



TECHNISCHE UNIVERSITÄT CHEMNITZ

---

Fakultät für Informatik  
Professur für Theoretische Informatik

# Diplomarbeit

Grenzen von Algorithmen zur Exploration  
sehr großer Graphen

Markus John

Chemnitz, den 22. August 2006

**Betreuer:** Prof. Dr. Andreas Goerdts

# Inhaltsverzeichnis

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Aufgabenstellung</b>                              | <b>1</b>  |
| <b>2</b> | <b>Einleitung</b>                                    | <b>2</b>  |
| 2.1      | Breitensuche auf Zufallsgraphen . . . . .            | 3         |
| 2.2      | Wahrscheinlichkeitsbegriffe . . . . .                | 5         |
| <b>3</b> | <b>Erwartungswert der Gradverteilung</b>             | <b>7</b>  |
| 3.1      | Dichte- und Verteilungsfunktion . . . . .            | 7         |
| 3.2      | Analyse der Breitensuche . . . . .                   | 8         |
| 3.3      | Schranken der Knoten- und Kopienanzahl . . . . .     | 9         |
| 3.4      | Wahrscheinlichkeit einer Suchbaumkante . . . . .     | 12        |
| 3.5      | Zusammenfassung . . . . .                            | 22        |
| <b>4</b> | <b>Konzentration um den Erwartungswert</b>           | <b>24</b> |
| 4.1      | Betrachtung der $\mathbf{B}_j(\mathbf{i})$ . . . . . | 28        |
| <b>5</b> | <b>Zusammenhang der Verteilungen</b>                 | <b>33</b> |
| 5.1      | Ausblick . . . . .                                   | 35        |
|          | <b>Literaturverzeichnis</b>                          | <b>37</b> |

# 1 Aufgabenstellung

*Name, Vorname des Diplomanden:* John, Markus

*Immatrikulations-Nr.:* XXXXX

**Thema:**  
**Grenzen von Algorithmen zur Exploration sehr großer Graphen**

## **Zielstellung:**

In vielen neueren Anwendungssituationen treten derart große Graphen auf, dass eine vollständige Kenntnis des Graphen, etwa durch Abspeicherung seiner Adjazenzlistendarstellung, nicht möglich ist. Ein Beispiel hier ist etwa der dem WWW zu Grunde liegende Graph. Neben der Größe des Graphen kann es auch andere Gründe dafür geben, dass dieser nicht abgespeichert werden kann; etwa einfach weil es zu lang dauern würde, ihn in seiner Gesamtheit zu ermitteln. Das ist zum Beispiel beim Internetgraphen (dieser unterscheidet sich vom WWW Graph) der Fall.

Trotzdem ist man an Informationen über derartige Graphen interessiert, und es gibt Algorithmen, die aus der partiellen Exploration des Graphen Schlüsse über seine Struktur ziehen. Der Algorithmus traceroute ist ein Verfahren, derartige Informationen zu ermitteln. Die Aufgabe der Diplomarbeit ist es, zur theoretischen Analyse dieses Algorithmus beizutragen.

Markus John soll die neueren Arbeiten von Achlioptas et al. zu dieser Thematik durcharbeiten und in eigenen Worten darstellen. Die in den Beweisen vorhandenen Lücken sollen geschlossen werden.

Anhand der bewiesenen Resultate soll auf die Qualität der von dem Algorithmus ermittelten Information, insbesondere über die Gradverteilung des betrachteten Graphen, geschlossen werden. Konkret soll die Frage beantwortet werden, inwieweit das allgemein beobachtete „power law“ der Gradverteilung großer Graphen eventuell nur die Konsequenz des verwendeten Explorationsalgorithmus ist.

## 2 Einleitung

Der hohe Abstraktionsgrad von Graphen erschließt diesen vielfältige Anwendungsmöglichkeiten. Ob soziale Strukturen, Unternehmensabläufe oder Verknüpfungen elektronischer Komponenten, durch Graphen lassen sich Zusammenhänge einfach ausdrücken und analysieren. Viele Algorithmen auf Graphen setzen jedoch vollständige Kenntnis der Knoten und Kanten voraus – ganz im Gegensatz zur Realität, in der oft nur begrenzt Informationen vorhanden sind.

Die Gründe für diese Einschränkungen sind zahlreich: Entweder übersteigt der Aufwand die zur Verfügung stehenden Mengen an Speicherplatz und Zeit oder es ist problembedingt sehr schwer bzw. überhaupt nicht möglich, den Graphen vollständig zu ermitteln. Ein Beispiel dafür ist der sogenannte Internetgraph. Betrachtet man alle an das Internet angeschlossenen Komponenten (Bürocomputer, Router, Webserver etc.), so lassen sich deren Verbindungen untereinander per Graph visualisieren [8]. In diesem Fall ist es aufgrund der großen Menge bereits vorhandener Teilnehmer und der enormen Fluktuation im Internet nicht möglich, eine lückenlose Momentaufnahme von diesem zu erhalten.

Je nach Fragestellung können aber auch unvollständige Informationen ausreichend sein. Ein interessanter und häufig untersuchter Aspekt des Internetgraphen ist dessen Gradverteilung. Wenn es möglich wäre, einen repräsentativen Ausschnitt zu untersuchen, so ließen sich eventuell Rückschlüsse auf die Verteilung der Knotengrade im gesamten Internet ziehen. Das in [8] dazu verwendete Verfahren wird als „traceroute sampling“ bezeichnet. Es basiert auf der Tatsache, dass eine Breitensuche auf einem ungewichteten Graphen einen Baum bestehend aus kürzesten Wegen vom Startknoten der Breitensuche zu allen erreichbaren Knoten liefert.

Für Netzwerke existiert ein Programm namens traceroute, welches den Weg einer Datenübertragung zwischen zwei Teilnehmern ermitteln und ausgeben kann. Aufgrund der Funktionsweise des Internets entsprechen diese Wege sehr häufig den jeweils kürzesten Verbindungen zwischen den betrachteten Komponenten. Bei mehrfacher Verwendung mit gleichem Start-, aber verschiedenen Zielpunkten erhält man als Vereinigung der dabei gefundenen Wege einen Ausschnitt eines Suchbaums oder zumindest eine sehr gute Näherung an diesen.

Die Anwendung von traceroute liefert somit unter diesen Bedingungen ein ähnliches Ergebnis wie eine auf dem Internetgraphen durchgeführte und vorzeitig abgebrochene Breitensuche. Weitere Details über den Zusammenhang von Breitensuche und traceroute sind der vorherigen Arbeit des Autors [6, Kapitel 2] zu entnehmen. Verbesserungen und Erweiterungen des „traceroute sampling“ wie Mercator [5] oder verteiltes traceroute [2] erreichen eine noch höhere Genauigkeit. Auch sie mussten jedoch zu dem Schluss kommen, dass eine vollständige Kenntnisnahme des Graphen bisher nicht möglich ist.

In Bezug auf [8] ergaben nachfolgende Untersuchungen eine power-law Verteilung für die Knotengrade im Internet [4]. Es folgten auf dem Wachstum des Netzes basierende Erklä-

rungsversuche [3], aber auch Untersuchungen, ob denn nicht die ermittelte Verteilung durch das eingesetzte Explorationsverfahren begünstigt oder gar verursacht wird [7].

Die vorliegende Arbeit basiert auf Achlioptas et al. [1] und setzt die vom Autor in [6] begonnenen Ausführungen fort. Es soll ein direkter Zusammenhang zwischen ursprünglicher und durch die Breitensuche ermittelter Gradverteilung zufälliger Graphen in Form einer Gleichung erarbeitet werden. Die erhaltenen Ergebnisse lassen sich aufgrund der erwähnten Verbindung zwischen Breitensuche und traceroute verwenden, um die Resultate von „traceroute sampling“ auf solchen randomisierten Graphen zu berechnen und qualitativ einzuschätzen.

## 2.1 Breitensuche auf Zufallsgraphen

*Definition 2.1 (Verteilung der Grade, Definition 3.1 aus [6])*

Ein Graph mit  $n$  Knoten besitzt die *Gradverteilung*  $A = (a_0, a_1, \dots)$ , wenn er genau  $a_k \cdot n$  Knoten vom Grad  $k$  enthält.

Die folgenden Betrachtungen beschränken sich auf zufällige Graphen mit festgelegten angemessenen Gradverteilungen nach Definition 2.2. So kann sichergestellt werden, dass der Graph mit hoher Wahrscheinlichkeit zusammenhängend [6, Lemma 3.3] und der Durchschnittsgrad endlich ist [6, Lemma 3.4] – beides Voraussetzungen für das Resultat dieser Arbeit, Satz 5.1.

*Definition 2.2 (angemessene Verteilung, Definition 3.2 aus [6])*

Seien  $\alpha > 2$  und  $C > 0$  konstant. Eine Gradverteilung  $A = (a_0, a_1, \dots)$  nennt man *angemessen*, wenn  $a_k = 0$  für  $k < 3$  und  $a_k < C \cdot k^{-\alpha}$  für alle  $k \geq 3$  gilt.

Ausgehend von einer solchen Gradverteilung wird eine Gradsequenz erstellt, die für jeden Knoten dessen exakten Grad vorgibt. Darüber lassen sich nun zufällige Kanten zwischen den Knoten ermitteln. Man ersetzt jeden Knoten durch eine Menge von Kopien, wobei die Anzahl der Kopien dem Knotengrad entspricht. Anstatt nun jedem Knoten eine variierende Anzahl von Kanten zuzuordnen zu müssen, ist es ausreichend, die Kanten so zu ermitteln, dass jede Kopie genau einer Kante zugeordnet wird. Die inzidenten Kanten eines Knotens entsprechen dann der Menge aller inzidenten Kanten seiner Kopien.

|                 | noch nicht in Q           | in Q                   | nicht mehr in Q            |
|-----------------|---------------------------|------------------------|----------------------------|
| Knoten (vertex) | unentdeckt<br>(untouched) | entdeckt<br>(pending)  | abgearbeitet<br>(explored) |
| Kopien (copy)   | unentdeckt<br>(untouched) | entdeckt<br>(enqueued) | abgearbeitet<br>(exposed)  |

Tabelle 2.1: Knotenbezeichnungen, Tabelle 3.1 aus [6]

Die Festlegung, welche Kopie in welcher Kante enthalten ist, lässt sich sehr einfach über ein perfektes Matching ermitteln. Die Kopien des Graphen werden uniform zufällig auf zwei gleichmächtige Mengen aufgeteilt, zwischen denen ein Matching gemäß der Gleichverteilung ermittelt wird. Das Matching selbst, d.h. die Paare von jeweils zwei Kopien, die aufeinander abgebildet werden, ergibt die Menge der Kanten. Durch dieses Verfahren ist garantiert, dass

jede Kopie exakt einer Kante zugeteilt wird. Es lässt sich auf diesem Wege nicht vermeiden, dass Schleifen oder Mehrfachkanten generiert werden. Für die meisten Knoten ist dies jedoch unwahrscheinlich und kann in den weiteren Betrachtungen vernachlässigt werden, wie durch Lemma 3.4 gezeigt wird.

Um die nachfolgenden Berechnungen zu vereinfachen, werden alle Schritte nach der Erstellung der Gradsequenz (Generierung der Kopien und Bestimmung des Matchings) in die Breitensuche integriert. Ausserdem wird das Matching nicht explizit vor Eintritt in die Schleife der Breitensuche ermittelt, sondern implizit über sogenannte Indices realisiert. Jede Kopie bekommt dafür einen zufälligen Index aus dem Intervall  $[0, 1] \subseteq \mathbb{R}$  zugeteilt. Bei der Betrachtung des Kopfs der Schlange in der Breitensuche wird als adjazente Kopie diejenige gewählt, die unter allen noch nicht abgearbeiteten Kopien den größten Index besitzt. Da die Indices zufällig gemäß Gleichverteilung bestimmt sind, unterscheidet sich die Verteilung des Ergebnisses nicht von der eines zu Anfang festgelegten Matchings.

---

**Algorithmus 1** zeitbasierte Breitensuche, Algorithmus 1 aus [6]

---

**Eingabe:** randomisierter Multigraph (Knotenmenge  $V$  mit Graden  $d_v$ ); Startknoten  $u_s \in V$

**Ausgabe:** Breitensuchbaum  $T = (V_T, E_T)$

```

1: function BFS
2:   for all  $u \in V$  do
3:     Erzeuge  $d_u$  viele Kopien  $u^{(1)}, u^{(2)}, \dots, u^{(d_u)}$ 
4:     Bestimme für jede Kopie  $u^{(i)}$  einen Index (zufällig, uniform)  $x_{u^{(i)}} \in [0, 1] \subseteq \mathbb{R}$ 
5:     Markiere alle Kopien als unentdeckt (ausgenommen Kopien des Startknotens  $u_s$ )
6:      $T = (\{u_s\}, \emptyset)$  // Breitensuchbaum anfangs nur der Startknoten
7:      $Q = (u_s^{(i)} \mid 1 \leq i \leq d_{u_s})$  // Markiere alle Kopien des Startknotens als entdeckt
8:     while  $Q \neq \emptyset$  do
9:       Entnimm die Kopie  $v^{(i)}$  vom Kopf der Schlange  $Q$ 
10:      Bestimme die Kopie mit dem größten Index unter allen nicht abgearbeiteten Kopien
      ohne  $v^{(i)}$ 
11:      Sei diese Kopie  $w^{(j)}$ . Diese wird adjazente Kopie von  $v^{(i)}$ 
12:      if  $w^{(j)}$  ist unentdeckt then //  $w^{(j)} \notin Q$ , d.h. Knoten  $w$  bisher nicht betreten
13:         $T = (V_T \cup \{w\}, E_T \cup \{(v, w)\})$  // Neuer Knoten und Kante zu  $T$  hinzufügen
14:        Füge alle Kopien von  $w$  außer  $w^{(j)}$  an das Ende von  $Q$  an
15:      else //  $w^{(j)} \in Q$ , d.h. Knoten  $w$  schon erreicht
16:        Entferne  $w^{(j)}$  aus  $Q$ 
17:      Markiere  $v^{(i)}$  und  $w^{(j)}$  als abgearbeitet
18:    return  $T$ 

```

---

Algorithmus 1 beschreibt den vollständigen Pseudocode der abgeänderten Breitensuche. Man kann sich diese fortlaufende Breitensuche als von einem Zeitpunkt  $t$  abhängig vorstellen: Der Algorithmus startet zum Zeitpunkt  $t = 1$  und endet zum Zeitpunkt  $t = 0$ . Dazwischen entspricht  $t$  dem Index der Kopie  $w^{(j)}$ , d.h. der zum Kopf der Schlange adjazenten Kopie mit dem größten Index aller verbleibenden Kopien.

Wie man an Lemma 3.1 sehen kann, ermöglicht dieser Standpunkt sehr elegante Angaben darüber, wieviele Kopien zu einem beliebigen Zeitpunkt bereits abgearbeitet oder entdeckt wurden und somit in der Schlange enthalten sind. Ohne diese Betrachtungsweise wäre es sehr

viel schwieriger, den Verlauf der Breitensuche in entsprechenden Gleichungen auszudrücken. Ausführlichere Beschreibungen über die veränderte Breitensuche und die zugehörigen Grundlagen bei der Erstellung zufälliger Graphen lassen sich in [6, Kapitel 3] nachlesen. Insbesondere das dort enthaltene Beispiel könnte für das Verständnis recht hilfreich sein.

## 2.2 Wahrscheinlichkeitsbegriffe

In den Beweisen ab Kapitel 3 werden sehr oft die Formulierungen „mit hoher Wahrscheinlichkeit“ und „mit sehr hoher Wahrscheinlichkeit“ verwendet. Diese drücken die Entwicklung der Wahrscheinlichkeit für das Eintreffen eines Ereignisses aus, wenn eine gewisse Bezugsgröße (in dieser Arbeit entspricht die Bezugsgröße immer der Anzahl Knoten  $n$ ) gegen Unendlich läuft.

*Definition 2.3 (hohe Wahrscheinlichkeit)*

Eine Folge von Ereignissen  $\mathcal{E}_n$  tritt mit *hoher Wahrscheinlichkeit* (m.h.W.) ein, falls für  $n \rightarrow \infty$  gilt

$$\Pr[\mathcal{E}_n] = 1 - o(1) .$$

Die Wahrscheinlichkeit eines solchen Ereignisses nähert sich für wachsende  $n$  beliebig gut der Wahrscheinlichkeit des sicheren Ereignisses an. Es lässt sich leicht zeigen, dass diese Schranke in einigen Fällen nicht ausreichend ist. Zuvor muss jedoch noch geklärt werden, was unter der Zusammenfassung mehrerer Ereignisse verstanden wird.

