\mathrm{m}})^{0.02}/5\bigr) \cdot \bigl|I^{t,j}\bigr| \\ &\leq 2 \cdot \exp\bigl(-(2^t \cdot d_{\mathrm{m}})^{0.02}/5\bigr) \cdot \bigl|I^{t,j}\bigr| \,.\end{aligned}$$

Die Tschebyscheff-Ungleichung zeigt: Mit Wahrscheinlichkeit $1-O(|I^{t,j}|/n^2)$ ist

$$\bigl|I^{t,j} \setminus R'\bigr| = \left|\sum_{u \in I^{t,j}} X_u\right| \leq 3 \cdot \exp\bigl(-(2^t \cdot d_{\mathrm{m}})^{0.02}/10\bigr) \cdot n\,. \tag{3.23}$$

Obige Gleichung gilt also für alle $0 \leq t < \log n$ simultan mit Wahrscheinlichkeit $1 - \sum_t O(|I^{t,j}|/n^2) = 1 - O(1/n)$. Summieren wir anschließend noch über $j = 1, 2$, sehen wir, dass (3.23) mit Wahrscheinlichkeit $1 - O(1/n)$ für alle t *und* für alle j gilt.

Falls $d_{\mathrm{m}} \geq \log^{51} n$ ist, zeigt (3.23), dass mit Wahrscheinlichkeit $1-O(1/n)$ alle Mengen $I^{t,j} \setminus R'$ leer sind, so dass $R' = V$ ist und sowohl $\overline{\mathcal{R}}$ als auch $\overline{\mathcal{C}}$ leer sind. Lemma 17 gilt also.

Sei nun für den Rest des Beweises $d_{\mathrm{m}} < \log^{51} n$. Jedes $u \in I^{t,j} \setminus R'$ trägt

$$\mathrm{Vol}(u, V) = O(\mathrm{Vol}(u, V_j)) = O(2^{t+1} \cdot d_{\mathrm{m}})$$

zu $\mathrm{Vol}(V \setminus R', V)$ bei. Wir summieren über t. Alle Knoten, deren Anzahl von Nachbarn in V_j nicht wie erwartet ist, tragen also zu $\mathrm{Vol}(V \setminus R', V)$ insgesamt

$$\begin{aligned}\sum_{t \geq 0} \mathrm{Vol}(I^{t,j} \setminus R', V) &\leq \sum_{t \geq 0} \bigl|I^{t,j} \setminus R'\bigr| \cdot O(2^{t+1} \cdot d_{\mathrm{m}}) \\ &\leq \sum_{t \geq 0} 3 \cdot \exp\bigl(-(2^t \cdot d_{\mathrm{m}})^{0.02}/10\bigr) \cdot n \cdot O(2^{t+1} \cdot d_{\mathrm{m}}) \\ &\leq \sum_{t \geq 0} n/(2^{t+3} \cdot {d_{\mathrm{m}}}^4) = n/(4 \cdot {d_{\mathrm{m}}}^4)\end{aligned}$$

bei. Wegen $\mathrm{Vol}(V \setminus R', V) \leq \sum_{j=1}^{2} \sum_{t \geq 0} \mathrm{Vol}(I^{t,j} \setminus R', V)$ gilt mit Wahrscheinlichkeit $1 - O(1/n)$

$$\mathrm{Vol}(V \setminus R', V) \leq \frac{n}{2 \cdot {d_{\mathrm{m}}}^4} \qquad \text{und} \qquad \mathrm{Vol}(V, V \setminus C') \leq \frac{n}{2 \cdot {d_{\mathrm{m}}}^4}\,, \tag{3.24}$$

wobei wir Letzteres auf dem gleichen Weg wie in der obigen Rechnung erhalten.

Um einen Widerspruch zu erzeugen, nehmen wir nun an, dass $\mathrm{Vol}(\overline{\mathcal{R}}, V_j)$ oder $\mathrm{Vol}(V_i, \overline{\mathcal{C}})$ größer als $n/{d_{\mathrm{m}}}^4$ ist. Aufgrund unserer Konstruktion von $\mathcal{R}$ und $\mathcal{C}$ erreichen wir zwangsläufig eine Situation, in der

$$\mathrm{Vol}(V_i \setminus R'_{ij}, V_j) > n/{d_{\mathrm{m}}}^4 \qquad \text{und} \qquad \mathrm{Vol}(V_i, V_j \setminus C'_{ij}) \leq n/{d_{\mathrm{m}}}^4 \tag{3.25}$$

oder

$$\mathrm{Vol}(V_i \setminus R'_{ij}, V_j) \leq n/{d_\mathrm{m}}^4 \qquad \text{und} \qquad \mathrm{Vol}(V_i, V_j \setminus C'_{ij}) > n/{d_\mathrm{m}}^4$$

gilt. Wir widerlegen (3.25) im Detail. Die zweite Möglichkeit können wir analog ausschließen.

Wir unterbrechen unseren Prozess zu dem Zeitpunkt, an dem (3.25) *erstmalig* gilt. Alle erwarteten Grade w'_u sind $O(n^{1-\varepsilon})$, für $\varepsilon > 0$ und konstant. Zusätzlich ist $d_\mathrm{m} \leq \log^{51} n$. Beides zusammen liefert

$$\frac{n}{{d_\mathrm{m}}^4} < \mathrm{Vol}(V_i \setminus R'_{ij}, V_j) \leq \frac{n}{{d_\mathrm{m}}^4} + O(n^{1-\varepsilon}) = \frac{n}{{d_\mathrm{m}}^4} \cdot (1 + o(1))\,. \tag{3.26}$$

Wegen $R'_{ij} \subseteq R'$ gilt

$$\begin{array}{rcccl} V_i \setminus R'_{ij} & = & (V_i \setminus R') & \cup & ((V_i \cap R') \setminus R'_{ij})\,, \quad \text{sowie} \\ \mathrm{Vol}(V_i \setminus R'_{ij}, V_j) & = & \mathrm{Vol}(V_i \setminus R', V_j) & + & \mathrm{Vol}((V_i \cap R') \setminus R'_{ij}, V_j) \\ & \leq & \mathrm{Vol}(V_i \setminus R'_{ij}, V) & + & \mathrm{Vol}((V_i \cap R') \setminus R'_{ij}, V_j) \\ & \leq & n/(2 \cdot {d_\mathrm{m}}^4) & + & \mathrm{Vol}((V_i \cap R') \setminus R'_{ij}, V_j) \end{array}$$

Wegen (3.26) folgt daraus wiederum

$$\frac{n}{{d_\mathrm{m}}^4} < \frac{n}{2{d_\mathrm{m}}^4} + \mathrm{Vol}((V_i \cap R') \setminus R'_{ij}, V_j)$$

also

$$\mathrm{Vol}((V_i \cap R') \setminus R'_{ij}, V_j) > \frac{n}{2 \cdot {d_\mathrm{m}}^4}\,. \tag{3.27}$$

Jeder Knoten $u \in V_i \cap R'$, den wir aus R'_{ij} in Schritt 4. entfernen, hat (zu diesem Zeitpunkt) mindestens $\mathrm{Vol}(u, V_j) \cdot D/d_\mathrm{m}$ Nachbarn in $V_j \setminus C'_{ij}$. Wir fügen niemals Knoten zu R'_{ij} oder C'_{ij} hinzu, sondern entfernen nur Knoten aus diesen Mengen. Damit verlieren wir keine der eben gezählten Kanten zwischen $(V_i \cap R') \setminus R'_{ij}$ und $V_j \setminus C'_{ij}$ zu einem späteren Zeitpunkt.

Wir erhalten

$$\begin{array}{rcl} s_A(V_i \setminus R'_{ij}, V_j \setminus C'_{ij}) & \geq & s_A((V_i \cap R') \setminus R'_{ij}, V_j \setminus C'_{ij}) \\ & \geq & \displaystyle\sum_{u \in (V_i \cap R') \setminus R'_{ij}} \frac{D \cdot \mathrm{Vol}(u, V_j)}{d_\mathrm{m}} \\ & = & \dfrac{D \cdot \mathrm{Vol}((V_i \cap R') \setminus R'_{ij}, V_j)}{d_\mathrm{m}} \overset{(3.27)}{>} \dfrac{D \cdot n}{2 \cdot {d_\mathrm{m}}^5}\,, \end{array}$$

dahingegen ist

$$\begin{array}{rcl} \mathrm{Vol}(V_i \setminus R'_{ij}, V_j \setminus C'_{ij}) & = & \dfrac{\mathrm{Vol}(V_i \setminus R'_{ij}, V_j) \cdot \mathrm{Vol}(V_i, V_j \setminus C'_{ij})}{\mathrm{Vol}(V_i, V_j)} \\ & \overset{(3.25)}{\underset{(3.26)}{\leq}} & \dfrac{n^2 \cdot (1 + o(1))}{{d_\mathrm{m}}^8 \cdot \mathrm{Vol}(V_i, V_j)} \leq \dfrac{n^2 \cdot (1 + o(1))}{{d_\mathrm{m}}^8 \cdot \delta n \cdot d_\mathrm{m}} \leq \dfrac{n}{{d_\mathrm{m}}^8}\,. \end{array}$$

Wir haben also zwei Mengen $U = V_i \setminus R'_{ij}$ und $U' = V_j \setminus C'_{ij}$ zwischen denen trotz des kleinen Volumens relativ viele Kanten verlaufen. Wegen

$$\begin{aligned} u &= \max\{|U|, |U'|\} \leq \max\left\{\frac{\mathrm{Vol}(U, V_j)}{d_\mathrm{m}}, \frac{\mathrm{Vol}(V_i, U')}{d_\mathrm{m}}\right\} \\ &\overset{(3.25),(3.26)}{\leq} \frac{n}{d_\mathrm{m}} \cdot (1 + o(1)) \leq \frac{n}{2} \end{aligned}$$

können wir diese Situation mithilfe von Lemma 22 widerlegen: Offensichtlich ist Punkt 1. verletzt. Angenommen, der zweite Punkt von Lemma 22 gilt. Dann ist

$$\begin{aligned} \frac{D \cdot n}{2 \cdot {d_\mathrm{m}}^5} \cdot \ln \frac{D \cdot {d_\mathrm{m}}^3}{2} &< s_A(U, U') \cdot \ln \frac{s_A(U, U')}{\mathrm{Vol}(U, U')} \leq 20 \cdot u \cdot \ln \frac{n}{u} \\ &\leq \frac{40 \cdot n}{{d_\mathrm{m}}^5} \cdot \ln \frac{{d_\mathrm{m}}^5}{2}. \end{aligned}$$

Das ist jedoch für $D > 140$ falsch. Also ist (3.25) mit Wahrscheinlichkeit $1 - O(1/n^4)$ unerfüllt und $\mathrm{Vol}(\overline{\mathcal{R}}, V_j) \leq n/{d_\mathrm{m}}^4$. Die gleiche Aussage gilt für $\mathrm{Vol}(V_i, \overline{\mathcal{C}})$.

Punkt 3. von Lemma 17 folgt direkt:

$$\mathrm{Vol}(\overline{\mathcal{R}}, \overline{\mathcal{C}}) = \frac{\mathrm{Vol}(\overline{\mathcal{R}}, V_j) \cdot \mathrm{Vol}(V_i, \overline{\mathcal{C}})}{\mathrm{Vol}(V_i, V_j)} \leq \frac{n^2}{{d_\mathrm{m}}^8 \cdot d_\mathrm{m} \cdot \delta n} \leq \frac{n}{{d_\mathrm{m}}^8}.$$

□

3.3 Erweiterungen

Das relativ große „spectral gap“ der Matrix M^* aus Algorithmus 13 von $\Omega({d_\mathrm{m}}^{0.49})$ lässt Raum für verschiedene Modellerweiterungen. Aus den gemachten Berechnungen ergibt sich beispielsweise, dass die Einschränkung $|V_i| \geq \delta \cdot n$ mit δ =konstant aufgehoben und stattdessen δ mit $\overline{w}$ parametrisiert werden kann, z. B. $\delta \geq 1/\ln \overline{w}$. Ebenso ist es möglich, $w_u \geq \varepsilon \cdot \overline{w}$ durch $w_u \geq \overline{w}^{1-\varepsilon}$ zu ersetzen, wenn $\varepsilon > 0$ eine (ausreichend) kleine Konstante ist. Ebenfalls könnte man zulassen, dass die Mengenzahl k nicht mehr konstant ist, sondern mit $\overline{w}$ wächst.

Die jedoch interessanteste Erweiterung betrifft die Matrix D. In Abschnitt 2.1 haben wir auf Seite 13 gefordert, dass D vollen Rang haben soll. Wir werden weiter unten sehen, dass es genügt, wenn D keine zwei Zeilen enthält, die linear abhängig sind. Dies ist eine deutlich schwächere Einschränkung, die auch notwendig ist, wie folgende Überlegung zeigt.

Angenommen, $D = (d_{ij})$ hat zwei linear abhängige Zeilen i und i'. Dann gilt $d_{i'j} = \alpha \cdot d_{ij}$ und ohne Beschränkung der Allgemeinheit $0 \leq \alpha \leq 1$. Beachte: Da D symmetrisch ist, gilt

$$d_{i'i'} = \alpha \cdot d_{i'i} = \alpha \cdot d_{ii'} = \alpha^2 \cdot d_{ii}.$$

Wir konstruieren jetzt andere Modellparameter, die zwar exakt die gleichen Kantenwahrscheinlichkeiten implizieren, jedoch nur $k - 1$ Mengen in den Graphen pflanzen. Die Idee dabei ist, V_i und $V_{i'}$ zu verschmelzen und den Faktor α in die Gewichte zu übertragen.

Zur besseren Unterscheidung der „alten“ und der „neuen“ Modellinstanz kennzeichnen wir Letztere mit einen Unterstrich. Sei also $\underline{V_i} = V_i \cup V_{i'}$ und $\underline{V_j} = V_j$ für $j \notin \{i, i'\}$. Für $u \in V_{i'}$ setzen wir $\underline{w_u} = \alpha \cdot c \cdot w_u$ und für $u \notin V_{i'}$ sei $\underline{w_u} = c \cdot w_u$, wobei

$$c = 1 + (\alpha - 1) \cdot \sum_{u \in V_{i'}} \frac{w_u}{\overline{w} \cdot n}$$

ein Korrekturfaktor ist, um den veränderten Durchschnittsgrad auszugleichen. $\underline{D}$ entsteht, indem wir aus D die Zeile i' und die Spalte i' löschen.

Es ist leicht nachzurechnen, dass die Kantenwahrscheinlichkeiten bei beiden Modellinstanzen gleich sind. Wir führen dies exemplarisch für $u \in V_i$ und $v \in V_{i'}$ vor. In der Ausgangsinstanz ist

$$\mathbf{Pr}\left[\{u, v\}\right\} \in E] = d_{ii'} \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot n} = d_{ii} \cdot \alpha \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot n}\,. \tag{3.28}$$

In der neuen Instanz gehören u und v zu $\underline{V_i}$, also ist

$$\mathbf{Pr}\left[\{u, v\} \in E\right] = \underline{d_{ii}} \cdot \frac{\underline{w_u} \cdot \underline{w_v}}{\underline{\overline{w}} \cdot n} = d_{ii} \cdot \frac{(c \cdot w_u) \cdot (\alpha \cdot c \cdot w_v)}{\underline{\overline{w}} \cdot n}\,.$$

Mit

$$\begin{aligned}
\underline{\overline{w}} \cdot n &= \sum_{u \in V} \underline{w_u} = \sum_{u \notin V_{i'}} \underline{w_u} + \sum_{u \in V_{i'}} \underline{w_u} = \sum_{u \notin V_{i'}} c \cdot w_u + \sum_{u \in V_{i'}} \alpha \cdot c \cdot w_u \\
&= \sum_{u \in V} c \cdot w_u + \sum_{u \in V_{i'}} (\alpha - 1) \cdot c \cdot w_u \\
&= c \cdot \overline{w} \cdot n \cdot \left(1 + (\alpha - 1) \cdot \sum_{u \in V_{i'}} \frac{w_u}{\overline{w} \cdot n}\right) = c^2 \cdot \overline{w} \cdot n
\end{aligned}$$

folgt

$$\mathbf{Pr}\left[\{u, v\} \in E\right] = d_{ii} \cdot \frac{w_u \cdot \alpha \cdot w_v}{\overline{w} \cdot n}\,,$$

was identisch zu (3.28) ist.

Es gibt also Modellinstanzen mit unterschiedlicher Mengenzahl, die dieselben Kantenwahrscheinlichkeiten implizieren. Eine Rekonstruktion der Mengen $V_1, \ldots, V_k$ ist somit – ohne weitere Zusatzinformationen – unmöglich. Vielmehr ist das Finden der Mengen $\underline{V_1}, \ldots, \underline{V_{k-1}}$ zu erwarten.

Aus den eben dargelegten Gründen beschränken wir uns auf Matrizen D mit Rank $k' \leq k$, die aber keine zwei linear abhängigen Zeilen enthalten.

