

FACHBEREICH INFORMATIK

BACHELORARBEIT

ANALYSE DES NORMENNETZWERKS DER INTERNET REQUESTS FOR COMMENTS

Autor:

Maciej Wiśnieszczak

Betreuer:

Prof. Dr. Robert Tolksdorf

1 Einleitung

Im Rahmen dieser Arbeit werden die Beziehungen zwischen den *Request for Comments* - Dokumenten (RFC) extrahiert, als ein Netzwerk dargestellt und dem entsprechend analysiert.

Die *Requests for Comments* sind von der *Internet Engineering Task Force* (IETF) publizierte, nummerierte Memoranda, die unterschiedliche Internet-bezogene Standards, Standardvorschläge, Verhalten und Forschung beschreiben. Die Erscheinung eines RFC-Dokuments bedeutet nicht, dass dessen Inhalt gleichzeitig als Standard anerkannt wird, hierzu ist ein weiterer Prozess notwendig. In seinem Rahmen muss ein RFC-Dokument unterschiedliche Reifegrade (Proposed Standard, Draft Standard und Internet Standard) erreichen, die von einer IETF-Untergruppe, *Internet Engineering Steering Group* (IESG) zugewiesen werden. Ein solcher Status ist u.a. von der allgemeinen Akzeptanz, der Anzahl von existierenden Implementierungen und von dem operationellen Erfolg abhängig [2]. Diese Dokumente spielen eine enorm wichtige Rolle bei der Entwicklung vom globalen Netzwerk, weil sie die grundlegenden Protokolle, wie TCP und IP und Dienste, wie DNS oder HTTP definieren.

Das erste RFC-Dokument erschien am 7. April 1969, wurde von Steve Crocker [3] geschrieben und handelt von der zur Netzwerkkommunikation notwendigen Software. Die Dokumente sind durch eine feste, formale Struktur charakterisiert, die eine einfache, automatisierte Verarbeitung des Inhalts ermöglicht. Außerdem sind die veröffentlichten RFCs fest, d.h. der Text darf unter keinen Umständen verändert oder korrigiert werden, eine Aktualisierung darf nur mittels eines neuen Dokuments erfolgen. Diese zwei Eigenschaften ermöglichen sowohl eine Untersuchung des aus den Beziehungen resultierenden Normennetzwerkes, als auch des Netzwerkwachstumsprozesses.

Die RFC-Dokumente sind keine autarken Objekte, sie beziehen sich aufeinander in unterschiedlicher Art und Weise. Neue Memoranda können die alten entweder komplett ersetzen oder ergänzen, es existieren auch Vorkommen von anderen Dokumenten direkt im Text eines Memorandums. Folgende Beziehungen werden aus den Texten extrahiert und verarbeitet:

- Referenzen/Zitierungen (Vorkommen von den anderen RFCs innerhalb des Dokuments)
- Ersetzungen (*Obsoletes/Obsoleted by*)
- Aktualisierungen (*Updates/Updated by*)

Merkwürdig ist hier, dass die Bezüge jeweils gerichtet sind, obwohl bei den Ersetzungen und Aktualisierungen zwei komplementäre, gerichtete Kanten erwartet werden. Die Dokumente werden als Netzwerkknoten interpretiert und zusätzlich wird ihr Status abgelesen, der folgende Werte annehmen kann:

- *Standards Track* (Im Rahmen des Standardisierungsprozesses)
 - *Proposed Standard*
 - *Draft Standard*
 - *Internet Standard*
- *Best Current Practice* (BCP)
- *Informational*
- *Experimental*
- *Historic*
- Unbekannt (*Unknown*)

Dieser gibt an, von welcher Bedeutung ein Dokument ist. Die wichtigsten Spezifikationen gehören zur Gruppe von Standards, wobei sich die *Best Current Practice*-Memoranda auf der Beschreibung von einer bestimmten Vorgehensweise oder Richtlinie konzentrieren. *Informational*-RFCs beinhalten ein breites Spektrum von Dokumenten, von einer Liste der zugewiesenen Portnummern bis zu Aprilscherze. Als experimentell werden die Spezifikationen bezeichnet, die entweder wirklich experimentelle Ansätze beschreiben oder aus prozedurellen Gründen nicht als Standard anerkannt werden.

Jedes Dokument besitzt auch ein Veröffentlichungsdatum, welches, zusammen mit der zeitlichen Invarianz (der Text darf nie verändert werden), erlaubt die Erzeugung von einem RFC-Netzwerk, das zu einem bestimmten Zeitpunkt galt. Dadurch kann das Wachstum beobachtet und analysiert werden.

2 Grundlagen und Aufbau der Arbeit

Ein Netzwerk ist ein System, das sich mathematisch als ein Graph modellieren lässt. Es verfügt über eine Menge von Objekten (Knoten) und Beziehungen zwischen denen (Kanten). Kanten können entweder eine gerichtete oder ungerichtete Relation widerspiegeln. Diese Art der Darstellung ermöglicht eine Anwendung von den unterschiedlichen, graphentheoretischen Algorithmen und Metriken auf das Netzwerk. Dadurch kann man verschiedene Eigenschaften und Charakteristiken von einem solchen Netzwerk untersuchen. Das Anwendungsgebiet ist sehr breit, da viele Prozesse und Zustände in einer Netzwerkform dargestellt werden können. Zu den Beispielen gehören nicht nur Sozialnetzwerke oder World Wide Web, sondern auch der Blutkreislauf oder neuronale Netzwerke.

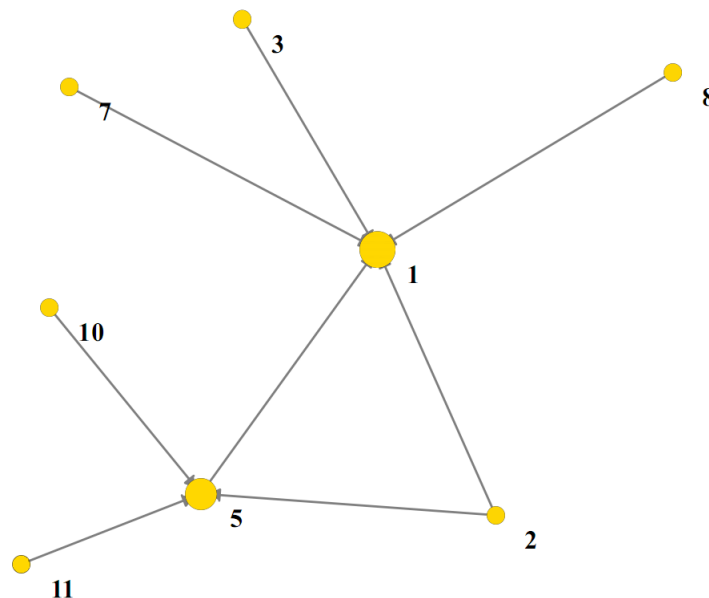


Abbildung 1: Kleines Netzwerk mit gerichteten Kanten

Die Entstehung der Graphentheorie ist mit der im Jahre 1736 von Leonhard Euler vorgeschlagener Lösung für das Königsberger Brückenproblem verbunden. Er hat mittels eines Graphen nachgewiesen, dass es keinen Rundgang (später: Eulerkreis) gibt, der jede Brücke genau einmal überquert. Später wurde die Graphentheorie z.B. auch in der Chemie eingesetzt. Ein weiteres, wichtiges Datum ist April 1933, als Jacob Moreno, ein US-Amerikanischer Psychologe rumänischer Herkunft das erste Soziogramm, das die studentischen Beziehungen widerspiegelte, vorgestellt hat. Dieses Ereignis zählt als Anfang der Forschung über die sozialen Netzwerke. Sie betrachten verschiedene Relationen, nicht

nur Freundschaften, sondern auch geschäftliche oder internationale Beziehungen. Aus dem Grund, dass Soziologie bereits in diesen Zeiten existierte und relativ gut entwickelt war, wurden die Sozialnetzwerke sehr häufig untersucht. Bei der Analyse wurden viele Metriken erfunden, die relevant für die allgemeine Netzwerkwissenschaft sind sowie Eigenschaften festgelegt.

Zu den wichtigsten Erkenntnissen gehört das Kleine-Welt-Phänomen. Der Psychologe Stanley Milgram hat untersucht, wieviele Zwischenschritte ein Brief braucht, um den Empfänger zu erreichen. Die Briefe wurden natürlich nicht direkt gesendet, sondern von zufällig gewählten Personen an ihre Bekannten weitergegeben, die eine höhere Wahrscheinlichkeit hatten, den Empfänger persönlich zu kennen [4]. Von den Sendungen, die das Ziel erreicht haben, ergab sich eine Entfernung von nur etwa 6 Schritten, obwohl die Sender und der Empfänger wurden so gewählt, dass die geographische Entfernung maximalisiert wurde. Diese Eigenschaft ist nicht nur für soziale Netzwerke charakteristisch, sondern auch kann sie z.B. in einem Netzwerk von miteinander agierenden Proteinen beobachtet werden. [5]

Eine weitere, wichtige Eigenschaft, die künstlich und natürlich erzeugte Netzwerke abgrenzt ist die Verteilung des Knotengrades. Im Jahre 1965 zeigte Derek de Solla Price, dass die Gradverteilung des Zitationsnetzwerkes von wissenschaftlichen Arbeiten ein Potenzgesetz folgt[6]. Das ist charakteristisch für natürlich wachsende Graphen, weil bei einer zufälligen Erzeugung hat jeder Knoten gleiche Chance, mit einem anderen verbunden zu sein. Ein Netzwerk, das diese Eigenschaft besitzt, wird skalenfrei genannt.

Die Untersuchung von großen sozialen Netzwerken wurde mit der Entwicklung von Rechnern deutlich vereinfacht. Händische Datenerfassung und Analyse ist bis einige Hundert Knoten beschränkt. Alles was darüber hinaus geht muss auf der Basis von einer statistischen Auswertung erfolgen. Optische Untersuchung des Netzwerks ist in diesem Fall nur dann möglich, wenn die Knoten Gruppen bilden, die viele gemeinsame Kanten besitzen. Zu den wichtigen Metriken gehört u.a. die bereits genannte Knotengradverteilung und durchschnittliche Länge des Weges oder der Clusterkoeffizient, der die Vernetzung von Knoten angibt.

Die RFC-Dokumente können auch als eine Art vom sozialen Netzwerk betrachtet werden. Sie beziehen sich lediglich auf ein bestimmtes Gebiet des Internets und agieren miteinander. Eine Referenz im Text kann als eine Bekanntschaft interpretiert werden, wobei zu beachten ist, dass im Gegensatz zur echten Welt, diese Relation gerichtet ist. Die Aktualisierungen und Ersetzungen von den Dokumenten liegen näher zur Mitarbeiterrotation in einer Abteilung als zur Eltern-Kind-Beziehung, da ein Memorandum gleichzeitig mehrere ältere Vorgänger ersetzen kann und nicht alle Dokumente haben einen Elternteil. Eine solche Aufspaltung führt zur Entstehung von zwei unterschiedlichen Netzwerkebenen mit verschiedenen erwarteten Strukturen. Die erste (Referenzebene) besitzt nur gerichtete Kanten und soll eine sternartige Organisation aufweisen, wobei die zweite (Aktualitätsebene) einen mehr linearen, bzw. baumartigen Verlauf hat und jede Beziehung, wie "aktualisiert" oder "ersetzt" auch eine entsprechende, entgegengerichtete Kante "aktualisiert von" oder "ersetzt von" haben soll. Um die Ergebnisse nicht zu verfälschen, wird die Auswertung immer auf diesen zwei Netzwerkebenen separat durchgeführt.

In dem ersten Teil der Arbeit werden die beiden Netzwerkebenen zum aktuellen Zeitpunkt (30.07.2010) analysiert. Die resultierenden Graphen werden mithilfe des Clusterkoeffizienten auf Clustering getestet und miteinander verglichen, um die Vernetzung von Knoten zu untersuchen. Zudem wird die mittlere Länge eines Weges zwischen allen Knoten errechnet. Dadurch wird zusätzlich überprüft, ob das RFC-Normennetzwerk dem für menschlich erzeugte Netzwerke charakteristischen Small-World-Network-Prinzip folgt.

Die Gradverteilung (sowohl In-Grad, als auch Aus-Grad) von den jeweiligen Dokumenten wird grundsätzlich für zwei Zwecke untersucht. Erstens, wenn sie dem Potenzgesetz folgt, ist es eine für menschlich erzeugte Netzwerke typische Eigenschaft und zweitens, sie ermöglicht die Bestimmung von den wichtigsten, entscheidenden RFCs, die auch explizit genannt werden. Zusätzlich, die Gradverteilung macht auch eine Aussage über die Robustheit des Netzwerks.

Ferner wird Die Robustheit und Ausfallsicherheit des Graphen explizit kontrolliert. Obwohl ein Normennetzwerk nicht z.B. mit einem Stromnetzwerk vergleichbar ist, kann eine Ersetzung vom bestimmten, oft referenzierten Dokument mit einem Ausfall gleichgesetzt werden, da alle weitere, auf ihm beruhende Standards auch von der Ersetzung betroffen werden und ggf. aktualisiert werden sollen. Mixing Patterns sind die Muster,

nach denen die Knoten miteinander verbunden sind. Diese werden auf zwei Arten untersucht. In dem ersten Ansatz wird überprüft, ob die unterschiedlichen Knotentypen sich lieber mit gleichartigen Dokumenten "mischen" als mit den anderen. Zusätzlich wird versucht, eine Korrelation zwischen dem Knotengrad und dem Grad von Nachbarn zu finden. Das Ergebnis wird mit der erwarteten Disassortativität von Informationsnetzwerken verglichen.

In dem zweiten Teil der Arbeit wird die zeitliche Entwicklung des Netzwerkes untersucht. Für jeweilige Zeitpunkte werden die charakteristischen Werte berechnet, miteinander verglichen und analysiert. Dadurch wird überprüft, ob das Wachstum den Prinzipien des für Zitationsnetzwerke entwickelten Modell vom Price entspricht. Zusätzlich wird die Entstehung von lokalen Zentren nach eventuellen Regelmäßigkeiten beobachtet, indem die Zunahme von Referenzen mit dem bereits existierenden Ingrad verglichen wird.

Zudem wird die Aktualität von Dokumenten erforscht und qualitativ ausgewertet, d.h. es werden die Dokumente durchgeleuchtet, die sich auf veraltete Quellen beziehen und anschließend wird ihre Bedeutung für das gesamte Normennetzwerk analysiert. Es wird ein daraus folgender Qualitätsfaktor für das gesamte Netzwerk errechnet und für bestimmte Zeitpunkte angegeben.

Die Auswertung von den in den obigen Schritten enthaltenen Metriken gewährt einen tiefen Einblick in die Eigenschaften des gesamten Normennetzwerks. Durch die Feststellung von bestimmten Eigenschaften wird entschieden, ob das Netz die für natürlich wachsende Netzwerke charakteristische Eigenschaften besitzt und es über eine bestimmte Struktur verfügt.

3 Beschreibung von verwendeten Werkzeugen

3.1 Extrahierung von Daten

Um die Beziehungen zwischen den RFC-Dokumenten zu extrahieren wurde ein speziell zu diesem Zweck entwickeltes Werkzeug eingesetzt. Diese Java-Anwendung verarbeitet das RFC-Verzeichnis im XML-Format [1], um die Bezüge der Aktualitätsebene zu extrahieren. Diese Datei enthält die Informationen über die Aktualisierungen und Ersetzungen von den

jeweiligen RFCs. Zusätzlich stehen die grundlegenden Informationen zu jedem Text, wie z.B. die RFC-Nummer, die Autoren, der Titel und das Erscheinungsdatum zur Verfügung. Das Datum der Veröffentlichung spielt hier eine besonders wichtige Rolle, weil sie es ermöglicht, die Netzwerkstruktur zu einem bestimmten Zeitpunkt darzustellen.

Listing 1: Eintrag aus dem RFC-Index

```
<rfc-entry>
  <doc-id>RFC0329</doc-id>
  <title>ARPA Network Mailing Lists</title>
4  <author>
    <name>Network Information Center. Stanford Research Institute</name>
  </author>
  <date>
8    <month>May</month>
    <year>1972</year>
  </date>
  <format>
12    <file-format>ASCII</file-format>
    <char-count>23861</char-count>
    <page-count>13</page-count>
  </format>
  <obsoletes>
16    <doc-id>RFC0303</doc-id>
  </obsoletes>
  <obsoleted-by>
20    <doc-id>RFC0363</doc-id>
  </obsoleted-by>
  <current-status>UNKNOWN</current-status>
  <publication-status>UNKNOWN</publication-status>
24  <stream>Legacy</stream>
</rfc-entry>
```

Diese Informationen spiegeln die Aktualitätsebene wider. Um die Referenzen aus einem Dokument zu extrahieren, muss der komplette Text durchgearbeitet werden. Trotz der starken Strukturierung von RFC-Dateien, ergab sich nach einer Untersuchung, dass die Referenzen im Text unterschiedliche Formen annehmen. Zusätzlich sind einige so gebrochen, dass das Wort "RFC" und die dazugehörige Nummer sich in unterschiedlichen Zeilen befinden. Um alle solche Fälle abzufangen, war der Einsatz von folgendem regulären Ausdruck notwendig (Listing 2).

Trotzdem werden einige Referenzen nicht extrahiert, diese kann man in zwei Kategorien unterteilen. Die erste beinhaltet die Dokumente, die keine Textform besitzen. Das betrifft nur einige RFCs mit sehr niedrigen Nummern, das sind RFC8, RFC9 und RFC51.

Da diese Dokumente bereits veraltet sind, ist das Fehlen von diesen Referenzen unerheblich. Die zweite Kategorie sind z.B. die *Internet Official Protocol Standards*, die als Verzeichnis für als Standard anerkannten Dokumenten dienen. Dortige Darstellung ist durch eine tabellarische Form charakterisiert, wo die Nummern ohne das Wort "RFC" vorkommen, was die Extrahierung einschränkt. Das ist von relativ wenig Bedeutung, weil diese Beziehungen keinen Einfluss auf die Wichtigkeit von Netzwerkknoten haben, da es sich lediglich um einen Index handelt, ohne beitragenden Inhalt.

Listing 2: Regulärer Ausdruck zum Herausfinden von Referenzen

```
(\\S*)RFC([\\ -]*)([0-9]+)(\\S*)
```

```
Akzeptiert u.a.: RFC1234,
```

```
4   [RFC1234],
    RFC 1234,
    RFC-1234,
    RFC
8   1234
```

Das Ergebnis wird zum Zweck der weiteren Bearbeitung als eine GEXF-Datei (*Graph Exchange XML Format*) gespeichert [7]. Dieses Format wurde aus zwei Gründen gewählt, Erstens, ist es die native Darstellung für das Program *Gephi* und zweitens, es ermöglicht eine Speicherung der Lebensdauer von einem Knoten oder einer Kante. Diese Fähigkeit ist kritisch für die vorgesehene Analyse von der zeitlichen Entwicklung des RFC-Normennetzwerks und fehlt bei dem anderen gängigen Datenformat, GraphML. Die resultierende Datei hat folgende Gestalt (Beispiel für die Referenzebene):

Listing 3: Ausgabeformat für die weitere Bearbeitung

```
<?xml version="1.0" encoding="UTF-8"?><gexf xmlns="http://www.gexf.net/1.1draft" version="1.1">
<graph edgedefault="directed" mode="dynamic" start="1969-01-01" end="2011-01-01">
  <nodes>
4    ...
    <node id="RFC3549" label="RFC3549" start="2003-07-01" end="2011-01-01">
      <edges>
        <edge id="3549e0" source="RFC3549" target="RFC2748"/>
8        <edge id="3549e1" source="RFC3549" target="RFC2475"/>
        <edge id="3549e2" source="RFC3549" target="RFC1157"/>
        <edge id="3549e3" source="RFC3549" target="RFC3036"/>
        ...
12      </edges>
    </node>
    ...
16  </nodes>
</graph>
```

3.2 Gephi

Modellierung und Auswertung des Graphes wurde mithilfe des Werkzeuges *Gephi* durchgeführt. Diese Open-Source Anwendung verwendet die auf Java basierende Entwicklungsumgebung NetBeans als Basis und verfügt über mehrere Module, die für die richtige Auswertung notwendig sind.