*Definition 2.4 (Zusammenfassung von Ereignissen)*

Seien  $\mathcal{E}^{(j)}$  mehrere in Abhängigkeit von  $j$  definierte Ereignisse. Die *Zusammenfassung über alle  $j$*  ist definiert als die Schnittmenge der Ereignisse  $\mathcal{E}^{(j)}$  für alle  $j$ .

$$\mathcal{E} = \bigcap_j \mathcal{E}^{(j)}$$

Das Ereignis  $\mathcal{E}$  tritt genau dann ein, wenn alle Ereignisse  $\mathcal{E}^{(j)}$  zugleich eintreffen.

Sind für alle  $\mathcal{E}^{(j)}$  untere Schranken an die Eintrittswahrscheinlichkeiten bekannt, so lässt sich auch für die Zusammenfassung  $\mathcal{E}$  eine solche Schranke ermitteln. Durch nachfolgende Betrachtung wird ersichtlich, weshalb die untere Schranke der hohen Wahrscheinlichkeit nicht immer geeignet ist.

Sei eine Anzahl von Ereignissen  $\mathcal{E}^{(j)}$  gegeben, die jeweils mit hoher Wahrscheinlichkeit eintreten. Die Wahrscheinlichkeit für  $\mathcal{E}$  ist bestimmt durch

$$\Pr[\mathcal{E}] = 1 - \Pr\left[\overline{\bigcap_j \mathcal{E}^{(j)}}\right] = 1 - \Pr\left[\bigcup_j \overline{\mathcal{E}^{(j)}}\right] ,$$

wobei  $\overline{\mathcal{E}^{(j)}}$  das komplementäre Ereignis zu  $\mathcal{E}^{(j)}$  bezeichnet.  $\mathcal{E}$  tritt ein, falls keines dieser komplementären Ereignisse eintritt.

Aus der Formel für die Wahrscheinlichkeit der Vereinigung zweier Ereignisse erhält man die Abschätzung

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B] \leq \Pr[A] + \Pr[B] .$$

Verallgemeinert auf mehrere Ereignisse resultiert die erste Bonferroni-Ungleichung, welche auch als Boolesche Ungleichung bezeichnet wird.

$$\Pr \left[ \bigcup A_i \right] \leq \sum \Pr[A_i]$$

Für die Wahrscheinlichkeit des Ereignisses  $\mathcal{E}$  ergibt sich deshalb die untere Abschätzung

$$\Pr[\mathcal{E}] \geq 1 - \sum_j \Pr \left[ \overline{\mathcal{E}^{(j)}} \right] ,$$

auch bekannt als zweite Bonferroni-Ungleichung. Die Wahrscheinlichkeiten der einzelnen  $\overline{\mathcal{E}^{(j)}}$  lassen sich nach oben begrenzen und so vereinfacht sich der Ausdruck zu

$$\Pr[\mathcal{E}] = 1 - \sum_j o(1) .$$

(An dieser Stelle sei darauf hingewiesen, dass auch für die asymptotischen Notationen das Gleichheitszeichen anstelle von  $\in$  verwendet wird. Die Gleichungen sind deshalb nur in der üblichen Leserichtung eines Mitteleuropäers, d.h. von links nach rechts und von oben nach unten, zu interpretieren.)

Für eine konstante Anzahl von Ereignissen (und somit Summanden) erhält man die aus Definition 2.3 bekannte Gleichung der hohen Wahrscheinlichkeit. Wenn die Anzahl jedoch von  $n$  abhängt, so ergibt sich in vielen Fällen für die Zusammenfassung  $\mathcal{E}$  nur eine triviale Aussage bezüglich der Eintrittswahrscheinlichkeit.

Um auch in diesen Fällen eine (hohe) Wahrscheinlichkeit ermitteln zu können, führt man den Begriff der „sehr hohen Wahrscheinlichkeit“ ein.

*Definition 2.5 (sehr hohe Wahrscheinlichkeit)*

Eine Folge von Ereignissen  $\mathcal{E}_n$  tritt mit *sehr hoher Wahrscheinlichkeit* (m.s.h.W.) ein, falls für  $n \rightarrow \infty$  und alle  $c$  gilt

$$\Pr[\mathcal{E}_n] = 1 - o(n^{-c}) .$$

Die Zusammenfassung von  $O(n)$  vielen Ereignissen, die alle mit sehr hoher Wahrscheinlichkeit eintreffen, besitzt gleichfalls eine sehr hohe Wahrscheinlichkeit. In den nachfolgenden Kapiteln bildet dieser Schritt häufig den Abschluss eines Beweises, da somit gezeigt werden kann, dass eine Aussage mit sehr hoher Wahrscheinlichkeit für z.B. alle Knoten oder Knotengrade simultan zutrifft.

### 3 Erwartungswert der Gradverteilung

Die Erstellung eines Graphen mit der vorgegebenen Gradverteilung – insbesondere das Matching der Knotenkopien – stellt einen zufälligen Prozess dar. Der Grad eines Knotens im resultierenden Breitensuchbaum kann deshalb als Zufallsvariable angesehen werden. Aufgrund der Linearität des Erwartungswerts ergibt sich für die erwartete Anzahl von Knoten  $A_j^{obs}$  mit dem Grad  $j$

$$E \left[ A_j^{obs} \right] = \sum_{v \in V} \Pr [v \text{ besitzt im Breitensuchbaum den Grad } j] .$$

Zur genaueren Bestimmung des Erwartungswerts betrachtet man einen Knoten  $v$  mit dem Grad  $i$  im Verlauf der zeitbasierten Breitensuche. Stimmt der Zeitpunkt mit dem maximalen Index des Knotens überein, so wird dieser entdeckt und  $i - 1$  seiner Kopien an das Ende der Schlange angefügt. Die Kopie mit dem maximalen Index erhält direkt eine Kante zur Kopie am Kopf der Schlange und wird von diesem Augenblick an nicht weiter beachtet. Jede der verbleibenden  $i - 1$  Kopien erhöht genau dann den Grad des Knotens im Suchbaum um eins, falls sie erst eine Kante zugeteilt bekommt, wenn sie den Kopf der Schlange erreicht. Angenommen, dies trifft auf  $m$  der verbleibenden  $i - 1$  Kopien zu. Der resultierende Grad des Knotens im Suchbaum beträgt somit  $m + 1$  und für den Erwartungswert  $E \left[ A_{m+1}^{obs} \right]$  gilt

$$E \left[ A_{m+1}^{obs} \right] = n \sum_i a_i p_{i,m} , \tag{3.1}$$

wobei  $n$  die Anzahl der Knoten,  $a_i$  die Gradverteilung des Ausgangsgraphen und  $p_{i,m}$  die Wahrscheinlichkeit für das Eintreten des oben formulierten Ereignisses beschreibt.

#### 3.1 Dichte- und Verteilungsfunktion

Sei  $p_{i,m}(t_{max})$  die Wahrscheinlichkeit dieses Ereignisses (ein Knoten  $v$  mit dem Grad  $i$  besitzt im resultierenden Breitensuchbaum den Grad  $m + 1$ ) unter der Voraussetzung, dass  $v$  den maximalen Index  $t_{max}$  besitzt. Der maximale Index ist eine stetige Zufallsgröße auf dem Intervall  $[0, 1] \subseteq \mathbb{R}$  und zugleich das Maximum von  $i$  unabhängigen uniformen Zufallsgrößen auf dem gleichen Intervall – die zufällig bestimmten Indices  $t_k$  der Knotenkopien.

Aufgrund der Gleichverteilung der Indices gilt für alle  $t_k$  und beliebige  $t \in [0, 1] \subseteq \mathbb{R}$

$$\Pr [t_k \leq t] = t .$$

Da die  $t_k$  paarweise unabhängig voneinander sind, ergibt sich die sogenannte (stetige) Verteilungsfunktion von  $t_{max}$  mit

$$F(t) := \Pr [t_{max} \leq t] = \prod_{k=1}^i \Pr [t_k \leq t] = t^i .$$

Für ihre Ableitung, auch als Dichtefunktion bezeichnet, erhält man

$$f(t) := \frac{d}{dt} F(t) = it^{i-1} .$$

Unter Verwendung von

$$\begin{aligned} \Pr [t < t_{max} \leq t + dt] &= \Pr [t_{max} \leq t + dt] - \Pr [t_{max} \leq t] \\ &= F(t + dt) - F(t) \\ &= f(t) dt \end{aligned}$$

für infinitesimale  $dt$  und dem Satz der vollständigen Wahrscheinlichkeit lässt sich  $p_{i,m}$  wie folgt schreiben

$$p_{i,m} = \int_0^1 f(t) p_{i,m}(t) dt = \int_0^1 it^{i-1} p_{i,m}(t) dt . \quad (3.2)$$

### 3.2 Analyse der Breitensuche

Zur Bestimmung von  $p_{i,m}(t)$  betrachtet man die Wahrscheinlichkeit  $P_{vis}(t)$ , dass, falls  $v$  den maximalen Index  $t$  besitzt, eine beliebige, aber feste Kopie von  $v$  mit einem Index ungleich  $t$  – also eine Kopie, die an die Schlange angefügt wird – zu einer im Suchbaum enthaltenen Kante führt. Im Folgenden sei diese Kopie mit  $u$  und die adjazente Kopie entsprechend der zugehörigen Kante (der Partner von  $u$ ) mit  $w$  bezeichnet.

Die Kante zwischen  $u$  und  $w$  kann nur genau dann im Suchbaum enthalten sein, falls die folgenden Bedingungen gelten:

- (a) Der Kopie  $u$  wird erst eine Kante zugeordnet, wenn  $u$  den Kopf der Schlange erreicht hat.
- (b) Bis zu diesem Zeitpunkt muss  $w$  unentdeckt bleiben.

Gleichbedeutend kann man sagen, dass

- (a)  $w$  zum Zeitpunkt  $t$  unentdeckt ist und
- (b) alle Partner der Geschwisterkopien von  $w$  (die Partner aller weiteren Kopien des Knotens) ebenfalls zum Zeitpunkt  $t$  noch unentdeckt sind.

Angenommen,  $w$  ist zum Zeitpunkt  $t$  bereits entdeckt und deshalb in der Schlange enthalten. In diesem Fall liegt  $w$  vor  $u$ , da  $u$  erst zum Zeitpunkt  $t$  eingefügt wird. Somit wird  $u$  schon vor Erreichen des Kopfes der Schlange durch eine Kante mit  $w$  verbunden und entfernt.

Geht man hingegen davon aus, dass ein beliebiger Partner einer Geschwisterkopie von  $w$  zum Zeitpunkt  $t$  schon in der Schlange enthalten ist, so erreicht dieser Partner vor  $u$  den Kopf der Schlange und wird abgearbeitet. Dabei wird eine Kopie von  $w$  und somit auch  $w$  selbst entdeckt.

Beide Annahmen führen dazu, dass der zur Kopie  $w$  gehörende Knoten vor dem Abarbeiten von  $u$  entdeckt und die betrachtete Kante nicht in den Suchbaum übernommen wird.

Sei  $C_{unto}(t)$  die Anzahl unentdeckter und  $C_{unex}(t)$  die Anzahl der noch nicht abgearbeiteten und somit entweder in der Schlange enthaltenen oder unentdeckten Kopien zum Zeitpunkt  $t$ . Die Wahrscheinlichkeit für Bedingung (a) – ein zufällig per Gleichverteilung ermittelter Partner  $w$  von  $u$  bleibt bis zu diesem Zeitpunkt unentdeckt – beträgt

$$P_{unto}(t) = \frac{C_{unto}(t)}{C_{unex}(t)} .$$

Die Kopie  $w$  kann zum Zeitpunkt  $t$  noch nicht abgearbeitet sein, da sonst der Knoten  $v$  über  $w$  erreicht worden wäre, was den Voraussetzungen widerspricht. Die Anzahl aller in Frage kommenden Kopien beträgt deshalb  $C_{unex}(t)$ , wobei von diesen genau  $C_{unto}(t)$  viele dem günstigen Fall einer unentdeckten Kopie entsprechen.

Bedingt auf diesem Ereignis errechnet sich die Wahrscheinlichkeit dafür, dass  $w$  zu einem Knoten mit dem Grad  $k$  gehört, durch

$$P_{unto,k}(t) = \frac{kV_{unto,k}(t)}{C_{unto}(t)} ,$$

wobei mit  $V_{unto,k}$  die Anzahl aller unentdeckten Knoten mit Grad  $k$  beschrieben wird. Zwischen  $C_{unto}(t)$  und  $V_{unto,k}(t)$  gilt der Zusammenhang  $C_{unto}(t) = \sum_k kV_{unto,k}(t)$ .

Die zweite Bedingung fordert, dass alle Partner der  $k - 1$  Geschwisterkopien von  $w$  ebenfalls unentdeckt geblieben sind. Dabei soll vorläufig vernachlässigt werden, dass sich die Anzahl unentdeckter Kopien nach jedem Festlegen einer Kopie verringert (Auswahl ohne Zurücklegen). Ferner wird davon ausgegangen, dass  $v$ , die Nachbarn von  $v$  und wiederum deren Nachbarn einen Baum bilden, d.h.  $v$  in keinem Kreis der Länge drei oder vier vorkommt oder Mehrfachkanten auftreten. Setzt man all dies voraus (Beweise erfolgen in Abschnitt 3.4), so ist die Wahrscheinlichkeit für  $k - 1$  unentdeckte Partnerkopien  $P_{unto}(t)^{k-1}$ . Dies multipliziert mit der Wahrscheinlichkeit für den Knotengrad  $k$  von  $w$  und aufsummiert über alle möglichen Knotengrade ergibt das Resultat für Bedingung (b).

Auch in diesem Fall können die betrachteten Kopien noch nicht abgearbeitet sein, denn sonst wäre der zu  $w$  gehörende Knoten und somit auch  $w$  zum Zeitpunkt  $t$  bereits entdeckt.

Zusammengenommen erhält man

$$\begin{aligned} P_{vis}(t) &= P_{unto}(t) \sum_k P_{unto,k}(t) P_{unto}(t)^{k-1} \\ &= \sum_k P_{unto,k}(t) P_{unto}(t)^k \\ &= \sum_k \frac{kV_{unto,k}(t)}{C_{unto}(t)} \left( \frac{C_{unto}(t)}{C_{unex}(t)} \right)^k . \end{aligned} \tag{3.3}$$

### 3.3 Schranken der Knoten- und Kopienanzahl

#### *Lemma 3.1*

Sei  $(a_0, a_1, \dots)$  eine angemessene Gradverteilung und der zugehörige Graph  $G$  zusammenhängend. Ausserdem seien  $\beta$  und  $\varepsilon$  beliebige Konstanten mit  $\beta < \min(\frac{1}{2}, \frac{\alpha-2}{2})$  und  $\varepsilon > 0$ .

Dann gilt für alle  $t \in [0, 1]$  und  $j < n$  m.s.h.W.

$$|V_{unto,j}(t) - a_j t^j \cdot n| < n^{1/2+\varepsilon} \quad (3.4)$$

$$|C_{unex}(t) - c_{unex}(t) \cdot n| < n^{1/2+\varepsilon} \quad (3.5)$$

$$|C_{unto}(t) - c_{unto}(t) \cdot n| < n^{1-\beta}, \quad (3.6)$$

unter Verwendung der Definitionen  $c_{unex}(t) := \delta t^2 = \sum_j j a_j t^2$  und  $c_{unto}(t) := \sum_j j a_j t^j$ .

*Beweis* Der Beweis gliedert sich in zwei Teile: Die Bestimmung der Erwartungswerte von  $V_{unto,j}(t)$ ,  $C_{unex}(t)$  und  $C_{unto}(t)$  und dem Nachweis, dass die Zufallsvariablen in einem Bereich von  $o(n)$  um ihre Erwartungswerte konzentriert sind.