Wir konstruieren die Matrix M^* genau wie in Algorithmus 13 angegeben. Trotz der Modifikation bezüglich D bleiben die Aussagen von Theorem 14 unberührt. Aus diesem Theorem erhalten wir wiederum

Lemma 23. *Wenn Theorem 14 für M^* gilt, so hat M^* genau k' Eigenwerte mit Absolutbetrag $\Theta(1/\overline{w})$, während alle anderen Eigenwerte betragsmäßig nur $O(1/(\overline{w} \cdot {d_{\mathrm{m}}}^{0.49}))$ sind.*

Der Beweis ist analog zu dem von Lemma 6, weswegen wir auf ihn verzichten. Beachte, dass jetzt nur noch k' Eigenvektoren zur Verfügung stehen, um sämtliche k Mengen zu identifizieren. Das folgende Lemma zeigt, dass dies prinzipiell möglich ist.

Lemma 24. *Seien $s_1, s_2, \ldots, s_{k'}$ die Eigenvektoren zu den k' betragsgrößten Eigenwerten von M^* mit $\|s_i\| = \sqrt{n}$. Seien $\chi_1, \ldots, \chi_k$ die charakteristischen Vektoren von $V_1, \ldots, V_k$. Falls Theorem 14 für M^* zutrifft, gilt für die (eindeutige) Zerlegung*

$$s_i = \alpha_{i1} \cdot \chi_1 + \alpha_{i2} \cdot \chi_2 + \ldots + \alpha_{ik} \cdot \chi_k + \gamma_i \cdot u_i$$

mit $u_i \perp \chi_1, \ldots, \chi_k$ und $\|u_i\| = \sqrt{n}$

1. *$|\gamma_i| = O({d_{\mathrm{m}}}^{-0.49})$ für $1 \leq i \leq k'$.*
2. *Für alle $1 \leq j \neq j' \leq k$ gibt es ein $i \in \{1, \ldots, k'\}$, so dass*

$$|\alpha_{ij} - \alpha_{ij'}| \geq 1/\sqrt{\ln d_{\mathrm{m}}}\,.$$

Beweis. Wir beginnen mit Punkt 1. Da s_i ein Eigenvektor zum Eigenwert mit Betrag $\Theta(1/\overline{w})$ ist, gilt

$$\left|u_i^t \cdot (M^* \cdot s_i)\right| = \Theta(1/\overline{w}) \cdot \left|u_i^t \cdot s_i\right| = \Theta(1/\overline{w}) \cdot \left|\gamma_i \cdot u_i^t \cdot u_i\right| = \Theta(n/\overline{w}) \cdot |\gamma_i|\ .$$

Andererseits folgt aus Punkt 2. von Theorem 14 wegen $u_i \perp \chi_1, \ldots, \chi_k$

$$\begin{aligned} \left|u_i^t \cdot M^* \cdot s_i\right| &= n \cdot \left| \frac{u_i^t}{\|u_i\|} \cdot M^* \cdot \frac{s_i}{\|s_i\|} \right| = n \cdot O\left(\overline{w}^{-1.49} + {d_{\mathrm{m}}}^{-1.5}\right) \\ &= O\left(n/\left(\overline{w} \cdot {d_{\mathrm{m}}}^{0.49}\right)\right)\ . \end{aligned}$$

Beide Gleichungen zusammen ergeben $|\gamma_i| = O({d_\mathrm{m}}^{-0.49})$.

Wir kommen zu Punkt 2. Angenommen, es gäbe $j, j' \in \{1, \ldots, k\}$ mit $j \neq j'$, so dass für alle $i \in \{1, \ldots, k'\}$ $|\alpha_{ij} - \alpha_{ij'}| \leq 1/\sqrt{\ln d_\mathrm{m}}$ wäre. Dann steht der Vektor $v = \chi_j / |V_j| - \chi_{j'} / |V_{j'}|$ „annähernd“ senkrecht auf allen s_i, denn

$$\left|v^t \cdot s_i\right| = |\alpha_{ij} - \alpha_{ij'}| \leq \frac{1}{\sqrt{\ln d_\mathrm{m}}} .$$

Sei nun

$$v' = v - \sum_{i=1}^{k'} \frac{v^t \cdot s_i}{n} \cdot s_i .$$

Man kann nachrechnen, dass v' senkrecht auf allen s_i steht und „annähernd“ parallel zu v ist. Beide Vektoren v und v' haben Norm $\Theta(1/\sqrt{n})$. Da v' senkrecht zu $s_1, \ldots, s_{k'}$ steht, liegt er vollständig in dem Raum, der von den Eigenvektoren aufgespannt wird, deren Eigenwerte einen Betrag von $O(1/(\overline{w} \cdot {d_\mathrm{m}}^{0.49}))$ haben. Somit muss

$$\|M^* \cdot v'\| = O\left(\frac{1}{\overline{w} \cdot {d_\mathrm{m}}^{0.49}}\right) \cdot \|v'\| = O\left(\frac{1}{\overline{w} \cdot \sqrt{n} \cdot {d_\mathrm{m}}^{0.49}}\right) \tag{3.29}$$

gelten. Nun ist

$$\begin{aligned} \|M^* \cdot v'\| &\geq \|M^* \cdot v\| - \sum_{i=1}^{k'} \left\| M^* \cdot \frac{v^t \cdot s_i}{n} \cdot s_i \right\| \\ &\geq \|M^* \cdot v\| - \frac{1}{n \cdot \sqrt{\ln d_\mathrm{m}}} \sum_{i=1}^{k'} \|M^* \cdot s_i\| \\ &\geq \|M^* \cdot v\| - O\left(\frac{1}{\sqrt{n} \cdot \sqrt{\ln d_\mathrm{m}} \cdot \overline{w}}\right) . \end{aligned} \tag{3.30}$$

Wir werden gleich sehen, dass $\|M^* \cdot v\| = \Theta(1/(\overline{w} \cdot \sqrt{n}))$ ist, womit sich die Abschätzungen (3.30) und (3.29) widersprechen.

Sei nun $\eta = M^* \cdot v$. Für jedes $i \in \{1, \ldots, k\}$ gilt

$$\begin{aligned} \sum_{u \in V_i} \eta(u) &= \sum_{u \in V_i} \left(\sum_{u_2 \in V_j} \frac{m_{uu_2}}{|V_j|} - \sum_{u_2 \in V_{j'}} \frac{m_{uu_2}}{|V_{j'}|} \right) \\ &= \frac{\mathbf{1}^t \cdot {M^*}_{V_i \times V_j} \cdot \mathbf{1}}{|V_j|} - \frac{\mathbf{1}^t \cdot {M^*}_{V_i \times V_{j'}} \cdot \mathbf{1}}{|V_{j'}|} . \end{aligned}$$

Mit Theorem 14 ergibt sich daraus

$$\begin{aligned}\sum_{u\in V_i}\eta(u) &= \left(\frac{d_{ij}\cdot W_i\cdot W_j\cdot|V_i|}{\overline{w}\cdot n}-\frac{d_{ij'}\cdot W_i\cdot W_{j'}\cdot|V_i|}{\overline{w}\cdot n}\right)\cdot\left(1\pm O\left({d_{\mathrm{m}}}^{-0.49}\right)\right)\\ &= \frac{W_i\cdot|V_i|}{\overline{w}\cdot n}\cdot(d_{ij}\cdot W_j-d_{ij'}\cdot W_{j'})\cdot\left(1\pm O\left({d_{\mathrm{m}}}^{-0.49}\right)\right).\end{aligned}$$

Aus der Jensen-Ungleichung für konvexe Funktionen (vgl. A.6) folgt, dass $\sum_{u\in V_i}\eta(u)^2$ minimal ist, wenn alle $\eta(u)$ gleich dem Mittelwert

$$\frac{\eta(u)}{|V_i|}=\frac{W_i}{\overline{w}\cdot n}\cdot(d_{ij}\cdot W_j-d_{ij'}\cdot W_{j'})\cdot\left(1\pm O\left({d_{\mathrm{m}}}^{-0.49}\right)\right)$$

sind. Es ergibt sich

$$\begin{aligned}\sum_{u\in V_i}\eta(u)^2 &\geq |V_i|\cdot\left(\frac{W_i}{\overline{w}\cdot n}\cdot(d_{ij}\cdot W_j-d_{ij'}\cdot W_{j'})\cdot\left(1\pm O\left({d_{\mathrm{m}}}^{-0.49}\right)\right)\right)^2\\ &\geq |V_i|\cdot\left(\frac{W_i\cdot(d_{ij}\cdot W_j-d_{ij'}\cdot W_{j'})}{2\cdot\overline{w}\cdot n}\right)^2.\end{aligned}$$

Der Term $d_{ij}\cdot W_j-d_{ij'}\cdot W_{j'}$ ist für wenigstens ein $i\in\{1,\ldots,k\}$ ungleich 0, sonst wären Zeile j und Zeile j' in D linear abhängig. Somit erhalten wir für dieses i

$$\sum_{u\in V_i}\eta(u)^2\geq\left(\frac{W_i\cdot\Theta(1)}{\overline{w}\cdot n}\right)^2\cdot|V_i|=\Theta\left(\frac{1}{\overline{w}^{\,2}\cdot n}\right)$$

und daraus

$$\|M^*\cdot v\|=\|\eta\|=\sqrt{\sum_{u\in V}\eta(u)^2}\geq\sqrt{\sum_{u\in V_i}\eta(u)^2}=\Omega\left(\frac{1}{\overline{w}\cdot\sqrt{n}}\right).$$

□

Lemma 24 zeigt uns mit Punkt 1., dass die Störung von s_i durch den Vektor u_i relativ klein ist und mit Punkt 2., dass es für je zwei Mengen V_j und $V_{j'}$ immer einen Vektor s_i gibt, in dem die „typischen“ Einträge der Elemente in V_j deutlich von den „typischen“ Einträgen der Elemente in $V_{j'}$ abweichen. Das folgende Lemma präzisiert dies.

Lemma 25. *Seien M^*, $s_1,\ldots,s_{k'}$ und α_{ij} wie in Lemma 24 beschrieben. Wenn Theorem 14 für M^* gilt, dann gilt für alle $i\in\{1,\ldots,k'\}$ und alle $j\in\{1,\ldots,k\}$: Die Anzahl der Koordinaten $v\in V_j$ in s_i mit $|\alpha_{ij}-s_i(v)|\geq 1/\ln^2 d_{\mathrm{m}}$ ist $O(n\cdot\ln^4 d_{\mathrm{m}}/{d_{\mathrm{m}}}^{0.98})$.*

Beweis. Sei $v \in V_j$, so dass $|\alpha_{ij} - s_i(v)| \geq 1/\ln^2 d_\mathrm{m}$. Anhand der Zerlegung von s_i wie in Lemma 24 sehen wir, dass $|\gamma_i \cdot u_i(v)| \geq 1/\ln^2 d_\mathrm{m}$ sein muss. Wegen $|\gamma_i| = O({d_\mathrm{m}}^{-0.49})$ folgt $u_i(v) = \Omega({d_\mathrm{m}}^{0.49}/\ln^2 d_\mathrm{m})$. Und da $u_i^t \cdot u_i = n$ ist, kann es maximal $n \cdot O(\ln^4 d_\mathrm{m}/{d_\mathrm{m}}^{0.98})$ solcher Koordinaten in u_i geben. □

Zur Rekonstruktion der k gepflanzten Mengen teilen wir nun die Einträge der Vektoren s_i in Intervalle der Länge $1/\ln^{1.5} d_\mathrm{m}$ ein. Diese Intervallgröße ist im Vergleich zum Wert von $1/\ln^2 d_\mathrm{m}$ in Lemma 25 sehr groß. Wir können also davon ausgehen, dass (abgesehen von den erwähnten $n \cdot O(\ln^4 d_\mathrm{m}/{d_\mathrm{m}}^{0.98})$ Koordinaten) alle Einträge, die zu V_j gehören, in zwei benachbarten Intervallen liegen. Wir nennen diese beiden Intervalle „Hauptintervalle“ von V_j bezüglich s_i.

Wir betrachten die Hauptintervalle von V_j und $V_{j'}$ mit $j \neq j'$. Für einige der s_i können diese beliebig nah beieinander liegen und sogar identisch sein. Lemma 24 zeigt uns aber, dass es mindestens ein s_i gibt, wo die Werte α_{ij} und $\alpha_{ij'}$ relativ weit auseinander liegen, so dass die Hauptintervalle von V_j weit von den Hauptintervallen von $V_{j'}$ entfernt sind. Dadurch ergeben sich Lücken, mit deren Hilfe wir V_j von $V_{j'}$ unterscheiden können.

Um die Erkennung zu vereinfachen, betrachten wir Intervalle, in denen weniger als $n/{d_\mathrm{m}}^{0.97}$ Einträge liegen, als *leer*. Innerhalb zweier Hauptintervalle liegen $\Theta(n)$ Einträge. Daher ist maximal eines dieser beiden Intervalle leer. Außerhalb der $2k$ Hauptintervalle bezüglich s_i liegen wegen Lemma 25 maximal $k \cdot n \cdot O(\ln^4 d_\mathrm{m}/{d_\mathrm{m}}^{0.98}) < n/{d_\mathrm{m}}^{0.97}$ Einträge. Demnach sind außer den Hauptintervallen alle anderen Intervalle leer. Es gibt also zu jedem s_i maximal $2 \cdot k$ Intervalle, die nicht leer sind.

Natürlich können wir nicht erwarten, alle Mengen an nur einem Vektor s_i erkennen zu können. Wir werden daher von der Partition $P = \{V\}$ ausgehen und diese iteriert mithilfe sämtlicher s_i verfeinern, bis wir alle k Mengen identifiziert haben.

Wir definieren dazu die Funktion $f_{V',s} : \mathbb{Z} \to \{0, 1\}$, die uns angibt, welche Intervalle bei dem Vektor s leer bzw. nicht leer sind, wenn wir nur die Koordinaten aus V' berücksichtigen. Sei also

$$f_{V',s}(l) = \begin{cases} 1 & \text{falls } \left|\left\{ v \in V' : \dfrac{l}{\ln^{1.5} d_\mathrm{m}} \leq s(v) < \dfrac{l+1}{\ln^{1.5} d_\mathrm{m}} \right\}\right| \geq \dfrac{n}{{d_\mathrm{m}}^{0.97}} \\ 0 & \text{sonst} \end{cases} .$$

Nun betrachten wir den folgenden Algorithmus, der Algorithmus 13 erweitert.

Algorithmus 26.

Eingabe: Die Adjazenzmatrix $A = (a_{uv})$ eines Graphen $G = (V, E)$.
Ausgabe: Eine Partition P von V.

1. Berechne den Durchschnittsgrad $\overline{d} = \sum_{u=1}^{n} d_u/n$ und setze $d_{\mathrm{m}} = \overline{d}\,/\ln \overline{d}\,$.
2. Bestimme $U = \{u \in V : d_u \geq d_{\mathrm{m}}\}$.
3. Konstruiere $M^* = (m_{uv})$ mit $m_{uv} = a_{uv}/(d_u \cdot d_v)$ für $u, v \in U$ und $m_{uv} = 0$ sonst.
4. Berechne die Eigenvektoren $s_1, s_2, \ldots$ von M^* zu den Eigenwerten, deren Betrag $\geq 1/(\,\overline{d}\,\cdot \ln d_{\mathrm{m}})$ ist. Skaliere alle s_i auf Länge $\sqrt{n}$.
5. $P := \{V\}$.
6. Solange es $V' \in P$ und $s \in \{s_1, s_2, \ldots\}$ gibt mit
   $f_{V',s}(l_1) = 1, \quad f_{V',s}(l_2) = f_{V',s}(l_2+1) = 0, \quad f_{V',s}(l_3) = 1,$
   für $l_1 < l_2 < l_3$, dann
   setze $V'' := \{v \in V' : s(v) < (l_2+1)/\ln^{1.5} d_{\mathrm{m}}\}$ und
   ersetze V' in P durch V'' und $V' \setminus V''$.

Da der Algorithmus weder k noch k' kennt, benutzen wir in Schritt 4. den Term $1/(\,\overline{d}\,\cdot \ln d_{\mathrm{m}})$ um die Eigenwerte der Größe $\Theta(1/\overline{w})$ von denen der Größe $O(1/(\overline{w} \cdot {d_{\mathrm{m}}}^{0.49}))$ zu unterscheiden.

In Schritt 6. wird die gewählte Menge V' aufgespalten. Dazu trennen wir die Einträge von s zwischen zwei leeren Intervallen. Von den beiden Hauptintervallen von V_j bezüglich s ist wenigstens eines nicht-leer. Somit ist garantiert, dass niemals zwei Hauptintervalle einer Menge V_j getrennt werden.

Also bleiben fast alle Einträge von V_j in V' oder in V''. Das garantiert (gemeinsam mit Lemma 25) zu jedem Zeitpunkt für alle Mengen V_j

$$\max_{V' \in P} |V' \cap V_j| \geq |V_j| - O\left(\frac{\ln^4 d_{\mathrm{m}}}{{d_{\mathrm{m}}}^{0.98}} \cdot n\right). \tag{3.31}$$

Solange es ein $V' \in P$ gibt, das den Großteil von zwei Mengen V_j und $V_{j'}$ beinhaltet, ist die Bedingung aus Schritt 6. erfüllt und V' wird weiter getrennt: Wegen Lemma 24 gibt es ein s_i, so dass $|\alpha_{ij} - \alpha_{ij'}| \geq 1/\sqrt{\ln d_{\mathrm{m}}}$. Die kleine Intervallgröße von $1/\ln^{1.5} d_{\mathrm{m}}$ sorgt dafür, dass zwischen den Hauptintervallen von V_j und $V_{j'}$ mindestens $\ln d_{\mathrm{m}} - 4$ Intervalle sind. Von denen sind mindestens $(\ln d_{\mathrm{m}} - 4) - 2k > 3k$ leer. Zwischen den Hauptintervallen von V_j und $V_{j'}$ bezüglich s_i muss es also zwei benachbarte leere Intervalle geben.