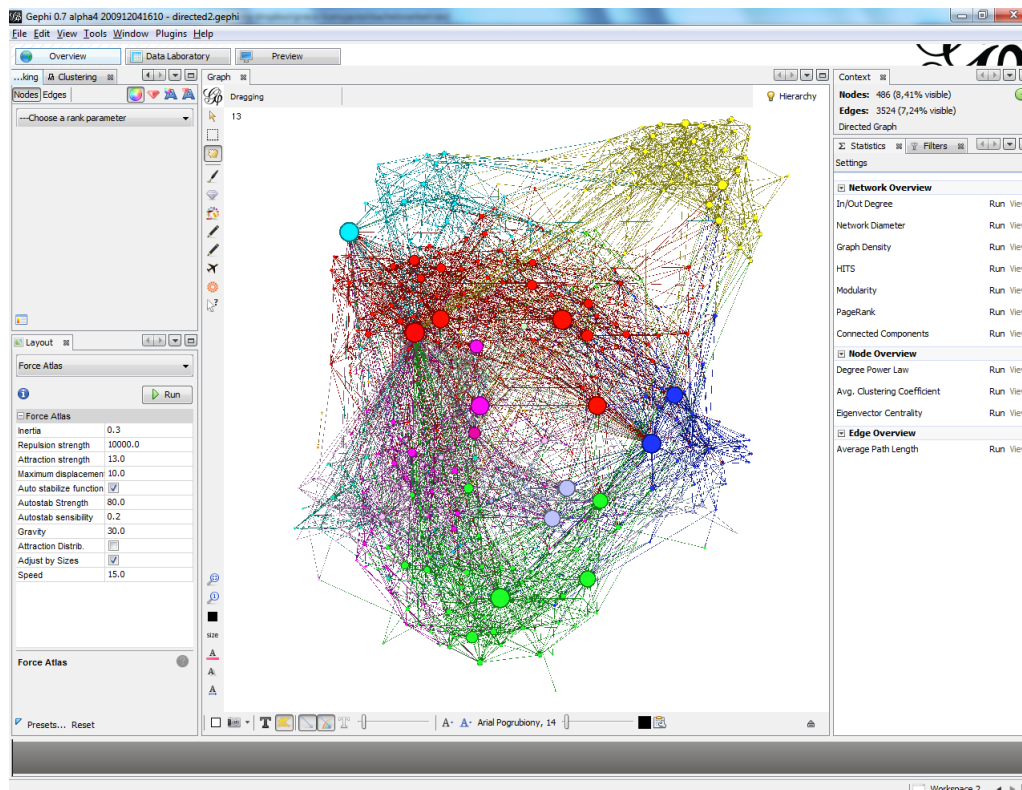


Abbildung 2: Oberfläche von Gephi

Visualisierung und die dazugehörigen Knotengruppierungsalgorithmen erlauben eine visuelle Analyse vom resultierenden Graphen. Da das Ergebnis 5778 Knoten und fast 50000 Kanten hat, müssen die Objekte entsprechend gruppiert werden, um die aus den Beziehungen resultierende Strukturen zu erkennen. Fünf eingebaute Algorithmen (*Force Atlas*, *Yifan Hu*, *Fruchterman Reingold*, *Yifan Hu Multilevel* und eine zufällige Verteilung) sorgen dafür, dass für verschiedene Graphenarten eine sinnvolle Darstellung gefunden werden kann. Die Algorithmen lassen mehrere Parameter beeinflussen, um einen Feinabgleich des Ergebnisses zu erlauben. Die Größe von Knoten kann anhand des Grades angepasst werden, um die Darstellung um diesen Faktor zu ergänzen.

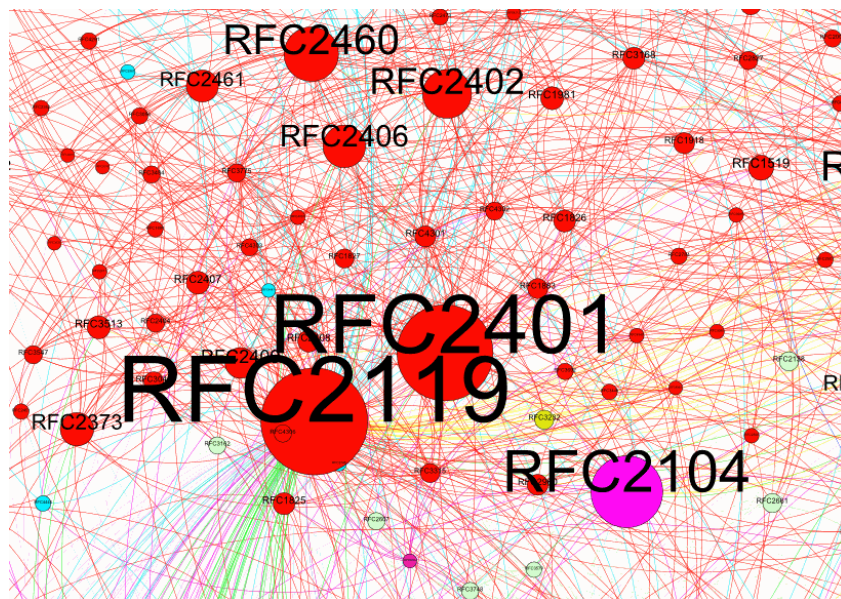


Abbildung 3: Ausschnitt aus der Graphenvisualisierung

Metriken ermöglichen eine Berechnung von typischen Netzwerkstatistiken, wie die Verteilung des Knotengrades, der Clusterkoeffizient oder die mittlere Weglänge. Es kann auch auf die Skalenfreiheit getestet werden und dies sogar mit grafischer Darstellung des Ergebnisses. Die Algorithmen können den Graphen wahlweise als einen gerichteten oder ungerichteten betrachten. Es fehlt leider die Exportfunktionalität, um z.B. die Knotenverteilung mit einer anderen Anwendung zu bearbeiten, die Diagramme lassen sich auch nicht speichern.

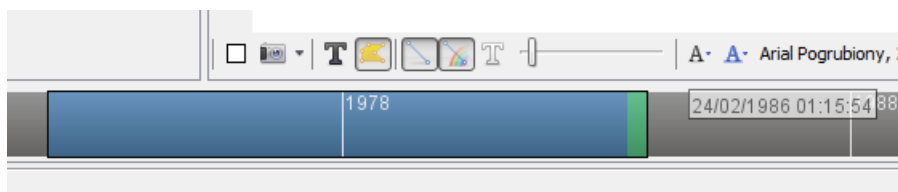


Abbildung 4: Auswahl der Zeitspanne

Die eingebauten **Filter** steuern die Darstellung anhand von verschiedenen Parametern, wie z.B. der Knotengrad oder auch die übergebene oder berechnete Eigenschaften. Alle Objekte, die in den spezifizierten Bereich nicht passen, werden ausgeblendet und werden

nicht z.B. bei den Platzierungsalgorithmen betrachtet. Das ist eine wichtige Funktionalität, insbesondere bei Netzwerken, die aus mehreren Tausend Knoten bestehen. Zusätzlich, kann die Filterung, falls die übergebene Daten auch eine Lebenszeit enthalten, in Abhängigkeit von diesem Parameter erfolgen. Zusätzlich können die Filter geschachtelt werden, so dass mehrere Bedingungen gleichzeitig eingehalten werden müssen.

3.3 Network Workbench Tool

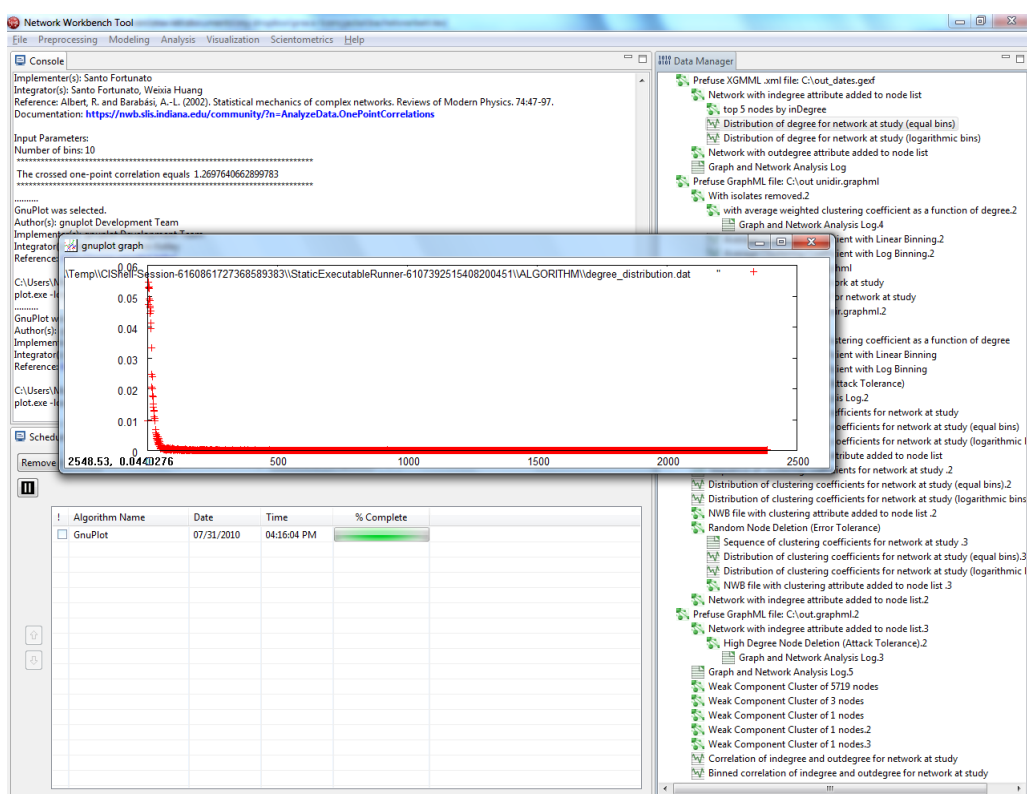


Abbildung 5: Oberfläche vom Network Workbench Tool

Das *Network Workbench Tool* ist eine an der University Indiana entwickelte Werkzeugensammlung, deren Hauptaufgabe eine numerische Analyse von Netzwerken ist. Diese Eclipse-basierte Anwendung akzeptiert als Eingabeformat unterschiedliche Dateiformate, darunter GraphML, Pajek NET, CSV und XGMML. Besonders das erste Format ist hier wichtig, da es einen reibungslosen Informationsaustausch mit dem Programm *Gephi* gewährleistet, das über eingebaute Konversionsmechanismen verfügt.

Die Funktionalitäten von NWB sind in fünf Gruppen geteilt:

- *Preprocessing*
- *Modeling*
- *Analysis*
- *Visualisation*
- *Scientometrics*

Unter "Vorverarbeitung" befinden sich die Funktionen, die die Struktur des Graphen beeinflussen. Es ist u.a. das Löschen von isolierten Knoten, Entfernung von Knoten außerhalb vom bestimmten Gradbereich, als auch Entfernung von zufälligen und am stärksten vernetzten Knoten möglich. Mit diesen Funktionen kann man das Netzwerk auf Robustheit testen.

Die als "Modellierung" bezeichneten Werkzeuge dienen dazu, neue Graphen mit bestimmten Eigenschaften zu generieren. Damit wurden die Zufallgraphen, mit denen das Normennetzwerk verglichen wurde, erzeugt. In der als "Analyse" gruppierten Funktionsammlung findet man u.a. die Berechnung von der Gradverteilung, unterschiedlichen Koeffizienten und Statistiken. Eine Übersicht über die Netzwerkkennzahlen, wie Anzahl von Knoten und Kanten, durchschnittlicher Knotengrad und die Anzahl von isolierten Knoten gewährleistet das *Network Analysis Toolkit*. Weitere Metriken sind in Kategorien unterteilt, die bezeichnen, auf welche Art von Graphen (gerichtet und gewichtet) sie sich anwenden lassen.

Das Ergebnis kann unterschiedliche Formen annehmen. Das sind nicht nur weitere Graphen, sondern auch einzelne Zahlen oder die ganzen Tabellen (z.B. bei der Gradverteilung). Diese können mit mehreren Visualisierungswerkzeugen beobachtet werden. Für Netzwerke stehen die auf dem *Prefuse Visualisation Toolkit* basierende Module und das *Graph Exploration System* GUESS. Diese sind relativ schlecht für die Visualisierung von großen Netzwerken geeignet, hier hat sich Gephi viel besser bewährt. Zur Darstellung von numerischen Daten wird das Open-Source Programm *GnuPlot* verwendet. Die Verarbeitung von bibliografischen Daten erleichtern die Funktionen aus der *Scientometrics*-Gruppe, sie fanden jedoch bei dem RFC-Netzwerk keine Anwendung.

4 Auswertung des RFC-Normennetzwerks

In diesem Teil der Arbeit wird das RFC-Normennetzwerk im gegenwärtigen Zustand ausgewertet, d.h. das letzte berücksichtigte RFC-Dokument hat die Nummer 5942. Die Berechnung und Auswertung des Graphen wird auf zwei bereits erwähnten Ebenen separat durchgeführt. Das ist die Referenzebene, die ausschließlich aus den im Text der RFCs enthaltenen Zitierungen besteht und die Aktualitätsebene, die lediglich die Aktualisierungen und Ersetzungen von den jeweiligen Dokumenten enthält. Das Ergebnis wird mit einem zufällig generierten Graphen mit gleicher Anzahl von Knoten und Kanten verglichen, um zu untersuchen, welche Charakteristiken das Normennetzwerk von ihm unterscheiden.

4.1 Small-World-Netzwerk

Der Kleine-Welt-Effekt besagt, dass in meisten, nicht zufällig erzeugten Netzwerken die mittlere Anzahl von Schritten zwischen den jeweiligen Knoten relativ niedrig ist. Er wurde in den sechszigen Jahren von Stanley Milgram [4] beschrieben, indem er ein Experiment durchgeführt hat, dessen Ziel war die Errechnung der mittleren Weglänge in sozialen Netzwerken.

1. Zuerst wurden zufällige Personen aus dem Bundesstaat Kansas ausgewählt, die als Sender dienen sollen. Sie haben ein Informationspaket erhalten, wo das Ziel des Experimentes beschrieben wurde und ein Empfänger in weit entferntem Boston angegeben wurde.
2. Jeder Teilnehmer soll den Brief entweder selbst an den Empfänger weitergeben (falls er ihm persönlich bekannt war), oder ihn einer Person überreichen, die eine höhere Wahrscheinlichkeit hatte, den Empfänger zu kennen.
3. Jede Übergabe soll auf der beigelegten Liste notiert werden. Zusätzlich wurde eine Antwortkarte an die Wissenschaftler gesandt, um die Verfolgung des Pakets zu erleichtern und um den eventuellen Unterbrechungspunkt zu identifizieren.

Von allen 296 Briefen, 64 wurden erfolgreich zugeteilt und die mittlere Weglänge betrug etwa 6 Schritte. [9] Das führte zur Schlussfolgerung, dass alle Einwohner von den Vereinigten Staaten um 6 Leute voneinander entfernt sind. Diese These wurde aus mehreren Gründen kritisiert, weil die Wissenschaftler die Länge der unterbrochenen Ketten

und Bereitschaft von den Teilnehmer nicht berücksichtigt haben. Diese Faktoren könnten das Endergebnis beeinflussen.

Eine ähnliche Studie wurde später als Zeichen des Respekts für dem bekannten ungarischen Mathematiker Paul Erdős von seinen Bekannten und Freunden durchgeführt. [10] Sie haben den Begriff "Erdős-Zahl" eingeführt. Paul Erdős selbst hat die Erdős-Zahl 0. Ein Wissenschaftler, der mit ihm eine Arbeit geschrieben hat - 1. Ein weiterer Koautor, der mit dem direkt mit Erdős kooperierenden veröffentlichte (aber nicht mit Erdős selbst), hatte die Erdős-Zahl 2 usw.

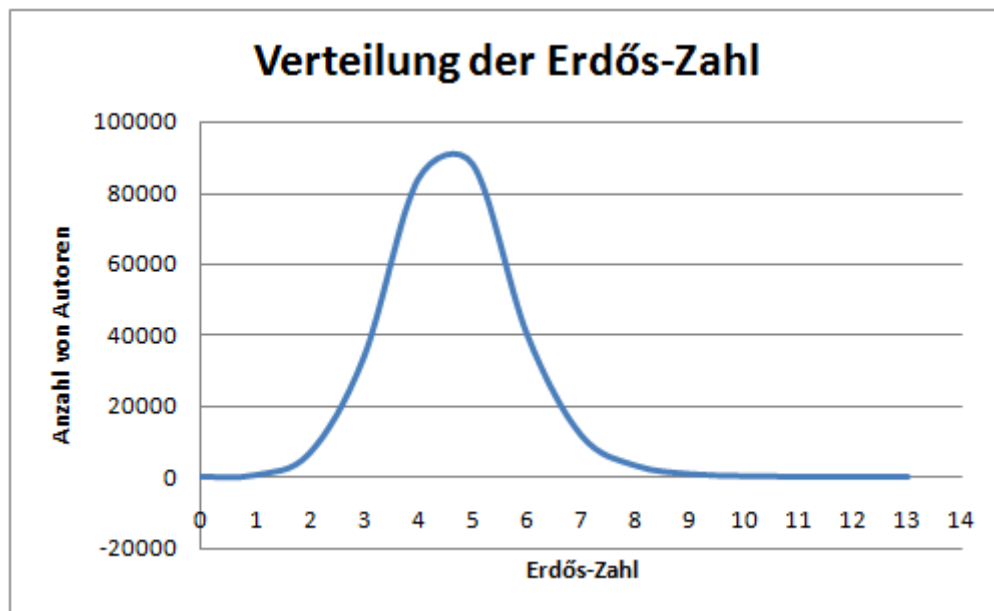


Abbildung 6: Verteilung der Erdős-Zahl

Die bisher erfassten Daten beinhalten etwa 1,9 Millionen Arbeiten von über 401 000 Autoren. Der höchste erreichte Wert der Erdős-Zahl liegt bei 13, was bedeutet, dass diese Person lediglich 13 Schritte vom Erdős in dem Zitationsnetzwerk entfernt ist. Im Durchschnitt beträgt die Weglänge lediglich 4,65.

Zudem existieren ähnliche, auf demselben Prinzip basierende Netzwerke, wie z.B. ein Netz von Schauspielern, die gemeinsam mit Kevin Bacon in einem Film gespielt haben. Insgesamt 1207404 Akteure wurden erfasst, die höchste Bacon-Zahl beträgt 8, die durchschnittliche - 2,981. [11]

4.1.1 Mittlere Länge des Weges

Der durchschnittliche Weg d zwischen allen Knoten in einem Graphen ist folgendermaßen definiert:

$$d = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \quad (1)$$

wobei n die Anzahl von Knoten angibt und d_{ij} der geodesische (d.h. - ohne Berücksichtigung der Eckengewichtung) Abstand zwischen Knoten i und j ist. Falls ein Knoten sich im isolierten Teil des Netzwerks befindet, werden nur die von ihm erreichbaren Knoten berücksichtigt.