Zu einem beliebigen Zeitpunkt  $t$  sind die unentdeckten Knoten genau diejenigen, deren maximaler Index kleiner als  $t$  ist. Die Wahrscheinlichkeit für einen Knoten mit dem Grad  $j$ , einen solchen maximalen Index kleiner als  $t$  zu besitzen, beträgt  $t^j$ . Der Erwartungswert für die Anzahl unentdeckter Knoten mit Grad  $j$  ist das Produkt aus dieser Wahrscheinlichkeit und der Anzahl aller Knoten dieses Grades. Somit ergibt sich

$$\begin{aligned} \mathbb{E}[V_{unto,j}(t)] &= a_j t^j \cdot n \\ \mathbb{E}[C_{unto}(t)] &= \sum_j j a_j t^j \cdot n = c_{unto}(t) \cdot n. \end{aligned}$$

Eine Kopie ist genau dann abgearbeitet, wenn sie durch eine Kante mit ihrem Partner verbunden wurde. Für den Endpunkt einer Kante verwendet die Breitensuche immer von allen verbleibenden der  $\sum_j j a_j n = \delta n$  Kopien diejenige mit dem größten Index. Deshalb sind zu einem beliebigen Zeitpunkt  $t$  genau die Kopien noch nicht abgearbeitet, bei denen der eigene und der Index des Partners kleiner als  $t$  sind. Da alle Indices der Kopien zufällig gemäß der Gleichverteilung aus dem Intervall  $[0, 1]$  bestimmt werden, liegt die Wahrscheinlichkeit hierfür bei  $t^2$  und der Erwartungswert ist

$$\mathbb{E}[C_{unex}(t)] = \delta t^2 \cdot n = c_{unex}(t) \cdot n.$$

Für den weiteren Beweis wird die Hoeffding Schranke in der folgenden Form benötigt.

*Satz 3.2*

Seien  $X_1, \dots, X_k$  paarweise unabhängige, nichtnegative Zufallsvariablen mit  $X_i \leq b_i$  für alle  $i$  und  $X = \sum_i X_i$ . Dann gilt für beliebige  $\Delta \geq 0$

$$\Pr[|X - \mathbb{E}[X]| \geq \Delta] \leq 2e^{-2\Delta^2 / \sum_i b_i^2}.$$

Die Verteilung der Zufallsvariable  $V_{unto,j}(t)$  ist identisch mit der Binomialverteilung  $\text{Bin}(a_j n, t^j)$ , da alle  $a_j n$  Knoten mit dem Grad  $j$  unabhängig voneinander mit der Wahrscheinlichkeit  $t^j$  bis zum Zeitpunkt  $t$  unentdeckt geblieben sind. Durch Anwendung der Hoeffding Schranke auf  $a_j n$  Zufallsvariablen, die jeweils durch  $b = 1$  beschränkt werden, erhält man für eine Mindestabweichung von  $\Delta = n^{1/2+\varepsilon}$  vom Erwartungswert die Wahrscheinlichkeit

$$\begin{aligned} \Pr\left[|V_{unto,j}(t) - a_j t^j \cdot n| \geq n^{1/2+\varepsilon}\right] &\leq 2e^{-2n^{1+2\varepsilon}/(a_j n)} \\ &= 2e^{-(2/a_j) \cdot n^{2\varepsilon}} \\ &= 2e^{-\Omega(n^{2\varepsilon})}. \end{aligned}$$

Zu jedem festen Zeitpunkt  $t$  und für ein beliebiges, aber festes  $j$  gilt deshalb (3.4) m.s.h.W. Um die Schranke für alle  $t$  und  $j$  nachzuweisen, genügt es aber nicht, die berechnete Wahrscheinlichkeit für die möglichen Werte von  $t$  und  $j$  zusammenzufassen, da die Menge aller  $t$  überabzählbar unendlich viele Elemente enthält.

Stattdessen wird das Intervall  $I = [0, 1]$  in ausreichend kleine Teilintervalle zerlegt, über welche anschliessend aufsummiert werden soll. Sei  $m = \sum_j j a_j n = \delta n$  die Anzahl aller Kopien. Da die Gradverteilung  $A = (a_0, a_1, \dots)$  laut Voraussetzung angemessen sein muss, ist der Durchschnittsgrad  $\delta$  endlich (Lemma 3.4 aus [6]). Das Intervall  $I$  wird in  $m^b$  viele Teilintervalle der Größe  $m^{-b}$  eingeteilt. Genauere Angaben über den Exponenten  $b$  können erst am Ende des Beweises gemacht werden.

Nun soll die Wahrscheinlichkeit für zwei feste Kopien, jeweils einen Index aus dem gleichen Intervall zugeteilt zu bekommen, betrachtet werden. Nimmt man das Intervall des ersten Index als beliebig an, so entfällt der uniform zufällig bestimmte Index der zweiten Kopie mit einer Wahrscheinlichkeit von  $m^{-b}$  ebenfalls auf dieses Intervall. Es existieren insgesamt  $\binom{m}{2} = \frac{1}{2}m(m-1) \leq m^2$  Paare von Kopien (ohne Beachtung der Reihenfolge). Für die untere Schranke der Wahrscheinlichkeit, jedes Intervall enthält maximal eine Kopie, ergibt sich daraus  $1 - m^2 \cdot m^{-b} = 1 - m^{2-b}$ . Dies entspricht auch der Wahrscheinlichkeit, dass in jedem dieser Zeitintervalle höchstens ein Schritt der Breitensuche ausgeführt wird und sich somit die Anzahl unentdeckter Knoten um maximal eins ändert.

Bedingt auf diesem Ereignis und vorausgesetzt,  $V_{unto,j}(t)$  weicht an den Intervallgrenzen m.s.h.W. um nicht mehr als die geforderten  $n^{1/2+\varepsilon}$  vom Erwartungswert ab, so gilt (3.4) auch für alle  $t$ , da innerhalb der Intervalle keine großen Änderungen von  $V_{unto,j}(t)$  möglich sind. Die Wahrscheinlichkeit, dass die genannte Schranke für irgendein  $t$  bei festem  $j$  verletzt wird, lässt sich also nach oben mit der Summe aus der Wahrscheinlichkeit für eine Verletzung an den Intervallgrenzen ( $2e^{-\Omega(n^{2\varepsilon})}$ ) multipliziert mit der Anzahl aller Grenzen ( $m^b$ ) und der Wahrscheinlichkeit dafür, dass zwei Indices aus dem gleichen Intervall stammen ( $m^{2-b}$ , in diesem Fall kann man über eine eventuell zu große Abweichung vom Erwartungswert nichts aussagen), beschränken. Aufsummiert über alle  $j < n$  ergibt sich

$$n \left( 2m^b e^{-\Omega(n^{2\varepsilon})} + m^{2-b} \right) = O(n^{3-b}) = o(n^{-c}) \quad \text{für } b > c + 3,$$

da  $m = \delta n < \frac{C}{\alpha-2}n$  (Lemma 3.4 aus [6]). Ungleichung (3.4) gilt deshalb m.s.h.W. gleichzeitig für alle  $t$  und  $j$ .

Nicht abgearbeitete Kopien treten, wie schon näher erläutert, immer paarweise auf. Die Zufallsvariable  $C_{unex}(t)$  genügt somit der Binomialverteilung  $\text{Bin}(\delta n/2, t^2)$ . Die Anwendung der Hoeffding Schranke mit  $\Delta = n^{1/2+\varepsilon}$  ergibt

$$\begin{aligned} \Pr \left[ |C_{unex}(t) - c_{unex}(t) \cdot n| \geq n^{1/2+\varepsilon} \right] &\leq 2e^{-2n^{1+2\varepsilon}/(\delta n/2)} \\ &= 2e^{-(4/\delta)n^{2\varepsilon}} \\ &= 2e^{-\Omega(n^{2\varepsilon})} \end{aligned}$$

und wie im vorherigen Fall erhält man durch anschliessendes Aufsummieren über Teilintervalle von  $t$  die Behauptung (3.5).

Für den Beweis, dass  $C_{unto}(t)$  um den Erwartungswert konzentriert ist, betrachtet man als Zufallsvariable  $X_i$  die Anzahl aller unentdeckten Kopien des Knotens  $i$ . O.B.d.A.

besitzt dieser Knoten den Grad  $b_i$ . Dann gilt  $X_i \leq b_i$  und  $C_{unto}(t) = \sum_i X_i$ . Der Nenner im Exponenten der Hoeffding Schranke ist die Summe der Quadrate aller Knotengrade und somit beschränkt durch

$$\sum_i b_i^2 = \sum_j j^2 a_j n < Cn \sum_j j^{2-\alpha} = \begin{cases} O(n^{4-\alpha}) & \alpha < 3 \\ O(n \log n) & \alpha = 3 \\ O(n) & \alpha > 3 \end{cases} .$$

Der Fall für  $\alpha > 3$  ergibt sich aus dem Integralkriterium unendlicher Reihen (siehe Lemma 3.4 aus [6]), der für  $\alpha = 3$  aus der Näherungsformel der harmonischen Reihe, während sich der letzte Fall mit  $s := 2 - \alpha$  und damit  $-1 < s < 0$  durch geeignetes Abschätzen der Folgeglieder bestimmen lässt:

$$\begin{aligned} \sum_j j^s &= 1^s + (2^s + 3^s) + (4^s + 5^s + 6^s + 7^s) + \dots \\ &\leq 1 + 2 \cdot 2^s + 4 \cdot 4^s + 8 \cdot 8^s + \dots \\ &= 1 + 2^{(1+s)} + 2^{(1+s) \cdot 2} + 2^{(1+s) \cdot 3} + \dots \\ &\leq \sum_{k=0}^{\lfloor \log n \rfloor} 2^{(1+s) \cdot k} \\ &= \frac{2^{(1+s) \cdot \lfloor \log n \rfloor + 1} - 1}{2^{(1+s)} - 1} \\ &= O(n^{1+s}) = O(n^{3-\alpha}) . \end{aligned}$$

Mit der oberen Abschätzung für  $\sum_i b_i^2$  liefert die Hoeffding Schranke bezüglich  $C_{unto}(t)$  für ein festes  $t$

$$\Pr \left[ |C_{unto}(t) - c_{unto}(t) \cdot n| \geq n^{1-\beta} \right] \leq \begin{cases} 2e^{-\Omega(n^{\alpha-2-2\beta})} & \alpha < 3 \\ 2e^{-\Omega(n^{1-2\beta}/\log n)} & \alpha = 3 \\ 2e^{-\Omega(n^{1-2\beta})} & \alpha > 3 \end{cases} .$$

Damit diese Wahrscheinlichkeiten exponentiell klein werden, muss  $\beta$  die Bedingungen  $(\alpha - 2) - 2\beta > 0$  und  $1 - 2\beta > 0$  erfüllen. Durch das schon beschriebene Aufsummieren über Teilintervalle von  $t$  mit  $\beta < \min(\frac{1}{2}, \frac{\alpha-2}{2})$  erhält man (3.6).  $\square$

### 3.4 Wahrscheinlichkeit einer Suchbaumkante

Da  $V_{unto,j}(t)$ ,  $C_{unex}(t)$  und  $C_{unto}(t)$  nach Lemma 3.1 m.s.h.W. um ihre Erwartungswerte konzentriert sind, lässt sich Gleichung (3.3) annäherungsweise schreiben als

$$p_{vis}(t) = \sum_k \frac{k a_k t^k}{c_{unto}(t)} \left( \frac{c_{unto}(t)}{c_{unex}(t)} \right)^k . \quad (3.7)$$

Erneut soll davon ausgegangen werden, dass keine Schleifen oder Mehrfachkanten auftreten und die Änderungen an  $V_{unto,j}(t)$ ,  $C_{unex}(t)$  und  $C_{unto}(t)$  durch die Auswahl von Kopien vernachlässigbar gering sind. Dann führt jede Kopie eines Knotens  $v$  mit dem Grad  $i$  (mit

Ausnahme der Kopie maximalem Index') unabhängig von den anderen Kopien jeweils mit der näherungsweise Wahrscheinlichkeit von  $p_{vis}(t)$  zu einer im Suchbaum enthaltenen Kante. Die Anzahl der Kopien mit einer solchen inzidenten Kante entspricht in etwa einer Binomialverteilung  $\text{Bin}(i - 1, p_{vis}(t))$ .

Diese Überlegungen sollen durch die Beweise, dass

- (a) der Knoten  $v$ , dessen Nachbarn und die Nachbarn der Nachbarn mit hoher Wahrscheinlichkeit einen Baum bilden und
- (b) die exakte Anzahl von Kopien, die zu Suchbaumkanten führen, nur sehr gering von der Binomialverteilung  $\text{Bin}(i - 1, p_{vis}(t))$  bzw.  $\text{Bin}(i - 1, p_{vis}(t))$  abweicht,

gestützt werden.

Punkt (a) lässt sich leicht zeigen, falls die Knotengrade des Graphen begrenzt sind. Das folgende Lemma verdeutlicht, dass in Graphen mit power-law Verteilungen nur sehr wenige Knoten mit sehr großen Graden auftreten. Für das anschließende Lemma kann man deshalb in den meisten Fällen von einem entsprechend kleinen Knotengrad ausgehen.

*Lemma 3.3*

Die Wahrscheinlichkeit für eine zufällig ausgewählte Kopie  $u$ , zu einem Knoten  $v$  mit einem Grad größer als  $k$  zu gehören, beträgt

$$\Pr [\text{deg}(v) > k] = o(k^{-2\beta}) .$$

*Beweis* Es gilt

$$\begin{aligned} \Pr [\text{deg}(v) > k] &= \frac{\sum_{j>k} j a_j}{\sum_j j a_j} \\ &= \frac{1}{\delta} \sum_{j>k} j a_j \\ &< \frac{1}{\delta} \sum_{j>k} j \cdot C j^{-\alpha} \quad (\text{Definition 2.2}) \\ &= \frac{C}{\delta} \sum_{j>k} j^{(1-\alpha)} \\ &< \frac{C}{\delta} \cdot \frac{k^{(2-\alpha)}}{\alpha - 2} \quad (\text{Integralkriterium, untere Grenze ist } k) \\ &= \frac{C}{\delta(\alpha - 2)} \cdot k^{-(\alpha-2)} = o(k^{-2\beta}) . \end{aligned}$$

□

*Lemma 3.4*

Es existieren zwei Konstanten  $\gamma > \eta > 0$  so, dass ein Knoten  $v$  mit dem Grad  $i < n^\eta$ , seine Nachbarn und deren Nachbarn mit einer Wahrscheinlichkeit von  $1 - o(n^{-\gamma})$  einen Baum darstellen. Die Wahrscheinlichkeit, dass der Knoten  $v$  oder dessen Nachbarn Schleifen oder Mehrfachkanten besitzen bzw. dass  $v$  in einem Dreieck oder Kreis der Länge 4 enthalten ist, beträgt somit  $o(n^{-\gamma})$ .

*Beweis* Die Wahrscheinlichkeit für einen Nachbarn von  $v$ , einen Knotengrad größer als  $n^\lambda$  – die Konstanten werden am Ende des Beweises festgelegt – zu besitzen, berechnet sich nach Lemma 3.3 zu  $o((n^\lambda)^{-2\beta}) = o(n^{-2\lambda\beta})$ . Summiert man diese über alle  $i < n^\eta$  Nachbarn von  $v$  auf, so erhält man als untere Schranke der Wahrscheinlichkeit, dass keiner der Nachbarn einen Knotengrad größer als  $n^\lambda$  besitzt,

$$1 - n^\eta \cdot o(n^{-2\lambda\beta}) = 1 - o(n^{\eta-2\lambda\beta}) . \quad (3.8)$$

Die folgenden Teile des Beweises bedingen auf diesem Ereignis, d.h. es wird vorausgesetzt, dass alle Nachbarn von  $v$  einen entsprechend kleinen Knotengrad besitzen.