So wird Schritt 6. so lange wiederholt, bis jede Menge V_j (bzw. der allergrößte Teil davon) in einer eigenen Menge in P separiert ist. Ungleichung (3.31) zeigt, dass sich der Algorithmus damit beendet und von jedem V_j nur ein unwesentlicher Anteil falsch eingruppiert wurde. Fassen wir alle Überlegungen dieses Kapitels zusammen, erhalten wir

Theorem 27. *Sei G ein Graph, der in unserem erweiterten Modell generiert wurde. Mit Wahrscheinlichkeit $1 - O(1/n)$ liefert Algorithmus 26 ei-*

ne Partition P, die von der gepflanzten Partition $V_1, \ldots, V_k$ in maximal $O\left(n \cdot \ln^4 d_{\mathrm{m}}/{d_{\mathrm{m}}}^{0.98}\right) = O\left(n/\,\overline{d}^{\,0.97}\right)$ Knoten abweicht.

Die oben erwähnten Modellerweiterungen bezüglich δ, k und ε sind zwar immer noch möglich, wir belassen es jedoch beim derzeitigen Stand und sehen uns einige experimentelle Ergebnisse an.

3.4 Experimente

Wir untersuchen exemplarisch zwei Partitionierungsprobleme: 3-Färbbarkeit und MinBisection. Beide unterscheiden sich wesentlich voneinander. Bei ersterem Problem wird versucht, die Anzahl der Kanten innerhalb der drei Mengen zu minimieren, bei zweiterem die Anzahl der Kanten zwischen den beiden Mengen. Weitere Partitionierungsprobleme wurden von Riediger in [Rie07] und [Rie08] eingehend untersucht. Wir nutzen Riedigers Implementierung aus [Rie08].

Um einen Einblick in die Leistungsfähigkeit von Algorithmus 26 zu erhalten, haben wir Graphen mit 1 000, 5 000 und 10 000 Knoten in unserem Modell generiert. Die Gewichtsverteilung entsprach einem power-law mit Exponenten 2.5. Das heißt, der Anteil der Knoten mit Gewicht d war proportional zu $d^{-2.5}$. Es gab also $\Theta(n/\log^{2.5} n)$ Knoten mit Gewicht $\log n$. Die Anzahl der Knoten mit Gewicht $n^{0.4}$ war $\Theta(n/(n^{0.4})^{2.5}) = \Theta(1)$. Das heißt, das maximale Gewicht lag in der Größenordnung $n^{0.4}$. Beachte, dass bei power-law-Verteilungen mit Exponenten ≥ 2 das mittlere Gewicht $\overline{w}$ unabhängig von n konstant ist.

Die Knotengrade zahlreicher sozialer und biologischer Netzwerke unterliegen der power-law-Verteilung mit Exponenten zwischen 2 und 3. Gleiches vermutet man z. B. vom Internet-Graphen (mit den Routern als Knoten) oder dem `www`-Graphen (mit den Seiten als Knoten). In [DM03] werden power-law-Verteilungen und ihr Vorkommen ausführlich diskutiert.

In solche Graphen haben wir nun 3-Färbungen gepflanzt. Dazu haben wir die Knotenmenge randomisiert in drei etwa gleich große Mengen V_1, V_2 und V_3 zerlegt. Für die D-Matrix wählten wir

$$D = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} .$$

Innerhalb der drei Mengen wurden also keine Kanten eingefügt. Eine Kante $\{u, v\}$ zwischen V_i und V_j $(i \neq j)$ wurde gemäß unseres Modells mit Wahrscheinlichkeit $w_u \cdot w_v/(\overline{w} \cdot n)$ generiert.

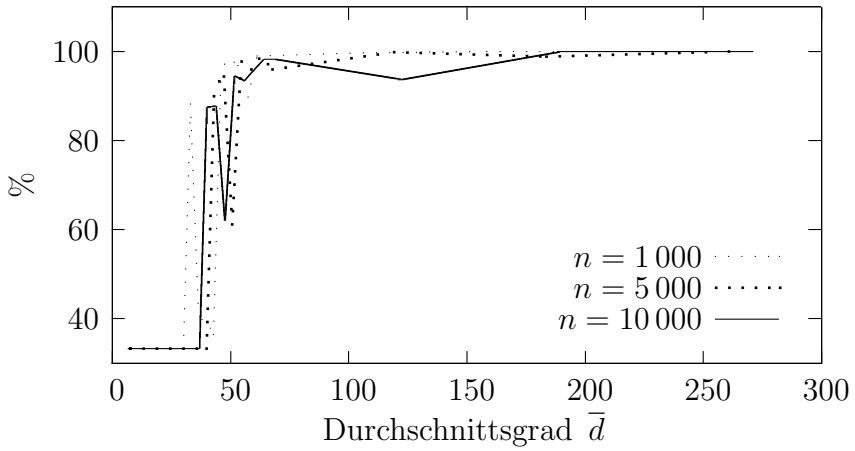


Abbildung 3.1: Übereinstimmung mit gepflanzter 3-Färbung

Auf die so erstellten Graphen wandten wir Algorithmus 26 an, der uns – zumindest für große Werte von $\overline{d}$ – eine Einteilung der Knoten liefert, die nahezu der gepflanzten Partition entspricht, vgl. Abbildung 3.1. Die Ausgabe des Algorithmus ist lediglich eine Annäherung, so dass nicht garantiert ist, dass es sich um eine 3-Färbung handelt. Innerhalb der gefundenen Mengen können noch Kanten verlaufen.

Natürlich ist es möglich, die gefundene Näherungslösung zu einer korrekten Färbung zu vervollkommnen. Wir verzichten jedoch darauf und belassen es bei der Lösung des Algorithmus.

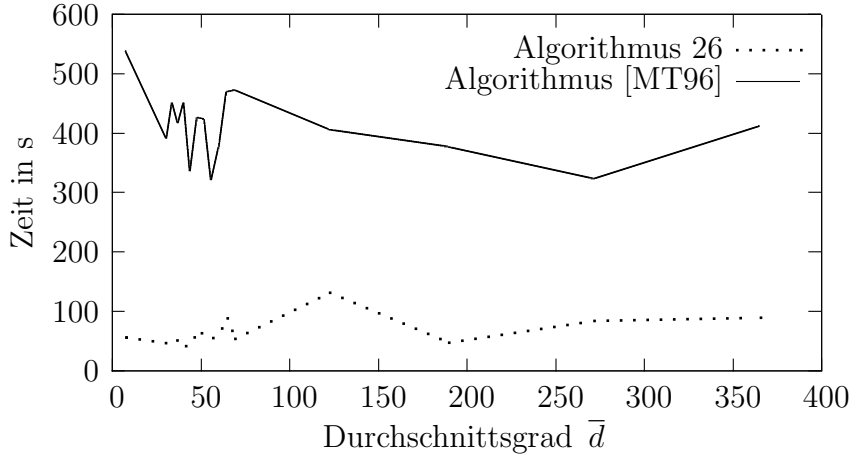


Abbildung 3.2: Laufzeit bei gepflanzter 3-Färbung und $n = 10\,000$

Um die Geschwindigkeit unseres Algorithmus einzuschätzen, verglichen

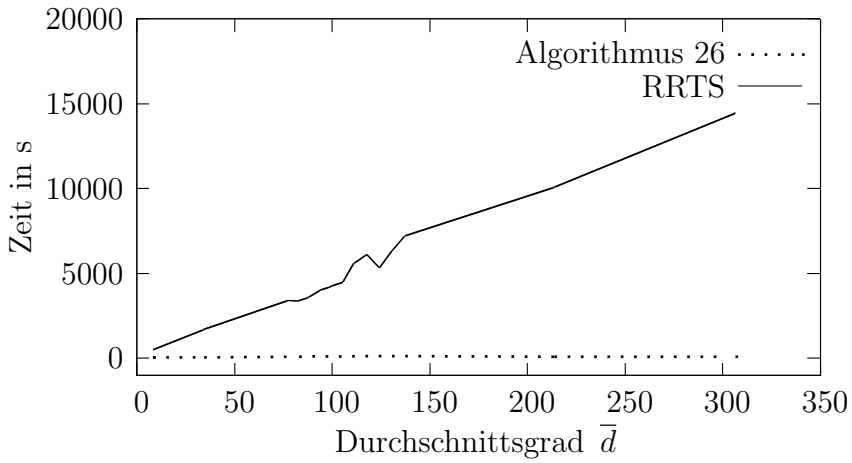


Abbildung 3.3: Laufzeit bei gepflanzter Bisektion und $n = 10\,000$

wir ihn mit einem sehr schnellen Färbungsalgorithmus aus [MT96], der auf DSATUR von Brélaz basiert [Bré79]. Die Implementierung stammt von Trick selbst. Die Zeitmessungen wurden auf einem AMD Athlon XP 2400+ mit 512 MB RAM durchgeführt. Es zeigt sich, dass unser Algorithmus vergleichsweise schnell agiert und nur einen Bruchteil der Zeit benötigt.

Ein ähnliches Bild zeigt sich bei der gepflanzten Bisektion. Zur Generierung teilten wir die Knotenmenge in zwei gleichgroße Mengen V_1 und V_2 und wählten

$$D = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} .$$

Die Wahrscheinlichkeiten für Kanten zwischen den beiden Teilen wurden also halbiert. Erwartungsgemäß liegen so etwa ein Drittel aller Kanten zwischen V_1 und V_2 und die restlichen zwei Drittel innerhalb der beiden Mengen.

Als Vergleichsalgorithmus benutzten wir „Reactive Randomized Tabu Search“ aus [BB99]. Wir verwendeten die sehr effiziente Implementierung von Dun, die für die Grid Challenge 2006 erstellt wurde [DTY06]. Die Hauptschleife von RRTS ließen wir nur einmal durchlaufen, da einerseits sehr gut partitioniert wurde, andererseits die Laufzeit bei Graphen mit 10 000 Knoten schon recht hoch war.

Bei den Experimenten ist aufgefallen, dass RRTS sehr schnell sehr gute Lösungen findet und diese in der verbleibenden Zeit nur wenig verbessern kann. Eine Lösung, die nach wenigen Sekunden gefunden wurde, war meistens nur wenige Prozent schlechter als die Lösung am Ende des Algorithmus. Es würde also genügen, RRTS nach wenigen Sekunden abzubrechen.

Die Laufzeiten unseres Algorithmus im Vergleich zu einer Iteration von

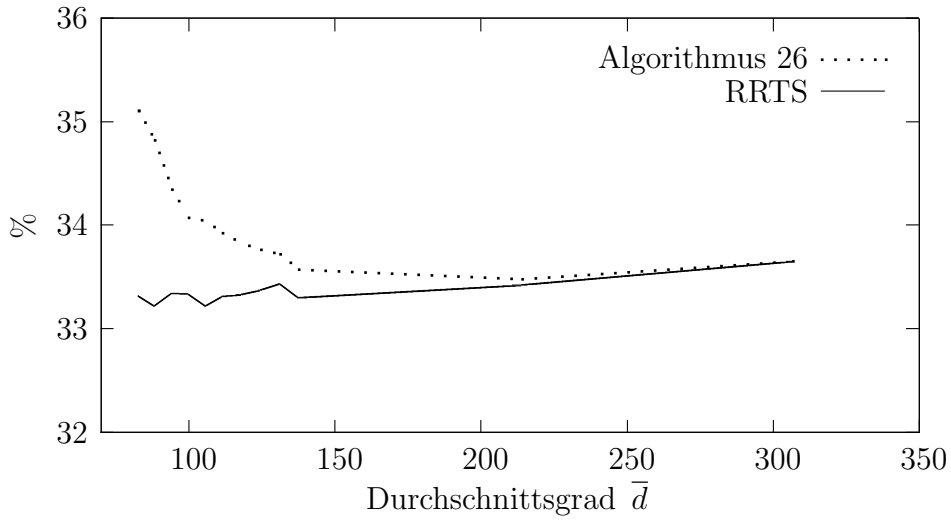


Abbildung 3.4: Anteil der Kanten zwischen den Mengen der Bisektion

RRTS sind in Abbildung 3.3 dargestellt. Wieder ist unser Algorithmus deutlich schneller als das Vergleichsverfahren.

Um die Qualität unserer Lösung zu bewerten, bestimmten wir den Anteil der Kanten zwischen den gefundenen Mengen V_1 und V_2. Da unser Algorithmus nicht speziell auf das Finden von Bisektionen ausgelegt ist, garantiert er nicht, dass beide Mengen gleich groß sind. Die Abweichung von diesem Idealwert betrug bei 10 000 Knoten ab einem Durchschnittsgrad von 80 maximal 2%. Wir störten uns nicht an dieser kleinen Ungenauigkeit und verglichen die gefundene Lösung in Abbildung 3.4 direkt mit der von RRTS.

Dort sehen wir für beide Algorithmen den relativen Anteil der Kanten zwischen den beiden Mengen der Partition. Je kleiner dieser ist, desto besser ist die Lösung. Es zeigt sich, dass die Verfahren ähnlich gut partitionieren. Vor allem bei den dichteren Instanzen ist der Unterschied zwischen beiden Methoden marginal.

Algorithmus 26 ist ein schnelles Verfahren, das zumindest auf zufälligen Graphen mit gepflanzten Partitionen sehr gute Ergebnisse liefert. Bemerkenswert ist, dass der Algorithmus keinerlei Zusatzinformationen über die verwendeten Modellparameter benötigt. Insbesondere löst er nahezu beliebige Partitionierungsaufgaben ohne Kenntnis des konkreten Problems.

Kapitel 4

Max3Sat

4.1 Das Modell

Zunächst spezifizieren wir noch einmal das verwendete Modell $\text{Form}_{n,3,p}$ und besprechen einige grundlegende Fakten. Die Menge der boole'schen Variablen sei $V = \{x_1, \ldots, x_n\}$, die Menge der Literale $L = \{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n\}$. Eine Klausel ist ein geordnetes Tripel von Literalen.

Es lassen sich demnach genau $(2n)^3$ solche Tripel bilden. Jedes Tripel fügen wir mit Wahrscheinlichkeit p unabhängig von den anderen in unsere Formel ein.

Beachte, dass wir auch Klauseln der Art $(x_1 \vee \neg x_1 \vee x_2)$ und $(x_1 \vee x_2 \vee x_1)$ zulassen. Es ist nur ein technisches Detail, die nachfolgenden Analysen auf Modelle zu übertragen, bei denen derartige Klauseln nicht generiert werden.

Wir benutzen in diesem Kapitel die Notation

$$(X, Y, Z) = (X, Y, Z)_F = \{(x \vee y \vee z) \in F : x \in X, y \in Y, z \in Z\}.$$

So beschreibt beispielsweise (L, L, L) alle Klauseln in F und $(L, L, \{z\})$ alle Klauseln aus F, die auf der dritten Position z haben.

Die Anzahl m der Klauseln ist binomialverteilt mit den Parametern $(2n)^3$ und p. So folgt aus Fakt 16, dass für $p \geq 1/n^{1.5}$ mit Wahrscheinlichkeit $1 - \exp(-n^{0.02})$

$$m = 8 \cdot n^3 \cdot p \pm (n^3 \cdot p)^{0.51} = 8n^3 \cdot p \cdot (1 + o(1)) \tag{4.1}$$

ist.

Ähnlich scharf ist die Anzahl der Vorkommen jedes Literals auf jeder der drei Positionen am Erwartungswert konzentriert: Für ein festes $z \in L$ gibt es $(2n)^2$ mögliche Klauseln, wo z auf einer festen Position $i = 1, 2, 3$ stehen kann. Wir erwarten in F somit $(2n)^2 \cdot p \geq 4 \cdot \sqrt{n}$ dieser Klauseln. Aus Fakt 16 können

wir schließen, dass mit Wahrscheinlichkeit $1 - \exp(-n^{0.24})$ jedes Literal auf jeder Position $4n^2 \cdot p \pm (n^2 p)^{0.75} = 4n^2 p \cdot (1 + o(1))$ oft vorkommt. Wir benutzen im Weiteren $\sim$, um die häufig auftretenden $(1 + o(1))$-Terme abzukürzen.

Während wir die beiden Tatsachen

$$|(L, L, L)| = m \sim 8n^3 \cdot p \qquad \text{und} \qquad |(L, L, \{z\})| \sim \frac{m}{2n} \sim 4n^2 \cdot p \tag{4.2}$$

an einer gegebenen Formel leicht überprüfen können, ist der folgende Fakt anspruchsvoller.

Unser Ziel ist es, algorithmisch zu zeigen, dass für jede Menge $X \subset L$ mit $|X| = n$

$$(X, X, X) \sim \frac{m}{8}$$

gilt.

Ist obige Gleichung nachgewiesen, lässt *jede* Belegung $\sim m/8$ Klauseln unerfüllt: Jede beliebige Belegung setzt genau n Literale auf „falsch“. Ist X diese Menge von Literalen, zeigt obige Gleichung, dass es $\sim m/8$ Klauseln in F gibt, die nur aus diesen Literalen bestehen, also unter dieser Belegung unerfüllt sind.

4.2 Idee und Algorithmus

Im Abschnitt 2.3 haben wir ausführlich diskutiert, warum spektrale Methoden beim Finden von Lösungen helfen. Es ist interessant, dass sie sich genauso eignen, Lösungen auszuschließen. Wir demonstrieren das an einem einfachen Beispiel.