Ein Netzwerk zeigt den Small-World-Effect, wenn d in Abhängigkeit von n logarithmisch oder langsamer wächst. [12] Der mit der Netzwerkentwicklung zunehmende Wachstum der Weglänge lässt sich dadurch erklären, dass jeder neu eingefügte Knoten als eine Abkürzung dienen kann, welche die bisher langen Wege zwischen bestimmten Netzwerk-bereichen signifikant reduziert. [13]

Dadurch, dass man die zeitliche Entwicklung des RFC-Normennetzwerks nachvollziehen kann, ist auch eine Untersuchung des Wachstums der mittleren Weglänge möglich, um festzustellen, ob das Netzwerk unter die Small-World-Kategorie fällt.

Referenzebene

Die Referenzebene ist ein gerichteter Graph, da die zitierten Artikel keinen Verweis auf die zitierenden enthalten. Um den Einfluss von dieser Eigenschaft zu untersuchen, jede Berechnung der mittleren Weglänge wurde für zwei Varianten durchgeführt: einmal wurde das Netzwerk als gerichtet und einmal als ungerichtet betrachtet. Zusätzlich wird es mit einem zufällig generierten Graphen verglichen. Um den Wachstumsprozess von diesem Zufallsgraphen zu simulieren, wurde in jedem Schritt eine entsprechende Anzahl von zufälligen Knoten mit NWB entfernt. Bei dem RFC-Netzwerk wurde stattdessen das Veröffentlichungsdatum berücksichtigt. Die Berechnung erfolgte mit *Gephi*.

Die mittlere Weglänge bei der Betrachtung als gerichtetes Netzwerk beträgt 5,215. Wenn die Richtung der Kanten vernachlässigt wird, liegt der Wert bei 3,035. Das Ergebnis bestätigt das Auftreten des Small-World-Effects, da die Entfernung von Knoten niedrig

ist. Die Nichtberücksichtigung von den Richtungen führt zu einer noch kleineren Welt, weil dadurch mehrere mögliche Pfade gleichzeitig zur Verfügung stehen.

Knoten	Gerichtet	Zufallsgraph gerichtet	Ungerichtet	Zufallsgraph gerichtet
5778	5,215	4,695	3,037	3,385
5200	5,257	4,783	3,074	3,463
4692	5,278	4,86	3,104	3,536
4194	5,292	4,949	3,132	3,615
3694	5,304	4,976	3,164	3,714
3198	5,276	4,992	3,158	3,85
2704	5,189	4,824	3,125	4,045
2215	4,794	4,452	3,13	4,325
1711	4,025	3,73	2,909	4,789
1209	3,417	2,624	2,453	5,713
699	2,162	1,836	4,53	8,3
197	1,191	1,217	2,809	2,024
102	1,074	1,133	1,977	1,333
50	1,3	1	2,157	1,25

Tabelle 1: Mittlere Weglänge in Abhängigkeit von Netzwerkgröße (Referenzebene)

Die mittlere Weglänge in Abhängigkeit von der Netzwerkgröße bei der Betrachtung als gerichtetes Netzwerk zeigt im Bereich von etwa 500 bis 5000 Knoten einen logarithmischen Verlauf, was durch den fast geradlinigen Verlauf der Kurve in diesem Teil (bei der logarithmierten Abszissenachse) gekennzeichnet ist. Merkwürdig ist, dass bei einem Netzwerk mit weniger als 500 Knoten das Wachstum deutlich schneller, als bei der weiteren Entwicklung ist. Ein weiteres Merkmal ist der Gleichgewichtszustand, der bei $n = 3500$ erreicht wird. Ab diesem Punkt wird der Wert von d niedriger. Der Verlauf von der Zufallsgraphenkurve stellt eine ähnliche Charakteristik dar, die Steigung ist jedoch höher und der Gleichgewichtszustand deutlicher.

Bei der Betrachtung des Netzwerks ohne Berücksichtigung der Kantenrichtung ist ein interessantes Phänomen zu beobachten. Das Verhältnis von n und d steigt logarithmisch von 100 bis etwa 700 Knoten, danach wird das Maximum erreicht und die mittlere Länge des Weges senkt sich deutlich. Bisher war der Verlauf von beiden Kurven sehr ähnlich,

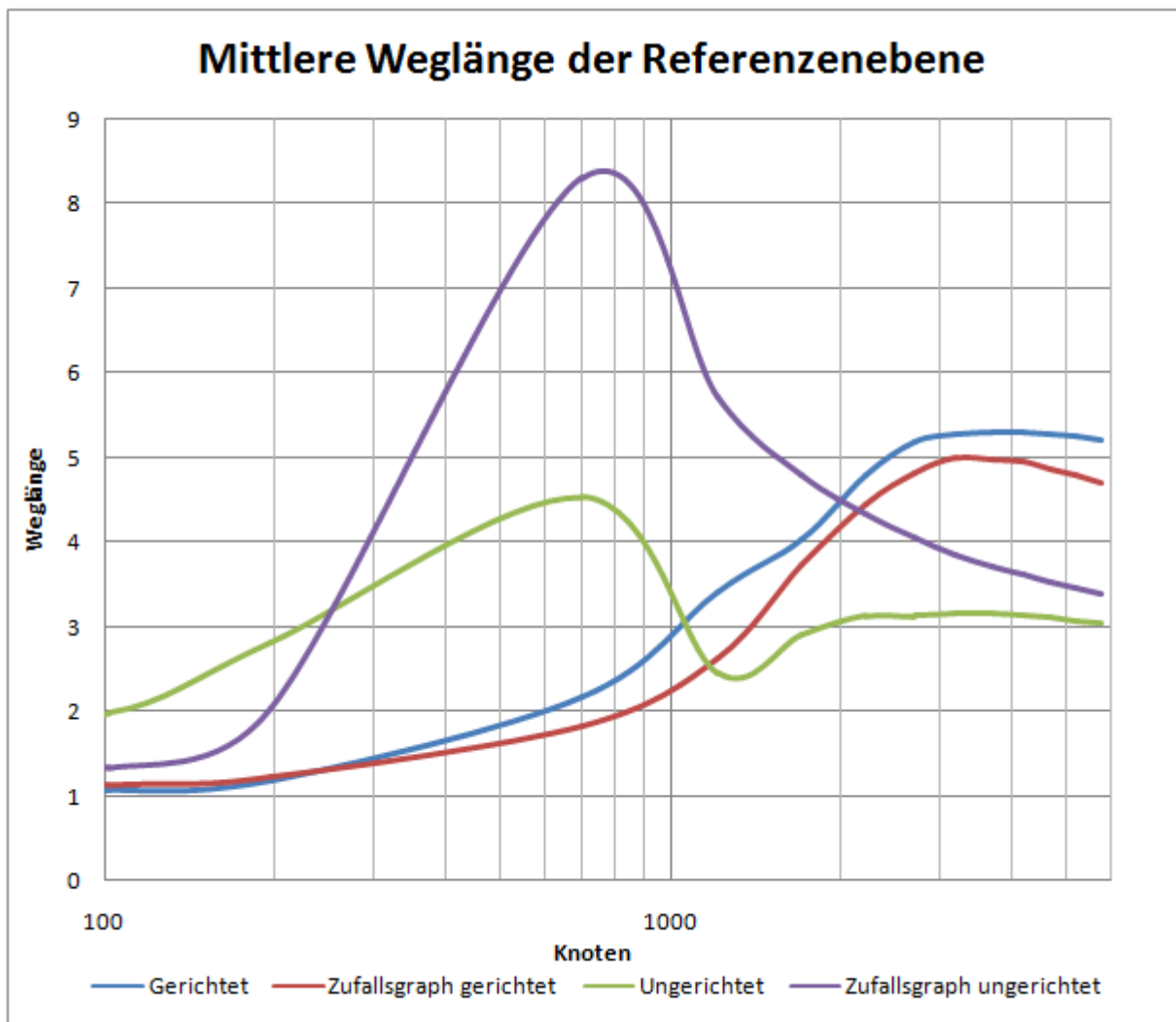


Abbildung 7: Mittlere Weglänge der Referenzebene

jedoch ab $n = 1000$ sind deutliche Unterschiede sichtbar. Das echte Netzwerk erreicht ein lokales Minimum, steigt leicht und schwebt danach im relativ kleinem Bereich, wobei der Zufallsgraph exponentiell abnimmt.

Aktualitätsebene

Die Aktualitätsebene enthält alle Aktualisierungen und Ersetzungen von den RFC-Dokumenten, aus diesem Grund ist ihre Struktur stark unterschiedlich von dem Referenzennetzwerk. Die Vernetzung ist deutlich geringer (1700 gegen etwa 50000 Kanten), außerdem existieren zahlreiche isolierte Untergraphen, die lediglich aus einem oder zwei Knoten bestehen und beeinflussen dadurch das Endergebnis. Auch die größeren Komponenten werden durch ihren linearen Aufbau aufgezeichnet. Falls ein Dokument A von

Dokument B und B von C ersetzt wird, ist kaum zu erwarten, dass zwischen A und C auch eine direkte Beziehung existiert - das erklärt den höheren Wert von d .

Knoten	RFC-Netzwerk	Zufallsgraph
5778	8,581	2,2
5200	8,591	1,878
4711	6,662	1,649
4194	6,895	1,546
3694	7,07	1,499
3198	7,144	1,412
2704	6,737	1,288
2205	6,916	1,196
1711	6,489	1,174
1198	4,936	1,094
699	4,396	1,069
197	1,895	1
102	1,5	1
50	1,6	0

Tabelle 2: Mittlere Weglänge in Abhängigkeit von der Netzwerkgröße (Aktualitätsebene)

Dieses Unternetzwerk wurde lediglich als ein ungerichteter Graph betrachtet, weil die jeweiligen Beziehungen immer in komplementären Paaren vorkommen. Die mittlere Länge des Weges ist bei gleicher Anzahl von Knoten deutlich höher als für einen zufällig generierten Graphen mit gleicher Anzahl von Knoten. Das liegt daran, dass die Anzahl von Kanten relativ gering ist und bei einer gleichen Wahrscheinlichkeit der Vernetzung, haben die Knoten des Zufallsgraphen nur kleine Komponenten erzeugt. Im Gegensatz, in dem strukturiert wachsenden RFC-Normennetzwerk sind mehrere solche Gruppierungen vorhanden, die durch die Entwicklung von unterschiedlichen Normen entstanden sind.

In einer halblogarithmischer Darstellung der Ergebnisse ist ein geradliniger Verlauf der mittleren Weglänge für das RFC-Netzwerk zu beobachten. Es gibt zwar kleine Schwankungen, die aber das Verhältnis $d \sim \log n$ im Allgemeinen nicht verletzen. In diesem Fall sind keine Sättigungspunkte erkennbar, ab denen der Wert von d sich senkt. Das betont die unterschiedliche Struktur von beiden Netzwerken. Ein weiteres solches Merkmal ist die Abwesenheit von einer nichtlogarithmischen Wachstumsphase - die Werte wachsen

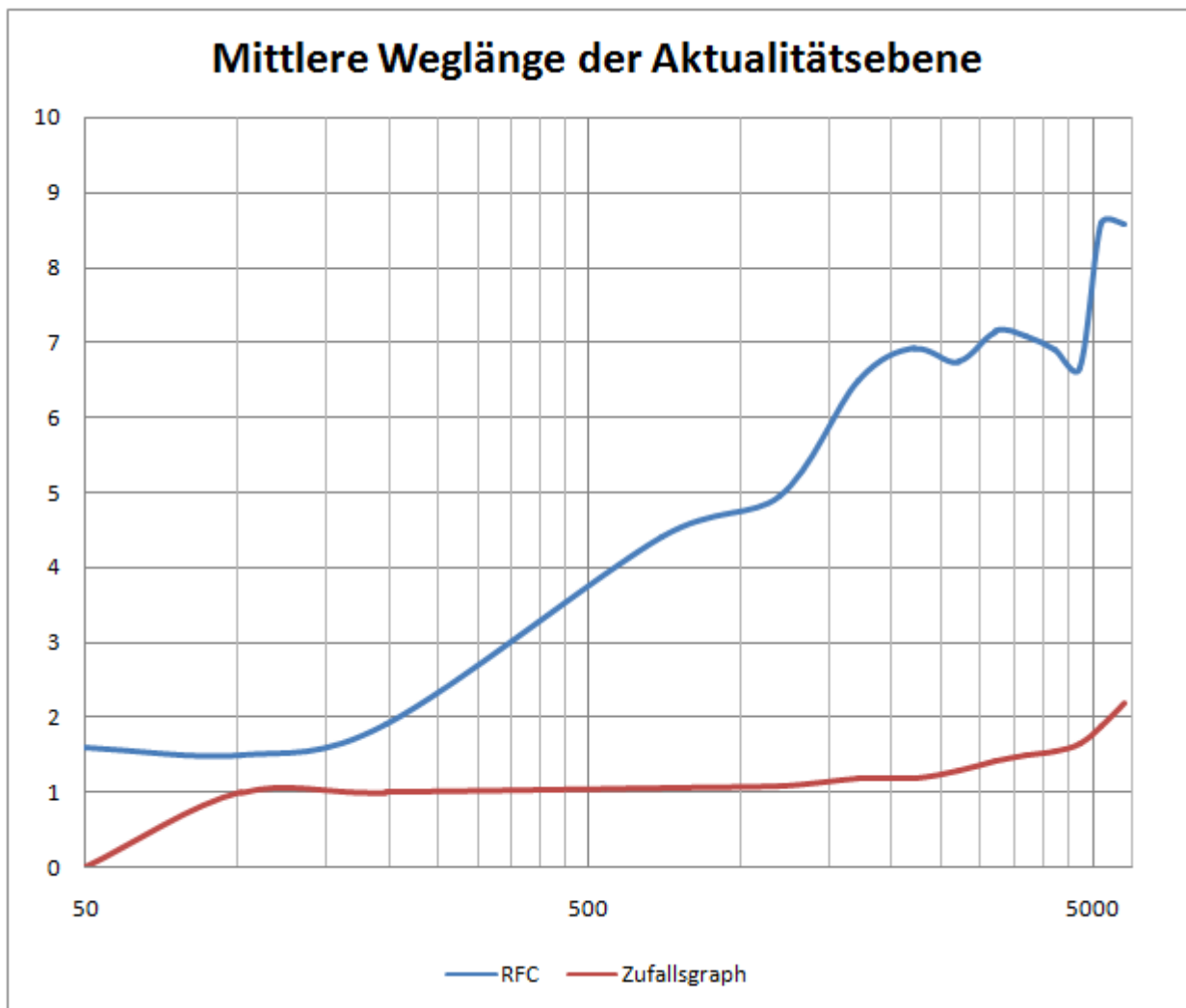


Abbildung 8: Mittlere Weglänge der Aktualitätsebene

vom Anfang an logarithmisch. Merkwürdig ist auch der schnelle Sprung bei $n = 4700$, der auf die Ersetzung bzw. Aktualisierung von vielen Dokumenten im Zeitraum von 2006 bis 2010 weist. Die Ergebnisse für den zufälligen Graphen lassen keine eindeutige Aussage machen, weil aufgrund des geringen Kanten/Knoten-Verhältnisses bewegen sie sich im sehr eingeschränkten Bereich.

Beide Ebenen des RFC-Netzwerks zeigen die charakteristischen Small-World-Effekte: Niedrige Entfernung zwischen allen Knoten und ein logarithmisches Wachstum der mittleren Weglänge in Abhängigkeit von der Knotenzahl, was sich auf der Aktualitätsebene von dem Zufallsgraphen deutlich unterscheidet. Die Untersuchung von diesen Eigenschaften hat auch eine abweichende Strukturierung der beiden Netzwerke verdeutlicht.

4.1.2 Clusterkoeffizient

Die Untersuchung von den natürlich entstandenen Netzwerken ergab eine charakteristische Eigenschaft - es entstehen Knotengruppen, innerhalb von denen man eine erhöhte Wahrscheinlichkeit (im Vergleich zu einem zufälligen Netzwerk) von der Verlinkung beobachtet [14]. Das bedeutet, wenn ein Knoten A mit Knoten B und C verbunden ist, ist die Wahrscheinlichkeit, dass eine Kante zwischen B und C existiert größer, als beim zufälligen Hinzufügen von den neuen Ecken.

Der Clusterkoeffizient C ist ein Maß für die Verlinkung eines Netzwerks, mit dem die o.g. Eigenschaft numerisch ausgewertet werden kann. Zur Berechnung des Wertes vom Koeffizienten werden zwei unterschiedliche Methoden angewendet.

- **Globaler Clusterkoeffizient**

In diesem Verfahren wird der Graph auf das Vorkommen von Tripeln untersucht. Ein Tripel besteht aus drei Knoten, die miteinander verbunden sind. Falls innerhalb von einem Tripel nur zwei Kanten vorhanden sind, spricht man von einem offenen Tripel. Analog, wenn 3 Kanten existieren, ist ein Tripel geschlossen und fällt unter den Begriff Clique, d.h. ein völlig vernetzter Untergraph.

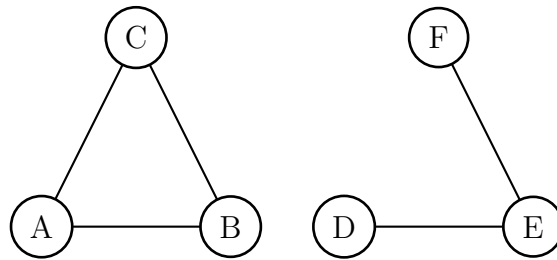


Abbildung 9: Geschlossenes und offenes Tripel

Der Globale Clusterkoeffizient gibt an, wieviele Tripel geschlossen sind und ist folgendermaßen definiert [12]:

$$C = \frac{\text{Anzahl von geschlossenen Tripeln}}{\text{Anzahl von allen Tripeln}} \quad (2)$$

- **Watts-Strogatz Clusterkoeffizient**

Der vom Duncan Watts und Steven Strogatz vorgeschlagene Verfahren zur Bestimmung des Vernetzungsgrades eines Netzwerk geht von dem lokalen Clusterkoeffizienten für die einzelnen Knoten heraus [15]. Der lokale Clusterkoeffizient gibt an, wie stark die Nachbarknoten N_i von einem bestimmten Knoten v_i miteinander verbunden sind. N_i ist die Menge von Knoten, die mit v_i mit einer Kante verbunden sind.

Sei k_i die Kardinalität von N_i . Der lokale Koeffizient C_i des Knoten v_i ergibt sich als die Anzahl von Ecken zwischen Knoten aus N_i durch die maximale Anzahl von Ecken für k_i Knoten. Hier unterscheidet man zwischen einem gerichteten und ungerichteten Graphen. In dem ersten Fall ist die Anzahl als $k_i(k_i - 1)$ gegeben, im zweiten - als $\frac{1}{2}k_i(k_i - 1)$, weil wenn keine Richtung angegeben wird, ist $e_{ij} = e_{ji}$. Falls der Knoten v_i isoliert ist, bzw. nur einen Nachbarn besitzt, wird $C_i = 0$ gesetzt.