Im ersten Schritt betrachtet man zunächst Schleifen und Mehrfachkanten. Es gibt hierfür vier Möglichkeiten:

- (a)  $v$  besitzt eine oder mehrere Schleifen
- (b)  $v$  besitzt Mehrfachkanten
- (c) ein Nachbar von  $v$  besitzt eine oder mehrere Schleifen
- (d) ein Nachbar von  $v$  besitzt Mehrfachkanten

Fall (a). Eine Schleife tritt auf, wenn zwei Kopien eines Knotens aufeinander abgebildet werden. Jede der  $i$  Kopien von  $v$  wird mit einer Wahrscheinlichkeit von höchstens  $(i-1)/(\delta n - 1)$  einer der anderen Kopien des Knotens zugeteilt. Die Wahrscheinlichkeit einer Schleife bei  $v$  ist somit begrenzt durch

$$O(i^2/(\delta n)) = O(n^{2\eta-1}) . \quad (3.9)$$

Fall (b). Hierbei beschränkt man sich anfangs auf einen beliebigen, aber festen Knoten mit dem Grad  $j \leq n^\lambda$ . Zwei unabhängig voneinander und uniform zufällig ausgewählte Kopien sind mit einer Wahrscheinlichkeit von höchstens  $j(j-1)/(\delta n)(\delta n - 1) = O(n^{2\lambda-2})$  Kopien dieses Knotens. Die Wahrscheinlichkeit dafür, dass beide die Kopien eines solchen Knotens (jetzt nicht mehr festgelegt) sind, beträgt somit höchstens  $n \cdot O(n^{2\lambda-2}) = O(n^{2\lambda-1})$ . Allgemein erhält man für die Wahrscheinlichkeit, dass zwei Kopien dem gleichen Knoten – unabhängig von dessen Grad – zugehören,

$$\begin{aligned} P_{sib} &= O\left(\frac{\sum_j j(j-1)a_j n}{(\sum_j j a_j n)^2}\right) \\ &= O\left(\frac{1}{\delta^2 n} \sum_j j^2 a_j\right) \\ &= \begin{cases} O(n^{2-\alpha}) & \alpha < 3 \\ O\left(\frac{\log n}{n}\right) & \alpha = 3 \\ O\left(\frac{1}{n}\right) & \alpha > 3 \end{cases} \\ &= o(n^{-2\beta}) . \end{aligned}$$

Zurück zu den Nachbarknoten von  $v$  mit einem maximalen Knotengrad  $n^\lambda$ . Falls  $v$  eine Mehrfachkante besitzt, so gibt es zwei Kopien von  $v$ , die auf Kopien des gleichen Knotens abgebildet werden. Es existieren  $\binom{i}{2}$  verschiedene Paarungen der Kopien von  $v$  und jedes dieser Paare wird mit einer Wahrscheinlichkeit von  $O(n^{2\lambda-1})$  auf nur einen Nachbarn von  $v$  abgebildet. Die obere Abschätzung der resultierenden Wahrscheinlichkeit einer Mehrfachkante lautet somit

$$\binom{i}{2} \cdot O(n^{2\lambda-1}) = O(n^{2\eta} \cdot n^{2\lambda-1}) = O(n^{2\eta+2\lambda-1}). \quad (3.10)$$

Fall (c). Analog zu Fall (a) erhält man bezüglich eines festen Nachbarknotens für die Wahrscheinlichkeit einer Schleife höchstens

$$O(n^{2\lambda}/(\delta n)) = O(n^{2\lambda-1}).$$

Fall (d). Analog zu Fall (b) ergibt sich für je ein Paar der maximal  $n^\lambda$  Kopien mit der Wahrscheinlichkeit  $P_{sib}$  und für den betrachteten Nachbarknoten insgesamt mit

$$O(n^{2\lambda}) \cdot P_{sib} = o(n^{2\lambda-2\beta})$$

eine Mehrfachkante.

Summiert man die Fälle (c) und (d) für alle  $i$  möglichen Knoten auf, so lässt sich die Wahrscheinlichkeit für das Auftreten von Schleifen oder Mehrfachkanten in der Nachbarschaft von  $v$  insgesamt nach oben hin durch

$$n^\eta \cdot o(n^{2\lambda-2\beta}) = o(n^{\eta+2\lambda-2\beta}) \quad (3.11)$$

begrenzen.

Nun sollen noch die Fälle betrachtet werden, in denen  $v$  Teil eines Dreiecks oder Vierecks ist. Ein solches Dreieck besteht immer aus zwei Kopien von  $v$  und je zwei Kopien von zwei verschiedenen Nachbarn – eine für die Kante zu  $v$  und eine weitere für die Kante zum anderen Nachbarn. Es gibt  $O(n^2)$  verschiedene Paare von Nachbarn, die in einem entsprechenden Dreieck enthalten sein können. Bezogen auf die beteiligten Kopien von  $v$  ergeben sich  $O(n^{2\eta})$  unterschiedliche Möglichkeiten, während es für die Kopien der beiden Nachbarn jeweils  $O(n^{2\lambda})$  Varianten sind. Betrachtet man ein gegebenes Paar von Kopien, so ist die Wahrscheinlichkeit einer Kante zwischen diesen  $O(1/(\delta n))$ . Zusammengefasst ergibt sich für die erwartete Anzahl an Dreiecken als obere Schranke

$$O(n^2 \cdot n^{2\eta} \cdot (n^{2\lambda})^2 / (\delta n)^3) = O(n^{2\eta+4\lambda-1}). \quad (3.12)$$

Die Überlegungen sind auf Vierecke bezogen recht ähnlich, nur dass nun die Kopien der beiden Nachbarn nicht aufeinander, sondern auf Kopien eines vierten Knotens abgebildet werden müssen. Die Wahrscheinlichkeit hierfür wurde schon als  $P_{sib}$  bestimmt. Als Ergebnis erhält man in diesem Fall

$$O(n^2 \cdot n^{2\eta} \cdot (n^{2\lambda})^2 \cdot P_{sib} / (\delta n)^2) = o(n^{2\eta+4\lambda-2\beta}). \quad (3.13)$$

Die Aussage des Lemmas ist genau dann verletzt, wenn eine dieser betrachteten Möglichkeiten eintritt. Dazu muss entweder ein Nachbarknoten von  $v$  mit einem Grad größer als  $n^\lambda$  existieren (siehe Gleichung 3.8), ein Knoten einem der Fälle (a) bis (d) entsprechen (siehe Gleichungen 3.9 - 3.11) oder  $v$  in einem Dreieck bzw. Viereck enthalten sein (siehe Gleichungen 3.12 - 3.13). Zusammengenommen ist die Wahrscheinlichkeit dafür, dass  $v$ , die Nachbarn von  $v$  und deren Nachbarn keinen Baum bilden,

$$\begin{aligned} & \max(o(n^{\eta-2\lambda\beta}), O(n^{2\eta-1}), O(n^{2\eta+2\lambda-1}), o(n^{\eta+2\lambda-2\beta}), O(n^{2\eta+4\lambda-1}), o(n^{2\eta+4\lambda-2\beta})) \\ & = \max(o(n^{\eta-2\lambda\beta}), o(n^{2\eta+4\lambda-2\beta})) = o(n^{-\gamma}) \end{aligned}$$

mit  $\gamma = -\max(\eta - 2\lambda\beta, 2\eta + 4\lambda - 2\beta)$ . Setzt man nun  $\eta = \beta^2/6$  und  $\lambda = \beta/4$ , so ergibt sich aus  $\eta - 2\lambda\beta = -\beta^2/3 < 0$  und  $2\eta + 4\lambda - 2\beta = \beta^2/3 - \beta < 0$  sowie  $\frac{\beta^2/3 - \beta}{-\beta^2/3} = \frac{3}{\beta} - 1 > 5$  das Ergebnis  $\gamma = \beta^2/3 > \eta > 0$ .  $\square$

Unter der Annahme, dass die Aussage von Lemma 3.4 auf den Graphen zutrifft, und noch eine genügend hohe Anzahl nicht abgearbeiteter und dadurch freier Kopien vorhanden ist, so kann die zufällige Festlegung einer Kopie als „Auswahl mit Zurücklegen“ betrachtet werden. Die zweite Bedingung fordert dabei, dass der Prozess der Kantenbestimmung noch nicht zu weit fortgeschritten und somit der Zeitpunkt  $t$  der Breitensuche ausreichend weit von Null entfernt ist. Die Anzahl im Suchbaum enthaltener Kanten eines Knotens ist unter diesen Voraussetzungen annähernd binomial verteilt.

*Lemma 3.5*

Seien  $\eta, \gamma$  wie in Lemma 3.4 definiert. Dann existiert eine Konstante  $\theta > 0$  so, dass für  $t \in [n^{-\theta}, 1]$  und  $i < n^\eta$  gilt

$$|p_{i,m}(t) - \Pr[\text{Bin}(i-1, P_{vis}(t)) = m]| < n^{-\gamma}.$$

*Beweis* Sei im Folgenden der minimale im Graph auftretende Knotengrad mit  $d_{min}$  bezeichnet, d.h.  $d_{min}$  ist das kleinste  $j$ , für welches  $a_j > 0$  gilt. Aufgrund der angemessenen Gradverteilung erhält man  $d_{min} \geq 3$  und für  $\theta = \beta/(2d_{min})$  somit  $\theta < 1/12$ . Laut Voraussetzung ist  $t \geq n^{-\theta}$ , daraus ergibt sich für den Erwartungswert der freien Kopien  $E[C_{unex}(t)] = \delta t^2 \cdot n = \Omega(n^{1-\beta/d_{min}})$ . Diese Schranke gilt nach Lemma 3.1 nicht nur für den Erwartungswert, sondern m.s.h.W. ebenso für die exakte Anzahl  $C_{unex}(t)$ .

Nun betrachtet man wieder einen Knoten  $v$  mit dem Grad  $i < n^\eta$ . Ausserdem soll der Knotengrad aller Nachbarn von  $v$  wie in Lemma 3.4 höchstens  $n^\lambda$  betragen. Die Anzahl sichtbarer und somit im Suchbaum enthaltener Kanten von  $v$  hängt von den Kopien der Nachbarn ab. Da  $v$  laut Voraussetzung weniger als  $n^\eta$  Nachbarn besitzt, die wiederum höchsten  $n^\lambda$  Kopien enthalten, ist die Anzahl relevanter Kopien nach oben durch  $n^{\eta+\lambda}$  beschränkt. Diese Kopien werden im Verlauf der Breitensuche zufällig ohne Zurücklegen bestimmt. Nimmt man hingegen an, dass die Kopien nach der Auswahl zurückgelegt und somit erneut ausgewählt werden können, und betrachtet man eine beliebige, aber feste Nachbarkopie, so wird diese mit einer Wahrscheinlichkeit von  $1/C_{unex}(t)$  in jedem weiteren der insgesamt  $n^{\eta+\lambda}$  Auswahlsschritte für die Kopien gezogen. Aufsummiert über

alle Nachbarkopien ist die Wahrscheinlichkeit einer Kollision somit höchstens

$$(n^{\eta+\lambda})^2 / C_{unex}(t) = O(n^{2\eta+2\lambda+\beta/d_{min}-1}) = O(n^{\frac{1}{3}\beta^2+\frac{5}{6}\beta-1}) = o(n^{-1/2}) .$$

Diese lässt sich in die Wahrscheinlichkeit  $o(n^{-\gamma})$ , mit der die Aussage von Lemma 3.4 nicht zutrifft, integrieren. Davon abgesehen können keine Kollisionen auftreten und der Prozess entspricht einer „Auswahl mit Zurücklegen“. Jede der  $i - 1$  von  $v$  ausgehenden Kanten ist deshalb mit der gleichen Wahrscheinlichkeit  $P_{vis}$  (siehe Gleichung 3.3) im Suchbaum enthalten. Für die Anzahl der sichtbaren Kanten ergibt sich die beschriebene Binomialverteilung.

Diese Betrachtungsweise liefert für  $p_{i,m}(t)$

$$p_{i,m}(t) = p_{Le3.4} \cdot q_{i,m}(t) + (1 - p_{Le3.4}) \cdot \Pr [\text{Bin}(i - 1, P_{vis}(t)) = m] .$$

Der erste Summand betrifft den Fall, dass die Aussage von Lemma 3.4 auf den betrachteten Knoten nicht zutrifft. Die resultierende Wahrscheinlichkeit  $q_{i,m}(t)$  muss für die Abschätzung nicht näher bestimmt werden. Für den zweiten Summanden kann, wie eben schon erläutert, davon ausgegangen werden, dass eine Binomialverteilung für die im Suchbaum enthaltenen Kanten vorliegt. Somit folgt direkt durch

$$\begin{aligned} & |p_{i,m}(t) - \Pr [\text{Bin}(i - 1, P_{vis}(t)) = m]| \\ &= |p_{Le3.4} \cdot q_{i,m}(t) - p_{Le3.4} \cdot \Pr [\text{Bin}(i - 1, P_{vis}(t)) = m]| \\ &= p_{Le3.4} \cdot |q_{i,m}(t) - \Pr [\text{Bin}(i - 1, P_{vis}(t)) = m]| \\ &\leq p_{Le3.4} < n^{-\gamma} \quad (\text{da } p_{Le3.4} = o(n^{-\gamma})) \end{aligned}$$

für  $n$  genügend groß die Behauptung. □

Nun sind alle Voraussetzungen geschaffen, um  $p_{i,m}(t)$  und dadurch auch  $p_{i,m}$  (siehe Gleichung 3.2) hinreichend genau bestimmen zu können.

*Lemma 3.6*

Sei  $(a_0, a_1, \dots)$  eine angemessene Gradverteilung und der zugehörige Graph  $G$  zusammenhängend. Dann existieren Konstanten  $\theta, \kappa, \eta, \mu > 0$  so, dass für alle  $t \in [n^{-\theta}, 1 - n^{-\kappa}]$  und alle  $i < n^\eta$  sowie genügend große  $n$  gilt

$$|p_{i,m}(t) - \Pr [\text{Bin}(i - 1, p_{vis}(t)) = m]| < n^{-\mu} .$$

*Beweis* Unter Verwendung von Lemma 3.5 und einer konstruktiven Null lässt sich die Behauptung mittels Dreiecksungleichung als

$$\begin{aligned} & |p_{i,m}(t) - \Pr [\text{Bin}(i - 1, p_{vis}(t)) = m]| \\ &= |p_{i,m}(t) - \Pr [\text{Bin}(i - 1, P_{vis}(t)) = m]| \\ &\quad + |\Pr [\text{Bin}(i - 1, P_{vis}(t)) = m] - \Pr [\text{Bin}(i - 1, p_{vis}(t)) = m]| \\ &\leq |p_{i,m}(t) - \Pr [\text{Bin}(i - 1, P_{vis}(t)) = m]| \\ &\quad + |\Pr [\text{Bin}(i - 1, P_{vis}(t)) = m] - \Pr [\text{Bin}(i - 1, p_{vis}(t)) = m]| \\ &< n^{-\gamma} + |\Pr [\text{Bin}(i - 1, P_{vis}(t)) = m] - \Pr [\text{Bin}(i - 1, p_{vis}(t)) = m]| \end{aligned}$$

schreiben. Die resultierende obere Abschätzung kann durch Einsetzen der Formel für die Binomialverteilung weiter vereinfacht werden.

$$\begin{aligned}
 & |\Pr [\text{Bin} (i-1, P_{vis}(t)) = m] - \Pr [\text{Bin} (i-1, p_{vis}(t)) = m]| \\
 = & \left| \binom{i-1}{m} P_{vis}(t)^m (1 - P_{vis}(t))^{i-m-1} - \binom{i-1}{m} p_{vis}(t)^m (1 - p_{vis}(t))^{i-m-1} \right| \\
 = & \binom{i-1}{m} p_{vis}(t)^m (1 - p_{vis}(t))^{i-m-1} \cdot \left| \left( \frac{P_{vis}(t)}{p_{vis}(t)} \right)^m \left( \frac{1 - P_{vis}(t)}{1 - p_{vis}(t)} \right)^{i-m-1} - 1 \right| \\
 \leq & \left| \left( \frac{P_{vis}(t)}{p_{vis}(t)} \right)^m \left( \frac{1 - P_{vis}(t)}{1 - p_{vis}(t)} \right)^{i-m-1} - 1 \right| \tag{3.14}
 \end{aligned}$$

Es wird später gezeigt, dass weder  $p_{vis}(t)$  noch  $1 - p_{vis}(t)$  den Wert Null erreichen und die durchgeführten Schritte deshalb korrekt sind. Um die Quotienten der Wahrscheinlichkeiten bestimmen zu können, soll  $P_{vis}$  durch die bekannte Größe  $p_{vis}$  und einen möglichst kleinen Fehlerterm ausgedrückt werden.

$$\begin{aligned}
 P_{vis}(t) &= \sum_k P_{unto,k}(t) P_{unto}(t)^k \\
 &= \sum_k \frac{k V_{unto,k}(t)}{C_{unto}(t)} \left( \frac{C_{unto}(t)}{C_{unex}(t)} \right)^k \\
 &= \sum_k \frac{k \frac{V_{unto,k}(t)}{n}}{\frac{C_{unto}(t)}{n}} \left( \frac{\frac{C_{unto}(t)}{n}}{\frac{C_{unex}(t)}{n}} \right)^k \tag{3.15}
 \end{aligned}$$

Die Ungleichungen (3.4) bis (3.6) lassen sich zu

$$\begin{aligned}
 \left| \frac{V_{unto,j}(t) - a_j t^j \cdot n}{n} \right| &< \frac{n^{1/2+\varepsilon}}{n} = n^{-1/2+\varepsilon} \\
 \left| \frac{C_{unex}(t) - c_{unex}(t) \cdot n}{n} \right| &< \frac{n^{1/2+\varepsilon}}{n} = n^{-1/2+\varepsilon} \\
 \left| \frac{C_{unto}(t) - c_{unto}(t) \cdot n}{n} \right| &< \frac{n^{1-\beta}}{n} = n^{-\beta}
 \end{aligned}$$

umformen und ergeben damit

$$\begin{aligned}
 \frac{V_{unto,j}(t)}{n} &= a_j t^j \pm O(n^{-1/2+\varepsilon}) \\
 \frac{C_{unex}(t)}{n} &= c_{unex}(t) \pm O(n^{-1/2+\varepsilon}) \\
 \frac{C_{unto}(t)}{n} &= c_{unto}(t) \pm O(n^{-\beta}) .
 \end{aligned}$$