Sei $G = (V, E)$ ein regulärer Graph mit Grad r und A seine Adjazenzmatrix. Wir interessieren uns dafür, ob es in G eine Bisektion gibt, die weniger als 49% aller Kanten enthält.

A habe die folgenden spektralen Eigenschaften:

1. Der Eigenvektor zum größten Eigenwert ist der Eins-Vektor $\mathbf{1}$.
2. Der zugehörige Eigenwert ist r, alle anderen Eigenwerte sind betragsmäßig $O(\sqrt{r})$.

Angenommen, es gäbe eine solche Bisektion. Dann müsste sich V in zwei gleich große Mengen V_1 und V_2 teilen lassen und $e(V_1, V_2)$ – die Anzahl der Kanten zwischen V_1 und V_2 – müsste $\leq 0.49 \cdot |E|$ sein.

Wir betrachten den Vektor $v = \mathbf{1}_{|V_1} - \mathbf{1}_{|V_2}$. Für die Elemente u aus V_1 ist also $v(u) = 1$ und für $u \notin X$ ist $v(u) = -1$ (vgl. Abschnitt 2.2). So gilt

$$\begin{aligned} v^t \cdot A \cdot v &= 2 \cdot e(V_1, V_1) - 2 \cdot e(V_1, V_2) + 2 \cdot e(V_2, V_2) \\ &= 2 \cdot (0.51 \cdot |E| - 0.49 \cdot |E|) = 0.04 \cdot |E| = \Theta(n \cdot r). \end{aligned}$$

Andererseits steht v senkrecht auf $\mathbf{1}$. Er liegt also vollständig in einem Unterraum, der von den restlichen Eigenvektoren aufgespannt wird. Alle zugehörigen Eigenwerte sind vom Betrag her $O(\sqrt{r})$. Deshalb ist

$$\left|v^t \cdot A \cdot v\right| = v^t \cdot v \cdot O(\sqrt{r}) = O(n \cdot \sqrt{r}).$$

Wir erhalten einen Widerspruch, wenn r groß genug ist. Also kann es eine derart kleine Bisektion *nicht* geben. Mithilfe der Eigenwerte von A konnten wir eine derart kleine Bisektion ausschließen, *ohne* eine einzige Bisektion direkt zu überprüfen. Allgemeiner gefasst, kann man sagen:

Lemma 28. *Seien $G = (V, E)$ ein Graph, $\lambda_1 \geq \ldots \geq \lambda_n$ die Eigenwerte seiner Adjazenzmatrix und e_1 der Eigenvektor zu λ_1 mit $\|e_1\| = \sqrt{n}$. Wenn für $0 < f = o(1)$ gilt*

1. $\max_{i>1} |\lambda_i| \leq f \cdot \lambda_1$.
2. $\mathbf{1}^t \cdot e_1 \geq n \cdot \sqrt{1 - f^2}$.

Dann gilt für jedes $X \subseteq V$ mit $|X| = n/2$: Innerhalb von X verlaufen $|E|/4 \pm n \cdot f \cdot \lambda_1$ Kanten.

Beweis. Seien $e_1, \ldots, e_n$ mit $\|e_i\| = \sqrt{n}$ die paarweise orthogonalen Eigenvektoren zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$.

Wir schreiben e_1 als $e_1 = a \cdot \mathbf{1} + b \cdot u'$ mit $u' \perp \mathbf{1}$ und $\|u'\| = \sqrt{n}$. Dann ist $a^2 + b^2 = 1$. Wegen $\mathbf{1}^t \cdot e_1 = a \cdot \mathbf{1}^t \cdot \mathbf{1} = a \cdot n$ folgt aus Punkt 2. $a \geq \sqrt{1 - f^2}$ ist somit $|b| \leq f$. Der größte Eigenvektor ist also annähernd der $\mathbf{1}$-Vektor.

So gilt für jedes $u \perp \mathbf{1}$ mit $\|u\| = \sqrt{n}$ die Ungleichung $|u^t \cdot e_1| = |b| \cdot u^t u' \leq |b| \cdot n \leq f \cdot n$. Schreiben wir u in der Eigenvektorbasis $e_1, \ldots, e_n$ als $u = \sum_{i=1}^n \alpha_i \cdot e_i$, so folgt $|\alpha_1| = |u^t \cdot e_1| / n \leq f$. Der Vektor e_1 hat also nur einen kleinen Anteil an $u \perp \mathbf{1}$.

Sei nun v mit $\|v\| = \sqrt{n}$ ein weiterer Vektor. Die Darstellung von v bezüglich der Eigenvektorbasis sei $v = \sum_{i=1}^n \beta_i \cdot e_i$. Beachte, dass $\sum_{i=1}^n \alpha_i^2 = \sum_{i=1}^n \beta_i^2 = 1$ gilt.

Dann ist

$$\begin{aligned}
\left|u^t \cdot A \cdot v\right| &= \left|\sum_{i=1}^n \alpha_i \cdot e_i \cdot A \cdot v\right| = \left|\sum_{i=1}^n \alpha_i \cdot \lambda_i \cdot e_i \cdot v\right| \\
&= \left|\sum_{i=1}^n \alpha_i \cdot \lambda_i \cdot e_i \sum_{j=1}^n \beta_j \cdot e_j\right| = \left|\sum_{i=1}^n \alpha_i \cdot \lambda_i \cdot \beta_i \cdot n\right| \\
&\leq |\alpha_1 \cdot \lambda_1 \cdot \beta_1 \cdot n| + \left|\sum_{i=2}^n \alpha_i \cdot \lambda_i \cdot \beta_i \cdot n\right| \\
&\leq f \cdot \lambda_1 \cdot n + f \cdot \lambda_1 \cdot \left|\sum_{i=2}^n \alpha_i \cdot \beta_i \cdot n\right| .
\end{aligned}$$

Aus der Ungleichung von Cauchy-Schwartz (vgl. (1.117b) in [BSMM05]) folgt $\sum_{i=2}^{n} \alpha_i \cdot \beta_i \leq \sum_{i=2}^{n} \alpha_i^2 \cdot \sum_{i=2}^{n} \beta_i^2 \leq 1$ und damit

$$\left|u^t \cdot A \cdot v\right| \leq 2 \cdot f \cdot \lambda_1 \cdot n$$

für alle u, v mit $\|u\| = \|v\| = \sqrt{n}$ und $u \perp \mathbf{1}$.

Sei nun $Y = V \setminus X$ und $u = \mathbf{1}_{|X} - \mathbf{1}_{|Y}$. Für v setzen wir $v = \sqrt{2} \cdot \mathbf{1}_{|X}$. Es gilt $\|u\| = \|v\| = \sqrt{n}$ und $u \perp \mathbf{1}$. Wir erhalten durch obige Überlegungen

$$2 \cdot f \cdot \lambda_1 \cdot n \geq \left|u^t \cdot A \cdot v\right| = \left|s_A(X,X) - s_A(Y,X)\right| .$$

Beachte, dass in $s_A(X,X)$ jede Kante, die innerhalb X liegt, zweimal gezählt wird.

Mit $v = \sqrt{2} \cdot \mathbf{1}_{|Y}$ erhalten wir analog

$$2 \cdot f \cdot \lambda_1 \cdot n \geq \left|s_A(Y,Y) - s_A(X,Y)\right| .$$

Beachte, $s_A(X,Y) = s_A(Y,X)$.

Insgesamt gilt

$$\begin{aligned} 2 \cdot |E| &= s_A(X,X) + s_A(X,Y) + s_A(Y,X) + s_A(Y,Y) \\ &= s_A(X,X) + 2 \cdot (s_A(X,X) \pm 2 \cdot f \cdot \lambda_1 \cdot n) + \\ &\qquad\qquad s_A(X,X) \pm 4 \cdot f \cdot \lambda_1 \cdot n \\ &= 4 \cdot s_A(X,X) \pm 8 \cdot f \cdot \lambda_1 \cdot n \,. \end{aligned}$$

Die Anzahl der Kanten innerhalb X entspricht $s_A(X,X)/2$, also $|E|/4 \pm f \cdot \lambda_1 \cdot n$. □

Die Adjazenzmatrix dichter zufälliger Graphen aus dem $G_{n,p}$-Modell hat typischerweise obige Eigenschaften mit $\lambda_1 \approx \overline{d}$ und $f = O(1/\sqrt{\overline{d}})$, wobei $\overline{d}$ der Durchschnittsgrad ist. So ergibt sich als Anzahl der Kanten innerhalb X der Wert $|E|/4 \cdot (1 \pm 1/\sqrt{\overline{d}})$.

Auf diese Weise können wir Mengen mit besonders vielen oder besonders wenigen Kanten ausschließen. Man kann Lemma 28 auch auf Mengen X mit $\alpha \cdot n$ Knoten ausweiten. Da wir diese stärkere Aussage nicht benötigen, bleiben wir bei obiger Variante.

Auf dem Weg zu unserem Ziel $|(X,X,X)| \sim m/8$ benötigen wir den „Projektionsgraphen“ $G_2 = (L, E)$. Dieser entsteht aus F, indem wir das letzte Literal jeder Klausel „vergessen“. Das heißt $\{b, c\} \in E$ genau dann, wenn für ein z gilt $(b \vee c \vee z) \in F$ oder $(c \vee b \vee z) \in F$. Wir lassen in G_2 auch Schlingen, also Kanten der Art $\{b, b\}$ zu, da die Analyse von G_2's Spektrum dadurch etwas einfacher wird.

Mithilfe von G_2 und obigem Lemma können wir überprüfen (Algorithmus 29), dass $|(X, X, L)| \sim m/4$ ist. Das genügt aber nicht, um eine Aussage über $|(X, X, X)|$ zu erhalten. Diese ergibt sich erst in Verbindung mit einem weiteren Graphen, dem „Produktgraphen“ $G_4 = (L \times L, E_4)$.

Wir bilden G_4 über das dritte Literal der Klauseln derart: Für zwei verschiedene Klauseln $(b_1 \vee b_2 \vee z)$ und $(c_1 \vee c_2 \vee z)$ aus F fügen wir die Kante $\{(b_1, c_1), (b_2, c_2)\}$ in E_4 ein. Beachte die Vertauschung der Literale bei den Knoten. Diese dient der besseren Verteilung der Kanten über die Knotenmenge.

Wegen (4.2) ist die Anzahl der Kanten in G_4 mit hoher Wahrscheinlichkeit

$$|E_4| \sim \sum_{z \in L} |(L, L, \{z\})|^2 \sim (4n^2p)^2 \cdot 2n = (2n)^5 p^2 .$$

Zählen wir nun die Kanten, die innerhalb $X \times X$ liegen, so ergibt sich ein Wert von

$$\sum_{z \in X} |(X, X, \{z\})|^2 + \sum_{z \notin X} |(X, X, \{z\})|^2 .$$

Jede der Summen ist minimiert, wenn deren Terme $|(X, X, \{z\})|$ am arithmetischen Mittelwert sind[1], also

$$\begin{aligned} &\geq \sum_{z \in X} \left(\frac{|(X, X, X)|}{n} \right)^2 + \sum_{z \notin X} \left(\frac{|(X, X, L \setminus X)|}{n} \right)^2 \\ &= \frac{|(X, X, X)|^2 + |(X, X, L \setminus X)|^2}{n} . \end{aligned}$$

Wieder ist die Summe minimiert, wenn $|(X, X, X)|$ und $|(X, X, L \setminus X)|$ beide den Wert

$$\frac{|(X, X, X)| + |(X, X, L \setminus X)|}{2} = \frac{|(X, X, L)|}{2} \sim \frac{m}{8}$$

haben. Also liegen innerhalb $X \times X$ in G_4 mindestens

$$\frac{2 \cdot (m/8)^2}{n} \sim 2p^2 n^5 \sim \frac{|E_4|}{16}$$

Kanten. Dies ist auch genau der Wert, den wir erwarten würden, da $X \times X$ ein Viertel aller Knoten von G_4 ausmacht und bei gleichmäßiger Verteilung $\sim 1/16$ aller Kanten innerhalb $X \times X$ liegen müssten.

[1] Dies folgt aus der Jensen-Ungleichung, siehe (A.6).

Andererseits zeigt obige Rechnung: Weicht $|(X,X,X)|$ wesentlich von $m/8$ ab, so liegen innerhalb $X \times X$ wesentlich mehr als $|E_4|/16$ Kanten. Wir präzisieren dies beim Beweis von Theorem 33.

Mithilfe der Eigenwerte der folgenden Matrix $\mathbf{A} = \mathbf{A}(F,p)$ können wir algorithmisch überprüfen, ob für jede Menge $X \subset L$ mit $|X| = n$ tatsächlich $\sim |E_4|/16$ Kanten innerhalb $X \times X$ liegen, woraus das gewünschte Resultat $|(X,X,X)| \sim m/8$ folgt.

Für $0 < p < 1$ und $b_1, b_2, z \in L$ sei

$$B_{b_1b_2z} = B_{b_1b_2z}(F,p) = \begin{cases} -1 & \text{falls } (b_1 \vee b_2 \vee z) \in F \\ p/(1-p) & \text{sonst} \end{cases}$$

$B_{b_1b_2z}$ ist also eine Art Indikator dafür, ob $(b_1 \vee b_2 \vee z) \in F$ ist oder nicht. Die Konstruktion ist so gewählt, dass der Erwartungswert von $B_{b_1b_2z}$ für $F \in \mathrm{Form}_{n,3,p}$ gerade 0 ist. Dies erleichtert uns die Analyse von $\mathbf{A}$.

Sei nun $\mathbf{A} = \mathbf{A}(F,p) = (\mathbf{a}_{b_1c_1,b_2c_2})$ die $|L| \times |L|$-Matrix mit

$$\mathbf{a}_{b_1c_1,b_2c_2} = \sum_{z \in V} (B_{b_1b_2z} \cdot B_{c_1c_2z} + B_{b_2b_1z} \cdot B_{c_2c_1z}) \qquad \text{für } (b_1,b_2) \neq (c_1,c_2)$$

und $\mathbf{a}_{b_1c_1,b_2c_2} = 0$ für $(b_1,b_2) = (c_1,c_2)$.

Die Konstruktion ist direkt G_4 nachempfunden: Jedes Klauselpaar $(b_1 \vee b_2 \vee z)$, $(c_1 \vee c_2 \vee z)$, das eine Kante $\{(b_1,c_1), (b_2,c_2)\}$ in G_4 erzeugt, trägt zu $\mathbf{a}_{b_1c_1,b_2c_2}$ eine 1 bei. Alle anderen Summanden sind im Betrag $O(p)$ (die meisten sogar $O(p^2)$), also nahe 0, da wir von $p = o(1)$ ausgehen.

$\mathbf{A}$ wirkt zunächst kompliziert, lässt sich aber in der späteren Analyse gut handhaben. Diese folgt der Idee von [FK81]. Wir können die Methoden aus [FKS89] hier nicht anwenden: Die Kanten des Produktgraphen sind *nicht* unabhängig voneinander, wenngleich die Abhängigkeiten sehr gering sind.

In [FG01] wurde die Adjazenzmatrix A eines Graphen analysiert, der G_4 sehr ähnelt. Zur Abschätzung von A's Eigenwerten wurde dort ebenfalls auf die Spurmethode zurückgegriffen. Die Spur selbst wurde aber auf andere Weise als in [FK81] abgeschätzt.

Die Spur einer quadratischen Matrix ist die Summe der Hauptdiagonalelemente. Diese entspricht gleichzeitig der Summe der Eigenwerte $\lambda_1, \ldots, \lambda_n$ der Matrix, vgl. Abschnitt 7.1.1. in [GV01]:

$$\mathrm{Spur}\,[A] = \sum_{i=1}^{n} \lambda_i \,.$$

Sei $k > 0$ eine natürliche Zahl. Dann kann man leicht nachrechnen, dass

$\lambda_1^k, \ldots, \lambda_n^k$ die Eigenwerte von A^k sind. Damit gilt

$$\text{Spur}\left[A^k\right] = \sum_{i=1}^{n} \lambda_i^k \,.$$

Ist A die Adjazenzmatrix eines Graphen G, entspricht Spur $\left[A^k\right]$ aber auch der Anzahl der geschlossenen Wege in G mit Länge k.

Durch Auszählen dieser Wege konnte in [FG01] gezeigt werden: Für konstantes k gilt

$$\text{Spur}\left[A^k\right] \sim 2^k \cdot n^{k \cdot 2\varepsilon} \,.$$

Für λ_1, den größten Eigenwert der Adjazenzmatrix von G_4, ergab sich $\lambda_1 \sim 2n^{2\varepsilon}$. Somit war

$$\text{Spur}\left[A^k\right] - \lambda_1^k = \sum_{i=2}^{n} \lambda_i^k = o(n^{k \cdot (2\varepsilon)}) \,.$$

Für gerades k gilt $\sum \lambda_i^k = \sum |\lambda_i|^k$. Somit waren alle anderen Eigenwerte $\lambda_2, \ldots, \lambda_n$ vom Betrag her $o(n^{2\varepsilon}) = o(\lambda_1)$. Dieses „spectral gap“ reichte aus, um die Existenz einer unabhängigen Menge mit Größe n^2 in dem – zu G_4 ähnlichen – Graphen zu widerlegen. Da eine erfüllbare Formel aber so eine unabhängige Menge induziert, kann so die Unerfüllbarkeit bewiesen werden.