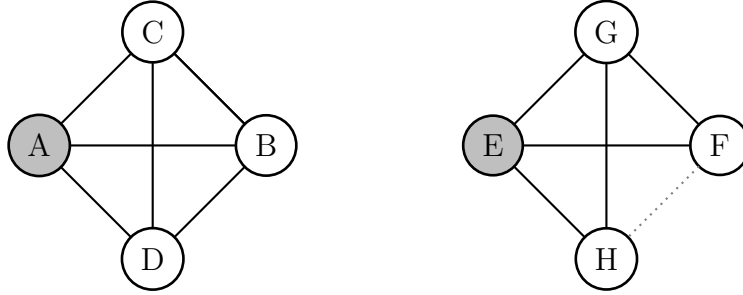


Abbildung 10: Lokaler Clusterkoeffizient: $C_A = 1$, $C_E = \frac{2}{3}$

Lokaler Clusterkoeffizient für gerichtete Graphen ist gegeben durch:

$$C_i = \frac{|\{e_{jk}\}|}{k_i(k_i - 1)}, v_j, v_k \in N_i, e_{jk} \in E \quad (3)$$

Für ungerichtete Graphen nimmt die Gleichung folgende Form an:

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)}, v_j, v_k \in N_i, e_{jk} \in E \quad (4)$$

Der Watts-Strogatz Clusterkoeffizient C des gesamten Netzwerks ist als ein Durchschnittswert von allen lokalen Koeffizienten definiert.

$$C = \frac{1}{|V|} \sum_i C_i \quad (5)$$

Die Berechnung von dem globalen Clusterkoeffizienten erfolgte mit dem *Network Workbench Tool*. Die Implementierung des Algorithmus setzt voraus, dass das Netzwerk keine isolierte Knoten besitzt und ungerichtet ist. Um das zu erreichen, wurden alle Knoten ohne Nachbarn mit der Funktion *Delete Isolates* entfernt. Ferner, die Referenzebene wurde als ungerichtet betrachtet. Der Watts-Strogatz Koeffizient wurde mit Gephi berechnet. In diesem Fall waren keine Randbedingungen spezifiziert und der Algorithmus ermöglichte eine Auswertung des Graphen als gerichtet und ungerichtet.

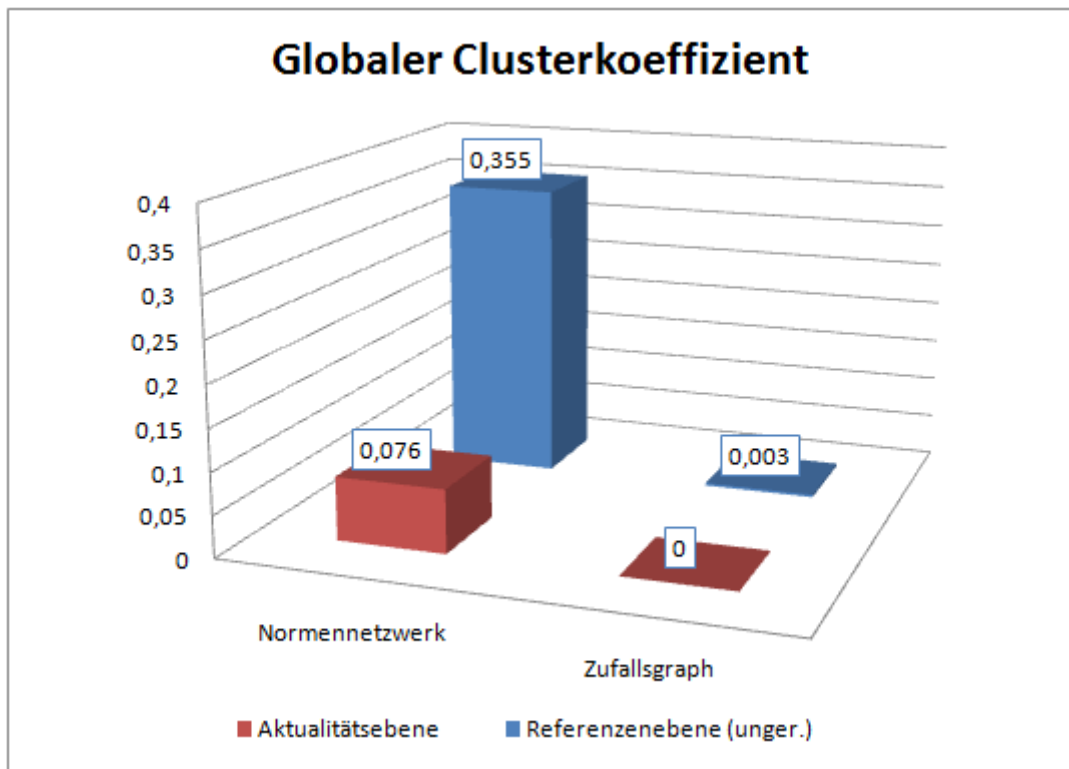


Abbildung 11: Globaler Clusterkoeffizient

Beide Ebenen des Normennetzwerkes weisen Clustering vor, sowohl bei der Berechnung des globalen, als auch des Watts-Strogatz Clusterkoeffizienten. Aufgrund von der viel höheren Anzahl von Kanten auf der Referenzebene, ist der Wert auch größer als

auf der Aktualitätsebene. Die Betrachtung des Zitierungnetzwerks als ungerichtet hat, wie erwartet, den Clusterkoeffizient verdoppelt. Die Ursache ist, dass sich in diesem Fall die im Nenner vorkommende maximale Anzahl von Kanten halbiert. Merkwürdig ist hier auch die Diskrepanz zwischen beiden Werten bei der Aktualitätsebene, wo der globale Koeffizient ein Dreifaches von $C_{(5)}$ ist. Dieses Verhalten lässt sich durch die Entfernung von isolierten Knoten erklären, die eine Voraussetzung des Algorithmus war. Da auf der Ebene 3769 (von insgesamt 5778) solche Knoten vorkommen, hat der Löschvorgang die Größe des Graphen um etwa 65% reduziert und dadurch auch die Anzahl von allen Tripeln signifikant verringert.

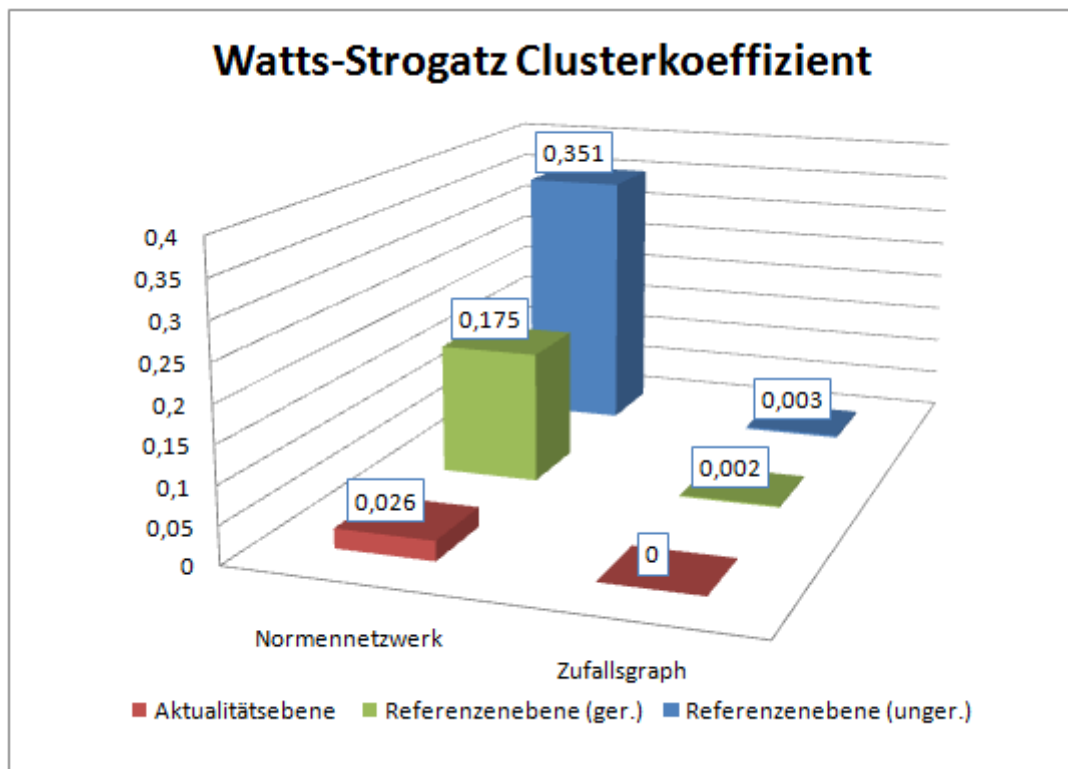


Abbildung 12: Watts-Strogatz Clusterkoeffizient

Im Vergleich zu den zufällig generierten Graphen mit derselben Anzahl von Knoten und Kanten sind die Ergebnisse für das Normennetzwerk signifikant höher. Das entspricht den Erwartungen - wenn ein RFC-Dokument zwei andere referenziert, ist es relativ wahrscheinlich, dass sie die Themenbereiche von ihnen überlappen. Daraus folgt, dass sie sich auch gegenseitig referenzieren können. Der niedrigere Wert für die Referenzebene liegt an der linearen Struktur des Graphen und an der Charakteristik von den dort enthalte-

nen Beziehungen. Zusätzlich, sind auf dieser Ebene viele Kanten vorhanden, die lediglich einen oder gar keinen Nachbarn besitzen. Das beeinflusst insbesondere den Watts-Strogatz Koeffizienten, weil in solchen Fällen der lokale Wert 0 beträgt.

Zusammenfassend, sowohl die Untersuchung von der mittleren Weglänge, als auch deren Wachstums in Abhängigkeit von der Anzahl der Knoten ergab, dass die beiden Teile des Normennetzwerks durch eine niedrige mittlere Entfernung von Knoten charakterisiert werden, die 5,215 für die Referenzebene und 8,581 für die Aktualitätsebene beträgt. Die beiden Werte liegen im Bereich, der für die thematisierten Dokumentennetzwerke, wie z.B. wissenschaftliche Publikationen charakteristisch ist. [12] Ferner, die Beobachtung der Abhängigkeit der mittleren Weglänge von der Netzwerkgröße führte zur Schlussfolgerung, dass die Entfernung nicht schneller als logarithmisch wächst und einen Sättigungspunkt erreicht, bei dem das Wachstum angehalten wird.

Zudem zeigen die beiden Netzwerkebenen Neigung zum Clustering - es entstehen Gruppen von Knoten, innerhalb von denen die gegenseitige Vernetzung stark ist. Der Watts-Strogatz Clusterkoeffizient beträgt 0,175 für das Referenzennetzwerk und 0,026 für die Aktualitätsebene. Der globale Clusterkoeffizient beträgt dementsprechend 0,355 und 0,076. Im Vergleich zu den zufälligen Graphen derselben Größe, weichen die Werte stark ab und zeigen, dass die Entstehung von Kanten kein zufälliger Prozess war.

Diese drei Merkmale - niedrige mittlere Weglänge, ihr maximal logarithmisches Wachstum und relativ höher Clusterkoeffizient weisen eindeutig darauf hin, dass die beiden Ebenen des RFC-Normennetzwerks, zusammen mit vielen menschlich oder natürlich erzeugten Netzen zu der Gruppe der Kleine-Welt-Netzwerken gehören und sich von den zufällig generierten Graphen derselben Größe unterscheiden.

4.2 Verteilung des Knotengrades

Jeder Knoten v_i eines Graphen wird durch seinen Grad d_i charakterisiert. Der Knotengrad gibt an, wieviele Kanten mit einem Knoten verbunden sind. Man unterscheidet zwischen dem Ingrad (nur hineingehende Kanten), dem Ausgrad (nur herausgehende Kanten) und dem gesamten Grad (beide Kantenarten). Bei ungerichteten Netzwerken sind alle drei Größen gleich, da die Ecken in beiden Richtungen durchgelaufen werden können. Falls ein Knoten isoliert ist, ist der Grad gleich 0. Da der Grad sich nur auf einen Knoten

bezieht, sind weitere Metriken notwendig, um eine Aussage über das gesamte Netzwerk zu machen. Das sind:

- **Durchschnittsgrad**

Diese Charakteristik gibt an, welchen durchschnittlichen Grad (bzw. Ingrad oder Ausgrad) das Netzwerk besitzt. Dieser Wert macht eine allgemeine Aussage über die Vernetzung eines Graphen. Er muss jedoch mit Vorsicht betrachtet werden, da es sich hier lediglich um ein arithmetisches Mittel handelt, das durch ein Vorkommen von einzelnen Knoten mit sehr hohem Grad leicht erhöht werden kann. Der Durchschnittsgrad $\bar{d}$ wird folgendermaßen berechnet:

$$\bar{d} = \frac{1}{|V|} \sum_i d_i, v_i \in V \quad (6)$$

- **Gradverteilung**

Die Gradverteilung ist ein Maß für die Wahrscheinlichkeit des Vorkommens eines Knoten v_i mit dem Grad k im gesamten Graphen. Für jeden bestimmten Knotengrad k ergibt sich die Wahrscheinlichkeit p_k als:

$$p_k = \frac{|\{v_i\}|}{|V|}, v_i \in V \wedge \deg(v_i) = k \quad (7)$$

Diese Darstellung ist aber nur dann geeignet, wenn der Unterschied zwischen dem maximalen und minimalen Grad relativ gering ist. In vielen Netzwerken ist er dagegen sehr breit. In solchen Fällen ist die Methode von der exponentiellen Einteilung (*Exponential Binning*) besser geeignet. Diese basiert darauf, dass statt die Wahrscheinlichkeit nur für einen bestimmten Grad zu berechnen, berücksichtigt man einen ganzen, exponentiell wachsenden Bereich (z.B. $k_1 = 1, k_2 = 2 - 3, k_3 = 4 - 7$ usw.) und dessen gesamte durchschnittliche Wahrscheinlichkeit wertet. Das verursacht zwar einen Verlust der Genauigkeit innerhalb eines Intervalls (man kann den einzelnen Graden keinen Wert zuweisen), glättet aber die Kurve und vereinfacht die Darstellung und Auswertung des Verlaufs. Bei der grafischen Darstellung wird für die jeweiligen Wahrscheinlichkeiten kein Bereich angegeben, sondern dessen Mitte.

4.2.1 Referenzebene

Die Referenzebene ist ein gerichtetes Netzwerk, deshalb ist eine separate Auswertung des In- und Ausgrads möglich. Dabei muss beachtet werden, dass eine Aussage über die Wichtigkeit von den bestimmten RFCs lediglich der Ingrad macht - er gibt die Anzahl der Memoranda an, die dieses Dokument zitieren. Der Ausgrad sagt nur aus, wieviele Referenzen im Text enthalten sind, was aber nicht die Bedeutung des Dokumentes widerspiegelt.

Die Berechnung von den Metriken erfolgte mit dem *Network Workbench Tool*. Die dort eingebauten Algorithmen ermöglichen nicht nur die Berechnung von dem Durchschnittgrad, sondern auch von der Gradverteilung (mit und ohne exponentielle Einteilung). Zusätzlich ist auch die Extrahierung von den Knoten im bestimmten Gradbereich möglich. Durchführung von den genannten Metriken erlaubt eine vollständige Analyse von diesem Aspekt.

Der durchschnittliche Ingrad beträgt 8,425 und der Ausgrad - 8,425. Diese Werte stimmen überein, weil bei der Betrachtung des ganzen Netzwerks ist die Anzahl der herausgehenden Kanten gleich der Anzahl der hineingehenden. Der gesamte Durchschnittsgrad ist 16,850. Die Ergebnisse für den Zufallsgraphen sind äquivalent, weil sie lediglich von der Kanten- und Knotenanzahl abhängig sind und diese sind in beiden Fällen gleich. Um den Einfluss von den stark vernetzten Knoten auf den mittleren Grad zu untersuchen und gleichzeitig die wichtigsten RFCs zu bestimmen, wurden 20 Dokumente mit dem maximalen Ingrad extrahiert.

Mit 2385 eingehenden Kanten liegt RFC2119 auf dem ersten Platz. Dieses Dokument spezifiziert die Begriffe, die in den anderen Texten vorkommen, um sie eindeutig innerhalb der gesamten Sammlung zu machen. Da die RFC-Nummern nach dem Erscheinungsdatum vergeben werden, ist leicht zu berechnen, dass etwa 62% von allen Texten, die nach der Veröffentlichung von RFC2119 erschienen, referenzieren dieses Memorandum. Die Erklärung von diesem Zustand ist einfach - die strikte Strukturierung von den RFC-Dokumenten setzt voraus, dass auch die dort enthaltenen Begriffe eindeutig definiert sind. Diese Definitionen sind bereichsübergreifend und werden nicht nur auf eine bestimmte Menge von Protokollen oder Lösungen eingeschränkt, was zur Entstehung von Referenzen aus vielen unabhängigen Dokumenten führt.

Ingrad	RFC	Titel	Status
2385	RFC2119	Key words for use in RFCs to Indicate Requirement Levels	BCP
468	RFC822	Standard for the format of ARPA internet text messages	BCP
372	RFC791	Internet Protocol	Standard
347	RFC2434	RTP Payload Format for Bundled MPEG	Experimental
295	RFC793	Transmission Control Protocol	Standard
290	RFC3261	SIP: Session Initiation Protocol	Proposed Standard
256	RFC2578	Structure of Management Information Version 2 (SMIv2)	Standard
254	RFC1157	Simple Network Management Protocol (SNMP)	Historic
248	RFC1034	Domain names - concepts and facilities	Standard
247	RFC1035	Domain names - implementation and specification	Standard
242	RFC2579	Textual Conventions for SMIv2	Standard
238	RFC2026	The Internet Standards Process – Revision 3	BCP
232	RFC2580	Conformance Statements for SMIv2	Standard
213	RFC2045	Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies	Draft Standard
202	RFC1213	Management Information Base for Network Management of TCP/IP-based internets:MIB-II	Standard
200	RFC1700	Assigned Numbers	Historic
199	RFC2401	Security Architecture for the Internet Protocol	Proposed Standard
196	RFC1155	Structure and identification of management information for TCP/IP-based internets	Standard
194	RFC2460	Internet Protocol, Version 6 (IPv6) Specification	Draft Standard
188	RFC768	User Datagram Protocol	Standard

Tabelle 3: 20 Dokumente mit höchstem Ingrad

Zu den weiteren, oft zitierten RFCs gehören die Spezifikationen von zahlreichen, gängigen Standards, wie E-Mail (RFC822), IP und TCP-Protokoll (RFC791, RFC793) SIP-Protokoll für die Internet-Telefonie (RFC3261) und das Simple Network Management Protocol (RFC1157). Sehr auffällig ist der Unterschied zwischen dem Grad von RFC2119 und RFC822 - das zweite Dokument besitzt nur ein fünftel eingehender Kanten des ersten. Diese Diskrepanz kann auf zwei Arten erklärt werden. Zuerst, RFC2119 ist, wie vorher beschrieben, bereichsübergreifend. Zweitens, nicht alle Dienste und Protokolle der höheren ISO/OSI Schichten beziehen sich direkt auf die grundlegenden Transport- und Netzwerk-Schichten, obwohl sie trotzdem bei der Übertragung verwendet werden. Merkwürdig ist auch die Unterteilung des Status von den jeweiligen Dokumenten - lediglich 6 davon wurden nicht auf den Standard-Track aufgenommen, alle anderen wurden bereits anerkannt oder befinden sich in der Testphase vor der Einführung als Internet Standard.