Die Konstanten  $\beta$  und  $\varepsilon$  seien dabei wie in Lemma 3.1 definiert. Eingesetzt entfernt dies die enthaltenen Zufallsvariablen  $V_{unto,j}(t)$ ,  $C_{unex}(t)$  und  $C_{unto}(t)$  aus Gleichung (3.15).

$$P_{vis}(t) = \sum_k \frac{k (a_k t^k \pm O(n^{-1/2+\varepsilon}))}{c_{unto}(t) \pm O(n^{-\beta})} \left( \frac{c_{unto}(t) \pm O(n^{-\beta})}{c_{unex}(t) \pm O(n^{-1/2+\varepsilon})} \right)^k \tag{3.16}$$

### 3. Erwartungswert der Gradverteilung

---

Da laut Voraussetzung  $t \in [n^{-\theta}, 1 - n^{-\kappa}]$  gelten soll, erhält man mit  $\sum_j j a_j t^j = \Theta(t^{d_{min}})$  – gilt, da  $d_{min} a_{d_{min}} t^{d_{min}} \leq \sum_j j a_j t^j \leq \sum_j j a_j t^{d_{min}} = \delta t^{d_{min}}$ , wobei  $d_{min}$  weiterhin den kleinsten auftretenden Knotengrad bezeichnet – die unteren Abschätzungen

$$\begin{aligned} c_{unex}(t) &= \delta t^2 = \Omega(n^{-2\theta}) \\ c_{unto}(t) &= \sum_j j a_j t^j = \Omega(n^{-d_{min}\theta}). \end{aligned}$$

In (3.16) lassen sich nun die Ausdrücke zu

$$\begin{aligned} P_{vis}(t) &= \sum_k \frac{k a_k t^k \pm k O(n^{-1/2+\varepsilon})}{c_{unto}(t) \left(1 \pm \frac{O(n^{-\beta})}{c_{unto}(t)}\right)} \left( \frac{c_{unto}(t) \left(1 \pm \frac{O(n^{-\beta})}{c_{unto}(t)}\right)}{c_{unex}(t) \left(1 \pm \frac{O(n^{-1/2+\varepsilon})}{c_{unex}(t)}\right)} \right)^k \\ &= \sum_k \frac{k a_k t^k \pm k O(n^{-1/2+\varepsilon})}{c_{unto}(t) \left(1 \pm \frac{O(n^{-\beta})}{c_{unto}(t)}\right)} \left( \frac{c_{unto}(t)}{c_{unex}(t)} \right)^k \left( \frac{1 \pm \frac{O(n^{-\beta})}{c_{unto}(t)}}{1 \pm \frac{O(n^{-1/2+\varepsilon})}{c_{unex}(t)}} \right)^k \\ &= \sum_k \frac{k a_k t^k \pm k O(n^{-1/2+\varepsilon})}{c_{unto}(t) (1 \pm O(n^{-\beta+d_{min}\theta}))} \left( \frac{c_{unto}(t)}{c_{unex}(t)} \right)^k \left( \frac{1 \pm O(n^{-\beta+d_{min}\theta})}{1 \pm O(n^{-1/2+\varepsilon+2\theta})} \right)^k \\ &= \sum_k \frac{k a_k t^k \pm k O(n^{-\xi})}{c_{unto}(t) (1 \pm O(n^{-\xi}))} \left( \frac{c_{unto}(t)}{c_{unex}(t)} \right)^k \left( \frac{1 \pm O(n^{-\xi})}{1 \pm O(n^{-\xi})} \right)^k \end{aligned}$$

umschreiben, wobei  $\xi = \min(\beta - d_{min}\theta, 1/2 - \varepsilon - 2\theta)$  und für ein geeignetes  $\varepsilon$  deshalb  $\xi = \beta - d_{min}\theta$  gilt. Wegen

$$1 + 2c \cdot n^{-\xi} = \frac{(1 + 2c \cdot n^{-\xi})(1 - c \cdot n^{-\xi})}{1 - c \cdot n^{-\xi}} = \frac{1 + c \cdot n^{-\xi} - 2c^2 \cdot n^{-2\xi}}{1 - c \cdot n^{-\xi}} \geq \frac{1}{1 - c \cdot n^{-\xi}}$$

und

$$1 - c \cdot n^{-\xi} = \frac{(1 - c \cdot n^{-\xi})(1 + c \cdot n^{-\xi})}{1 + c \cdot n^{-\xi}} = \frac{1 - c^2 \cdot n^{-2\xi}}{1 + c \cdot n^{-\xi}} \leq \frac{1}{1 + c \cdot n^{-\xi}}$$

mit einer Konstanten  $c > 0$ ,  $c \neq n^\xi$  und  $n$  genügend groß ergibt sich unter Verwendung von  $(1 \pm x)^y = 1 \pm O(xy)$  für  $x, y \geq 0$  und  $xy < 1$

$$\begin{aligned} P_{vis}(t) &= \sum_k \frac{k a_k t^k \pm k O(n^{-\xi})}{c_{unto}(t)} \left( \frac{c_{unto}(t)}{c_{unex}(t)} \right)^k \left(1 \pm O(n^{-\xi})\right)^{2k+1} \\ &= \sum_k \frac{c_{unto}(t)^k}{c_{unto}(t) c_{unex}(t)^k} \left( k a_k t^k \pm k O(n^{-\xi}) \right) \left(1 \pm k O(n^{-\xi})\right) \\ &= \sum_k \frac{c_{unto}(t)^k}{c_{unto}(t) c_{unex}(t)^k} \left( k a_k t^k \pm k^2 O(n^{-\xi}) \right). \end{aligned}$$

Diese Umformungen sind jedoch nur bei Termen der Summe mit  $2k + 1 < C \cdot n^\xi$  für eine Konstante  $C$  korrekt. In Summanden mit größerem  $k$  lässt sich die Abschätzung von  $(1 + x)^y$  nicht mehr anwenden. Da  $t \leq 1 - n^{-\kappa}$  laut Voraussetzung gegeben ist, folgt für diese Fälle

$$t^k \leq e^{-k \cdot n^{-\kappa}} = e^{-\Omega(n^\xi - \kappa)}.$$

Durch entsprechende Wahl der Konstanten wird dieser Ausdruck exponentiell klein. M.s.h.W. gibt es also keine Knoten mit so hohen Graden, die zu Beginn des betrachteten Zeitraums noch nicht entdeckt sind. Es folgt hieraus  $P_{unto,k}(t) = 0$  und damit auch für die Terme in der Summe der Wert Null.

Einsetzen der Definition von  $p_{vis}$  (Gleichung 3.7) und wiederholtes Abschätzen von  $c_{unto}(t)$  sowie  $c_{unex}(t)$  vereinfachen den Ausdruck zu

$$\begin{aligned} P_{vis}(t) &= p_{vis}(t) \pm O(n^{-\xi}) \sum_k \frac{k^2 \cdot c_{unto}(t)^k}{c_{unto}(t)c_{unex}(t)^k} \\ &= p_{vis}(t) \pm O(n^{-\xi}) \cdot \frac{1}{c_{unto}(t)} \cdot \sum_k k^2 \cdot O\left(t^{k(d_{min}-2)}\right) \\ &= p_{vis}(t) \pm O(n^{-\xi}) \cdot O\left(\frac{1}{t^{d_{min}}}\right) \cdot O\left(\sum_k k^2 t^k\right). \end{aligned}$$

Die Summe kann man dank geeigneter Umformungen durch die Formel der geometrischen Reihe nach oben begrenzen.

$$\begin{aligned} \sum_k k^2 t^k &\leq \sum_k (k+2)(k+1)t^k \\ &= \frac{d^2}{dt^2} \sum_k t^{k+2} \\ &\leq \frac{d^2}{dt^2} \frac{t^2}{1-t} \\ &= \frac{1}{(1-t)^3} \end{aligned}$$

Für die Wahrscheinlichkeit  $P_{vis}$  erhält man dadurch

$$P_{vis}(t) = p_{vis}(t) \pm O(n^{-\xi}) \cdot O\left(\frac{1}{t^{d_{min}}}\right) \cdot O\left(\frac{1}{(1-t)^3}\right).$$

Setzt man für das erste in der Gleichung auftretende  $t$  die untere und für das zweite die obere Grenze des Intervalls  $[n^{-\theta}, 1 - n^{-\kappa}]$  ein, so liefert dies

$$\begin{aligned} P_{vis}(t) &= p_{vis}(t) \pm O(n^{-\xi}) \cdot O(n^{d_{min}\theta}) \cdot O(n^{3\kappa}) \\ &= p_{vis}(t) \pm O(n^{-\xi+d_{min}\theta+3\kappa}) \\ &= p_{vis}(t) \pm O(n^{-\beta+2d_{min}\theta+3\kappa}). \end{aligned}$$

Es fehlt an dieser Stelle von  $p_{vis}$  noch eine untere

$$\begin{aligned}
 p_{vis} &= \sum_k k a_k t^k \cdot (c_{unto})^{k-1} \left( \frac{1}{c_{unex}} \right)^k \\
 &= \sum_k k a_k \cdot (c_{unto})^{k-1} \cdot \delta^{-k} t^{-k} \\
 &\geq d_{min} a_{d_{min}} \cdot (c_{unto})^{d_{min}-1} \cdot \delta^{-d_{min}} t^{-d_{min}} \\
 &= \Omega \left( t^{d_{min} \cdot (d_{min}-1)} \cdot t^{-d_{min}} \right) \\
 &= \Omega \left( t^{d_{min} \cdot (d_{min}-2)} \right) \\
 &= \Omega \left( n^{-\theta \cdot d_{min} \cdot (d_{min}-2)} \right)
 \end{aligned}$$

sowie obere Abschätzung

$$\begin{aligned}
 p_{vis} &= \sum_k k a_k t^k \cdot (c_{unto})^{k-1} \left( \frac{1}{c_{unex}} \right)^k \\
 &\leq \sum_k k a_k \cdot \delta^{k-1} t^{(k-1) \cdot d_{min}} \cdot \delta^{-k} t^{-k} \\
 &= \frac{1}{\delta} \sum_k k a_k \cdot t^{k \cdot (d_{min}-1) - d_{min}} \\
 &\leq t^{d_{min} \cdot (d_{min}-1) - d_{min}} \cdot \frac{1}{\delta} \sum_k k a_k \\
 &= (1 - n^{-\kappa})^{d_{min} \cdot (d_{min}-2)} \\
 &= 1 - \Omega(n^{-\kappa}),
 \end{aligned}$$

mit deren Hilfe der Beweis des Lemmas vollendet werden kann. Die Resultate für  $P_{vis}$  und  $p_{vis}$  lassen sich auf Ungleichung (3.14) anwenden.

$$\begin{aligned}
 \left( \frac{P_{vis}(t)}{p_{vis}(t)} \right)^m &= \left( 1 \pm \frac{O(n^{-\beta+2d_{min}\theta+3\kappa})}{p_{vis}(t)} \right)^m \\
 &= \left( 1 \pm O(n^{-\beta+3\kappa+\theta \cdot d_{min} \cdot d_{min}}) \right)^m \\
 &= 1 \pm m \cdot O(n^{-\beta+3\kappa+\theta \cdot d_{min} \cdot d_{min}}) \\
 \\
 \left( \frac{1 - P_{vis}(t)}{1 - p_{vis}(t)} \right)^{i-1-m} &= \left( 1 \pm \frac{O(n^{-\beta+2d_{min}\theta+3\kappa})}{1 - p_{vis}(t)} \right)^{i-1-m} \\
 &= \left( 1 \pm O(n^{-\beta+2d_{min}\theta+4\kappa}) \right)^{i-1-m} \\
 &= 1 \pm (i-1-m) \cdot O(n^{-\beta+2d_{min}\theta+4\kappa})
 \end{aligned}$$

Zusammen mit  $0 \leq m \leq i-1 < i < n^\eta$  ergibt sich unter der Bedingung, dass der

Exponent von  $n$  negativ wird (vgl. Abschätzung von  $(1+x)^y$ ),

$$\begin{aligned} \left(\frac{P_{vis}(t)}{p_{vis}(t)}\right)^m &= 1 \pm O(n^{-\beta+3\kappa+\theta \cdot d_{min} \cdot d_{min}+\eta}) \\ &= 1 \pm O(n^{-\beta+4\kappa+\theta \cdot d_{min} \cdot d_{min}+\eta}) \\ \left(\frac{1-P_{vis}(t)}{1-p_{vis}(t)}\right)^{i-1-m} &= 1 \pm O(n^{-\beta+2d_{min}\theta+4\kappa+\eta}) \\ &= 1 \pm O(n^{-\beta+4\kappa+\theta \cdot d_{min} \cdot d_{min}+\eta}) \end{aligned}$$

und somit insgesamt

$$\begin{aligned} &|p_{i,m}(t) - \Pr[\text{Bin}(i-1, p_{vis}(t)) = m]| \\ &< n^{-\gamma} + \left| \left(\frac{P_{vis}(t)}{p_{vis}(t)}\right)^m \left(\frac{1-P_{vis}(t)}{1-p_{vis}(t)}\right)^{i-m-1} - 1 \right| \\ &= n^{-\gamma} + O(n^{-\beta+4\kappa+\theta \cdot d_{min} \cdot d_{min}+\eta}). \end{aligned}$$

Nach Lemma 3.4 ist  $\gamma = 2\eta = \beta^2/3$ . Sei ausserdem  $\kappa = \beta^2/6$ ,  $\theta = \beta^2/(6d_{min}^2)$  (erfüllt auch die Bedingung  $\theta \leq \beta/(2d_{min})$  aus Lemma 3.5) und  $\varepsilon \leq 1/2 - \beta + (d_{min} - 2)\theta$  (siehe Definition von  $\xi$ ). Dann folgt daraus  $\beta - 4\kappa - \theta \cdot d_{min}^2 - \eta = \beta - \beta^2 > 0$  und für  $\mu < \min(\gamma, \beta - \beta^2)$  gilt  $|p_{i,m}(t) - \Pr[\text{Bin}(i-1, p_{vis}(t)) = m]| = o(n^{-\mu})$ .  $\square$

### 3.5 Zusammenfassung

Das Ergebnis von Lemma 3.6 ermöglicht es, den Erwartungswert der Gradverteilung im resultierenden Suchbaum der Breitensuche (siehe Gleichung 3.1) mit ausreichender Genauigkeit abzuschätzen. Aus Gleichung (3.7) kombiniert mit den Definitionen von  $c_{unto}(t)$  und  $c_{unex}(t)$  in Lemma 3.1 erhält man

$$p_{vis}(t) = \frac{1}{\sum_j j a_j t^j} \sum_k k a_k t^k \left( \frac{\sum_j j a_j t^j}{\delta t^2} \right)^k \quad (3.17)$$

und die Gleichungen (3.1) und (3.2) führen mit

$$a_{m+1}^{obs} = \sum_i a_i \left[ \int_0^1 i t^{i-1} \binom{i-1}{m} p_{vis}(t)^m (1-p_{vis}(t))^{i-m-1} dt \right] \quad (3.18)$$

zu folgendem Lemma.

*Lemma 3.7*

Sei  $(a_0, a_1, \dots)$  eine angemessene Gradverteilung und der zugehörige Graph  $G$  zusammenhängend. Dann existiert eine Konstante  $\zeta > 0$  so, dass für alle  $j < n$  und genügend große  $n$  gilt

$$\left| \mathbb{E} [A_j^{obs}] - a_j^{obs} \cdot n \right| < n^{1-\zeta}.$$

*Beweis* Es existieren drei Fehlerquellen bei der näherungsweise Bestimmung von  $E[A_j^{obs}]$ . Zum einen liefert Lemma 3.6 bei der Abschätzung von  $p_{i,m}(t)$  durch  $p_{vis}(t)$  einen maximalen Fehler von  $n^{-\mu}$ . Desweiteren sagt Lemma 3.6 über zwei Typen von Knoten nichts aus: Knoten mit einem Grad größer oder gleich  $n^\eta$  und Knoten, die zu einem Zeitpunkt  $t \notin [n^{-\theta}, 1 - n^{-\kappa}]$  in die Schlange eingefügt werden.