Da die Analyse verlangte, dass $k =$ konstant ist, war $p = \Omega(n^\varepsilon / n^{1.5})$ mit $\varepsilon = \varepsilon(k) =$ konstant nötig. Um diese Klauselwahrscheinlichkeit verringern zu können, wurde in [GL03] erstmals die Matrix $\mathbf{A}$ benutzt, um die Größe der größten unabhängigen Menge zu beschränken.

Wir vereinfachen die Analyse von $\mathbf{A}$ und erhalten dabei stärkere Aussagen über G_4: Nicht nur, dass in jeder Menge der Größe n^2 mindestens eine Kante liegt, sondern sogar $\sim |E_4|/16$. Daraus folgern wir – gemeinsam mit dem Projektionsgraphen – wie eingangs besprochen, dass F nur zu $\sim 7/8 \cdot m$ erfüllbar ist.

Der folgende Algorithmus überprüft, ob $|(X, X, L)| \sim m/4$ ist.

Algorithmus 29.
Eingabe: F in 3KNF mit den $2n$ Literalen L und m Klauseln.

1. Setze $d := m/(2n)$.
2. Konstruiere $G_2 = (L, E)$.
3. Prüfe, ob $|E| = m \pm 20 \cdot d^2$.
4. Prüfe, ob G_2 Lemma 28 mit $f \cdot \lambda_1 = d^{0.6}$ erfüllt.
5. Prüfe, ob $5 \cdot d^2 + 2n \cdot d^{0.6} \leq m/(4 \cdot n^{0.1})$ ist.

6. Falls alle drei Tests erfolgreich sind, gib aus

$$\begin{aligned} |(X,X,L)| &= |E|/4 \pm 2n \cdot d^{0.6} = m/4 \pm (5 \cdot d^2 + 2n \cdot d^{0.6}) \\ &= m/4 \cdot (1 \pm 1/n^{0.1})\,. \end{aligned}$$

7. Ansonsten gib „weiß nicht“ aus.

Die Korrektheit des Algorithmus folgt unmittelbar aus Lemma 28.

Lemma 30. *Sei $F \in \mathrm{Form}_{n,3,p}$ mit $1/n^{1.5} \leq p \leq 1/n^{1.4}$. Dann liefert Algorithmus 29 bei Eingabe F mit Wahrscheinlichkeit $1 - O(1/n^4)$ die Antwort aus Schritt 6.*

Die untere Schranke an p stellt keine Einschränkung dar, da wir ohnehin von $p \geq \ln^4 n/n^{1.5}$ ausgehen. Der Grund für die obere Schranke ist folgender. Bei der Konstruktion von G_2 werden mitunter verschiedene Klauseln auf dieselbe Kante projiziert. Je größer p ist, desto häufiger passiert dies und desto schlechter kann man aus der Kantenverteilung in G_2 Rückschlüsse auf $|(X,X,L)|$ ziehen.

Für $F \in \mathrm{Form}_{n,3,p}$ mit $p > 1/n^{1.4}$ kann man F in Teilformeln zerlegen und diese separat prüfen. Dazu könnte man beispielsweise die Literale in k feste Gruppen $L_1, \ldots, L_k$ einteilen. Die Teilformeln F_i mit $1 \leq i \leq k$ enthalten dann genau die Klauseln der Form $(\cdot \vee \cdot \vee z)$ mit $z \in L_i$. Da die L_i feste Gruppen sind, ist F_i eine zufällige Formel über $L \times L \times L_i$ und der Projektionsgraph ein zufälliger Graph über $L \times L$. Dabei sollte k nicht zu groß gewählt werden, da die Projektionsgraphen sonst zu dünn werden. Eine „vernünftige“ Wahl für die Gruppenzahl k ist $k = \lfloor n^{1.5} \cdot p \rfloor \geq 1$. So gilt Lemma 30 für jedes einzelne F_i und mit Wahrscheinlichkeit $1 - O(1/n^3)$ für alle F_i gleichzeitig.

Mit dieser Idee erreichen wir Werte für p bis zu $1/\sqrt{n}$. Falls p noch größer ist, lässt sich das Problem kombinatorisch lösen: Mit hoher Wahrscheinlichkeit ist dann für jedes Paar $b, c \in L$ die Anzahl der Klauseln von Typ $(b \vee c \vee \cdot)$ am Erwartungswert $p \cdot 2n \geq \sqrt{n}$ konzentriert, was sich leicht überprüfen lässt. Daraus folgt dann sofort, dass $|(X,X,L)| \sim n \cdot n \cdot (p \cdot 2n) = 4pn^3 \sim m/4$ ist.

Zusammen mit dem Beweis von Lemma 30, den wir in Abschnitt 4.3.1 führen, folgt

Lemma 31. *Sei $F \in \mathrm{Form}_{n,3,p}$ mit $p \geq 1/n^{1.5}$. Dann existiert ein Polynomialzeitalgorithmus, der bei Eingabe F mit Wahrscheinlichkeit $1 - O(1/n^3)$ zertifiziert, dass*

$$|(X,X,L)|_F = m/4 \cdot (1 \pm 1/n^{0.1})$$

ist.

Mit diesem Wissen können wir unseren Algorithmus für das Max3Sat-Problem formulieren:

Algorithmus 32.
Eingabe: $F \in \text{Form}_{n,3,p}$ und p mit $p \leq 1/\ln n$.

1. Prüfe, ob n groß genug ist.
2. Prüfe, ob $m = 8pn^3 \cdot (1 \pm 1/\sqrt{n})$.
3. Prüfe, ob für alle $z \in L$ $|(L, L, \{z\})| \leq 5pn^2$ ist.
4. Prüfe, ob $|(X, X, L)_F| = m/4 \cdot (1 \pm 1/n^{0.1})$.
5. Prüfe, ob $\|\mathbf{A}(F, p)\| \leq \ln^{3.5} n \cdot pn^{1.5}$ ist.
6. Falls alle fünf Tests erfolgreich sind, gib $7/8 \cdot m \cdot (1 + 4/\ln^{0.25} n)$ aus.
7. Ansonsten gib m aus.

Die Korrektheit des Algorithmus, also warum aus den Schritten 1. – 5. der sechste Schritt folgt, sehen wir in Abschnitt 4.3.3. Dort besprechen wir auch Schritt 1. und nennen alle Bedingungen, die n erfüllen muss, um „groß genug“ zu sein.

Den Beweis des folgenden Theorems führen wir in Abschnitt 4.3.2.

Theorem 33. *Sei $F \in \text{Form}_{n,3,p}$ mit $p \geq \ln^4 n/n^{1.5}$, dann gibt Algorithmus 32 mit Wahrscheinlichkeit $1 - O(1/n^3)$ die Antwort aus Schritt 6.*

Je größer p ist, desto mehr kann der Wert $4/\ln^{0.25} n$ aus Schritt 6. verbessert werden. Wir verzichten in der Analyse jedoch darauf, diese Möglichkeit zu besprechen.

Beachte, dass die Klauselwahrscheinlichkeit p zur Eingabe von Algorithmus 32 gehört. Es ist jedoch möglich, darauf zu verzichten und p durch $p' = m/(8 \cdot n^3)$ anzunähern. Dann gilt wegen (4.1) mit hoher Wahrscheinlichkeit $p' = p \pm O(p^{0.51}/n^{1.47})$.

Die Matrix $\mathbf{A}' = \mathbf{A}(F, p')$ unterscheidet sich in ihren Einträgen etwas von $\mathbf{A} = \mathbf{A}(F, p)$. Die Abweichung bei jedem B-Wert ist jedoch durch $O(p^{0.51}/n^{1.47})$ beschränkt. Das Gleiche gilt auch für die Abweichung der Erwartungswerte. Die Analyse von $\mathbf{A}$'s Norm in Abschnitt 4.3.2 lässt sich augenscheinlich auch auf $\mathbf{A}'$ übertragen, so dass die gleichen Ergebnisse auch mit $\mathbf{A}'$ erzielt werden können und p nicht zur Eingabe gehören muss.

4.3 Beweise

4.3.1 Beweis von Lemma 30

Aus (4.2) folgt $d \sim 4n^2p$ mit Wahrscheinlichkeit $1 - \exp(-n^{0.24})$. Jede Kante in G_2 ist auf mindestens eine Klausel in F zurückzuführen, also gilt $|E| \leq m$.

$|E|$ kann aber kleiner als m sein, wenn verschiedene Klauseln auf die gleiche Kante abgebildet werden.

Wir halten die Kante $\{b, c\}$ mit $b \neq c$ fest. Auf diese Kante werden Klauseln der Art

$$(b \vee c \vee \cdot) \qquad \text{und} \qquad (c \vee b \vee \cdot) \tag{$*$}$$

abgebildet. Die Wahrscheinlichkeit, dass von diesen $4n$ Klauseln mehr als 16 in F vorhanden sind, ist

$$\leq \binom{4n}{17} \cdot p^{17} \leq (4np)^{17} \leq \left(\frac{4}{n^{0.4}}\right)^{17} \leq \frac{1}{3 \cdot n^6} \,.$$

Für den Fall $b = c$ ist diese Wahrscheinlichkeit noch geringer. Summieren wir diesen Wert über alle möglichen $\binom{2n+1}{2} \leq 3n^2$ Kanten (inkl. Schlingen), stellen wir fest: Mit Wahrscheinlichkeit $1-1/n^4$ verlieren wir an keiner Kante mehr als 15 Klauseln aus F.

Sei $X_{\{b,c\}}$ die Indikatorvariable für das Ereignis „Bei Kante $\{b, c\}$ gehen Klauseln verloren“ und $X = \sum X_{\{b,c\}}$ die Anzahl der Kanten, wo wir Klauseln verlieren. Dann ist wegen obiger Überlegung $|E| \geq m - 15 \cdot X$ mit Wahrscheinlichkeit $1 - 1/n^4$.

$X_{\{b,c\}}$ ist 1, wenn zwei der Klauseln aus $(*)$ in F existieren. Die Wahrscheinlichkeit dafür ist $\leq \binom{4n}{2} \cdot p^2 \sim 8n^2p^2$. Summieren wir über die $\binom{2n+1}{2}$ möglichen Kanten in G_2, erhalten wir

$$\mathbf{E}[X] \leq 8n^2p^2 \cdot (1 + o(1)) \cdot \binom{2n+1}{2} \sim 16n^4p^2 \sim d^2 \,.$$

Mit Fakt 16 ergibt sich

$$\mathbf{Pr}\left[X \geq 1.1 \cdot d^2\right] \leq \exp\bigl(-\Omega(d^2)\bigr) \leq \exp(-\Omega(n)) \;.$$

Also ist mit Wahrscheinlichkeit $1 - 1/n^4 - \exp(-\Omega(n)) \geq 1 - 2/n^4$

$$|E| \geq m - 15 \cdot 1.1 \cdot d^2 \geq m - 20 \cdot d^2 \,.$$

Als nächstes untersuchen wir, warum G_2 mhW. den Test aus Schritt 4. besteht. Das dafür nötige „spectral gap“ beziehen wir aus Theorem 2. Dazu konstruieren wir geeignete Modellparameter, so dass ${M^*}_{V_1 \times V_1}$ genau der Adjazenzmatrix von G_2 entspricht.

Sei dazu $k = 2$, $D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ und $V_i = \{i\} \times L$. Die beiden gepflanzten Mengen entsprechen also Kopien von L. Die Kante $\{b, c\}$ wird genau dann *nicht* in G_2 eingefügt, wenn keine der $4n$ Klauseln aus $(*)$ in F existiert. Dies

geschieht mit Wahrscheinlichkeit $(1-p)^{4n}$. Damit ist die Kantenwahrscheinlichkeit in G_2 $p' = 1-(1-p)^{4n} \sim 4np$.

Wir setzen alle Gewichte w_u auf $\overline{w} = 2n \cdot p'$ und erhalten für $u, v \in V_1$

$$\mathbf{Pr}\left[a_{uv} = 1\right] = d_{11} \cdot \frac{w_u \cdot w_v}{\overline{w} \cdot 2n} = 1 \cdot \frac{(2n \cdot p') \cdot (2n \cdot p')}{(2n \cdot p') \cdot 2n} = p'$$

und als erwarteten Grad $w'_u = w'_v = p' \cdot 2n = \overline{w}$. Da die gleiche Situation für $u, v \in V_2$ gilt, ist der erwartete Durchschnittsgrad $\overline{w}' = \overline{w} = 2np'$. Die Normierung von A zu M hat damit keinen Effekt.

Wegen $w'_u = 2n \cdot p' \sim 8n^2 p > \sqrt{n}$ sind mit Wahrscheinlichkeit $1 - \exp(-n^{0.25})$ alle Knotengrade maximal zweimal so groß wie deren Erwartungswert $\overline{w}'$.

Mit der Wahl $C_1 = 5$ geht dann aus den Ausführungen ab Seite 30 $\mathbf{E}\left[|V \setminus U|\right] \leq \exp(-\Omega(\sqrt{n}))$ hervor. Die Markov-Ungleichung zeigt sofort: Mit Wahrscheinlichkeit $1 - \exp(-\Omega(\sqrt{n})) \cdot n$ ist $V = U$. Es müssen also keine Zeilen oder Spalten aus M gelöscht werden und $M^* = M$.

${M^*}_{V_1 \times V_1}$ und die Adjazenzmatrix von G_2 unterliegen demnach dem gleichen Zufallsprozess und wir können die Erkenntnisse aus Theorem 2 auf G_2 übertragen.

Beachte, dass Theorem 2 mit Wahrscheinlichkeit $1 - O(1/n)$ gilt. Dieser Wert rührt von der dort berechneten Schranke an $|V \setminus U|$, die auch in dünnen Zufallsgraphen des Modells gilt. In unserem (dichteren) Fall gilt $|V \setminus U| = 0$ aber mit der höhereren Wahrscheinlichkeit $1 - \exp(-\Omega(\sqrt{n})) \cdot n$. Dadurch wird die Restwahrscheinlichkeit von Theorem 2 nun von anderen Teilen des Beweises bestimmt und verringert sich deshalb von $O(1/n)$ auf $O(1/n^4)$.

Es gilt also für A, die Adjazenzmatrix von G_2 mit Wahrscheinlichkeit $1 - O(1/n^4)$:

1. $\mathbf{1}^t \cdot A \cdot \mathbf{1} = \overline{w}' \cdot 2n \cdot (1 \pm O(1/\sqrt{\overline{w}'}))$.
2. Für u, v mit $\|u\| = \|v\| = 1$ sowie $u \perp \mathbf{1}$ oder $v \perp \mathbf{1}$ gilt

$$\left|u^t \cdot A \cdot v\right| = O(\sqrt{\overline{w}'})\,.$$

Aus der Courant-Fischer-Charakterisierung der Eigenwerte (Fakt 5) erhalten wir sofort $\lambda_1 = \overline{w}' \cdot \left(1 \pm O\left(1/\sqrt{\overline{w}'}\right)\right)$ und $\max_{i>1} |\lambda_i| = O\left(\sqrt{\overline{w}'}\right)$. Mit $\overline{w}' = 2n \cdot p' \sim 8n^2 p \sim 2d$ folgt

$$\max_{i>1} |\lambda_i| \leq O\left(\sqrt{d}\right) \leq d^{0.6} = f \cdot \lambda_1\,.$$

Damit erfüllt G_2 die erste Bedingung des Lemmas 28.

Für die zweite Bedingung müssen wir zeigen, dass $\mathbf{1}^t \cdot e_1 \geq (2n) \cdot \sqrt{1-f^2}$, wobei $f = d^{0.6}/\lambda_1$ nach obiger Überlegung $\Theta(d^{-0.4})$ ist. Sei $e_1 = a \cdot \mathbf{1} + b \cdot u$ mit $\|u\| = \sqrt{2n}$, $u \perp \mathbf{1}$ und $a^2 + b^2 = 1$. Dann gilt wegen Theorem 2

$$\begin{aligned} e_1^t \cdot A \cdot e_1 &= a^2 \cdot \mathbf{1}^t \cdot A \cdot \mathbf{1} \pm O(b \cdot \sqrt{d} \cdot n) \\ &= \mathbf{1}^t \cdot A \cdot \mathbf{1} - b^2 \cdot \mathbf{1}^t \cdot A \cdot \mathbf{1} \pm O(b \cdot \sqrt{d} \cdot n) \\ &= \mathbf{1}^t \cdot A \cdot \mathbf{1} - \Theta(b^2 \cdot d \cdot n) \pm O(b \cdot \sqrt{d} \cdot n)\,. \end{aligned}$$

Da e_1 der Eigenvektor zum größten Eigenwert ist, muss $e_1^t \cdot A \cdot e_1 \geq \mathbf{1}^t \cdot A \cdot \mathbf{1}$ sein. Also müssen die beiden letzten Summanden insgesamt ≥ 0 sein und somit $\Theta(b^2 \cdot d \cdot n) = O(|b| \cdot \sqrt{d} \cdot n)$. Dies ist nur dann erfüllt, wenn $|b| = O(1/\sqrt{d})$ ist, also $a = \sqrt{1-b^2} = \sqrt{1 - O(1/d)}$. Wegen $\mathbf{1}^t \cdot e_1 = 2n \cdot a$ folgt Punkt 2. von Lemma 28 mit $f = d^{-0.4}$.