Bei der Untersuchung von den Knoten mit größter Anzahl von herausgehenden Kanten lässt sich zwar keine Aussage über die Bedeutung von jeweiligen Knoten machen, das Ergebnis ist aber trotzdem aufgrund der komplett unterschiedlichen Stati von den Dokumenten interessant. Sechs von zehn ausgewählten RFCs, die durch einen hohen Ausgrad charakterisiert werden, sind als *Informational* bezeichnet, zwei als *Historic* und zwei als unbekannt. Im Gegensatz zu den Knoten mit großem Ingrad, gehört kein Memorandum zum Standards-Track. Der am häufigsten referenzierende Text ist RFC1012. Dieses Dokument ist einfach ein Inhaltsverzeichnis von den RFCs von 1 bis 999. Auf dem zweiten Platz befindet sich RFC2626, wo das Jahre-2000-Problem im Bezug auf das Internet beschrieben wurde.

Zusätzlich sind unter Top-10 drei Texte aus dem Bereich 3790-3799 vorhanden, die die Umstellung von IPv4 auf IPv6 in als Standard anerkannten RFCs untersuchen. Zudem sind auf der Liste auch zwei *Assigned Numbers*-Dokumente, die als Verzeichnis von vergebenen Portnummern dienen und dadurch zahlreiche andere RFCs referenzieren. Im Gegenteil zu den Knoten mit höchstem Ingrad, sind die hier aufgelisteten Dokumente von relativ wenig Bedeutung für die Entwicklung des Normennetzwerks. Es darf jedoch nicht vernachlässigt werden, dass solche verzeichnisartige Dokumente zur deutlichen Kürzung von der durchschnittlichen Weglänge führen können. Dieser Aspekt wird bei der Untersuchung von der Netzwerkrobustheit genauer betrachtet.

Ausgrad	RFC	Titel	Status
920	RFC1012	Bibliography of Request For Comments 1 through 999	Informational
413	RFC2626	The Internet and the Millennium Problem (Year 2000)	Informational
267	RFC3795	Survey of IPv4 Addresses in Currently Deployed IETF Application Area Standards Track and Experimental Documents	Informational
201	RFC3790	Survey of IPv4 Addresses in Currently Deployed IETF Internet Area Standards Track and Experimental Documents	Informational
198	RFC1000	Request For Comments reference guide	Unknown
184	RFC1700	Assigned Numbers	Historic
162	RFC3796	Survey of IPv4 Addresses in Currently Deployed IETF Operations & Management Area Standards Track and Experimental Documents	Informational
144	RFC1340	Assigned Numbers	Historic
143	RFC2896	Remote Network Monitoring MIB Protocol Identifier Macros	Informational
133	RFC1011	Official Internet protocols	Unknown

Tabelle 4: 10 Dokumente mit höchstem Ausgrad

Wenn die Verteilung des Knotengrades dem Potenzgesetz folgt, wird ein Netzwerk mit dieser Eigenschaft skalenfrei genannt. [18] Diese Einteilung setzt jedoch nicht voraus, dass weitere Eigenschaften und Metriken, wie z.B. mittlere Weglänge in Abhängigkeit von der Netzwerkgröße auch skalenfrei sind.

$$p_k \sim \frac{1}{k^\alpha} \Leftrightarrow p_k \sim k^{-\alpha} \quad (8)$$

In skalenfreien Netzen ist die Wahrscheinlichkeit, dass ein Knoten einen bestimmten Grad k besitzt, proportional zu $k^{-\alpha}$, wobei α ein konstanter Exponent ist. Dieser asymptotische Verlauf bedeutet, dass sehr viele Knoten mit einem geringen Grad existieren. Zusätzlich sind auch wenige Knoten mit sehr hohem Grad vorhanden, die als lokale Zentren, mit denen die anderen Knoten verbunden sind, dienen.

Diese Verteilung unterscheidet sich deutlich von dem Verhalten eines Zufallsgraphen. Wenn die Kanten mit gleicher Wahrscheinlichkeit eingefügt werden, enthält man eine Glockenkurve - es gibt sehr viele Knoten mit einem ähnlichen Grad im mittleren Bereich, die Streuung von den Werten ist auch relativ gering.

Viele Netzwerke weisen diese Eigenschaft auf - darunter das Internet, Netzwerk von Schauspielern und sogar das Netzwerk von sexuellen Kontakten. [12] Für die Untersuchung des RFC-Normennetzwerks sind als Vergleichswerte die von Price angegebenen Koeffizienten für das wissenschaftliche Zitationsnetzwerk, das zur RFC-Referenzebene vergleichbar ist, geeignet. [19] Interessanterweise, folgt die Knotengradverteilung im World Wide Web dem Potenzgesetz auch. Diese wurde von Barabási untersucht und trotz der Erwartung, dass bei der Vielzahl von Themen, Interessengebieten und Anzahl der existierenden Seiten eher ein zufälliger Graph beobachtet wird, ist das WWW skalenfrei. Obwohl eine gravierende Mehrheit von Seiten (80%) weniger als vier Links hat, 0,001% hat mehr als 1000 Verknüpfungen - diese halten das Netz zusammen. [17].

Mit derartiger Verteilung ist der sog. *Preferential Attachment Process* eng verbunden. Dieser Prozess besagt, dass die Knoten, die einen höheren Grad haben, auch eine erhöhte Wahrscheinlichkeit, wieder verlinkt zu werden haben. Dieses Phänomen kommt in der Natur vor, z.B. bei der Entstehung von neuen Arten innerhalb von Gattungen - die, die bereits viele Arten haben, wachsen auch schneller. [20]. Ein ähnlicher Prozess lässt sich bei den wissenschaftlichen Arbeiten beobachten. Die, die oft zitiert wurden, gewinnen an Bedeutung und werden Grundlagen vom bestimmten Themenbereich - was wiederum zur Entstehung von mehreren Referenzen führt. Die Skalenfreiheit lässt sich auch bei der wissenschaftlichen Leistungsfähigkeit (gemessen in veröffentlichten Artikeln) beobachten. Die Anzahl von Personen, die n Artikel geschrieben haben, ist proportional zu n^{-2} . Spätere Untersuchungen ergaben, dass der Koeffizient niedriger ist und etwa 1,7 beträgt. [22] Diese Proportionalität wurde bereits im Jahre 1926 von A. J. Lotka beobachtet und nach ihm - Lotkas Gesetz - benannt. [21]

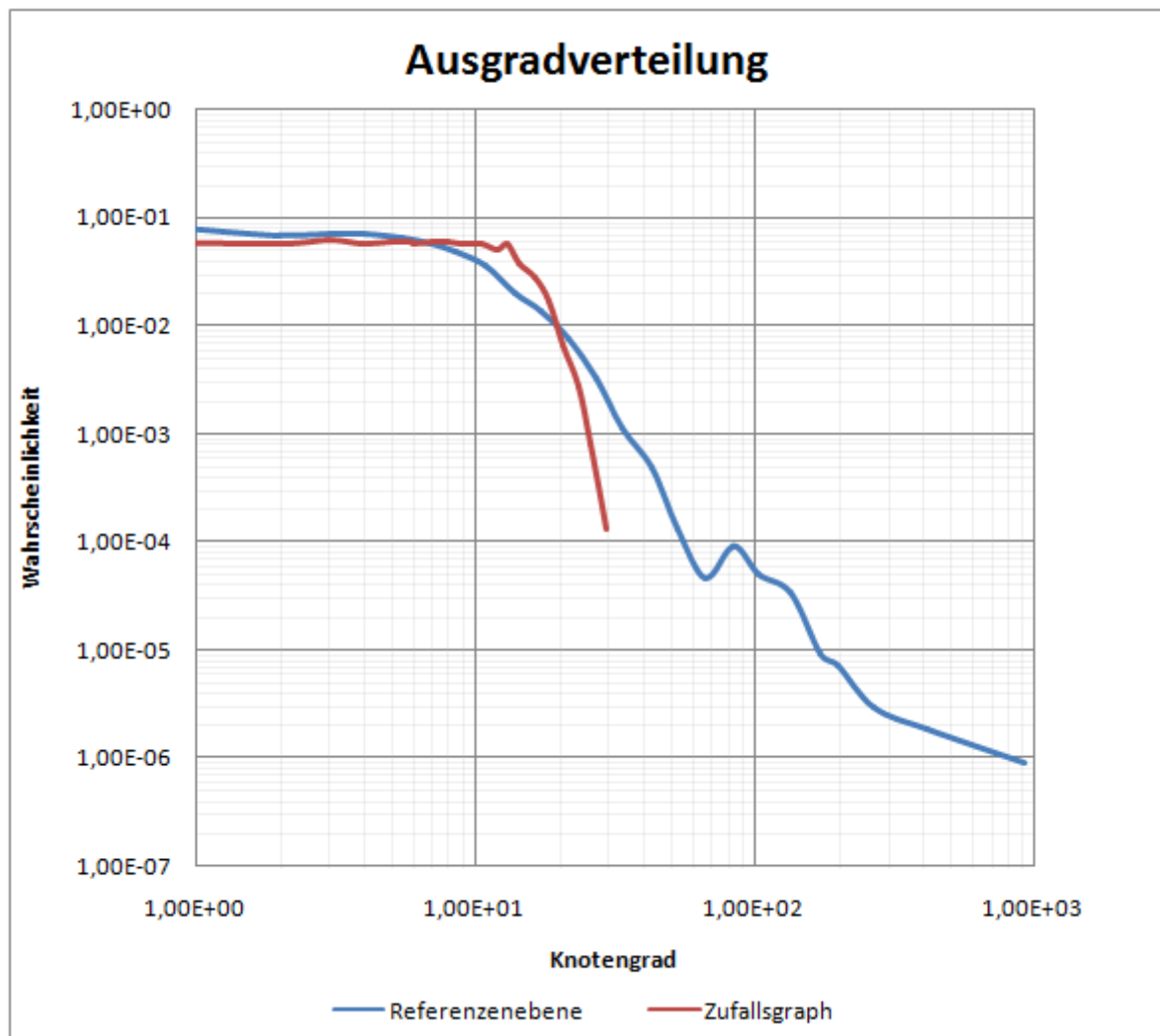


Abbildung 13: Ausgradverteilung (Referenzebene)

Die Ausgradverteilung der Referenzebene zeigt eine interessante Charakteristik. Die Wahrscheinlichkeit, dass der Grad im Bereich von etwa 1 bis 6 liegt, ist fast gleich. Erst ab dem Wert 10 ist eine deutliche Senkung der Knotenanzahl zu beobachten. Knoten, die den Ausgrad > 10 haben, bilden etwa 26% des Netzwerks. Im Vergleich zu einem Zufallsgraphen ist eine höhere Streuung von den Ausgraden sichtbar. Außerdem ist der Verlauf der Kurve weniger steil und sogar im ersten Abschnitt ist die zunehmende Tendenz zu beobachten, wobei der zufällig erzeugte Graph in diesem Bereich fast konstante Werte annimmt. Die mit Excel berechnete Ausgleichsgerade hat folgende Form:

$$f(x) = 0,782x^{-2}$$

Der α -Exponent für die Ausgradverteilung der Referenzebene beträgt also 2.

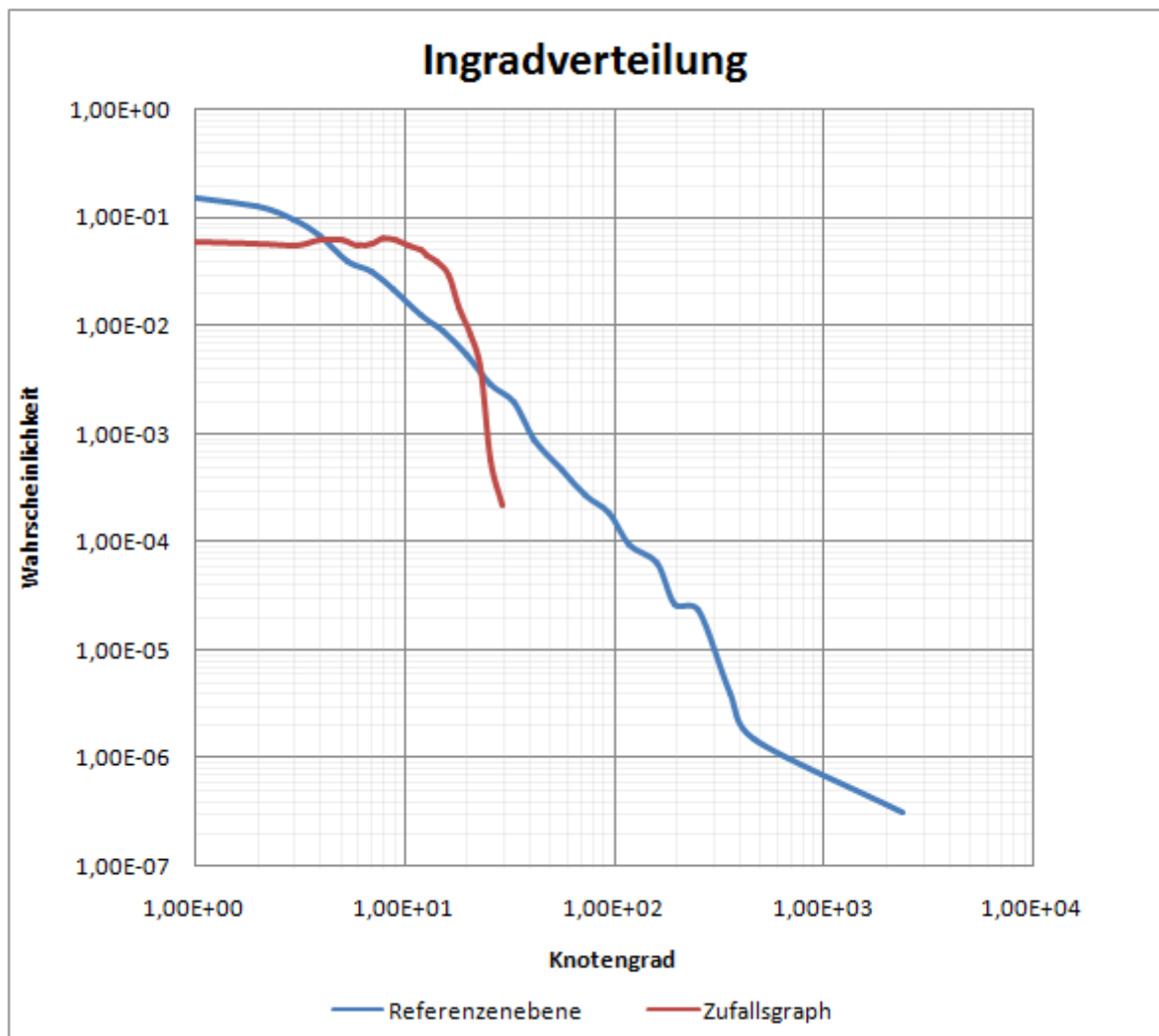


Abbildung 14: Ingradverteilung (Referenzebene)

Die Verteilung von eingehenden Kanten zeigt noch offensichtlicher ihre Skalenfreiheit. Es existiert kein Bereich, in dem die Knoten eine ähnliche Gradverteilung haben - der Verlauf nimmt ständig und mit konstanter Geschwindigkeit ab. Das bedeutet, dass man keine Klassen von Knoten, die mit gleicher Wahrscheinlichkeit vorkommen, bestimmen kann. Der Unterschied zwischen dem Normennetzwerk und dem zufälligen Graphen ist sehr auffällig - sowohl der Argumentenbereich, als auch die Steigung der Kurven stellen diverse Charakteristiken dar. Der α -Exponent beträgt 1,86 und die Ausgleichsgerade ist durch folgende Funktion gegeben:

$$f(x) = 0,688x^{-1,86}$$

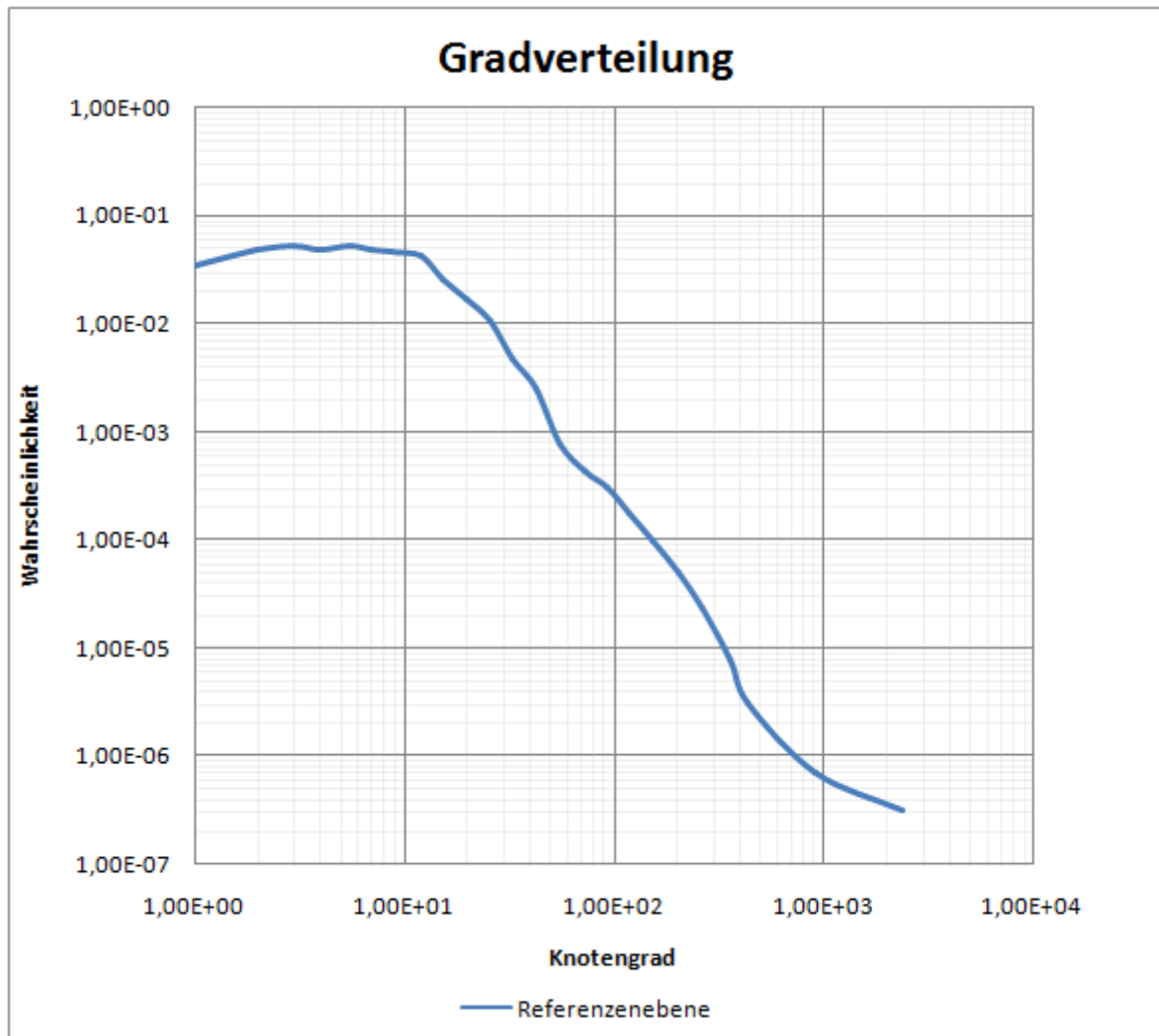


Abbildung 15: Gradverteilung (Referenzebene ungerichtet)

Bei der Betrachtung der Referenzebene als ungerichtetes Netzwerk, ist auch eine skalenfreie Verteilung des Knotengrades bemerkbar, jedoch nur ab $d = 10$. Für $d \leq 10$ bleibt die Wahrscheinlichkeit der Vernetzung im relativ engen Bereich. Dieser flache Kurventeil ist interessant, da ein ähnlicher Verlauf in der Verteilung von den wissenschaftlichen Zitierungen auch bemerkbar ist. [23] Das kann eventuell als ein Merkmal für verschiedene Zitationsnetzwerke betrachtet werden. Ein Vergleich mit dem Zufallsgraphen derselben Größe war wegen für die angewendete Software zu geringer Gradstreuung (Voraussetzung für die exponentielle Einteilung) unmöglich. α beträgt 1,82 und die Ausgleichsgerade hat folgende Form:

$$f(x) = 0,861x^{-1,82}$$

4.2.2 Aktualitätsbene

Bei der Auswertung der Knotengradverteilung von der Aktualitätsebene muss beachtet werden, dass die Kanten ungerichtet sind, weil die Beziehungen zwischen den Knoten immer komplementär sind. Diese Netzwerkebene ist auch unterschiedlich aufgebaut. Trotzdem kann man zwischen Ereignissen, die von einem Dokument selbst ausgelöst werden (*updates*, *obsoletes*) und denen, die von extern kommen (*updated by*, *obsoleted by*) unterscheiden. Außerdem ist nicht nur die Anzahl von Kanten viel geringer, sondern auch haben die Verläufe der Vernetzung eine andere Form - sie sind nicht mehr sternartig, sondern baumartig - es gibt einen Anfangspunkt, der als eine Wurzel für die weiteren Aktualisierungen und Ersetzungen dient.