Die erste Fehlerursache pflanzt sich unverändert bis zum Wert  $a_j^{obs}$  fort und ergibt dann durch den Faktor  $n$  einen maximalen Fehler von  $n^{1-\mu}$ . Für die Abschätzung kann man davon ausgehen, dass die Abweichung  $\Delta = \Pr[\text{Bin}(i-1, p_{vis}(t)) = m] - p_{i,m}(t)$  von  $t$  und  $i$  unabhängig ist, z.B. indem man sie durch eine obere Schranke wie  $n^{-\mu}$  ersetzt.

$$\begin{aligned}
 a_j^{obs} \cdot n &= n \sum_i a_i \left[ \int_0^1 it^{i-1} (p_{i,m}(t) + \Delta) dt \right] & |\Delta| < n^{-\mu} \\
 &= n \sum_i a_i \left[ \int_0^1 it^{i-1} p_{i,m}(t) dt \right] + n \sum_i a_i \left[ \int_0^1 it^{i-1} \Delta dt \right] \\
 &= E[A_j^{obs}] + n\Delta \cdot \sum_i a_i \left[ \int_0^1 it^{i-1} dt \right] \\
 &= E[A_j^{obs}] + n\Delta
 \end{aligned}$$

Die Fehlerbeträge der anderen Ursachen sind nach oben durch die Anzahl betroffener Knoten beschränkt, da diese Knoten fälschlicherweise im Faktor  $n$  enthalten sind und deren Wahrscheinlichkeiten bzw. sichtbare Kanten nicht getrennt betrachtet werden.

Die Anzahl von Knoten mit einem Grad größer oder gleich  $n^\eta$  ist

$$n \sum_{j \geq n^\eta} a_j < Cn \sum_{j \geq n^\eta} j^{-\alpha} < \frac{C}{\alpha-1} \cdot n \cdot n^{\eta(1-\alpha)} = O\left(n^{1-(\alpha-1)\eta}\right).$$

Für die Anzahl  $X$  der Knoten, die zu einem nicht von Lemma 3.6 abgedeckten Zeitpunkt in die Schlange eingefügt werden, ergeben sich  $n$  unabhängige Indikatorzufallsvariablen  $X_i$  – pro möglichem Knoten ausserhalb des Intervalls eine. Jeder dieser Indikatoren besitzt die gleiche Wahrscheinlichkeit von  $n^{-\theta} + n^{-\kappa}$ . Durch Anwenden der Chernoff-Schranke auf das binomialverteilte  $X$  mit einem Erwartungswert von  $z = n^{1-\theta} + n^{1-\kappa}$  erhält man

$$\Pr[X > [1 + (e-1)] \cdot z] < \left[ \frac{e^{e-1}}{(1+e-1)^{(1+e-1)}} \right]^z = e^{-z}.$$

M.s.h.W. gilt deshalb  $X \leq e \cdot z < n^{1-\zeta}$  für  $\zeta < \min(\theta, \kappa)$  und genügend große  $n$ , wobei die (exponentiell kleine) Fehlerwahrscheinlichkeit der Chernoff-Schranke in die Abweichung  $n^{1-\zeta}$  integriert wird ( $X$  ist nach oben durch die Anzahl aller Knoten, also  $n$ , beschränkt):

$$e^{-z} \cdot n + (1 - e^{-z}) \cdot e \cdot z = o(n^{1-\zeta}).$$

Durch die erweiterte Bedingung  $\zeta < \min(\mu, (\alpha-1)\eta, \theta, \kappa)$  werden auch die schon betrachteten Fehlerquellen einbezogen und der Beweis ist abgeschlossen.  $\square$

## 4 Konzentration um den Erwartungswert

Das Ergebnis aus Kapitel 3, insbesondere Lemma 3.7, ist nicht ausreichend, um genauere Aussagen über  $A_j^{obs}$ , also die Anzahl von Knoten eines bestimmten Grades im resultierenden Suchbaum der Breitensuche, zu treffen. Zuvor muss noch ermittelt werden, wie stark die exakte Gradverteilung von dem in Lemma 3.7 bestimmten Erwartungswert abweicht.

*Satz 4.1*

Es existiert eine Konstante  $\rho$  so, dass m.s.h.W. für alle  $j$  gilt

$$\left| A_j^{obs} - \mathbb{E} \left[ A_j^{obs} \right] \right| \leq O(n^{1-\rho}) .$$

*Beweis* Es wird schnell deutlich, dass das bisherige Verfahren - die Fälle, die in den betrachteten Situationen auftreten können, werden getrennt voneinander untersucht (vgl. Beweis Lemma 3.4 und 3.7) - nicht ohne weiteres anwendbar ist. Der Grund dafür ist intuitiv klar: Die Änderung einer einzigen Kante im Ursprungsgraphen kann enorme Auswirkungen auf den resultierenden Suchbaum und damit auch auf  $A_j^{obs}$  besitzen.

Die Lösung des Problems liegt in der Möglichkeit, die Schritte der zeitbasierten Breitensuche geschickt in „Paketen“ zusammenzufassen und den Zustand des Suchbaums nach Ausführung eines solchen „Paketes“ zu untersuchen. Im Detail soll davon ausgegangen werden, dass der Suchbaum bis zu einer festgelegten Tiefe  $r$  ermittelt wurde, d.h. alle Knoten mit einer maximalen Distanz von  $r$  zur Wurzel sind erreicht. Für diesen Zeitpunkt sei die Anzahl von Kopien in der Schlange  $\mathcal{Q}$  der Breitensuche sowie die Anzahl unentdeckter Kopien bekannt. Nun lässt sich zeigen, dass die Gradverteilung der Knoten in Tiefe  $r+1$  entsprechend der Aussage des Lemmas konzentriert ist. Allgemein gilt diese Aussage über die Konzentration der Gradverteilung für die Nachbarn eines „Bündels“ von Kopien, falls dieses „Bündel“ aus in der Schlange  $\mathcal{Q}(t)$  enthaltenen Kopien für einen Zeitpunkt  $t$  besteht.

Die Aufteilung aller Kopien des Graphen in solche „Bündel“ erfolgt implizit, indem vor Beginn der Breitensuche die zu betrachtenden Zeitpunkte festgelegt werden. Summiert man die Knotengrade der Nachbarn von  $\mathcal{Q}(t)$  über alle festgelegten  $t$  auf, so ergibt dies unter gewissen Bedingungen an die Zeitpunkte  $t$  eine gute Abschätzung von  $A_j^{obs}$ . Die Begriffe „Nachbar“ und „adjazent“ beziehen sich dabei – wenn nicht anders erwähnt – immer auf den Suchbaum, wobei der Knoten, über den Kopien entdeckt und in die Schlange eingefügt wurden, ausgenommen ist. Im Folgenden soll gezeigt werden, dass die Terme der Summe um ihren Erwartungswert konzentriert sind und der Erwartungswert  $\mathbb{E}[A_j^{obs}]$  nur sehr gering vom Erwartungswert dieser Summe abweicht.

Sei  $Q(t) := |\mathcal{Q}(t)|$  die Anzahl der zum Zeitpunkt  $t$  in der Schlange enthaltenen Kopien und

$$q(t) := \mathbb{E}[Q(t)] = (c_{unex}(t) - c_{unto}(t)) \cdot n \tag{4.1}$$

der Erwartungswert dieser Anzahl. Nun werden  $r \leq \log^2 n$  viele Zeitpunkte fixiert, für welche die Schlange und die daraus resultierenden Nachbarkopien untersucht werden. Beginnend mit  $t_1 = 1$  ergibt sich für alle  $i$  der nächste Zeitpunkt  $t_{i+1} \geq 0$  als das Maximum aller Zeitpunkte, für die gilt

$$c_{unex}(t_i) - c_{unex}(t_{i+1}) \geq \frac{q(t_i)}{n} + 2n^{-\beta}, \quad (4.2)$$

wobei  $\beta$  durch Lemma 3.1 definiert ist. Falls für ein  $i < \log^2 n$  kein solches  $t_{i+1}$  existiert (folgt z.B. aus  $c_{unex}(t_i) < 2n^{-\beta}$ ), dann sei  $r$  dieses  $i$ . Ansonsten gilt  $r = \log^2 n$ .

Für jeden Knotengrad  $j$  wird  $B_j(i)$  als die Anzahl der zu  $\mathcal{Q}(t_i)$  adjazenter Knoten mit dem resultierenden Grad  $j$  im Suchbaum definiert. Unter Verwendung von Lemma 4.2, welches zeigt, dass die  $B_j(i)$  jeweils um ihre Erwartungswerte konzentriert sind, lässt sich auch die Konzentration von  $A_j^{obs}$  beweisen. Dafür müssen zwei Bedingungen zutreffen: (1) kein Knoten wird mehrfach gezählt und (2) fast alle Knoten werden durch einen der gewählten Zeitpunkte erfasst.

Die erste Bedingung trifft zu, wenn zwischen den Zeitpunkten  $t_i$  und  $t_{i+1}$  mindestens alle zum Zeitpunkt  $t_i$  in der Schlange enthaltenen Kopien abgearbeitet werden. Es wird daher niemals ein Knoten mehrfach gezählt, falls für alle  $i$  die Ungleichung

$$C_{unex}(t_i) - C_{unex}(t_{i+1}) \geq Q(t_i) \quad (4.3)$$

erfüllt ist und somit die  $\mathcal{Q}(t_i)$  paarweise disjunkt sind.

*Die Ungleichung 4.3 ist aus [1] übernommen wurden. Der Autor der vorliegenden Arbeit ist jedoch der Meinung, dass die Ungleichung nicht korrekt ist. Pro Ausführungsschritt der Breitensuche sind zwei Kopien betroffen, die somit als abgearbeitet markiert werden. Nur eine dieser Kopien – der Kopf der Schlange – muss jedoch in  $\mathcal{Q}$  enthalten sein. Die zweite Kopie könnte bisher unentdeckt geblieben sein, wäre somit noch nicht Element der Schlange und dürfte deshalb auch nicht gezählt werden. In der simplen Differenz der  $C_{unex}$  wird diese Möglichkeit jedoch nicht beachtet.*

*Die Autoren von [1] wurden per E-Mail darüber in Kenntnis gesetzt. In einer Antwort bestätigt David Kempe das Problem und verspricht eine Lösung für die endgültige Version ihrer Arbeit.*

*Als möglicher Lösungsansatz kommt ein korrigierender Faktor in Frage. Ändert man die Ungleichung in*

$$C_{unex}(t_i) - C_{unex}(t_{i+1}) \geq 2Q(t_i) \quad (4.4)$$

*um, so ist sichergestellt, dass selbst im worst-case alle Kopien der Schlange abgearbeitet sind. Es wird sich jedoch zeigen, dass dadurch neue Probleme aufgeworfen werden, die sich evtl. nicht so schnell lösen lassen.*

Nach Lemma 3.1 und Ungleichung (4.2) gilt für ein geeignet gewähltes  $\varepsilon$  m.s.h.W.

$$\begin{aligned} C_{unex}(t_i) - C_{unex}(t_{i+1}) &\geq n \cdot (c_{unex}(t_i) - c_{unex}(t_{i+1})) - 2n^{1/2+\varepsilon} \\ &\geq q(t_i) + 2n^{1-\beta} - 2n^{1/2+\varepsilon} \\ &\geq q(t_i) + \frac{3}{2}n^{1-\beta} \end{aligned}$$

und

$$\begin{aligned}
 q(t_i) + \frac{3}{2}n^{1-\beta} &= n \cdot (c_{unex}(t_i) - c_{unto}(t_i)) + \frac{3}{2}n^{1-\beta} \\
 &\geq C_{unex}(t_i) - C_{unto}(t_i) - n^{1/2+\varepsilon} - n^{1-\beta} + \frac{3}{2}n^{1-\beta} \\
 &\geq C_{unex}(t_i) - C_{unto}(t_i) \\
 &= Q(t_i). \tag{4.5}
 \end{aligned}$$

Das Zusammenfassen von maximal  $\log^2 n$  vielen Zeitpunkte zeigt, dass dies m.s.h.W. gleichzeitig für alle  $t_i$  gilt. Durch die Ungleichung (4.3) erschliesst sich auch die Bedeutung der gewählten Zeitpunkte: Alle zum Zeitpunkt  $t_i$  in der Schlange enthaltenen Kopien sind zum Zeitpunkt  $t_{i+1}$  m.s.h.W. abgearbeitet. Dies entspricht auf die Breitensuche übertragen der Erweiterung des Suchbaums um mindestens eine Ebene.

*Für den Nachweis der geänderten Ungleichung 4.4 ist es notwendig, die Festlegung der Zeitpunkte  $t_i$  anzupassen. Die einfachste Methode ist sicher, den neuen Faktor 2 auch für Ungleichung 4.2 zu übernehmen:*

$$c_{unex}(t_i) - c_{unex}(t_{i+1}) \geq 2 \frac{q(t_i)}{n} + 4n^{-\beta}. \tag{4.6}$$

*Der weitere Beweis verläuft analog.*

Nun soll die Anzahl der nicht erfassten Knoten nach oben abgeschätzt werden. Damit ein Knoten in keinem  $B_j(i)$  enthalten ist, darf seine Kopie mit dem größten Index zu keiner Kopie aus  $\mathcal{Q}(t_i)$  für alle  $i$  adjazent sein. Die Anzahl nicht in der Vereinigung aller  $\mathcal{Q}(t_i)$  enthaltenen Kopien ist somit gleichzeitig eine obere Abschätzung der nicht erfassten Knoten.

Diese Kopien könnten zum einen erst nach dem Zeitpunkt  $t_r$  durch die Breitensuche in die Schlange  $\mathcal{Q}$  eingefügt werden. Falls die Breitensuche vorzeitig beendet wird, d.h.  $r < \log^2 n$ , so folgt aus  $c_{unex}(t_r) - c_{unex}(0) < q(t_r)/n + 2n^{-\beta}$  (es existiert kein passendes  $t_{r+1} \geq 0$ ) und  $c_{unex}(0) = 0$  (am Ende sind alle Kopien abgearbeitet) mit Gleichung (4.1) die Schranke  $c_{unto}(t_r) < 2n^{-\beta}$ . Durch Lemma 3.1 gelangt man zu der Aussage, dass die Anzahl der unentdeckten Kopien m.s.h.W. durch  $O(n^{1-\beta})$  begrenzt wird.

*An dieser Stelle tritt nun das erste ernsthafte Problem auf. Da Ungleichung 4.2 durch Ungleichung 4.6 ersetzt werden sollte, ändert sich auch die Begrenzung für die Anzahl nicht erfasster Knoten im Falle einer vorzeitig beendeten Breitensuche. Es ergibt sich  $c_{unex}(t_r) < 2(c_{unex}(t) - c_{unto}(t)) + 4n^{-\beta}$  und deshalb  $c_{unto}(t_r) < \frac{1}{2}c_{unex}(t_r) + 2n^{-\beta}$ . Ohne genauere Informationen über  $c_{unex}$  zu einem nicht explizit bestimmten Zeitpunkt  $t_r$  lassen sich als keine Aussagen über die Anzahl der verbleibenden Kopien treffen.*

*Leider reicht die noch zur Verfügung stehende Zeit – diese Zeilen wurden nur wenige Wochen vor dem Abgabetermin der Arbeit verfasst – nicht aus, um genauere Überlegungen zu diesem Problem anzustellen oder alternative Lösungswege zu suchen. So verbleibt dem Autor an dieser Stelle nur der Verweis auf die korrigierte, endgültige Fassung von [1].*

Für den Fall  $r = \log^2 n$  benötigt man die in der endgültigen Version von [1] bewiesene Aussage, dass der maximale Abstand zweier Knoten (Diameter) der hier betrachteten Zufallsgraphen mit einer Wahrscheinlichkeit von mindestens  $1 - n^{-1/2}$  durch  $\log^2 n$

beschränkt ist. Selbst wenn dann  $C_{unto}(t_r) = \Omega(n)$  gilt, d.h. keine hinreichend kleine Schranke an die verbleibende Kopienzahl existiert, so tritt dieses Ereignis jedoch nur mit einer Wahrscheinlichkeit von höchstens  $n^{-1/2}$  ein. Im Falle einer Beschränkung des Diameters sind selbst die am weitesten vom Startknoten entfernten Knoten in einem  $B_j(i)$  enthalten (oder zählen zu dem Typ Knoten, der im nächsten Absatz betrachtet wird), da zum Zeitpunkt  $t_r$  alle Knoten bis zu einer Entfernung von  $r - 1$  erreicht werden. Es ergibt sich  $c_{unto}(t_r) \cdot n = O(n^{1/2}) = O(n^{1-\beta})$ .