Zuletzt sehen wir uns an, warum F mhW. Schritt 5. besteht. Wegen (4.2) ist mit Wahrscheinlichkeit $1 - \exp(-\sqrt{n})$ die Zahl der Klauseln $m \sim 8n^3 p \geq 8 \cdot n^{1.5}$. Nun gilt

$$d^2 = \left(\frac{m}{2n}\right)^2 = \frac{m^2}{4n^2} = m \cdot O\left(\frac{n^{1.6}}{n^2}\right) = O\left(\frac{m}{n^{0.4}}\right) = o\left(\frac{m}{n^{0.1}}\right)$$

und

$$n \cdot d^{0.6} = n \cdot \left(\frac{m}{2n}\right)^{0.6} = O\left(\frac{m^{0.6}}{n^{0.1}}\right) \cdot n^{0.5} = O\left(\frac{m^{0.6}}{n^{0.1}}\right) \cdot m^{1/3} = o\left(\frac{m}{n^{0.1}}\right).$$

Also besteht F mit Wahrscheinlichkeit $\geq 1 - O(1/n^4)$ jeden der drei Tests. □

4.3.2 Beweis von Theorem 33

Schritt 2. und 3. von Algorithmus 32 lassen sich leicht durchführen. Unsere Überlegungen bei der Modellbesprechung in Abschnitt 4.1 zeigen, dass beide Gleichungen mit Wahrscheinlichkeit $1 - O(1/n^3)$ korrekt sind.

Laut Lemma 31 lässt sich Schritt 3. so implementieren, dass die Prüfung mit der gleichen Wahrscheinlichkeit erfolgreich ist. Die untere Schranke an n in Schritt 1. ergibt sich während der folgenden Herleitung.

Es bleibt zu zeigen, dass $\|\mathbf{A}(F,p)\|$ mit Wahrscheinlichkeit $1 - O(1/n^3)$ maximal $\ln^{3.5}(n) \cdot pn^{1.5}$ ist. Der Beweis basiert auf der Technik aus [FK81].

Sei also $F \in \mathrm{Form}_{n,3,p}$ mit $p \geq \ln^4(n)/n^{1.5}$. In den folgenden Rechnungen tritt sehr häufig der Term $p \cdot n^{1.5}$ auf. Zugunsten der Übersicht kürzen wir diesen durch $f = p \cdot n^{1.5}$ ab und halten fest, dass $f \geq \ln^4 n$ ist.

Die Matrix $\mathbf{A} = \mathbf{A}(F,p)$ wurde folgendermaßen definiert: Für $b_1, b_2, z \in L$ sei

$$B_{b_1b_2z} = B_{b_1b_2z}(F,p) = \begin{cases} -1 & \text{falls } (b_1 \vee b_2 \vee z) \in F \\ p/(1-p) & \text{sonst} \end{cases}.$$

Dann ist die $(2n)^2 \times (2n)^2$-Matrix $\mathbf{A} = \mathbf{A}(F,p) = (\mathbf{a}_{b_1c_1,b_2c_2})_{(b_1,c_1),(b_2,c_2)\in L^2}$ gegeben durch

$$\mathbf{a}_{b_1c_1,b_2c_2} = \begin{cases} 0 & \text{falls } (b_1,b_2) = (c_1,c_2) \\ \sum_{z\in L}(B_{b_1b_2z}\cdot B_{c_1c_2z} + B_{b_2b_1z}\cdot B_{c_2c_1z}) & \text{sonst} \end{cases}.$$

Beachte, dass $\mathbf{A}$ symmetrisch ist.

Da jede mögliche Klausel mit Wahrscheinlichkeit p in F eingefügt wird, gilt für alle $b_1, b_2, z \in L$

$$\mathbf{E}[B_{b_1b_2z}] = p\cdot(-1) + (1-p)\cdot\frac{p}{1-p} = -p+p = 0\,. \tag{4.3}$$

Mit der Linearität des Erwartungswertes haben damit auch alle Einträge von $\mathbf{A}$ den Erwartungswert 0.

Seien $\lambda_1 \geq \ldots \geq \lambda_{(2n)^2}$ die (reellen) Eigenwerte von $\mathbf{A}$. Wir setzen $\lambda = \|\mathbf{A}\| = \max\{\lambda_1, -\lambda_{(2n)^2}\}$. Für $k =$ gerade gilt

$$\lambda^k \leq \text{Spur}\left[\mathbf{A}^k\right] = \sum_{i=1}^{(2n)^2}\lambda_i^k$$

und damit $\mathbf{E}\left[\lambda^k\right] \leq \mathbf{E}\left[\text{Spur}\left[\mathbf{A}^k\right]\right]$. Wir zeigen weiter unten, dass für $k \sim \ln n$

$$\mathbf{E}\left[\text{Spur}\left[\mathbf{A}^k\right]\right] \leq (\ln^{3.3} n\cdot f)^k \tag{4.4}$$

ist. Daraus folgt mit der Markov-Ungleichung

$$\begin{aligned}\mathbf{Pr}\left[\lambda \geq \ln^{3.5} n\cdot f\right] &= \mathbf{Pr}\left[\lambda^k \geq (\ln^{3.5} n\cdot f)^k\right] \leq \frac{\mathbf{E}\left[\lambda^k\right]}{(\ln^{3.5} n\cdot f)^k} \\ &\leq \frac{\mathbf{E}\left[\text{Spur}\left[\mathbf{A}^k\right]\right]}{(\ln^{3.5} n\cdot f)^k} \leq \frac{(\ln^{3.3} n\cdot f)^k}{(\ln^{3.5} n\cdot f)^k} = o(1/n^{10}).\end{aligned}$$

Es bleibt (4.4) zu zeigen. Aus der Definition von $\mathbf{A}^k$ folgt unmittelbar

$$\text{Spur}\left[\mathbf{A}^k\right] = \sum_{b_1=1}^{2n}\sum_{c_1=1}^{2n}\ldots\sum_{b_k=1}^{2n}\sum_{c_k=1}^{2n}\mathbf{a}_{b_1c_1,b_2c_2}\cdot\mathbf{a}_{b_2c_2,b_3c_3}\cdots\mathbf{a}_{b_kc_k,b_1c_1} \tag{4.5}$$

Falls $(b_k, b_1) = (c_k, c_1)$ oder für ein i $(b_i, b_{i+1}) = (c_i, c_{i+1})$ gilt, ist wegen der Definition der $\mathbf{a}$'s das gesamte Produkt $\mathbf{a}_{b_1c_1,b_2c_2} \cdot \mathbf{a}_{b_2c_2,b_3c_3} \cdot \ldots \cdot \mathbf{a}_{b_kc_k,b_1c_1} = 0$. Daher können wir im Weiteren davon ausgehen, dass $(b_i, b_{i+1}) \neq (c_i, c_{i+1})$ für $1 \leq i < k$ und $(b_k, b_1) \neq (c_k, c_1)$. Setzen wir die Definition der $\mathbf{a}$'s in (4.5) ein, erhalten wir

$$\begin{aligned}
\text{Spur}\left[\mathbf{A}^k\right] &= \sum_{b_1,\ldots,b_k} \sum_{c_1,\ldots,c_k} \left(\sum_{z_1 \in L} (B_{b_1b_2z_1} \cdot B_{c_1c_2z_1} + B_{b_2b_1z_1} \cdot B_{c_2c_1z_1}) \right) \cdot \ldots \\
&\qquad \cdot \left(\sum_{z_k \in L} (B_{b_kb_1z_k} \cdot B_{c_kc_1z_k} + B_{b_1b_kz_k} \cdot B_{c_1c_kz_k}) \right) \\
&= \sum_{b_1,\ldots,b_k} \sum_{c_1,\ldots,c_k} \sum_{z_1,\ldots,z_k} (B_{b_1b_2z_1} \cdot B_{c_1c_2z_1} + B_{b_2b_1z_1} \cdot B_{c_2c_1z_1}) \cdot \ldots \\
&\qquad \cdot (B_{b_kb_1z_k} \cdot B_{c_kc_1z_k} + B_{b_1b_kz_k} \cdot B_{c_1c_kz_k}) \,.
\end{aligned}$$

Multiplizieren wir jetzt noch alle Faktoren aus, entstehen 2^k Terme

$$X_j = X_j(b_1, \ldots, b_k, c_1, \ldots, c_k, z_1, \ldots, z_k) \,,$$

so dass

$$\text{Spur}\left[\mathbf{A}^k\right] = \sum_{b_1,\ldots,b_k} \sum_{c_1,\ldots,c_k} \sum_{z_1,\ldots,z_k} \sum_{j=1}^{2^k} X_j \,.$$

Dabei hat jedes X_j die Form

$$X_j = B_{\beta_1} \cdot B_{\gamma_1} \cdot B_{\beta_2} \cdot B_{\gamma_2} \cdot \ldots \cdot B_{\beta_k} \cdot B_{\gamma_k} \,, \tag{4.6}$$

mit $\beta_i = b_ib_{i+1}z_i$ und $\gamma_i = c_ic_{i+1}z_i$ oder $\beta_i = b_{i+1}b_iz_i$ und $\gamma_i = c_{i+1}c_iz_i$ für $1 \leq i < k$ und analog mit 1 statt $i+1$ für $i = k$. Beachte: Aufgrund obiger Annahme gilt $\beta_i \neq \gamma_i$.

Sei $C = (b_1, \ldots, b_k, c_1, \ldots, c_k)$ und $Z = (z_1, \ldots, z_k)$. Dann sei $|C| = |\{b_1, \ldots, b_k, c_1, \ldots, c_k\}|$ und $|Z| = |\{z_1, \ldots, z_k\}|$ die Anzahl der verschiedenen Elemente in C und Z. Wir werden zeigen

$$\mathbf{E}\left[\text{Spur}\left[\mathbf{A}^k\right]\right] = \sum_{c=1}^{2k} \sum_{z=1}^{k} \sum_{\substack{C \\ |C|=c}} \sum_{\substack{Z \\ |Z|=z}} \sum_{j=1}^{2^k} \mathbf{E}[X_j] \leq (\ln^{3.3} n \cdot f)^k.$$

Zunächst beweisen wir, dass für große Werte von c und z $\mathbf{E}[X_j] = 0$ ist. Dadurch müssen wir im Weiteren nur noch

$$\mathbf{E}\left[\text{Spur}\left[\mathbf{A}^k\right]\right] = \sum_{c=1}^{k+2} \sum_{z=1}^{k/2} \sum_{\substack{C \\ |C|=c}} \sum_{\substack{Z \\ |Z|=z}} \sum_{j=1}^{2^k} \mathbf{E}[X_j] \tag{4.7}$$

abschätzen.

Seien also C und Z feste Folgen mit $|C| = c$ und $|Z| = z$. Dann sei $X_j = B_{\beta_1} \cdot B_{\gamma_1} \cdot \ldots \cdot B_{\beta_k} \cdot B_{\gamma_k}$ ein beliebiger der 2^k zugehörigen Terme. Wenn $z > k/2$ oder $c > k + 2$ ist, gibt es einen Faktor B_δ in X_j, der nur einmal auftritt. Da $\mathbf{E}[B_\delta] = 0$ ist und B_δ von allen anderen Faktoren in X_j unabhängig ist, gilt $\mathbf{E}[X_j] = 0$.

Angenommen, es gibt keinen solchen Faktor B_δ. Wir gehen von links nach rechts entlang des Ausdrucks (4.6). Es gibt exakt z Plätze, an denen ein Element aus Z erstmalig auftritt. An jedem solchen Platz entstehen so 2 B-Faktoren, die zuvor nicht auftraten. Also enthält X_j mindestens $2z$ verschiedene B-Faktoren. Da jeder dieser $2z$ Faktoren mindestens zweimal in X_j auftreten soll, muss $2k \geq 4z$ gelten, also $z \leq k/2$.

Wir gehen erneut in X_j von links nach rechts, um $c \leq k + 2$ zu zeigen. Die beiden ersten B-Faktoren nehmen höchstens vier verschiedene Elemente von C auf. In allen restlichen B-Faktoren tritt maximal ein Element aus C erstmalig auf, da der jeweils andere Platz durch ein Element aus einem vorherigen B-Faktor belegt ist. Also brauchen wir (abgesehen von den beiden ersten B-Faktoren) mindestens $c-4$ weitere Faktoren, um die restlichen $c-4$ Elemente zu verteilen. Es gibt also $2+(c-4) = c-2$ paarweise verschiedene Faktoren. Wenn jeder davon zweimal auftreten soll, muss $2(c-2) \leq 2k$, also $c \leq k+2$ sein. Also gilt (4.7).

Sei B_α ein beliebiger Faktor, der genau r-mal, $r \geq 2$, in X_j auftritt. Da wir von $p \leq \frac{1}{2}$ ausgehen können, gilt

$$\mathbf{E}[B_\alpha^r] = p \cdot (-1)^r + (1-p) \cdot \left(\frac{p}{1-p}\right)^r \leq p + \frac{p^r}{(1-p)^{r-1}} \leq 2p\,.$$

Wie wir oben gesehen haben, existieren mindestens $\max\{2z, c-2\}$ verschiedene B-Faktoren in X_j. So können wir abschätzen

$$\mathbf{E}[X_j] \leq (2p)^{\max\{2z,c-2\}}\,.$$

Beachte, dass die rechte Seite nur noch von c und z abhängt. Wir haben

$$\mathbf{E}\left[\text{Spur}\left[\mathbf{A}^k\right]\right] \leq \sum_{c=1}^{k+2} \sum_{z=1}^{k/2} \sum_{\substack{C \\ |C|=c}} \sum_{\substack{Z \\ |Z|=z}} 2^k \cdot (2p)^{\max\{2z,c-2\}}\,.$$

Für ein festes c können wir jedes C mit $|C| = c$ durch folgenden Prozess konstruieren: Als ersten wählen wir aus den L Elementen c aus, die wir dann über die $2k$ möglichen Plätze verteilen. Also gibt es maximal $(2n)^c \cdot c^{2k} \leq (2n)^c \cdot (2k)^{2k}$ mögliche Folgen C.

Analog schätzen wir ab, dass es maximal $(2n)^z \cdot z^k \leq (2n)^z \cdot k^k$ mögliche Folgen Z mit $|Z| = z$ gibt. Damit gilt

$$\sum_{c=1}^{k+2} \sum_{z=1}^{k/2} \sum_{\substack{C \\ |C|=c}} \sum_{\substack{Z \\ |Z|=z}} 2^k \cdot (2p)^{\max\{2z,c-2\}} \leq \sum_{c=1}^{k+2} \sum_{z=1}^{k/2} 2^{4k} \cdot n^{c+z} \cdot k^{3k} \cdot p^{\max\{2z,c-2\}}.$$

Als nächstes schätzen wir $n^{c+z} \cdot p^{\max\{2z,c-2\}}$ nach oben ab. Falls $2z > c-2$ gilt, ist $c \leq 2z+1$ und mit $p = f/n^{1.5}$

$$n^{c+z} \cdot p^{\max\{2z,c-2\}} \leq n^{3z+1} \cdot (f/n^{1.5})^{2z} = n \cdot f^{2z} \leq n \cdot f^k \,.$$

Im Fall $2z \leq c-2$ ist $z \leq c/2 - 1$ und

$$n^{c+z} \cdot p^{\max\{2z,c-2\}} \leq n^{1.5 \cdot c - 1} \cdot (f/n^{1.5})^{c-2} = n^2 \cdot f^{c-2} \leq n^2 \cdot f^k \,.$$

Also gilt

$$\mathbf{E}\left[\text{Spur}\left[\mathbf{A}^k\right]\right] \leq \sum_{b=1}^{k+2} \sum_{z=1}^{k/2} 2^{4k} \cdot k^{3k} \cdot n^2 \cdot f^k \leq (k+2) \cdot \frac{k}{2} \cdot 2^{4k} \cdot k^{3k} \cdot n^2 \cdot f^k \,.$$

Sei nun k die kleinste gerade Zahl $\geq \ln n$, dann können wir die rechte Seite für ausreichend große n durch $(O(\ln^3 n) \cdot f)^k$ abschätzen. Damit gilt

$$\mathbf{E}\left[\text{Spur}\left[\mathbf{A}^k\right]\right] \leq (\ln^{3.3} n \cdot f)^k \,.$$

□

4.3.3 Korrektheit von Algorithmus 32

Sei $F \in \text{Form}_{n,3,p}$ mit $\ln^4 n/n^{1.5} \leq p \leq 1/\ln n$ und $X \subset L$ mit $|X| = n$ beliebig. Für jedes $z \in L$ sei $d_z = |(X, X, \{z\})|$ und

$$d = \sum_{z \in L} d_z \cdot (d_z - 1) \,.$$

Der Wert von d entspricht also genau der Anzahl der Kanten, die in G_4 innerhalb $X \times X$ liegen.