Aufgrund der kleineren Kantenzahl ist der maximale vorkommende Grad viel niedriger als auf der Referenzebene und beträgt 24. Der Durchschnittliche Grad beträgt lediglich 0,596. Auffällig ist die hohe Anzahl von isolierten Knoten, die überhaupt keine Ecken besitzen - 65,2%. Die ersten drei Plätze auf der Liste von den am häufigsten verlinkten Knoten sind von drei Dokumenten belegt die das *Domain Name System* betreffen. Das erste - RFC1035 wird von einer Reihe von neuen Memoranda aktualisiert, wobei die zwei weiteren Texte - RFC4035 und RFC4034 sich mit der Sicherheit des Domainsystems beschäftigen. Zudem kommt noch RFC2181 mit 9 Kanten, wo die DNS-Spezifikation erklärt wird. Die zahlreichen Veränderungen innerhalb vom DNS weisen darauf hin, dass dieses System eine wesentliche Erweiterung erlebt hat. Anhand von den Kantentypen (extern und intern) kann man auch erkennen, dass RFC1035 aktualisiert und erweitert wurde und die zwei übrigen RFCs haben selbst mehrere Memoranda verändert.

Interessanterweise sind auf den weiteren Plätzen auch Dokumente vorhanden, die paarweise bestimmte Themenbereiche beschreiben - wie LDAP (RFC4510, RFC3377) und MPLS (RFC5462, RFC3473). Außerdem besteht bei den zwei letzten Dokumenten eine starke Diskrepanz zwischen den externen und internen Kantenzahl, zusätzlich ist die zweite Zahl häufiger größer. Dieser Sachverhalt ist durch die Natur der Beziehungen bedingt - ein Dokument kann oft mehrere ältere aktualisieren, bzw. zusammenfassen. Teilweise wird ein bestimmter Text nur von einem anderen ersetzt. Große und wichtige Dokumente (wie RFC1035) sind die Ausnahme von dieser Regel - aufgrund ihrer Komplexität werden sie von mehreren Texten aktualisiert. Auffällig ist hier auch, dass alle bis auf ein Dokument aus dem *Standards Track* stammen.

Grad	RFC	Titel	Status	Extern	Intern
24	RFC1035	Domain names - implementation and specification	Standard	21	3
19	RFC4035	Protocol Modifications for the DNS Security Extensions	Proposed Standard	1	18
19	RFC4034	Resource Records for the DNS Security Extensions	Proposed Standard	1	18
13	RFC5462	Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field	Proposed Standard	0	13
10	RFC2026	The Internet Standards Process – Revision 3	BCP	8	2
10	RFC3473	Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions	Proposed Standard	10	0
10	RFC4510	Lightweight Directory Access Protocol (LDAP): Technical Specification Road Map	Proposed Standard	0	10
9	RFC3377	Lightweight Directory Access Protocol (v3): Technical Specification	Proposed Standard	1	8
9	RFC2181	Clarifications to the DNS Specification	Proposed Standard	6	3
8	RFC1350	The TFTP Protocol (Revision 2)	Standard	7	1
8	RFC1349	Type of Service in the Internet Protocol Suite	Proposed Standard	1	7
8	RFC3658	Delegation Signer (DS) Resource Record (RR)	Proposed Standard	4	4
8	RFC123	Proffered Official ICP	Unknown	4	4

Tabelle 5: Dokumente mit höchstem Grad (Aktualitätsebene)

Daraus folgt, dass sich möglicherweise die von der IETF bestimmte Bedeutung der unterschiedlichen RFC-Kategorien in der Knotengradverteilung der Aktualitätsebenewiderspiegelt. Zusätzlich haben die als *Proposed Standard* kategorisierten Dokumente am häufigsten höhere Anzahl der internen Kanten. Im Gegensatz, die *Standard*-RFCs haben mehr externe Kanten. Das entspricht dem *Standards Track*-Anerkennungsprozess. Die alten Standards werden von den neueren, noch mit dem Status *Proposed Standard* RFCs abgelöst.

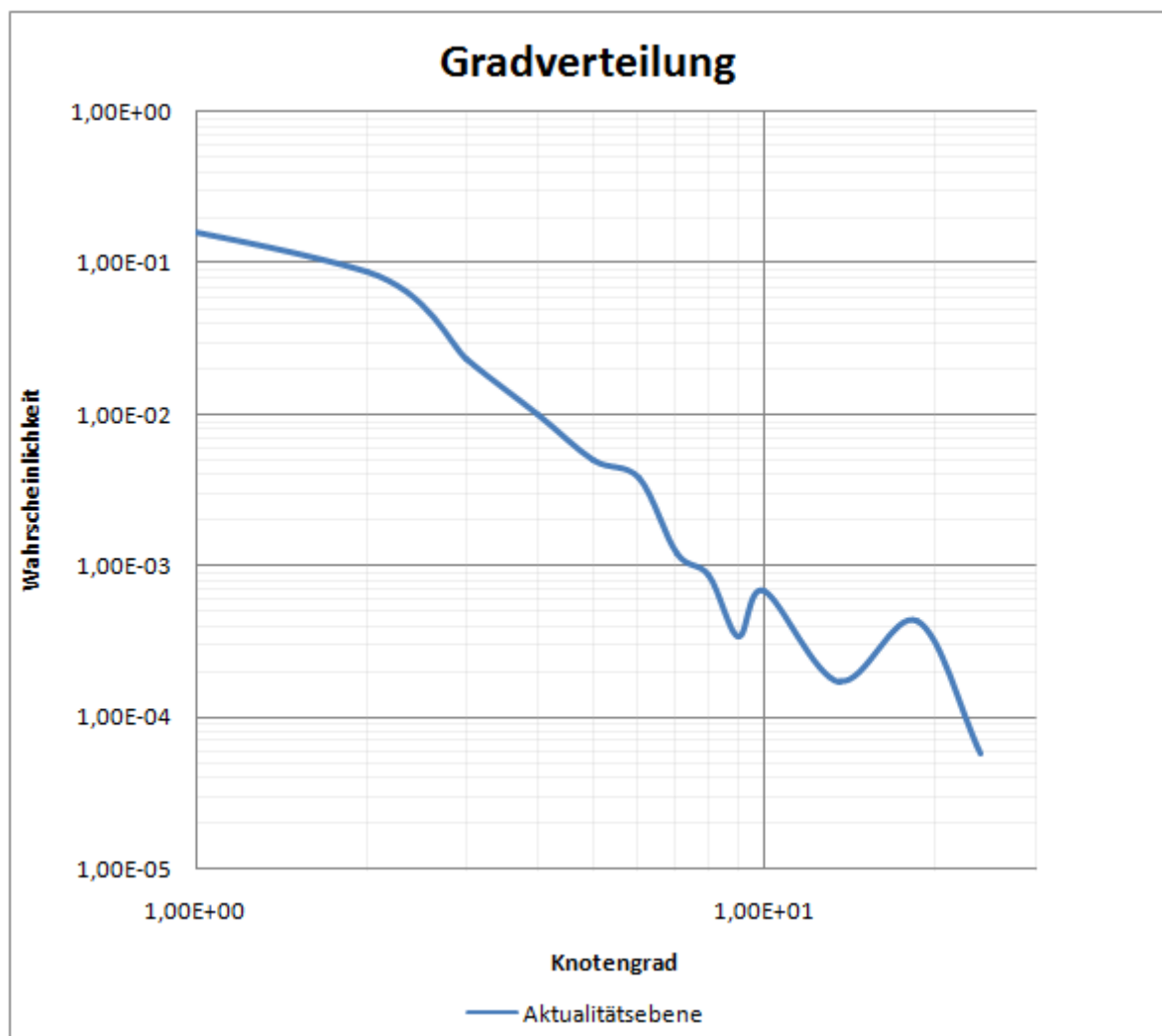


Abbildung 16: Gradverteilung (Aktualitätsebene)

Trotz der geringen Anzahl von Kanten, zeigt die Verteilung des Knotengrades einen annähernd geradelinigen Verlauf. Das Rauschen auf der rechten Seite der Kurve liegt an dem niedrigen durchschnittlichen Grad - es sind zu wenig Messwerte vorhanden, um ein

vernünftiges Histogramm zu erzeugen. Obwohl sich die Struktur der Aktualitätsebene von der Referenzebene stark unterscheidet, zeigt die grafische Darstellung eindeutig, dass dieses Netzwerk eine skalenfreie Gradverteilung hat. Diese Differenzen verursachen jedoch nicht, dass der *Preferential Attachment* Prozess verhindert wird. Die Ursache kann daran liegen, dass die bereits mit höherem Grad bezeichnete Dokumente (die z.B. viele andere RFCs verändern) von größeren Bedeutung sind und daher von den weiteren Memoranden öfter ersetzt bzw. aktualisiert werden. Aufgrund von der zu geringen Streuung von dem Zufallsgraph, wurde hier kein Vergleich durchgeführt. Der α -Exponent beträgt 2,0635 und die Ausgleichsgerade wird durch folgende Gleichung definiert:

$$f(x) = 0,0891x^{-2,0635}$$

Zusammenfassend, die Untersuchung der Gradverteilung von beiden Ebenen des Normennetzwerks ergab, dass sie dem Potenzgesetz folgt und als skalenfrei kategorisiert werden kann. Das bedeutet, dass ein Großteil von den RFC-Dokumentennur wenige Verknüpfungen mit den anderen Knoten des Graphen besitzt. Im Gegensatz zu den zufälligen Graphen, kommen in dem RFC-Normennetzwerk Memoranda mit einem sehr hohen Grad vor, die bei der ausschließlich auf der Wahrscheinlichkeitsrechnung basierenden Erzeugung nicht vorkommen würden. Nach dem *Preferential Attachment* Prinzip, mit dem Netzwerkwachstum gewinnen sie zusätzlich an Bedeutung und werden noch öfter verknüpft. Zusätzlich dienen solche Dokumente als *Hubs*, die die nicht miteinander verbundenen Knoten verlinken und dadurch die mittlere Weglänge des Netzwerks reduzieren.

Die Analyse von den wichtigsten Dokumenten aus beiden Ebenen ergab, dass die Sortierung nach dem Ingrad auf der Referenzebene und nach dem gesamten Grad auf der Aktualitätsebene die richtige Bedeutung von den RFCs wiedergibt. Diese Aussage trifft nicht zu, wenn die Knoten nach dem Ausgrad (Referenzebene) geordnet werden - ohne RFC-Verzeichnisse würde sich vernachlässigbar wenig im Internet ändern, ohne TCP oder IP - sehr viel. Die berechneten α -Exponenten, die die Steigung von den Verteilungskurven approximieren, liegen im Bereich von 1,8 bis 2,0635. Diese Werte liegen in derselben Größenordnung wie bei den anderen, in der Literatur vorkommenden Netzwerken, wie z.B. das Schauspielernetz (2,3), Internet (2,3) oder Zitationsnetzwerk(3,0). [12]

4.3 Ausfallsicherheit und Robustheit

In diesem Teil der Arbeit wird die Ausfallsicherheit des RFC-Normennetzwerks untersucht. In einem Stromnetzwerk verursacht ein Ausfall eines wichtigen Knoten erhebliche Probleme - möglicherweise bis zu einer Trennung vom kompletten Netzwerkteil. Analog, wenn ein wichtiger Bahnhof nicht durchgefahren werden kann, entfallen mehrere Verbindungen. Eine ähnliche Situation kann auch in dem RFC-Netzwerk vorkommen. Es kann zwar kein Dokument ausfallen, da im Internet mehrere offizielle und inoffizielle Mirrors von <http://rfc-editor.org> vorhanden sind, es kann auch keine bereits veröffentlichte Spezifikation verändert werden - die RFC-Dokumente sind durch zeitliche Invarianz charakterisiert. Was aber vorkommen kann ist die Einführung von den neuen Standards, die die alten und wichtigen (wie z.B. DNS, TCP, IP oder HTTP) ersetzen, ohne Rückwärtskompatibilität zu gewährleisten. In diesem Fall sollen alle auf dem veralteten Dokument basierende Texte aktualisiert bzw. angepasst werden. Der Standardisierungsprozess der IETF ist zwar so aufgebaut, dass die möglichen Auswirkungen von jeder Ersetzung berücksichtigt werden, jedoch in Zukunft kann man nicht ausschließen, dass die Rückwärtskompatibilität berücksichtigt wird.

Die Robustheit eines Netzwerks lässt sich auf zwei unterschiedliche Arten messen. Erstens, können die zufällig gewählten Knoten entfernt werden. Von einem skalenfreien Netzwerk kann man in solchen Fall erwarten, dass die durchschnittliche Weglänge zwischen Knoten davon unberührt bleibt. Derartige Ausfallsicherheit liegt an der Gradverteilung - die Mehrheit von Knoten in derartigen Graphen hat einen geringen Grad und ist daher von wenig Bedeutung für die gesamte Netzwerkstruktur. Bei der Entfernung von Knoten aus einem Zufallsgraphen kann man ein ähnliches Ergebnis erwarten, da dort alle Knoten von gleicher Wichtigkeit sind.

Der zweite Ansatz basiert auf der Entfernung von Knoten mit höchstem In- bzw. Ausgrad. Derartiges Vorgehen kann zu gravierenden Auswirkungen für ein skalenfreies Netzwerk führen, da sein Integrität auf den wichtigsten Knoten basiert. Wenn diese vernichtet werden, verlieren mehrere Netzwerkteile die Verbindung zueinander, was eine wesentliche Steigung der mittleren Weglänge mit sich bringt. Bei dem zufällig erzeugten Graphen soll im Gegensatz keine bemerkbare Auswirkung beobachtet werden, da die Bedeutung von den jeweiligen Knoten gleich ist.

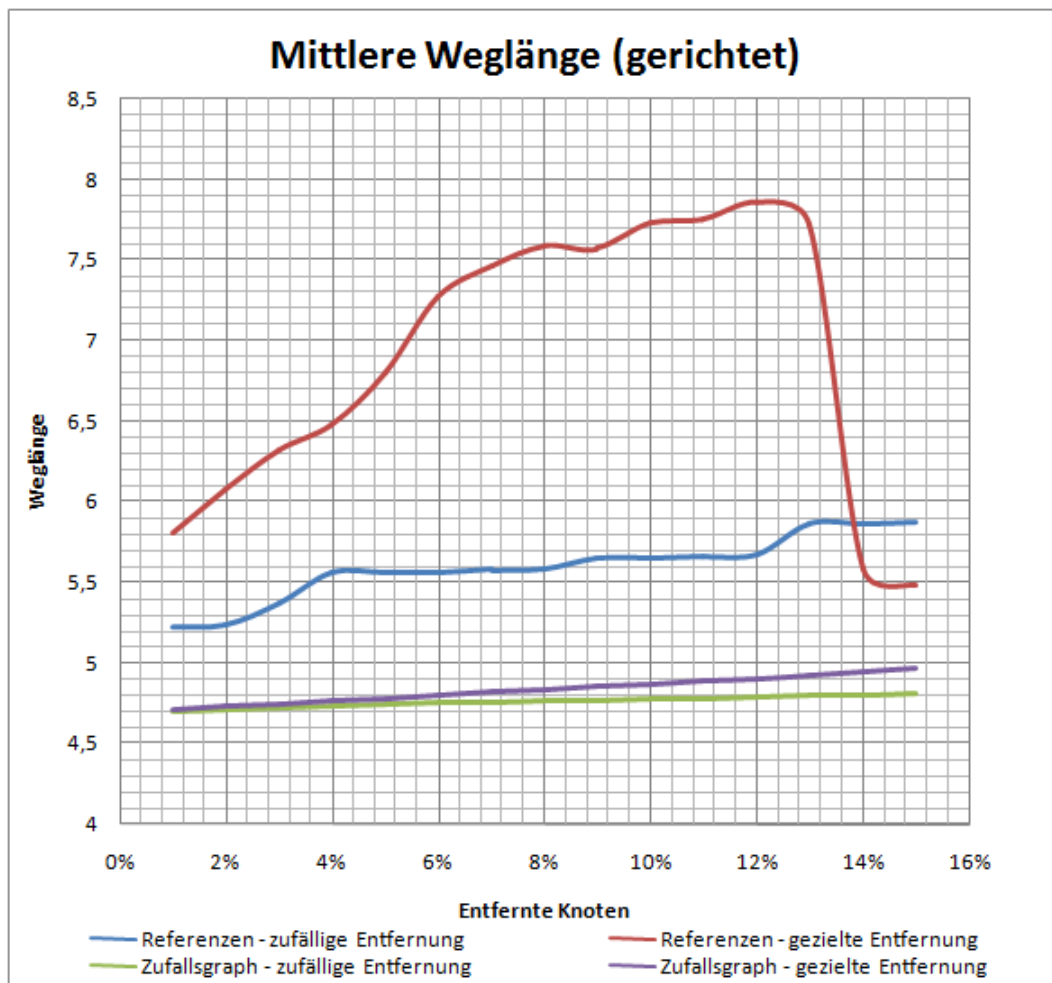


Abbildung 17: Mittlere Weglänge (Referenzebene gerichtet)

Um den Einfluß von der Knotenentfernung auf die Netzwerkstruktur zu untersuchen, wurden aus der jeweiligen Netzwerkebene immer 58 Knoten ($\sim 1\%$ der Netzwerkgröße) in insgesamt 15 Schritten gelöscht. Es wurden beide Löschrategien durchgeführt, d.h. sowohl die zufällige, als auch die auf die Knoten mit höchstem Grad gezielte Entfernung. Als Maß für die Änderungen von der Netzwerkstruktur wurde die mittlere Weglänge und der Watts-Strogatz Clusterkoeffizient gewählt. Diese geben an, wie stark die Dokumente vernetzt sind. Außerdem wurde die Referenzebene als gerichteter und ungerichteter Graph betrachtet, um die Wirkung der Kantenrichtung auf die Metriken zu analysieren. Die Ergebnisse wurden mit einem zufällig erzeugten Netzwerk verglichen, um den Unterschied zwischen den skalenfreien und zufälligen Graphen zu beobachten. Die Entfernung von Knoten erfolgte mit dem *Network Workbench Tool*, Berechnung - mit *Gephi*.