Andererseits ist es auch möglich, dass Kopien nach einem Zeitpunkt  $t_i$  in die Schlange  $\mathcal{Q}$  eingefügt, aber vor dem Zeitpunkt  $t_{i+1}$  abgearbeitet werden. Deren adjazente Kopien sind somit auch durch kein  $B_j(i)$  erfasst. Zur Bestimmung dieser Anzahl muss beachtet werden, dass  $c_{unex}(t)$  eine reellwertige, stetige Funktion darstellt. Falls  $c_{unex}(t_{i+1})$  existiert, so gilt in (4.2) die Gleichheit. Nach Lemma 3.1 folgt für einen festen Zeitpunkt – und nach Zusammenfassen über  $i$  für alle Zeitpunkte  $t_i$  simultan – m.s.h.W.

$$\begin{aligned}
 [C_{unex}(t_i) - C_{unex}(t_{i+1})] - Q(t_i) &\leq (c_{unex}(t_i) - c_{unex}(t_{i+1})) \cdot n + 2n^{1/2+\varepsilon} - Q(t_i) \\
 &= q(t_i) + 2n^{1-\beta} + 2n^{1/2+\varepsilon} - Q(t_i) \\
 &= \mathbb{E}[Q(t_i)] - Q(t_i) + 2n^{1-\beta} + 2n^{1/2+\varepsilon} \\
 &\leq 3n^{1-\beta} + 3n^{1/2+\varepsilon} \\
 &= O(n^{1-\beta}).
 \end{aligned}$$

Es ergibt sich m.s.h.W. also, dass (1) alle  $Q(t_i)$  paarweise disjunkt und (2) in der Vereinigung aller  $Q(t_i)$  höchstens  $r \cdot O(n^{1-\beta}) + O(n^{1-\beta}) = \tilde{O}(n^{1-\beta})$  viele Kopien nicht enthalten sind. Der Ausdruck  $\tilde{O}(n^{1-\beta})$  umfasst dabei als Faktor ein endliches Polynom in  $\log n$ .

Deshalb gilt für ein festes  $j$  m.s.h.W.

$$\left| A_j^{obs} - \sum_{i=1}^r B_j(i) \right| = \tilde{O}(n^{1-\beta}) \quad (4.7)$$

und somit aufgrund der Linearität des Erwartungswerts auch

$$\begin{aligned}
 \left| \mathbb{E} \left[ A_j^{obs} \right] - \sum_{i=1}^r \mathbb{E} [B_j(i)] \right| &= \left| \mathbb{E} \left[ A_j^{obs} - \sum_{i=1}^r B_j(i) \right] \right| \\
 &\leq \mathbb{E} \left[ \left| A_j^{obs} - \sum_{i=1}^r B_j(i) \right| \right] \\
 &= \tilde{O}(n^{1-\beta}) \cdot (1 - o(n^{-c})) + n \cdot o(n^{-c}) \\
 &= \tilde{O}(n^{1-\beta}) \quad (\text{setzen } c = \beta), \quad (4.8)
 \end{aligned}$$

da der Ausdruck  $\left| A_j^{obs} - \sum_{i=1}^r B_j(i) \right|$  nach oben durch die Anzahl der Knoten beschränkt ist.

Nach Lemma 4.2 existiert ein  $\tau > 0$  so, dass  $|B_j(i) - \mathbb{E}[B_j(i)]| = O(n^{1-\tau})$  m.s.h.W. erfüllt wird. Abschliessend ist es wieder notwendig, die Aussagen für alle  $i$  und  $j$  zu

kombinieren, um mit der Dreiecksungleichung und m.s.h.W.

$$\begin{aligned}
 \left| A_j^{obs} - \mathbb{E} \left[ A_j^{obs} \right] \right| &= \left| A_j^{obs} - \sum_{i=1}^r B_j(i) + \sum_{i=1}^r B_j(i) - \sum_{i=1}^r \mathbb{E} [B_j(i)] \right. \\
 &\quad \left. + \sum_{i=1}^r \mathbb{E} [B_j(i)] - \mathbb{E} \left[ A_j^{obs} \right] \right| \\
 &\leq \left| A_j^{obs} - \sum_{i=1}^r B_j(i) \right| + \sum_{i=1}^r |B_j(i) - \mathbb{E} [B_j(i)]| \\
 &\quad + \left| \mathbb{E} \left[ A_j^{obs} \right] - \sum_{i=1}^r \mathbb{E} [B_j(i)] \right| \\
 &= \tilde{O}(n^{1-\beta}) + r \cdot O(n^{1-\tau}) \\
 &= \tilde{O}(n^{1-\beta}) + \tilde{O}(n^{1-\tau}) \\
 &= O(n^{1-\rho}) \quad \text{mit } \rho < \min(\beta, \tau)
 \end{aligned}$$

für alle Knotengrade  $j$  erhalten zu können. □

## 4.1 Betrachtung der $B_j(i)$

Wie anfangs bereits erwähnt, so besteht das Hauptproblem dieses Kapitels darin, dass minimale Änderungen am Zufallsgraphen enorme Auswirkungen auf den resultierenden Suchbaum haben können. Im Beweis des Satzes 4.1 wurde dieses Problem auf die Betrachtung der  $B_j(i)$ , also die Anzahl adjazenter Knoten eines vorgegeben Grades bezüglich der Arbeitsschlange zu einem festen Zeitpunkt, und deren Erwartungswerte reduziert.

Der Vorteil dabei ist, dass nun Änderungen durch Vertauschen zweier Kanten überschaubare Auswirkungen besitzen. Bei diesen Vertauschungen werden zwei Kanten  $\{p_1, p_2\}, \{p_3, p_4\}$  durch  $\{p_1, p_3\}, \{p_2, p_4\}$  ersetzt, d.h. die beteiligten Kopien über Kreuz neu verbunden. Sowohl vor als auch nach dem Austausch gibt es für die Kopien einer Kante drei Möglichkeiten: Keine der beiden, genau eine oder alle Kopien können in  $\mathcal{Q}(t_i)$  enthalten sein.

Unabhängig davon ändert sich der Wert von  $B_j(i)$  um höchstens 2, wobei dieses Maximum genau dann erreicht wird, wenn anfangs beide Kanten genau eine Kopie aus  $\mathcal{Q}(t_i)$  enthalten und dies nach dem Austausch jedoch für keine Kante mehr zutrifft oder im entgegengesetzten Fall.

*Lemma 4.2*

Es existiert eine Konstante  $\tau > 0$  so, dass für ein beliebiges, aber festes  $i$  m.s.h.W.

$$|B_j(i) - \mathbb{E} [B_j(i)]| = O(n^{1-\tau})$$

gleichzeitig für alle  $j$  gilt.

*Beweis* Das folgende Theorem aus der Arbeit von Wormald [10, Theorem 2.19] bildet die Grundlage für den Beweis.

*Theorem 4.3*

Sei  $X_k$  eine Zufallsvariable, die für gleichförmig zufällige Konfigurationen  $M, M'$  von  $k$  Kopien definiert ist. Weiterhin gelte für eine Konstante  $c$

$$|X_k(M) - X_k(M')| \leq c,$$

falls  $M$  und  $M'$  durch genau eine Vertauschung ineinander überführt werden können. Dann gilt für alle  $\Delta > 0$

$$\Pr [|X_k - \mathbb{E}[X_k]| \geq \Delta] < 2e^{-\Delta^2/(kc^2)}.$$

Im Folgenden bezeichne  $\mathcal{E}_{q,\mathbf{b}}$  für feste Werte  $q$  und  $\mathbf{b} = (b_1, \dots, b_n)$  das Ereignis  $Q(t_i) = q$  und  $V_{unto,j}(t_i) = b_j$  für alle  $j$ . Bedingt auf diesem Ereignis ist das durch die Kanten festgelegte Matching  $M$  der  $k = q + \sum_j j b_j$  Kopien, die noch nicht abgearbeitet wurden, gleichförmig zufällig. Sei nun  $B_j(i)$  die Zufallsvariable  $X_k$  aus Theorem 4.3. Die oben ausgeführte Betrachtung einer Vertauschung von genau zwei Kanten liefert  $c = 2$  als Konstante des Theorems. Es ergibt sich mit  $\Delta = n^{1/2+\varepsilon}$  aus

$$\begin{aligned} \Pr [|B_j(i) - \mathbb{E}[B_j(i) | \mathcal{E}_{q,\mathbf{b}}]| \geq \Delta] &< 2e^{-\Delta^2/(kc^2)} \\ &= 2e^{-n^{1+2\varepsilon}/O(n)} \\ &= 2e^{-\Omega(n^{2\varepsilon})}, \end{aligned}$$

dass m.s.h.W.

$$|B_j(i) - \mathbb{E}[B_j(i) | \mathcal{E}_{q,\mathbf{b}}]| < n^{1/2+\varepsilon} \tag{4.9}$$

für alle  $\varepsilon > 0$  erfüllt ist. Falls die Länge der Schlange und die Anzahl unentdeckter Knoten für jeden auftretenden Knotengrad bekannt sind, kann das Theorem 4.3 somit direkt angewendet werden. Da sich diese Werte normalerweise nicht exakt bestimmen lassen, müssen Abweichungen hinsichtlich ihrer Auswirkungen auf die oben ermittelte Schranke betrachtet werden. Lemma 4.4 zeigt, dass sich der bedingte Erwartungswert  $\mathbb{E}[B_j(i) | \mathcal{E}_{q,\mathbf{b}}]$  nur gering vom tatsächlichen Erwartungswert  $\mathbb{E}[B_j(i)]$  unterscheidet. Als Schlussfolgerung führt eine Konzentration um den bedingten Erwartungswert auch zu der zu beweisenden Konzentration um den tatsächlichen Erwartungswert.

Die möglichen Werte von  $q$  und  $\mathbf{b}$  werden hierbei auf die Intervalle

$$I^q := \left[ q(t_i) - 2n^{1-\beta}, q(t_i) + 2n^{1-\beta} \right]$$

bzw.

$$\begin{aligned} I_j^b &:= \left[ a_j t_i^j \cdot n - n^{1/2+\varepsilon}, a_j t_i^j \cdot n + n^{1/2+\varepsilon} \right] \\ I^b &:= I_1^b \times \dots \times I_n^b \end{aligned}$$

für eine Konstante  $0 < \varepsilon < 1/2$  eingeschränkt. Nun sei  $\mathcal{E}_{\leq}$  das Ereignis  $Q(t_i) \in I^q$  und  $V_{unto,j}(t_i) \in I_j^b$ . Nach Lemma 3.1 trifft dieses Ereignis m.s.h.W. ein (vgl. Gleichung 4.5).

Durch Lemma 4.4 wird bewiesen, dass der bedingte Erwartungswert um den tatsächlichen konzentriert ist, falls die Voraussetzungen  $q \in I^q$  und  $\mathbf{b} \in I^b$  erfüllt sind. Für ein konstantes  $\tau > 0$  gilt dann

$$|\mathbb{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}] - \mathbb{E}[B_j(i)]| = O(n^{1-\tau}).$$

Zusammen mit Ungleichung (4.9) erhält man durch die Dreiecksungleichung für das Ereignis  $\mathcal{E}_{\leq}$  m.s.h.W.

$$\begin{aligned} |B_j(i) - \mathbb{E}[B_j(i)]| &\leq |B_j(i) - \mathbb{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}]| + |\mathbb{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}] - \mathbb{E}[B_j(i)]| \\ &= O(n^{1-\tau}). \end{aligned} \quad (4.10)$$

Da die Differenz von  $B_j(i)$  und  $\mathbb{E}[B_j(i)]$  vom Betrag her nach oben durch die Anzahl der Knoten beschränkt ist, ändert sich die Schranke aus Gleichung (4.10) auch dann nicht, wenn das Ereignis  $\mathcal{E}_{\leq}$  nicht mehr vorausgesetzt wird (vgl. Gleichung 4.8). Der Beweis ist vollständig, wenn die erhaltene (sehr hohe) Wahrscheinlichkeit der Schranke für alle Werte von  $j$  zusammengefasst wird.  $\square$

Zum endgültigen Abschluss der Beweisführung wird nur noch ein weiteres Lemma über die Konzentration des bedingten Erwartungswertes von  $B_j(i)$  benötigt. Intuitiv dabei erscheint, dass diese Konzentration zu erwarten ist, da die beteiligten Variablen  $q$  und  $\mathbf{b}$  nach Voraussetzung nur sehr geringen Abweichungen unterworfen sind. Es stellt sich jedoch als erstaunlich aufwendig heraus, diesen Sachverhalt exakt nachzuweisen.

*Lemma 4.4*

Es existiert eine Konstante  $\tau > 0$  so, dass für beliebige  $q \in I^q$  und  $\mathbf{b} \in I^b$

$$|\mathbb{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}] - \mathbb{E}[B_j(i)]| = O(n^{1-\tau})$$

erfüllt ist.

*Beweis* Für den Beweis sollen zwei Situationen betrachtet werden, wobei die Länge der Schlangen bzw. die Anzahl der unentdeckten Knoten in diesen Fällen nicht mehr als um einen festgelegten Betrag voneinander abweichen. Es lässt sich zeigen, dass unter diesen Bedingungen die Erwartungswerte nahe beieinander liegen, wodurch sich Schlussfolgerungen für die bedingten Erwartungswerte ergeben.

Seien  $q, q'$  und  $\mathbf{b}, \mathbf{b}'$  mit  $|q - q'| \leq 4n^{1-\beta}$  und  $|b_j - b'_j| \leq 2n^{1/2+\varepsilon}$  für alle  $j$  gegeben. Ausserdem werden die Bezeichnungen  $\hat{q} = \min(q, q')$  und  $\hat{b}_j = \min(b_j, b'_j)$  sowie  $\mathcal{E} := \mathcal{E}_{q,\mathbf{b}}$ ,  $\mathcal{E}' := \mathcal{E}_{q',\mathbf{b}'}$  und  $\hat{\mathcal{E}} := \mathcal{E}_{\hat{q},\hat{\mathbf{b}}}$  eingeführt. Aus der Behauptung, für ein  $\tau > 0$  und alle  $j$  gelte

$$\left| \mathbb{E}[V_{\text{unto},j}(t_i) \mid \mathcal{E}] - \mathbb{E}[V_{\text{unto},j}(t_i) \mid \hat{\mathcal{E}}] \right| = O(n^{1-\tau})$$

und

$$\left| \mathbb{E}[V_{\text{unto},j}(t_i) \mid \mathcal{E}'] - \mathbb{E}[V_{\text{unto},j}(t_i) \mid \hat{\mathcal{E}}] \right| = O(n^{1-\tau}),$$

folgt mittels Dreiecksungleichung direkt

$$\left| \mathbb{E}[V_{\text{unto},j}(t_i) \mid \mathcal{E}] - \mathbb{E}[V_{\text{unto},j}(t_i) \mid \mathcal{E}'] \right| = O(n^{1-\tau}). \quad (4.11)$$

Nun wird zuerst der  $(q, \mathbf{b})$ -Fall betrachtet: Eine beliebige, aber feste Menge bestehend aus  $q - \hat{q}$  vielen Kopien der Schlange wird ebenso wie die Kopien einer beliebigen Menge von jeweils  $b_j - \hat{b}_j$  unentdeckten Knoten für alle Knotengrade  $j$  schwarz gefärbt. Zur Ermittlung des Matchings, d.h. der Kanten, werden zunächst die Nachbarn der schwarzgefärbten Kopien gemäß Gleichverteilung unter allen Kopien bestimmt und – so sie denn noch ungefärbt sind – blau gefärbt. Auf den restlichen Kopien, die von nun an als weiß angesehen werden, wird ebenfalls ein gleichförmiges Matching generiert, um die noch fehlenden Kanten zu ergänzen. Die Anzahl blauer Kopien entspricht einer nicht näher bekannten Verteilung  $D_{q, \mathbf{b}}$ , aber übersteigt in keinem Fall die Anzahl ihrer Nachbarn, der schwarzen Kopien. Diese Anzahl ist – da die Differenz  $b_j - \hat{b}_j$  durch die Gesamtzahl von Kopien, die Knoten mit dem Grad  $j$  zugeordnet sind, beschränkt ist – für alle  $\nu > 0$  höchstens

$$\begin{aligned}
 (q - \hat{q}) + \sum_j j \cdot (b_j - \hat{b}_j) &\leq 4n^{1-\beta} + \sum_{j \leq n^\nu} j \cdot (b_j - \hat{b}_j) + \sum_{j > n^\nu} j \cdot (b_j - \hat{b}_j) \\
 &\leq 4n^{1-\beta} + \sum_{j \leq n^\nu} j \cdot 2n^{1/2+\varepsilon} + \sum_{j > n^\nu} j \cdot a_j n \\
 &= 4n^{1-\beta} + O(n^{1/2+\varepsilon+2\nu}) + O(n^{1-(\alpha-2)\nu}) \\
 &= O(n^{1-\tau}),
 \end{aligned}$$

wobei  $\tau < \min(\beta, 1/2 - \varepsilon - 2\nu, (\alpha - 2)\nu)$  gilt. Für  $\tau > 0$  muss  $1/2 - \varepsilon - 2\nu > 0$  erfüllt sein. Da  $\varepsilon < 1/2$  ist (s. Definition von  $I^q$  und  $I^b$ , vorheriges Lemma), verbleibt für die Wahl von  $\nu$  als Einschränkung  $\nu < (1/2 - \varepsilon)/2$ .