Aus Schritt 2. und 4. erhalten wir für ausreichend große n

$$\begin{aligned} \sum_{z \in L} d_z &= |(X, X, L)| = m/4 \cdot (1 \pm 1/n^{0.1}) \\ &= 2pn^3 \cdot \left(1 \pm \frac{1}{\sqrt{n}}\right) \cdot \left(1 \pm \frac{1}{n^{0.1}}\right) = 2pn^3 \cdot \left(1 \pm \frac{2}{n^{0.1}}\right) . \end{aligned}$$

Sei χ der n^2-dimensionale charakteristische Vektor von $X \times X$. Wegen $\chi^t\chi = |X \times X| = n^2$ gilt $|\chi^t \mathbf{A}\chi| \leq n^2 \cdot \|\mathbf{A}\|$. Mit der Definition von $\mathbf{A}$ erhalten wir

$$\begin{aligned} n^2 \cdot \|\mathbf{A}\| &\geq |\chi^T \mathbf{A}\chi| = \left| \sum_{(b_1,b_2,c_1,c_2)\in X^4} \mathbf{a}_{b_1c_1,b_2c_2} \right| \\ &= \left| \sum_{\substack{(b_1,b_2,c_1,c_2)\in X^4 \\ (b_1,b_2)\neq(c_1,c_2)}} \sum_{z\in V} (B_{b_1b_2z} \cdot B_{c_1c_2z} + B_{b_2b_1z} \cdot B_{c_2c_1z}) \right| \\ &= 2 \cdot \left| \sum_{\substack{(b_1,b_2,c_1,c_2)\in X^4 \\ (b_1,b_2)\neq(c_1,c_2)}} \sum_{z\in V} B_{b_1b_2z} \cdot B_{c_1c_2z} \right|. \end{aligned}$$

In obiger Doppelsumme können drei verschiedene Produkte auftreten, je nachdem welche Klauseln in F vorhanden sind und welche nicht. Sind $(b_1 \vee b_2 \vee z)$ *und* $(c_1 \vee c_2 \vee z)$ vorhanden, ist das Produkt laut der Definition der B's $(-1) \cdot (-1) = 1$. Dieses Produkt tritt genau d mal auf.

Fehlt die Klausel $(b_1 \vee b_2 \vee z)$, während $(c_1 \vee c_2 \vee z)$ vorhanden ist, so ist das Produkt $-p/(1-p)$. Für $(b_1 \vee b_2 \vee z)$ bleiben also bei festem z von den n^2 möglichen Klauseln der Art $(X \vee X \vee \{z\})$ noch $n^2 - d_z$ zur Auswahl. Wegen Schritt 3. ist dies $n^2 \cdot (1 \pm 5p)$. Für $(c_1 \vee c_2 \vee z)$ stehen genau d_z Klauseln zur Verfügung. Insgesamt tragen diese Produkte zu obiger Summe also

$$\begin{aligned} \sum_{z\in L} -\frac{p}{1-p} \cdot n^2 \cdot (1 \pm 5p) \cdot d_z &= -\frac{pn^2}{1-p} \cdot (1 \pm 5p) \cdot \sum_{z\in L} d_z \\ &= -p \cdot n^2 \cdot \frac{1 \pm 5p}{1-p} \cdot 2pn^3 \cdot \left(1 \pm \frac{2}{n^{0.1}}\right) \\ &= -2p^2n^5 \cdot \left(1 \pm \frac{1}{\sqrt{\ln n}}\right) \end{aligned}$$

bei. Der letzte Schritt gilt für n groß genug. Das Gleiche gilt für den Fall $(b_1 \vee b_2 \vee z) \in F$ und $(c_1 \vee c_2 \vee z) \notin F$.

Für den Fall, dass weder $(b_1 \vee b_2 \vee z)$ noch $(c_1 \vee c_2 \vee z)$ in F enthalten sind, hat das Produkt den Wert $p^2/(1-p)^2$. Dies trifft auf

$$\sum_{z\in L} (n^2 - d_z) \cdot (n^2 - d_z - 1) = 2n \cdot n^2 \cdot (1 \pm 5p) \cdot n^2 \cdot (1 \pm 6p)$$

Kombinationen zu. Demnach steuern diese Produkte einen Wert von

$$2p^2n^5 \cdot \frac{(1 \pm 5p) \cdot (1 \pm 6p)}{(1-p)^2} = 2p^2n^5 \cdot \left(1 \pm \frac{1}{\sqrt{\ln n}}\right)$$

bei, wenn n groß genug ist.

Insgesamt erhalten wir

$$\begin{aligned} n^2 \cdot \|\mathbf{A}\| &\geq 2 \cdot \left| d - 4p^2n^5 \cdot \left(1 \pm \frac{1}{\sqrt{\ln n}}\right) + 2p^2n^5 \cdot \left(1 \pm \frac{1}{\sqrt{\ln n}}\right) \right| \\ &= 2 \cdot \left| d - 2p^2n^5 \cdot \left(1 \pm \frac{3}{\sqrt{\ln n}}\right) \right| . \end{aligned}$$

Wegen $\|\mathbf{A}\| \leq \ln^{3.5} n \cdot p \cdot n^{1.5} \leq p^2n^3/\sqrt{\ln n}$ folgt

$$d = 2p^2n^5 \cdot \left(1 \pm \frac{4}{\sqrt{\ln n}}\right) . \tag{4.8}$$

Beachte, dass dieser Wert $\sim |E_4|/16$ ist.

Daraus wollen wir nun $|(X,X,X)| \sim pn^3$ folgern. Es ist

$$\begin{aligned} d &= \sum_{z \in L} d_z(d_z - 1) = \sum_{z \in L} d_z^2 - \sum_{z \in L} d_z \\ &= \sum_{z \in X} d_z^2 + \sum_{z \notin X} d_z^2 - 2pn^3 \cdot \left(1 \pm \frac{2}{n^{0.1}}\right) . \end{aligned} \tag{4.9}$$

Wegen der Jensen-Ungleichung (siehe A.6) ist $\sum_{z \in X} d_z^2$ minimiert, wenn alle d_z an ihrem arithmetischen Mittelwert $|(X,X,X)|/n$ sind. Demnach ist

$$\sum_{z \in X} d_z^2 \geq n \cdot \left(\frac{|(X,X,X)|}{n}\right)^2 = \frac{|(X,X,X)|^2}{n} . \tag{4.10}$$

Genauso gilt

$$\sum_{z \notin X} d_z^2 \geq \frac{|(X,X,L \setminus X)|^2}{n} . \tag{4.11}$$

Sei nun $|(X,X,X)| = pn^3 \cdot (1+\delta)$. Dann ist

$$\begin{aligned} |(X,X,L \setminus X)| &= |(X,X,L)| - |(X,X,X)| \\ &= 2pn^3 \cdot \left(1 \pm \frac{2}{n^{0.1}}\right) - pn^3 \cdot (1+\delta) \\ &= pn^3 \cdot (1-\delta) \cdot \left(1 \pm \frac{4}{n^{0.1}}\right) . \end{aligned}$$

Beim Einsetzen von (4.10) und (4.11) in (4.9) ergibt sich so

$$\begin{aligned} d &\geq \frac{(pn^3(1+\delta))^2}{n} + \frac{(pn^3 \cdot (1-\delta) \cdot (1 \pm 4/n^{0.1}))^2}{n} - 2pn^3 \cdot \left(1 \pm \frac{2}{n^{0.1}}\right) \\ &\geq p^2n^5 \cdot (1+\delta)^2 + p^2n^5 \cdot (1-\delta)^2 \pm O(p^2n^5/n^{0.1}) \\ &= 2p^2n^5 \cdot (1+\delta^2 \pm O(n^{-0.1}))\,, \end{aligned}$$

da $p \geq \ln^4 n/n^{1.5}$ ist. Mit (4.8) folgt

$$2p^2n^5 \cdot \left(1 \pm \frac{4}{\sqrt{\ln n}}\right) \geq 2p^2n^5 \cdot (1+\delta^2 \pm O(n^{-0.1}))$$

bzw. $4/\sqrt{\ln n} \geq \delta^2 - O(n^{-0.1})$, also $|\delta| \leq 3/\ln^{0.25} n$ für n groß genug. Mit $m = 8n^3p \cdot (1 \pm 1/n^{0.1})$ folgt $|(X,X,X)| \geq m/8 \cdot (1 - 4/\ln^{0.25} n)$ für ausreichend große n und damit die Korrektheit des Algorithmus. □

Abbildungsverzeichnis

Literaturverzeichnis

[AK97] Alon, Noga; Kahale, Nabil: A Spectral Technique for Coloring Random 3-Colorable Graphs. In: *SIAM Journal of Computation* 26 (1997), Nr. 6, S. 1733 – 1748

[BB99] Battiti, Roberto; Bertossi, Alan A.: Greedy, Prohibition, and Reactive Heuristics for Graph Partitioning. In: *IEEE Transactions on Computers* 48 (1999), Nr. 4, S. 361 – 385

[BKPS98] Beame, Paul; Karp, Richard M.; Pitassi, Toniann; Saks, Michael E.: On the Complexity of Unsatisfiability Proofs for Random k-CNF Formulas. In: *ACM Symposium on Theory of Computing*, 1998, S. 561 – 571

[BMPW99] Brin, Sergey; Motwani, Rajeev; Page, Lawrence; Winograd, Terry: The PageRank Citation Ranking: Bringing Order to the Web / Stanford University. 1999 (1999-66). – Forschungsbericht

[Bop87] Boppana, Ravi B.: Eigenvalues and Graph Bisection: An Average-Case Analysis (Extended Abstract). In: *IEEE Symposium on Foundations of Computer Science*, 1987, S. 280–285

[Bré79] Brélaz, Daniel: New methods to color the vertices of a graph. In: *Communications of the ACM* 22 (1979), Nr. 4, S. 251 – 256

[BS95] Blum, Avrim; Spencer, Joel: Coloring Random and Semi-Random k-Colorable Graphs. In: *Journal of Algorithms* 19 (1995), Nr. 2, S. 204 – 234

[BSMM05] Bronstein, Ilja N.; Semendjajew, Konstantin A.; Musiol, Gerhard; Mühlig, Heiner: *Taschenbuch der Mathematik*. 6. vollständig überarbeitete und ergänzte Auflage. Deutsch (Harri), 2005. – ISBN 3–817–12006–0

[Chu97] Chung, Fan R. K.: *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society, 1997. – ISBN 0–821–80315–8

[CK01] Condon, Anne; Karp, Richard M.: Algorithms for graph partitioning on the planted partition model. In: *Random Structures and Algorithms* 18 (2001), Nr. 2, S. 116 – 140

[CL02] Chung, Fan R. K.; Lu, Linyuan: Connected Components in Random Graphs with Given Expected Degree Sequences. In: *Annals of Combinatorics* 6 (2002), Nr. 2, S. 125–145

[CLV03] Chung, Fan R. K.; Lu, Linyuan; Vu, Van: The spectra of random graphs with given expected degrees. In: *Proceedings of the National Academy of Sciences of the USA* 100 (2003), Nr. 11, S. 6313 – 6318

[CO05] Coja-Oghlan, Amin: A spectral heuristic for bisecting random graphs. In: *ACM-SIAM Symposium on Discrete Algorithms*, 2005, S. 850–859

[CO06] Coja-Oghlan, Amin: An Adaptive Spectral Heuristic for Partitioning Random Graphs. In: *International Colloquium on Automata, Languages and Programming*, 2006, S. 691 – 702

[COL06] Coja-Oghlan, Amin; Lanka, André: The Spectral Gap of Random Graphs with Given Expected Degrees. In: *International Colloquium on Automata, Languages and Programming* Bd. 4051, 2006 (LNCS), S. 15 – 26

[DBM00] Dubois, Olivier; Boufkhad, Yacine; Mandler, Jacques: Typical random 3-SAT formulae and the satisfiability threshold. In: *ACM-SIAM Symposium on Discrete Algorithms*, 2000, S. 126 – 127

[DHM04] Dasgupta, Anirban; Hopcroft, John E.; McSherry, Frank: Spectral Analysis of Random Graphs with Skewed Degree Distributions. In: *IEEE Symposium on Foundations of Computer Science*, 2004, S. 602 – 610

[DM03] Dorogovtsev, Sergey N.; Mendes, José F.: *Evolution of Networks: From Biological Nets to the Internet and WWW*. New York, NY, USA : Oxford University Press, Inc., 2003. – ISBN 0–198–51590–1

[DTY06] Dun, Nan; Taura, Kenjiro; Yonezawa, Akinori: A Parallelization of State-of-the-Art Graph Bisection Algorithms on the Grid. In: *Summer United Workshops on Parallel, Distributed, and Cooprative Processing*, 2006, S. 13 – 18

[FG01] Friedman, Joel; Goerdt, Andreas: Recognizing More Unsatisfiable Random 3-SAT Instances Efficiently. In: *International Colloquium on Automata, Languages and Programming*, 2001, S. 310 – 321

[FK81] Füredi, Zoltán; Komlós, János: The eigenvalues of random symmetric matrices. In: *Combinatorica* 1 (1981), Nr. 3, S. 233 – 241

[FKS89] Friedman, Joel; Kahn, Jeff; Szemerédi, Endre: On the second eigenvalue of random regular graphs. In: *ACM Symposium on Theory of Computing*, 1989, S. 587 – 598

[Fla03] Flaxman, Abraham: A spectral technique for random satisfiable 3CNF formulas. In: *ACM-SIAM Symposium on Discrete Algorithms*, 2003, S. 357 – 363

[FO04] Feige, Uriel; Ofek, Eran: Easily Refutable Subformulas of Large Random 3CNF Formulas. In: *International Colloquium on Automata, Languages and Programming*, 2004, S. 519 – 530

[Fri99] Friedgut, Ehud: Necessary and sufficient conditions for sharp thresholds of graph properties and the k-SAT problem. In: *Journal of the American Mathematical Society* 12 (1999), S. 1017 – 1054

[Fu95] Fu, Xudong: *On the Complexity of Proof Systems*, University of Toronto, Diss., 1995

[GL03] Goerdt, Andreas; Lanka, André: Recognizing more random unsatisfiable 3-SAT instances efficiently. In: *Electronic Notes in Discrete Mathematics* 16 (2003), S. 21 – 46

[GV01] Golub, Gene H.; Van Loan, Charles F.: *Matrix Computations*. 3. John Hopkins University Press, 2001. – ISBN 0–801–95414–8

[Hås01] Håstad, Johan: Some optimal inapproximability results. In: *Journal of the Association for Computing Machinery* 48 (2001), Nr. 4, S. 798 – 859

[HS03] Hajiaghayi, MohammadTaghi; Sorkin, Gregory B.: The Satisfiability Threshold of Random 3-SAT Is at Least 3.52 / IBM Research Report. 2003 (RC22942). – Forschungsbericht

[Int04] Interian, Yannet: Approximation Algorithm for Random MAX-kSAT. In: *International Conference on Theory and Applications of Satisfiability Testing*, 2004, S. 173 – 182

[Jen06] Jensen, Johan L.: Sur les fonctions convexes et les inegalites entre les valeurs moyennes. In: *Acta Mathematica* 30 (1906), S. 175 – 193

[JLR00] Janson, Svante; Luczak, Tomasz; Rucinski, Andrzej: *Random Graphs*. John Wiley and Sons, Inc., 2000. – ISBN 0–471–17541–2

[KKL03] Kaporis, Alexis C.; Kirousis, Lefteris M.; Lalas, Efthimios G.: Selecting complementary pairs of literals. In: *Electronic Notes in Discrete Mathematics* 16 (2003), S. 47 – 70

[KS92] Kautz, Henry A.; Selman, Bart: Planning as Satisfiability. In: *European Conference on Artificial Intelligence*, 1992, S. 359 – 363

[KS03] Krivelevich, Michael; Sudakov, Benny: The Largest Eigenvalue Of Sparse Random Graphs. In: *Combinatorics, Probability & Computing* 12 (2003), Nr. 1, S. 61 – 72

[Kuč77] Kučera, Luděk: Expected Behavior of Graph Coloring Algorithms. In: *Fundamentels of Computation Theory* Bd. 56, Springer, 1977 (Lecture Notes in Computer Science), S. 447 – 451

[McS01] McSherry, Frank: Spectral Partitioning of Random Graphs. In: *IEEE Symposium on Foundations of Computer Science*, 2001, S. 529 – 537

[MP02] Mihail, Milena; Papadimitriou, Christos: On the Eigenvalue Power Law. In: *Randomization and Approximation Techniques*, 2002, S. 254 – 262

[MT96] Mehrotra, Anuj; Trick, Michael A.: A Column Generation Approach for Graph Coloring. In: *INFORMS Journal on Computing* 8 (1996), S. 344 – 354

[Rie07] Riediger, Steffen: Implementierung eines Algorithmus zur Partitionierung von Graphen. 2007. – Studienarbeit

[Rie08] Riediger, Steffen: Schnelle Partitionierung von real-world- und Zufallsgraphen. 2008. – Diplomarbeit

[VK02] Vega, Wenceslas F. l.; Karpinski, Marek: 9/8-Approximation Algorithm for Random MAX-3SAT. In: *Electronic Colloquium on Computational Complexity* (2002), Nr. 70

[Zwi03] Zwillinger, Daniel: *Standard Mathematical Tables and Formulae*. 31. Chapman & Hall, 2003. – ISBN 1–584–88291–3

Anhang A

Elementare Abschätzungen

A.1. $\binom{n}{k} \leq \left(\frac{\mathrm{e} \cdot n}{k}\right)^k$ für k groß genug

Stirlings Formel (siehe z. B. (8.103h) in [BSMM05]) besagt $k! = \left(\frac{k}{\mathrm{e}}\right)^k \cdot \sqrt{2\pi k} \cdot (1 + O(1/k)) \geq (k/\mathrm{e})^k$. Es ergibt sich für ausreichend große k

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \frac{n^k}{(k/\mathrm{e})^k} = \left(\frac{\mathrm{e} \cdot n}{k}\right)^k .$$

A.2. $1 + x \leq \mathrm{e}^x$ für alle $x \in \mathbb{R}$

Wir betrachten die Differenzfunktion $f(x) = \mathrm{e}^x - x - 1$. Dann ist

$$\begin{aligned} & f'(x) = \mathrm{e}^x - 1 \\ \text{und} \qquad & f''(x) = \mathrm{e}^x . \end{aligned}$$

Wegen $f''(x) = \mathrm{e}^x > 0$ ist f konvex. Da $f'(0) = 0$ ist, ist das Minimum von f bei $f(0) = 0$. Daher ist f für kein $x \in \mathbb{R}$ negativ, also $1 + x \leq \mathrm{e}^x$.