Das Löschen von zufälligen Knoten aus der Referenzebene führt lediglich zu ei-

ner geringen Steigerung der mittleren Weglänge. Merkwürdig sind die Sprünge, der Wert wächst nicht linear, sondern verschiebt sich schnell und bleibt für mehrere Löschvorgänge fast durchgehend konstant. Ein derartiger Verlauf wird wahrscheinlich dadurch verursacht, dass in bestimmten Schritten die wichtigen Knoten mit höherem Grad zufällig entfernt werden, was eine Verlängerung der mittleren Weglänge begründet. Das gezielte Löschen von Knoten mit dem höchsten Grad bringt einen viel steileren Verlauf. Bis zu etwa 13% kann man ein Wachstum der durchschnittlichen Weglänge beobachten, danach passiert ein unerwarteter, extrem schneller Rückgang.

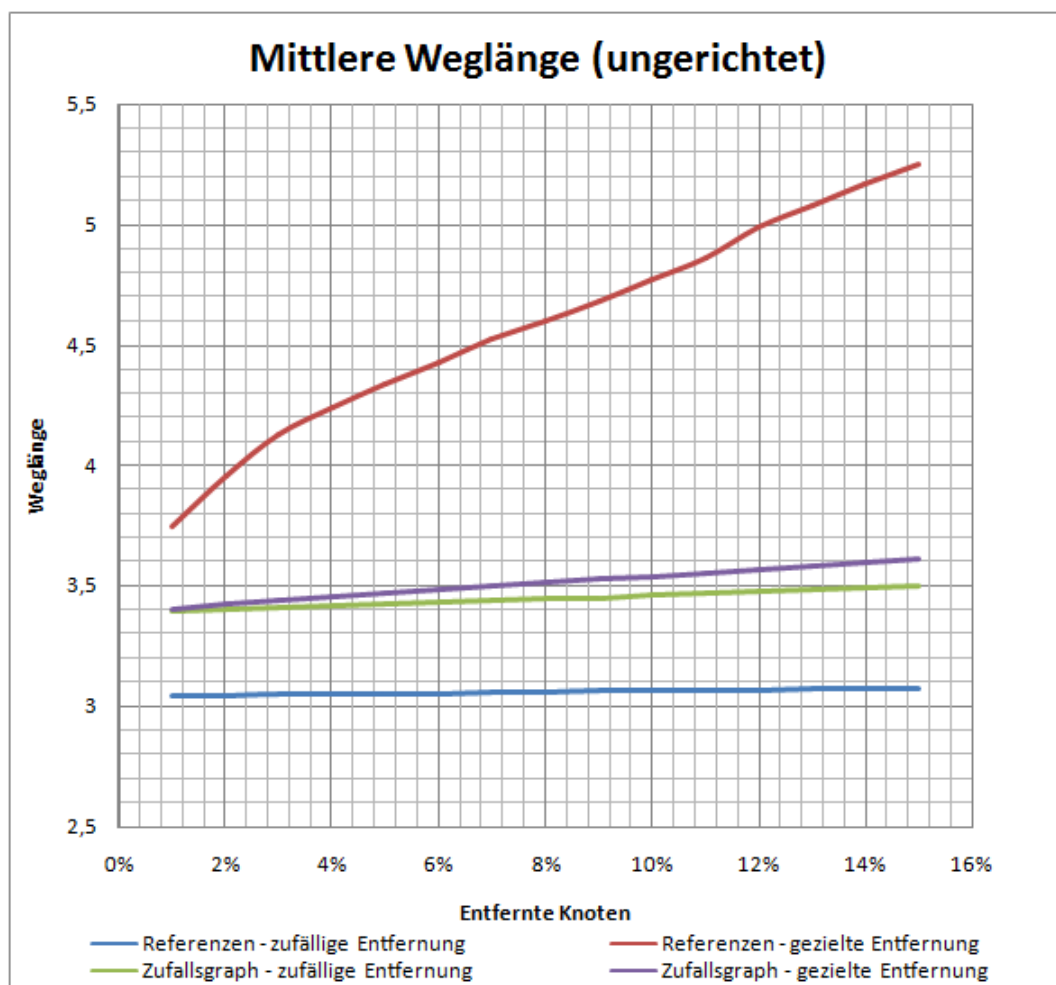


Abbildung 18: Mittlere Weglänge (Referenzebene ungerichtet)

Dieses Verhalten resultiert daraus, dass an dieser Stelle die wichtigen Knoten, die zwei Netzwerkteile verbinden, gelöscht wurden. Dadurch sind zwei separate Untergraphen entstanden, zwischen denen kein Pfad existiert. Das bedeutet aber nicht, dass sie keine gemeinsamen Knoten haben - das ist durchaus möglich, die Ecken sind jedoch so ausgerichtet, dass kein Weg durchgelaufen werden kann und wenn ein Graph in zwei getrennte Komponenten geteilt wird, senkt die mittlere Weglänge, da bei der Berechnung werden die Knoten des anderen Unternetzwerks nicht mehr berücksichtigt. Im Vergleich zu einem Zufallsgraphen wächst die durchschnittliche Entfernung zwischen Knoten viel schneller. Es liegt an der Gradverteilung, die beim zufälligen Einfügen von Kanten keine Entstehung von bevorzugten, wichtigen Knoten verursacht. Deshalb sind auch die Unterschiede zwischen beiden Löschvorgängen in diesem Fall gering.

Ohne Berücksichtigung der Kantenrichtung werden ähnliche Ergebnisse erhalten. Das gezielte Löschen von entscheidenden Knoten führt wieder zur deutlichen Verlängerung der mittleren Weglänge, wobei eine zufällige Entfernung kaum Einfluß auf diesen Wert hat - die Steigung der Gerade ist niedriger als bei dem gerichteten Graphen. Bei solcher Betrachtung kommt auch kein plötzlicher Abfall der Weglänge vor, da die Entfernung von 15% der wichtigsten Knoten nicht ausreichend ist, um das Normennetzwerk in zwei komplett separate Teile zu splitten. Das ist ein wichtiges Merkmal - es zeigt, dass die Verbindungen zwischen unterschiedlichen Dokumenten bzw. Gebieten zahlreich sind und nicht ausschließlich über die Knoten mit höherem Grad verlaufen - das weist auf die hohe Integrität des Normennetzwerks hin. Ein Vergleich mit dem Zufallsgraphen zeigt zwar, dass ein gezieltes Löschen zur schnelleren Steigung der mittleren Pfadlänge führt, aber das Normennetzwerk ist dafür weniger für die zufällige Entfernung von Knoten empfindlich - das ist auch eine Folge der Skalenfreiheit - im Zufallsgraphen sind alle Knoten vergleichbar wichtig, im derartigen Netzwerk - haben die meisten einen niedrigen Grad.

Der Clusterkoeffizient zeigt, wie stark die Knoten miteinander verbunden sind und ob dadurch die Cliques entstehen. Die natürlich wachsenden Netzwerke werden oft durch ein hohes Clustering bezeichnet, im Gegensatz besitzen die zufälligen Graphen deutlich kleineren Clusterkoeffizienten, der sich an die Null annähert. Die Entfernung von den zufälligen Knoten beeinflusst diesen Wert nur geringfügig, erst das Löschen von den wichtigen Dokumenten verursacht einen spürbaren Rückgang von dieser Metrik. Die stärkste Abnahme findet im Anfangsabschnitt der Kurve statt, also dort, wo die RFCs mit al-

lerhöchstem Grad aus dem Netzwerk gelöscht wurden. Dieser Sachverhalt lässt sich so interpretieren, dass die Knoten mit höherem Grad sich bei der Entstehung von Cliquen in dem Normennetzwerk beteiligen und eine wesentliche Rolle in der Netzwerkstruktur spielen. Trotzdem ist der Wert des Clusterkoeffizienten nach der Entfernung von 15% der wichtigsten Dokumente viel höher als der von einem Zufallsgraphen mit entsprechender Anzahl von Knoten. Das zeigt, dass die Small-World-Eigenschaften von dem Netzwerk sehr stark eingepreßt sind und nicht nur von den wichtigen Dokumenten initiiert werden, sondern auch auf den unteren Ebenen des Netzwerks ersichtlich sind. Dieser Fakt, in der Verbindung mit der zwar wachsenden, aber trotzdem niedriger mittleren Weglänge zeigt, dass auch nach der Entfernung von mehreren, entscheidenden Knoten das resultierende Netzwerk der Referenzen als *Small-World-Network* klassifiziert werden kann.

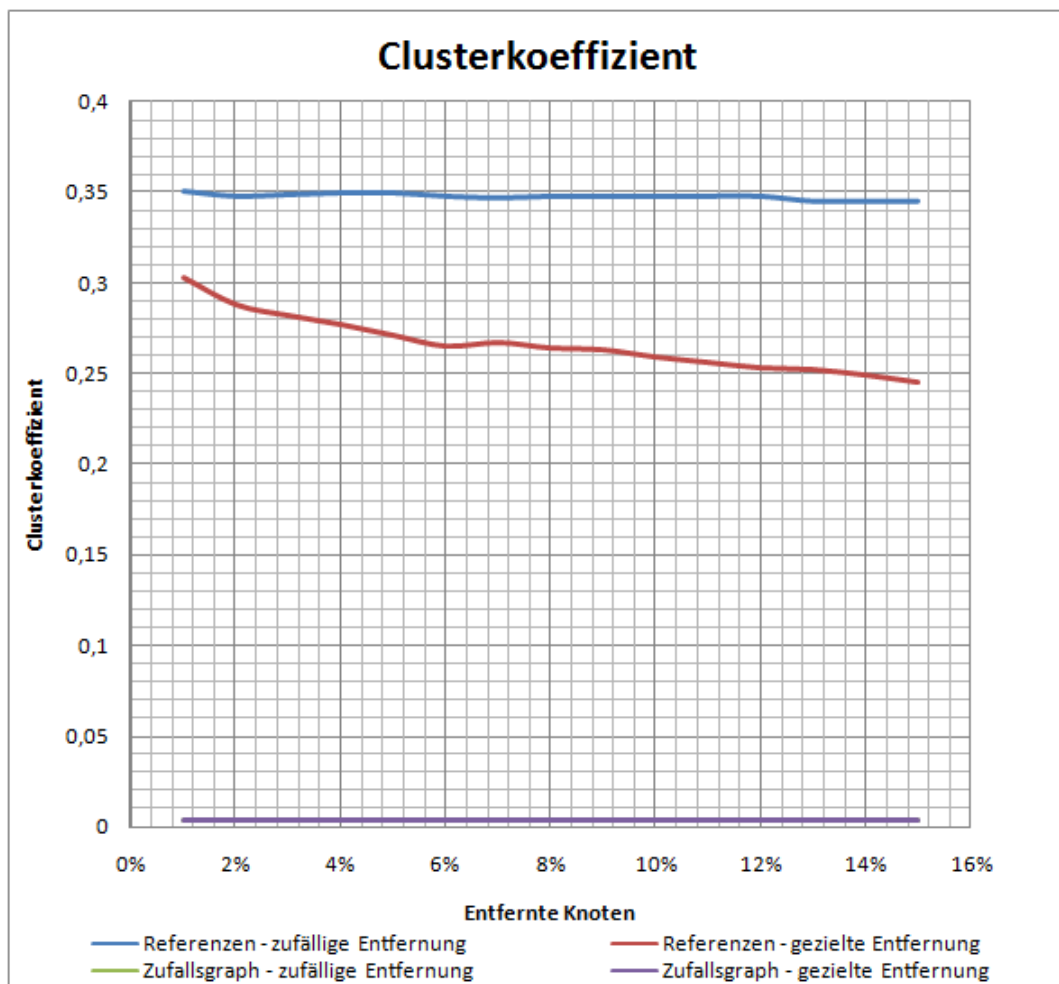


Abbildung 19: Clusterkoeffizient (Referenzebene)

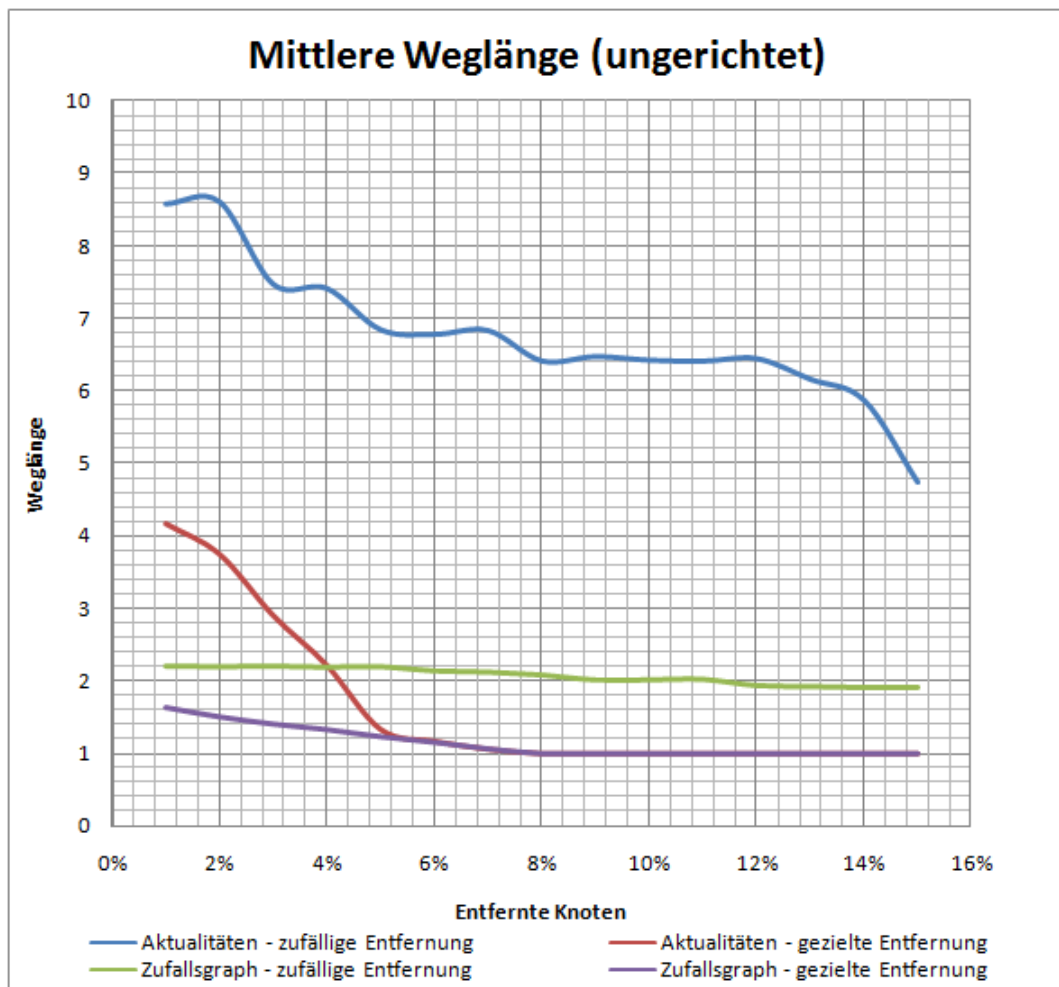


Abbildung 20: Mittlere Weglänge (Aktualitätsebene)

Der stark abweichende Aufbau der Aktualitätsebene wird in den Ergebnissen widerspiegelt. Wegen des geringen Kanten/Knoten-Verhältnisses, wird die mittlere Weglänge nach der Entfernung von Knoten immer geringer. Diese Tendenz kann man sowohl bei dem gezielten, als auch bei dem zufälligen Löschen beobachten. Der Grund ist, dass die isolierten Knoten, die bereits am Anfang ein Großteil dieser Netzwerkebene bilden, bei der Berechnung nicht berücksichtigt werden, da sie nicht erreichbar sind. Werden viele Knoten mit höherem Grad entfernt - wird das Netzwerk in mehrere unabhängige Komponenten geteilt, innerhalb von denen die Weglänge immer geringer wird, da die Knotenmenge eingeschränkt wird. Gezielte Entfernung hat sehr schwerwiegende Folgen für das Netzwerk. Bereits nach vier Schritten befindet sich die mittlere Weglänge auf dem Niveau des Zufallsgraphen, nach acht Schritten beträgt sie 1. Das bedeutet, dass lediglich die einzelnen, isolierten Knoten im Graphen geblieben sind. Diese Tendenz lässt sich auch gut beim ak-

zidentellen Löschen beobachten, dort wird jedoch nicht ein so niedriger Wert erreicht. Den Erwartungen zufolge, zeigen die Zufallsgraphen zwar eine Unempfindlichkeit für zufälliges Löschen, jedoch führt aufgrund der niedriger Kantenanzahl die Entfernung von den wichtigsten Knoten zum Erreichen eines völlig isolierten Netzwerks.

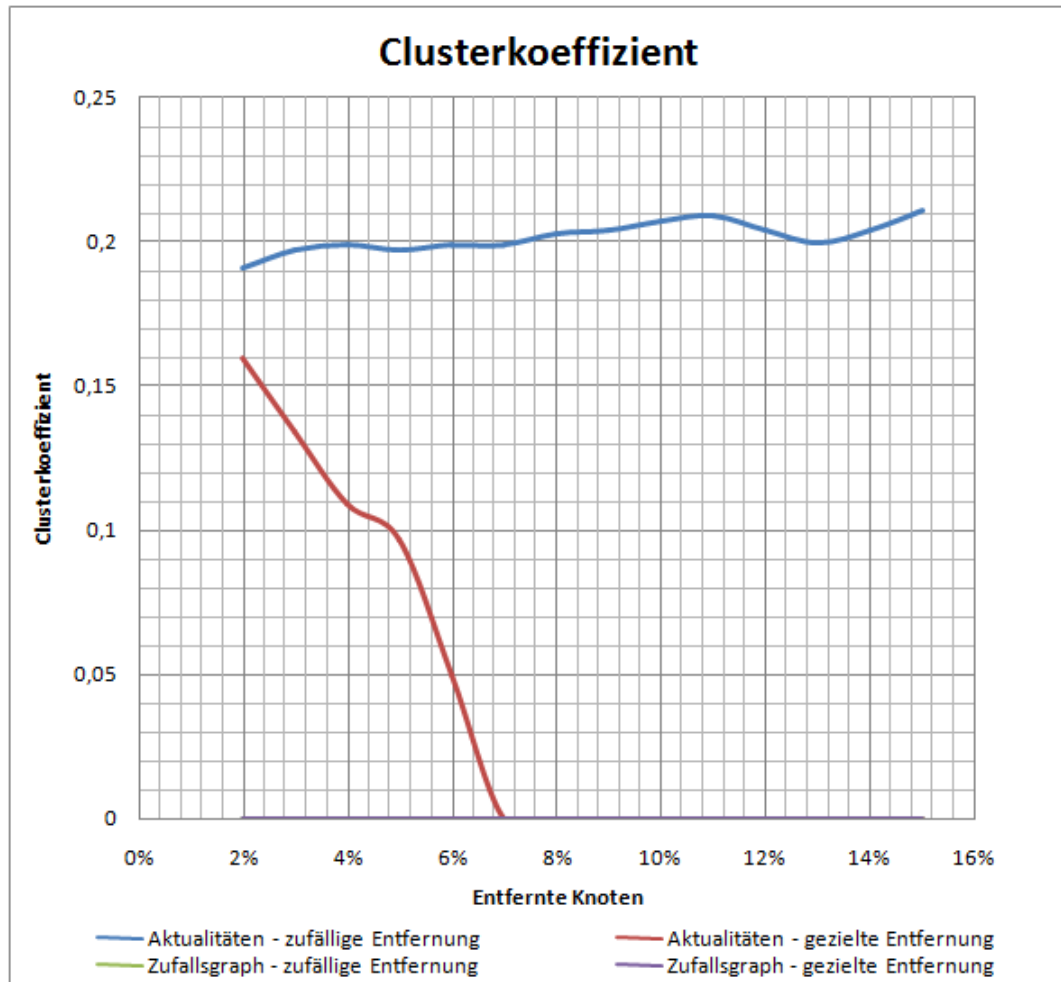


Abbildung 21: Clusterkoeffizient (Aktualitätsebene)

Die Auswertung des Clusterkoeffizienten führt zu einer interessanten Schlussfolgerung. Der Wert fällt zwar, wenn die wichtigen RFCs entfernt werden, wächst aber, wenn Knoten durch Zufall aus dem Graphen ausgenommen werden. Die Ursache liegt wieder an dem geringen durchschnittlichen Grad - Entfernung von einem Knoten, der zwischen zwei anderen lag, erhöht eventuell ihren Clusterkoeffizient, da im Nenner weniger mögliche Kanten stehen. Wenn die meisten Knoten bereits im Anfangszustand des Netzwerks isoliert sind, führt ihre Entfernung zum größeren gesamten Clusterkoeffizienten, weil ansonsten müssten derartige Dokumente mit $C_i = 0$ bezeichnet werden, was den Durchschnitt nach unten

zieht. Beim gezielten Löschen fällt der Wert des Clusterkoeffizienten schnell und deutlich, da die Knoten mit höherem Grad, die die lokalen Werte für ihre Nachbarn erhöhen, aus dem Graphen entnommen werden. Bereits bei 7% beträgt der Clusterkoeffizient 0 und gleicht sich somit mit dem Ergebnis für den Zufallsgraphen aus. In diesem Fall darf das Netzwerk nicht mehr als *Small-World* bezeichnet werden, da es keine Kanten besitzt und somit werden die grundlegenden Eigenschaften nicht eingehalten.