Im Vergleich dazu wird der  $(\hat{q}, \hat{\mathbf{b}})$ -Fall untersucht, wobei ein gleichförmiges Matching wie folgt gefunden werden kann. Nach der Verteilung  $D_{q, \mathbf{b}}$  wird eine Zahl  $k$  gewählt und entsprechend viele zufällig gemäß Gleichverteilung ermittelte Kopien blau gefärbt. Alle verbleibenden Kopien erhalten eine weiße Färbung. Wie im  $(q, \mathbf{b})$ -Fall wird das Matching getrennt für die blauen und weißen Kopien durchgeführt. Nun bezeichnet man alle Knoten, die mindestens eine schwarze Kopie enthalten, als schwarz, ansonsten blau, falls mindestens eine so gefärbte Kopie vorhanden ist, und weiß in allen anderen Fällen.

Nach Definition von  $\hat{q}$  und  $\hat{\mathbf{b}}$  ist die Anzahl der nicht schwarz gefärbten unentdeckten Knoten in beiden Fällen identisch ( $\hat{b}_j$  für jeden Knotengrad  $j$ ). Analoges gilt für die Kopien in der Schlange, wobei in beiden Fällen jeweils  $\hat{q}$  Kopien nicht schwarz gefärbt sind. Desweiteren ist die Wahrscheinlichkeitsverteilung blauer Knoten wie gefordert ebenfalls gleich, sodass sich in beiden Fällen der gleiche Erwartungswert für die Anzahl weißer Knoten ergibt.

Der Erwartungswert von  $B_j(i)$  kann auf die beiden Fälle bezogen also nur um die Anzahl blauer und schwarzer Knoten variieren. Da die Anzahl entsprechend gefärbter Kopien wie gezeigt deterministisch durch  $O(n^{1-\tau})$  beschränkt ist, stellt dies auch bei ungünstiger Wahl der einzufärbenden Kopien eine obere Grenze für die Anzahl blauer oder schwarzer Knoten dar. Diese Aussage trifft für alle nach Voraussetzung möglichen Ereignisse  $\mathcal{E}$  und  $\hat{\mathcal{E}}$  und nach (4.11) somit auch für  $\mathcal{E}$  und  $\mathcal{E}'$  zu.

Schränkt man  $q$  und  $q'$  auf das Intervall  $I^q$  sowie  $\mathbf{b}$  und  $\mathbf{b}'$  auf  $I^b$  ein, so sind die Voraussetzungen für Gleichung (4.11) immer erfüllt. Nun entspricht die Vereinigung aller

Ereignisse  $\mathcal{E}'$  dem im vorherigen Lemma definierten Ereignis  $\mathcal{E}_{\leq}$  und da für alle  $\mathcal{E}'$  die gleiche Schranke  $O(n^{1-\tau})$  nachgewiesen werden konnte, erhält man

$$|\mathbb{E}[B_j(i) \mid \mathcal{E}_{q,\mathbf{b}}] - \mathbb{E}[B_j(i) \mid \mathcal{E}_{\leq}]| = O(n^{1-\tau}) .$$

Aus der sehr hohen Wahrscheinlichkeit von  $\mathcal{E}_{\leq}$  und der Beschränktheit von  $B_j(i)$  ergibt sich ausserdem für alle  $c$  (vgl. Gleichung 4.8)

$$|\mathbb{E}[B_j(i) \mid \mathcal{E}_{\leq}] - \mathbb{E}[B_j(i)]| = O(n^{-c})$$

und letztendlich verhilft die Dreiecksungleichung zum Abschluss des Beweises. □

## 5 Zusammenhang der Verteilungen

In den Kapiteln 3 und 4 wurde getrennt voneinander bewiesen, dass die zu beobachtende Gradverteilung im Breitensuchbaum um deren Erwartungswert und dieser wiederum um den durch Gleichung (3.18) bestimmten Wert konzentriert ist. Kombiniert man beide Resultate, so ergibt sich mit folgendem Satz die Kernaussage dieser Arbeit.

### Satz 5.1

Sei  $G$  ein zufälliger Multigraph mit der angemessenen Gradverteilung  $(a_0, a_1, \dots)$ , wobei vorausgesetzt wird, dass  $G$  zusammenhängend ist. Sei ausserdem  $T$  ein Breitensuchbaum von  $G$  und  $A_j^{obs}$  die Anzahl von Knoten mit dem Grad  $j$  in  $T$ . Dann existiert eine Konstante  $\zeta > 0$  so, dass mit hoher Wahrscheinlichkeit  $|A_j^{obs} - a_j^{obs} \cdot n| < n^{1-\zeta}$  für alle  $j$  und

$$a_{m+1}^{obs} = \sum_i a_i \left[ \int_0^1 it^{i-1} \binom{i-1}{m} p_{vis}(t)^m (1 - p_{vis}(t))^{i-1-m} dt \right],$$

$$p_{vis}(t) = \frac{1}{\sum_j j a_j t^j} \sum_k k a_k t^k \left( \frac{\sum_j j a_j t^j}{\delta t^2} \right)^k$$

erfüllt ist. Unter Verwendung von Generierungsfunktionen [9] erhält man als elegantere Schreibweise folgenden Ausdruck: Mit  $g(z) = \sum_{j=0}^{\infty} a_j z^j$  ergibt sich  $a_j^{obs}$  als Koeffizient von  $z_j$  in

$$g^{obs}(z) = z \int_0^1 g' \left[ t - \frac{(1-z)}{g'(1)} g' \left( \frac{g'(t)}{g'(1)} \right) \right] dt.$$

*Beweis* Der erste Teil des Satzes ist eine direkte Folgerung aus den Kapiteln 3 und 4 und ergibt sich insbesondere aus Lemma 3.7 (unter Einbezug der Gleichungen 3.18 und 3.17) und Satz 4.1.

Für den weiteren Beweis betrachte man die Generierungsfunktion der ursprünglichen Gradverteilung

$$g(z) = \sum_i a_i z^i.$$

Dieser Ausdruck vereint die Informationen über die vollständige Gradverteilung und ermöglicht es, benötigte Größen zur Berechnung der Verteilung im Breitensuchbaum kompakt als

$$c_{unto}(t) = t g'(t), \quad \delta = g'(1), \quad c_{unex}(t) = t^2 g'(1)$$

auszudrücken. Damit vereinfacht sich auch der Ausdruck für  $p_{vis}(t)$  aus Gleichung (3.7) zu

$$\begin{aligned}
 p_{vis}(t) &= \sum_k \frac{ka_k t^k}{tg'(t)} \left( \frac{g'(t)}{tg'(1)} \right)^k \\
 &= \frac{1}{tg'(t)} \sum_k ka_k \left( \frac{g'(t)}{g'(1)} \right)^k \\
 &= \frac{1}{tg'(t)} \cdot \frac{g'(t)}{g'(1)} g' \left( \frac{g'(t)}{g'(1)} \right) \\
 &= \frac{1}{tg'(1)} g' \left( \frac{g'(t)}{g'(1)} \right). \tag{5.1}
 \end{aligned}$$

Gesucht ist nun die entsprechende Approximation an die Generierungsfunktion der Gradverteilung im Suchbaum nach Ablauf der Breitensuche

$$g^{obs}(z) = \sum_i a_i^{obs} z^i.$$

Die Koeffizienten dieses Polynoms entsprechen dabei den bekannten Werten  $a_{m+1}^{obs}$  nach Gleichung (3.18)

$$\begin{aligned}
 g^{obs}(z) &= \sum_m a_{m+1}^{obs} \cdot z^{m+1} \\
 &= \sum_m \sum_{i>m} a_i \cdot z^{m+1} \left[ \int_0^1 it^{i-1} \binom{i-1}{m} p_{vis}(t)^m (1-p_{vis}(t))^{i-m-1} dt \right] \\
 &= \sum_i \sum_{m<i} a_i \cdot z^{m+1} \left[ \int_0^1 it^{i-1} \binom{i-1}{m} p_{vis}(t)^m (1-p_{vis}(t))^{i-m-1} dt \right].
 \end{aligned}$$

Durch das Aufteilen von  $z^{m+1}$  und Umsortieren von Summe und Integral erhält man

$$\begin{aligned}
 g^{obs}(z) &= z \sum_i \sum_{m<i} a_i \left[ \int_0^1 it^{i-1} \binom{i-1}{m} (z \cdot p_{vis}(t))^m (1-p_{vis}(t))^{i-m-1} dt \right] \\
 &= z \sum_i a_i \left[ \int_0^1 it^{i-1} \sum_{m<i} \binom{i-1}{m} (z \cdot p_{vis}(t))^m (1-p_{vis}(t))^{i-m-1} dt \right]
 \end{aligned}$$

und nach dem binomischen Lehrsatz ist dies gleich

$$\begin{aligned}
 g^{obs}(z) &= z \sum_i a_i \int_0^1 it^{i-1} [z \cdot p_{vis}(t) + (1-p_{vis}(t))]^{i-1} dt \\
 &= z \sum_i a_i \int_0^1 it^{i-1} (1 - (1-z) \cdot p_{vis}(t))^{i-1} dt.
 \end{aligned}$$

Mit Anwendung der Definition von  $g(z)$  und Gleichung (5.1) resultiert daher

$$\begin{aligned} g^{obs}(z) &= z \int_0^1 \sum_i a_i i \cdot [t (1 - (1 - z) \cdot p_{vis}(t))]^{i-1} dt \\ &= z \int_0^1 g' [t (1 - (1 - z) \cdot p_{vis}(t))] dt \\ &= z \int_0^1 g' \left[ t - \frac{1-z}{g'(1)} \cdot g' \left( \frac{g'(t)}{g'(1)} \right) \right] dt, \end{aligned}$$

weshalb dank der Eindeutigkeit eines Polynoms bezüglich seiner Koeffizienten der Beweis vollendet ist. □

## 5.1 Ausblick

Ein Ergebnis des Satzes 5.1 ist der Zusammenhang zwischen der Gradverteilung des ursprünglichen Graphen und der des ermittelten Suchbaums. Mit Kenntnis der Gradverteilung des Ausgangsgraphen lässt sich - bis auf Abweichungen der Größenordnung  $o(n)$  - das Ergebnis der Breitensuche vorherbestimmen. In [1, Kapitel 6] wenden die Autoren den Satz auf Graphen mit identischen ( $\delta$ -regulär) und solche mit poissonverteilten Knotengraden an, dessen Aussagen an dieser Stelle kurz zusammengefasst werden sollen.

Im Fall regulärer Graphen besitzen alle Knoten den Grad  $\delta$ . Die zugehörige Generierungsfunktion vereinfacht sich zu  $g(z) = z^\delta$ , da für alle  $a_k$  mit  $k \neq \delta$  gilt  $a_k = 0$ . Es ergibt sich als asymptotisches Verhalten für alle  $a_{m+1}^{obs}$  mit  $m \geq 2$

$$\frac{m^{-1}}{\delta^{1/(\delta-2)}(\delta-2)} < a_{m+1}^{obs} < \frac{m^{-1+1/(\delta-2)}}{\delta-2}.$$

Für beliebige, aber feste  $\delta$  erhält man durch die Breitensuche immer eine power-law Verteilung, deren Exponent  $\alpha$  für  $\delta \rightarrow \infty$  gegen Eins strebt.

Eine Poissonverteilung mit der Generierungsfunktion  $g(z) = e^{-\delta(1-z)}$  führt im Grenzwert für große  $\delta$  und  $m < \delta - \delta^\kappa$  mit  $\kappa > 1/2$  zu

$$a_{m+1}^{obs} = (1 - o(1)) \frac{\Gamma(m)}{\delta m!} \sim \frac{1}{\delta m},$$

wobei  $\Gamma(m)$  die Gammafunktion – eine Erweiterung der Fakultätsfunktion auf positive reelle Zahlen – bezeichnet. Bis zu einem Knotengrad von  $m \sim \delta$  ergibt sich somit auch für Poissonverteilungen als Resultat der im Suchbaum auftretenden Grade eine power-law Verteilung.

Die Beispiele zeigen deutlich, dass die Gradverteilung des Suchbaums nicht direkt der ursprünglichen Verteilung entspricht. Es stellt sich daher die Frage, ob aus der ermittelten Gradverteilung ein Rückschluss auf die originale Verteilung möglich ist. Aufgrund der hohen Komplexität des Zusammenhangs, insbesondere der in Satz 5.1 angegebenen Gleichungen,

bleibt dieses Problem vorerst ungelöst. Es ist bisher nicht einmal geklärt, ob eine solche Umkehrung der Gleichung überhaupt existiert.

Dank des anfangs beschriebenen und in [6, Kapitel 2] ausführlicher erläuterten Zusammenhangs zwischen Breitensuchbaum eines Graphen und der durch traceroute gewonnenen Informationen in Netzwerken lassen sich Rückschlüsse über die Qualität des traceroute samplings ziehen. Dessen Ergebnisse sind wie die der Breitensuche auch mit verfahrensbedingten Fehlern behaftet, die sich nicht vermeiden lassen. Desweiteren ist nicht bekannt, ob und gegebenenfalls wie diese Fehler aus den Ergebnissen herausgerechnet werden können. So gewonnene Informationen über die Eigenschaften der untersuchten Netzwerke (z.B. des Internetgraphen) und dabei insbesondere die erhaltene Gradverteilung müssen deshalb mit gebührender Skepsis betrachtet werden. In [6] werden diese Schlüsse durch vom Autor beschriebene und durchgeführte Simulationen bestätigt. Unabhängig von der ursprünglichen Gradverteilung und der Knotenanzahl erhält man nur in den wenigsten Fällen eine identische Verteilung im Breitensuchbaum.

Die für den Internetgraphen so oft beobachtete power-law Verteilung kann durch den Explorationsalgorithmus traceroute sampling verursacht sein. Selbst wenn der Internetgraph wirklich einer solchen Verteilung genügt, so ist anzunehmen, dass der ermittelte Exponent  $\alpha$  nicht mit dem realen übereinstimmen muss. Für die Zukunft wäre es interessant zu untersuchen, ob die angesprochene Umkehrung der Gleichung aus Satz 5.1 existiert und bestimmt werden kann. Damit wäre die Möglichkeit gegeben, die erwiesenermaßen fehlerhaften Ergebnisse des Explorationsverfahrens zu korrigieren und der Lösung der ursprünglichen Aufgabe – die Ermittlung von charakteristischen Informationen über nur teilweise erforschbare Graphen – einen Schritt näher zu kommen.

# Literaturverzeichnis

- [1] Dimitris Achlioptas, Aaron Clauset, David Kempe und Cristopher Moore: *On the Bias of Traceroute Sampling*, 2005.
- [2] Benoit Donnet, Timur Friedman und Mark Crovella: *Improved Algorithms for Network Topology Discovery*, 2005.
- [3] Alex Fabrikant, Elias Koutsoupias und Christos H. Papadimitriou: *Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet*, 2002.
- [4] C. Faloutsos, M. Faloutsos und P. Faloutsos: *On power-law relationships of the internet topology*, 1999.
- [5] Ramesh Govindan und Hongsuda Tangmunarunkit: *Heuristics for Internet Map Discovery*, 2000.
- [6] Markus John: *Implementierung und Evaluierung eines Algorithmus zur Exploration sehr großer Graphen*, 2005.
- [7] Anukool Lakhina, John W. Byers, Mark Crovella und Peng Xie: *Sampling Biases in IP Topology Measurements*, 2003.
- [8] Jean Jacques Pansiot und Dominique Grad: *On Routes and Multicast Trees in the Internet*, 1998.
- [9] H. S. Wilf: *generatingfunctionology*, 1994.
- [10] N. C. Wormald: *Models of random regular graphs*, 1999.



# Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig angefertigt, nicht anderweitig zu Prüfungszwecken vorgelegt und keine anderen als die angegebenen Hilfsmittel verwendet habe. Sämtliche wissentlich verwendete Textausschnitte, Zitate oder Inhalte anderer Verfasser wurden ausdrücklich als solche gekennzeichnet.

Chemnitz, den 22. August 2006

---

Markus John