A.3. $\mathrm{e}^x - 1 \leq \mathrm{e}^c \cdot x^2/2 + x$ für $x \leq c$

Erneut untersuchen wir die Differenzfunktion $f(x) = e^c \cdot x^2/2 + x - (e^x - 1)$. Deren Ableitungen

$$\begin{aligned} & f'(x) = \mathrm{e}^c \cdot x + 1 - \mathrm{e}^x \\ \text{und} \qquad & f''(x) = \mathrm{e}^c - \mathrm{e}^x \end{aligned}$$

zeigen, dass f für $x \leq c$ konvex ist und das (wegen der Konvexität einzige) Minimum im Intervall $(-\infty : c]$ bei $f(0) = 0$ liegt. Daher ist f in diesem Intervall ≥ 0.

A.4. $k \cdot x \cdot \ln(k/x)$ ist für $0 < x < k/\mathrm{e}$ monoton wachsend in x
Die erste Ableitung nach x

$$\frac{\partial(k \cdot x \cdot \ln(k/x))}{\partial x} = k \cdot \ln(k/x) + k \cdot x \cdot \frac{1}{k/x} \cdot \frac{-k}{x^2} = k \cdot (\ln(k/x) - 1)$$

ist für $0 < x < k/\mathrm{e}$ positiv.

A.5. Für $x \in (0:1)$ ist $-\log x < 4/\sqrt{x}$.
Die erste Ableitung von $f(x) = 4/\sqrt{x} + \log x$ ist

$$f'(x) = -\frac{2}{\sqrt{x^3}} + \frac{1}{\ln 2} \cdot \frac{1}{x} = \frac{1}{x} \cdot \left(\frac{1}{\ln 2} - \frac{2}{\sqrt{x}} \right).$$

Für $x \leq 1$ ist der letzte Faktor negativ und wegen $x > 0$ das gesamte Produkt. Also ist f in diesem Intervall monoton fallend. Zusammen mit $f(1) = 4 > 0$ ist f folglich im ganzen Intervall $(0:1)$ positiv, also ist $4/\sqrt{x} > -\log x$.

A.6. Seien $x_1, x_2, \ldots, x_k$ reelle Zahlen. Dann gilt für jede konvexe Funktion $f : \mathbb{R} \to \mathbb{R}$:

$$\sum_{i=1}^{k} f(x_i) \geq \sum_{i=1}^{k} f\left(\frac{\sum_{i=1}^{k} x_i}{k} \right) = k \cdot f\left(\frac{\sum_{i=1}^{k} x_i}{k} \right).$$

Diese Beziehung stellt einen Spezialfall (alle $\lambda_i = 1/k$) der folgenden Jensen-Ungleichung ([Jen06]) für konvexe Funktionen dar.

Fakt 34. *Sei $f : \mathbb{R} \to \mathbb{R}$ eine konvexe Funktion und $x_1, x_2, \ldots, x_k$ beliebige reelle Zahlen. Seien außerdem $\lambda_1, \lambda_2, \ldots, \lambda_k$ positive reelle Zahlen, deren Summe 1 ist. Dann gilt*

$$f\left(\sum_{i=1}^{k} \lambda_i \cdot x_i \right) \leq \sum_{i=1}^{k} \lambda_i \cdot f(x_i).$$

Beweis. Für $k = 1$, folgt sofort $\lambda_1 = 1$ und damit $f(1 \cdot x_1) \leq 1 \cdot f(x_1)$, was immer wahr ist. Sei nun $k = 2$. Da f konvex ist, liegt die Gerade g durch die Punkte $(x_1, f(x_1))$ und $(x_2, f(x_2))$ im Intervall $[x_1, x_2]$ nie unter der Funktion f. Mit der Darstellung

$$g(x) = \frac{f(x_2) - f(x_1)}{x_1 - x_2} \cdot (x_1 - x) + f(x_1)$$

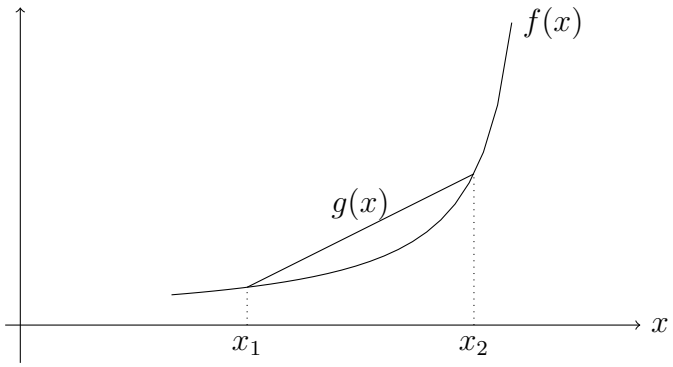


Abbildung A.1: Lagebeziehung zwischen f und g.

von g erhalten wir $g(\lambda_1 \cdot x_1 + \lambda_2 \cdot x_2) = \lambda_1 \cdot f(x_1) + \lambda_2 \cdot f(x_2)$:

$$\begin{aligned} g(\lambda_1 \cdot x_1 + \lambda_2 \cdot x_2) &= \frac{f(x_2) - f(x_1)}{x_1 - x_2} \cdot (x_1 - (\lambda_1 \cdot x_1 + \lambda_2 \cdot x_2)) + f(x_1) \\ &= \frac{f(x_2) - f(x_1)}{x_1 - x_2} \cdot ((1 - \lambda_1) \cdot x_1 - \lambda_2 \cdot x_2) + f(x_1) \\ &= \frac{f(x_2) - f(x_1)}{x_1 - x_2} \cdot (\lambda_2 \cdot x_1 - \lambda_2 \cdot x_2) + f(x_1) \\ &= (f(x_2) - f(x_1)) \cdot \lambda_2 + f(x_1) \\ &= f(x_1) \cdot (1 - \lambda_2) + \lambda_2 \cdot f(x_2) = \lambda_1 \cdot f(x_1) + \lambda_2 \cdot f(x_2)\,. \end{aligned}$$

Damit folgt die Behauptung, weil

$$f(\lambda_1 \cdot x_1 + \lambda_2 \cdot x_2) \leq g(\lambda_1 \cdot x_1 + \lambda_2 \cdot x_2) = \lambda_1 \cdot f(x_1) + \lambda_2 \cdot f(x_2)$$

ist.

Den Fall $k > 2$ beweisen wir induktiv. Den nötigen Induktionsanfang haben wir für $k = 2$ schon bewiesen. Sei nun also $k > 2$. Wir zerlegen die Summe

$$\sum_{i=1}^{k} \lambda_i \cdot x_i = (1 - \lambda_k) \left(\sum_{i=1}^{t-1} \frac{\lambda_i}{1 - \lambda_k} \cdot x_i \right) + \lambda_k \cdot x_k\,.$$

Mit dem Induktionsanfang ergibt sich

$$f\left(\sum_{i=1}^{k} \lambda_i \cdot x_i \right) \leq (1 - \lambda_k) \cdot f\left(\sum_{i=1}^{t-1} \frac{\lambda_i}{1 - \lambda_k} \cdot x_i \right) + \lambda_k \cdot f(x_k)\,.$$

Wir schließen nochmals induktiv

$$f\left(\sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} \cdot x_i \right) \leq \sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} \cdot f(x_i)\,.$$

Die letzten beiden Ungleichungen ergeben zusammen

$$f\left(\sum_{i=1}^{k} \lambda_i \cdot x_i\right) \leq (1-\lambda_k) \cdot \sum_{i=1}^{k-1} \frac{\lambda_i}{1-\lambda_k} \cdot f(x_i) + \lambda_k \cdot f(x_k) = \sum_{i=1}^{k} \lambda_i \cdot f(x_i)\,.$$

□

Thesen

Bei der Partitionierung von Graphen versucht man, die Knoten anhand gemeinsamer Merkmale zu gruppieren. Die Gemeinsamkeiten ergeben sich aus der Verteilung der Kanten. Man kann zum Beispiel versuchen, eine 3-Färbung (keine Kanten innerhalb der drei Mengen) zu finden. Da bereits die Entscheidungsvariante dieses Problems $\mathcal{NP}$-vollständig ist, sind keine Polynomialzeitalgorithmen für dieses Problem zu erwarten.

Um average-case-Aussagen zu erhalten, pflanzt man in einen – ansonsten zufälligen – Graphen eine 3-Färbung ein. Nun analysiert man, inwieweit Algorithmen in der Lage sind, die gepflanzte 3-Färbung zu rekonstruieren. Alon und Kahale [AK97] konnten zeigen, dass dies in Polynomialzeit gelingt, wenn das $G_{n,p}$-Modell zu Grunde liegt. Dabei bediente sich ihr Algorithmus spektraler Methoden.

Das $G_{n,p}$-Modell repräsentiert reale Graphen unzureichend, da alle Knoten den gleichen erwarteten Grad haben. Natürliche Graphen wie der des `www` oder auch soziale Netzwerke haben eine sehr irreguläre Gradverteilung (z. B. power-law). Man geht daher auf allgemeinere Modelle über. Zum Beispiel auf das Chung-Lu-Modell, wo der erwartete Grad jedes Knotens frei festgelegt werden kann.

These 1 Spektrale Methoden helfen, das Problem auch im Chung-Lu-Modell mit gepflanzter 3-Färbung in Polynomialzeit zu lösen. Die Erfolgswahrscheinlichkeit geht mit wachsender Knotenzahl gegen 1.

Es gibt noch zahllose andere Partitionierungsprobleme. Daher ist man an Algorithmen interessiert, die nicht wissen, welche Struktur in den Graphen gepflanzt wurde, sondern sich an unterschiedliche Situationen anpassen. Man spricht vom allgemeinen Partitionierungsproblem.

These 2 Mit spektralen Methoden lässt sich eine (fast) beliebige gepflanzte Partition in Polynomialzeit wiederfinden. Die Erfolgswahrscheinlichkeit geht mit wachsender Knotenzahl gegen 1.

Für das Chung-Lu-Modell mit gepflanzter Partition wurden in [DHM04] erste Ergebnisse erzielt, die jedoch recht dichte Graphen voraussetzen.

These 3 Um das allgemeine Partitionierungsproblem im Chung-Lu-Modell mit gepflanzter Partition zu lösen, genügt es, dass die Kantenzahl linear in der Knotenzahl ist.

Die Thesen 1-3 wurden in Kapitel 2 der Arbeit bestätigt.

Das Verfahren aus [DHM04] ist unpraktikabel, da zur Eingabe Informationen gehören, die bei realen Graphen nicht zur Verfügung stehen. Die folgende These (die alle vorherigen erweitert) konnte in Kapitel 3 bestätigt werden.

These 4 Es gibt einen Algorithmus für das allgemeine Partitionierungsproblem im Chung-Lu-Modell mit gepflanzter Partition mit den Eigenschaften:

- Die Eingabe besteht ausschließlich aus dem Graphen.
- Der Algorithmus hat polynomielle Laufzeit.
- Der Anteil der falsch gruppierten Knoten geht mit wachsendem Durchschnittsgrad gegen 0.
- Die Erfolgswahrscheinlichkeit geht mit wachsender Knotenzahl gegen 1.
- Es genügt, dass die Kantenzahl linear in der Knotenzahl ist.

Ein anderes fundamentales Problem ist Max3Sat. Bei einer gegebenen aussagenlogischen Formel in 3KNF ist das Ziel, eine Belegung zu finden, die möglichst viele Klauseln wahr macht. Genauso interessant ist es, eine obere Schranke für die Anzahl der gleichzeitig erfüllbaren Klauseln zu finden. Da dies co-$\mathcal{NP}$-schwer ist, untersucht man auch hier, wie gut das bei zufälligen Instanzen möglich ist.

These 5 Es gibt einen Algorithmus für Formeln, generiert im $\text{Form}_{n,3,p}$-Modell, mit den Eigenschaften:

- Die Eingabe besteht aus der Formel und p.
- Der Algorithmus hat polynomielle Laufzeit.
- Der Algorithmus gibt stets eine obere Schranke.
- Die Wahrscheinlichkeit für eine nahezu optimale Antwort geht mit wachsender Knotenzahl gegen 1.
- Es genügt, dass die Klauselzahl $\geq \ln^4 n \cdot n^{1.5}$ ist.

Kapitel 4 zeigt, dass auch diese These korrekt ist.

Publikationen

Zeitschriftenbeiträge

[1] Coja-Oghlan, Amin; Goerdt, Andreas; Lanka, André: *Spectral Partitioning of Random Graphs with Given Expected Degrees.* 2007. – Eingereicht

[2] Coja-Oghlan, Amin; Lanka, André: *Finding Planted Partitions in Random Graphs with General Degree Distributions.* 2007. – Eingereicht

[3] Coja-oghlan, Amin; Goerdt, Andreas; Lanka, André: Strong Refutation Heuristics for Random k-SAT. In: *Combinatorics, Probability & Computing* 16 (2007), Nr. 1, S. 5–28

[4] Coja-Oghlan, Amin; Goerdt, Andreas; Lanka, André; Schädlich, Frank: Techniques from combinatorial approximation algorithms yield efficient algorithms for random 2k-SAT. In: *Theoretical Computer Science* 329 (2004), Nr. 1-3, S. 1–45

[5] Goerdt, Andreas; Lanka, André: Recognizing more random unsatisfiable 3-SAT instances efficiently. In: *Electronic Notes in Discrete Mathematics* 16 (2003), S. 21 – 46

Konferenzbeiträge

[1] Coja-Oghlan, Amin; Lanka, André: *Partitioning Random Graphs with General Degree Distributions.* 2008. – Eingereicht

[2] Coja-Oghlan, Amin; Goerdt, Andreas; Lanka, André: Spectral partitioning of random graphs with given expected degrees. In: *4th IFIP International Conference on Theoretical Computer Science* Bd. 209, 2006 (IFIP), S. 271–282

[3] Coja-Oghlan, Amin; Lanka, André: The Spectral Gap of Random Graphs with Given Expected Degrees. In: *International Colloquium on*

Automata, Languages and Programming Bd. 4051, 2006 (LNCS), S. 15 – 26

[4] Goerdt, Andreas; Lanka, André: On the Hardness and Easiness of Random 4-SAT Formulas. In: *ISAAC* Bd. 3341, 2004 (LNCS), S. 470–483

[5] Coja-Oghlan, Amin; Goerdt, Andreas; Lanka, André: Strong Refutation Heuristics for Random k-SAT. In: *RANDOM* Bd. 3122, 2004 (LNCS), S. 310–321

[6] Goerdt, Andreas; Lanka, André: Recognizing more Random Unsatisfiable 3-SAT Instances efficiently. In: *LICS Workshop Typical Case Complexity and Phase Transitions*, 2003

[7] Coja-Oghlan, Amin; Goerdt, Andreas; Lanka, André; Schädlich, Frank: Certifying Unsatisfiability of Random 2k-SAT Formulas using Approximation Techniques. In: *FCT* Bd. 2751, 2003 (LNCS), S. 15–26

Lebenslauf

Persönliche Angaben

geboren am:	25.09.1978
in:	Lutherstadt Eisleben
Familienstand:	ledig
Kinder:	Helena Sophie *2002 Dora Marie *2004

Ausbildung und berufliche Tätigkeit

05/2002 – 03/2008	wissenschaftlicher Mitarbeiter an der TU Chemnitz Professur „Theoretische Informatik“ Lehre: Theoretische Informatik I-III Datenschutz und Datensicherheit Maschinenorientierte Programmierung Rechnerarchitektur Softwarepraktikum Höhere Programmiersprachen (LWB) Multimedia (LWB)
10/1998 – 04/2002	Informatikstudium an der TU Chemnitz Vertiefungsgebiet Theoretische Informatik Nebenfach Mathematik Abschluss: „Mit Auszeichnung“ Durchführung Übungen Theoretische Informatik I,II
10/1997 – 06/1998	Grundwehrdienst als Fernmelder
09/1991 – 07/1997	Gymnasium an der Bergmannsallee Eisleben Abiturnote 1.3
09/1985 – 07/1991	POS „Ernst Thälmann“ Eisleben

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne unzulässige Hilfe angefertigt habe. Die Arbeit wurde nicht anderweitig (auch nicht in ähnlicher Form) zu Prüfungszwecken vorgelegt oder veröffentlicht. Es wurden keine anderen als die angegebenen Hilfsmittel verwendet. Gedanken aus fremden Quellen, die direkt oder indirekt übernommen wurden, sind als solche kenntlich gemacht.

Bei der geistigen Herstellung der Arbeit waren keine weiteren Personen beteiligt, insbesondere wurde auch nicht die Hilfe eines Promotionsberaters in Anspruch genommen. Dritte haben weder mittelbar noch unmittelbar geldwerte Leistungen für Arbeiten erhalten, die mit dem Inhalt der vorgelegten Dissertation in Zusammenhang stehen.

Chemnitz, den 18.01.2008

André Lanka