Die Untersuchung der Auswirkung von dem zufälligen Löschen von Knoten auf beiden Netzwerkebenen ergab, dass es nur einen sehr geringen Einfluss auf die Netzwerkeigenschaften hat. Dieses Verhalten bestätigt, dass das RFC-Normennetzwerk zu den skalenfreien Netzen gehört. Zahlreiches Vorkommen von Knoten mit niedrigem Grad verursacht, dass derartige Dokumente auch öfter beim Löschen gewählt werden und ihre Entfernung keine gravierende Folgen für das gesamte Netzwerk hat. Im Gegensatz, ein auf die wichtigsten Knoten gezielter Angriff verursacht eine deutliche Veränderung der Parameter. Die durchschnittliche Weglänge wächst (auf der Referenzebene) oder sinkt (auf der Aktualitätsebene) signifikant, die Kurve wird jedoch durch einen relativ linearen Verlauf charakterisiert. Daraus folgt, dass die durchschnittliche Weglänge proportional zur Anzahl der entfernten Knoten ist. In Abhängigkeit von dem Kanten/Knoten-Verhältnis kann der Vorfaktor positiv oder negativ sein.

Der Clusterkoeffizient der Referenzebene wird von der nicht gezielten Entfernung kaum beeinflusst, eine stärkere, aber nicht gravierende Auswirkung ist bei der anderen Eingriffsart bemerkbar. Der Wert sinkt im weiten Bereich proportional zur Anzahl der entfernten Knoten, die Geschwindigkeit der Abnahme ist aber ziemlich gering. Die größten Folgen hat der erste Löschvorgang, da er die am stärksten verlinkten Knoten als Ziel nimmt. Trotzdem besitzt diese Netzwerkebene einen viel höheren Koeffizient als ein vergleichbarer Zufallsgraph. Das trifft aber nicht zu, wenn die Aktualitätsebene betrachtet wird, dort hat die gezielte Entfernung drastische Folgen - der Wert nimmt so stark ab, dass der Nullpunkt erreicht wird. Das Netzwerk verliert dadurch seine *Small-World*-Eigenschaften und von einem zufälligen Graphen nicht unterschieden werden kann.

4.4 *Mixing Patterns*

Der Begriff *Mixing Patterns* stammt aus der Analyse von Sozialnetzwerken und gibt an, nach welchen Mustern sich unterschiedliche Knotenarten mit den anderen verbinden.

Diese können in vielen Netzwerkarten beobachtet werden und geben Auskunft über die Struktur des Netzwerks. Bei einem zufällig generierten Graphen ist die Häufigkeit von Verbindungen zwischen verschiedenen Knotentypen proportional zu dem prozentuellen Anteil von diesen. Ein solches Verhalten resultiert daraus, dass die Wahrscheinlichkeit von der Entstehung einer Kante zwischen einem Knotentyp A und B folgendem Wert entspricht,

$$P(A \rightarrow B) = \frac{|A|}{n} \cdot \frac{|B|}{n} \quad (9)$$

weil die Auswahl von Knoten durch Zufall passiert. Im Gegensatz, bei den Netzwerken aus der echten Welt ist es oft möglich, eine Korrelation zwischen Knotenarten zu beobachten. Ein gutes Beispiel dafür ist ein von M. Morris untersuchte Netzwerk von Ehepaaren, wo die ethnische Herkunft als Knotentyp dient. [25] In diesem Graphen ist eine Tendenz, dass die Knoten sich viel häufiger mit anderen derselben Art verbinden, zu beobachten. Ein ähnlicher Vergleich kann auch mit dem Normennetzwerk durchgeführt werden, indem man den Status von den jeweiligen Dokumenten als Knotentyp interpretiert.

4.4.1 Knotenartkorrelation

Als Maß für die Korrelationen, wo keine numerische Bezeichnung des Typs vorhanden ist, wird der Assortativitätskoeffizient r verwendet. [12] Der Wertebereich von r liegt zwischen 0 und 1, je höher das Ergebnis, desto stärker neigen sich die Knoten dazu, mit den anderen von derselben Art zu verbinden. Der Wert von r ist folgendermaßen definiert:

$$r = \frac{\text{spur}(e) - \|e^2\|}{1 - \|e^2\|} \quad (10)$$

e ist eine normierte Matrix von Ergebnissen, in der die jeweiligen Anzahlen von Ecken bestimmter Art in der Ergebnismatrix E durch die Summe von allen Elementen geteilt werden. spur ist die Spurfunktion, die alle Elemente auf der Diagonale einer Matrix summiert.

$$e = \frac{E}{\|E\|} \quad (11)$$

Typ	Anteil
<i>Best Current Practice</i>	3,1%
<i>Draft Standard</i>	2,4%
<i>Experimental</i>	5,9%
<i>Historic</i>	3,9%
<i>Informational</i>	31,4%
<i>Proposed Standard</i>	36,1%
<i>Standard</i>	1,6%
<i>Unknown</i>	15,7%

Tabelle 6: Verteilung von Knotenarten

Aus der Aufteilung folgt, dass das RFC-Normennetzwerk zwei führende Knotenarten besitzt, *Informational*- und *Proposed Standard*- Memoranda bilden über 67% des gesamten Graphen. Falls keine Korrelation zwischen den Knotenarten existiert, sollen sie auch zu den am häufigsten verlinkten Typen gehören. Der Anteil von Knoten mit einem unbekannten Status ist relativ hoch. Das kann die Genauigkeit von den Ergebnissen negativ beeinflussen, da die als *Unknown* bezeichnete Dokumente können eventuell als jeder von den anderen Typen klassifiziert werden. Merkwürdig ist auch die hohe Diskrepanz innerhalb vom *Standard Track*. Obwohl die Standardvorschläge selbst 36,1% des Normennetzwerks bilden, lediglich 2,4% sind in eine weitere Anerkennungsphase übergegangen und nur 1,6% zum offiziellen Standard wurden. Das zeigt wie häufig sich die Standardvorschläge nicht durchsetzen können.

Diese Untersuchung wird auf beiden Netzwerkebenen durchgeführt. Von dem Referenznetzwerk soll keine starke Korrelation erwartet werden, da es skalenfrei ist und viele relativ unwichtige Dokumente mit unterschiedlichen Stati zahlreiche die relativ kleine Gruppe von den Knoten mit höherem Grad referenzieren. Außerdem, der hohe Clusterkoeffizient weist darauf hin, dass die Knoten in zahlreiche, thematisierte Gruppen aufgeteilt sind, wo sich möglicherweise unterschiedliche Dokumententypen gegenseitig zitieren, weil ihr Themenbereich sich überschneidet. Die Aktualitätsebene soll eine höhere Korrelation aufweisen, weil falls die Dokumente vom bestimmten Typ ersetzt werden, kann man erwarten, dass das neue Memorandum möglicherweise von gleicher Art ist - es ist z.B. wahrscheinlich, dass wenn ein Text aus dem *Standard Track* aktualisiert wird, wird das neue Dokument auch dort aufgenommen.

zu / von	<i>BCP</i>	<i>Draft Std.</i>	<i>Experimental</i>	<i>Historic</i>	<i>Informational</i>	<i>Proposed Std.</i>	<i>Standard</i>	<i>Unknown</i>
<i>BCP</i>	0,79%	0,21%	0,60%	0,07%	2,52%	5,40%	0,08%	0,00%
<i>Draft Standard</i>	0,24%	0,68%	0,42%	0,19%	2,29%	4,50%	0,15%	0,02%
<i>Experimental</i>	0,12%	0,06%	0,55%	0,13%	1,33%	0,72%	0,01%	0,08%
<i>Historic</i>	0,12%	0,24%	0,21%	1,08%	1,75%	1,27%	0,14%	0,49%
<i>Informational</i>	0,66%	0,27%	0,70%	0,50%	7,34%	5,35%	0,09%	0,02%
<i>Proposed Standard</i>	1,27%	0,83%	1,72%	0,91%	12,38%	20,72%	0,23%	0,03%
<i>Standard</i>	0,36%	0,64%	0,75%	1,30%	3,11%	5,36%	0,49%	0,71%
<i>Unknown</i>	0,01%	0,14%	0,14%	1,24%	2,51%	0,37%	0,34%	3,05%

Tabelle 7: Normierte Mischmatrix (Referenzebene)

	<i>BCP</i>	<i>Draft Std.</i>	<i>Experimental</i>	<i>Historic</i>	<i>Informational</i>	<i>Proposed Std.</i>	<i>Standard</i>	<i>Unknown</i>
<i>BCP</i>	2,54%	0,14%	0,09%	0,06%	0,61%	0,38%	0,00%	0,00%
<i>Draft Standard</i>	0,14%	1,56%	0,17%	0,06%	0,20%	4,59%	0,98%	0,43%
<i>Experimental</i>	0,09%	0,17%	0,87%	0,14%	0,49%	1,56%	0,26%	0,17%
<i>Historic</i>	0,06%	0,06%	0,14%	3,58%	0,87%	0,81%	0,06%	0,55%
<i>Informational</i>	0,61%	0,20%	0,49%	0,87%	6,98%	1,59%	0,12%	0,29%
<i>Proposed Standard</i>	0,38%	4,59%	1,56%	0,81%	34,51%	2,28%	2,31%	0,46%
<i>Standard</i>	0,00%	0,98%	0,26%	0,06%	0,12%	0,46%	0,35%	0,63%
<i>Unknown</i>	0,00%	0,43%	0,17%	0,55%	0,29%	0,46%	0,63%	13,65%

Tabelle 8: Normierte Mischmatrix (Aktualitätsebene)

Aus der normierten Mischmatrix für die Referenzebene ergibt sich, dass die am häufigsten im gesamten Netzwerk vorkommenden *Proposed Standard*-Dokumente auch sehr oft zitiert werden - in 5 von 8 Fällen befinden sie sich auf dem ersten Platz. Die historischen Memoranda beziehen sich häufig auf die Standards, was durch ihr Alter erklärt werden kann - die dort zitierten Texte wurden im Laufe der Zeit anerkannt. Die Standards referenzieren oft sich selbst, was möglicherweise daran liegt, dass die dort enthaltenen Thesen eine begründete Position haben müssen. Dadurch, dass die unbekannten Memoranda praktisch nur auf sich selbst beziehen, kann ihr negativer Einfluss auf die Ergebnisse ausgeschlossen werden. Außerdem ist die Diskrepanz zwischen *Proposed Standard* und Informationale Dokumenten merkwürdig. Obwohl sie einen sehr ähnlichen, prozentuellen Anteil besitzen, wird die zweite Dokumentenart um Faktor 2 bis 4 seltener zitiert. Daraus folgt, dass die informationellen Memoranda, trotz ihrer zahlreichen Vorkommen weniger Bedeutung für das gesamte Netz haben. Der Assortativitätskoeffizient beträgt für diese Netzwerkebene $r = 0,15$. Das ist ein weiterer Hinweis, dass keine Korrelation zwischen Knotenarten im Referenznetzwerk vorhanden ist.

Eine völlig andere Situation kann man auf der Aktualitätsebene beobachten. Dort zeigen 5 von 8 Knotenarten eine Korrelation und verbinden sich am häufigsten mit Dokumenten derselben Art. Diese auf der Diagonale liegende Werte sind deutlich höher als die, die bei den anderen Typen vorkommen. Die Ausnahmen lassen sich auch einfach begründen - *Draft Standards* und *Proposed Standards* sind nebeneinanderliegende Stufen des *Standard Tracks* und die häufige Mischung von diesen zwei Arten kann durch eine Weiterentwicklung erklärt werden. Dieselbe Situation kommt bei dem Paar *Experimental* und *Proposed Standard* - die Verbesserungen in den experimentellen Spezifikationen werden in den *Standards Track* aufgenommen. In diesem Netzwerk tendieren die unbekannten Dokumente auch dazu, sich mit gleichartigen zu verbinden, die Anzahl von Kanten zu anderen Typen ist marginal. Dadurch kann hier auch ein negativer Einfluss ausgeschlossen werden. Der Assortativitätskoeffizient beträgt $r = 0,509$. Dieser Wert ist viel höher als für das Zitationsnetzwerk und weist eindeutig auf eine Korrelation zwischen Knotenarten auf dieser Netzwerkebene hin. Somit wird die Interpretation der Ergebnistabelle bestätigt.

4.4.2 Knotengradkorrelation

Ein weiterer Aspekt der Abhängigkeit zwischen Netzwerkknoten ist die Knotengradkorrelation. Hier wird untersucht, ob sich die Dokumente mit einem bestimmten In-, bzw.

Ausgrad häufiger mit denen, die durch einen ähnlichen Grad charakterisiert werden, verbinden. Dieser Wert wird mit dem in der Statistik üblichen Pearson-Koeffizient ρ bemessen. Er nimmt die Werte von -1 bis 1 an. Ein negatives Ergebnis bedeutet eine inverse Abhängigkeit, in diesem Fall - Knoten mit höherem Grad verbinden sich mit Knoten mit niedrigem Grad. Positiver Wert zeigt eine einfache Korrelation. Je höheren Wert $|\rho|$ annimmt, desto stärker ist die Abhängigkeit von den Werten.

Eine starke Knotengradkorrelation bedeutet, dass in dem Netzwerk relativ unabhängige und kaum verbundene Knotenebenen vorhanden sind, die aus den Dokumenten in einem bestimmten, kleinen Gradbereich bestehen. Aufgrund der unterschiedlichen Struktur von beiden Netzwerkebenen können auch diverse Varianten untersucht werden. Für das Zitationsnetzwerk sind das:

- **Ingrad** → **Ingrad**-Korrelation - hier wird überprüft, ob der Ingrad von Knoten ein Faktor ist, der die Entstehung von den eingehenden Kanten beim anderen Knoten beeinflusst.
- **Ausgrad** → **Ingrad**-Korrelation - Abhängigkeit vom Ausgrad eines Knoten mit einer herausgehenden Kante mit dem Ingrad des Knoten mit einer hineingehenden Kante.
- **Ingrad** → **Ausgrad**-Korrelation - wie oben, aber es wird der Einfluss vom Ingrad des Startknoten auf den Ausgrad des Endknoten überprüft. Da das Netzwerk gerichtet ist, unterscheidet sich diese Korrelation von der obigen.
- **Ausgrad** → **Ausgrad**-Korrelation überprüft, inwieweit sich die Ausgrade von den jeweiligen Knoten gegenseitig beeinflussen.

Da die Aktualitätsebene lediglich aus ungerichteten Kanten besteht, können dort keine von den genannten Kombinationen angewendet werden und der einzige Korrelationskoeffizient, der dort berechnet werden kann, ist die Abhängigkeit vom gesamten Grad zwischen den Knoten.

Da die angewendete Software nicht über eine Funktion zur Berechnung von unterschiedlichen Knotengradkorrelationen verfügt, wurden die jeweiligen Grade anhand von den extrahierten Daten berechnet und in einer tabellarischen Form zusammengefasst.

Zum Beispiel: für einen Knoten A mit Grad 3, der mit B , C und D verbunden ist (Grade: 1,2,4) nimmt die Ergebnistabelle folgende Form:

Grad 1	Grad 2
3	1
3	2
3	4

Tabelle 9: Beispiel der Ergebnistabelle

Eine solche Darstellung ermöglicht die spätere Auswertung von den Ergebnissen mit der Funktion `CORREL()` der Anwendung Microsoft Excel.

Referenzebene	
Korrelation	ρ
Ingrad $\rightarrow$ Ingrad	-0,0205
Ingrad $\rightarrow$ Ausgrad	-0,0085
Ausgrad $\rightarrow$ Ingrad	-0,0773
Ausgrad $\rightarrow$ Ausgrad	-0,0811
Aktualitätsebene	
Grad $\rightarrow$ Grad	0,301

Tabelle 10: Pearson-Koeffizienten des Normennetzwerks

Die Knotengrade auf der Referenzebene zeigen keine Korrelation, unabhängig von der gewählten Variante der Überprüfung. Dieses Verhalten lässt sich durch die Skalenfreiheit und die Small-World-Eigenschaften erklären - die Knoten mit höherem Grad dienen als lokale Zentren, an die, nach dem *Preferential Attachment*-Prinzip, neue Dokumente mit niedriger Anzahl von hinaus-, bzw. hineingehenden Kanten angebunden werden. Das selbst würde jedoch zu einem negativen Wert des Pearson-Koeffizientes führen. Dieser Einfluss wird durch den relativ hohen Wert des Clusterkoeffizientes (auch nach der gezielten Entfernung von den wichtigsten Knoten) ausgeglichen. Diese Metrik weist darauf hin, dass die Dokumente sich nicht nur mit den entscheidenden RFCs verbinden, sondern auch mit denen, die von einem niedrigeren Grad charakterisiert werden. Die beiden Faktoren verursachen, dass ein Dokument gemeinsame Ecken mit den anderen Texten von unterschiedlichen Graden besitzt. Diese Unregelmäßigkeit spiegelt sich in den Werten des Pearson-Koeffizienten wider.

Im Vergleich zur Referenzebene hat die Aktualitätsebene einen höheren Pearson-Koeffizient, der jedoch auch auf keinen eindeutigen Zusammenhang zwischen den Knotengraden hinweist. Als Erklärung für diesen Unterschied können zwei Gründe genannt werden. Erstens, dieses Netzwerk besitzt deutlich weniger Kanten als die andere Ebene und hat eine andere, baumförmige Struktur. Zweitens, die Streuung der Werten ist um etwa den Faktor 100 geringer - eine Verbindung zwischen Knoten mit Graden 20 und 10 wirkt auf den Pearson-Koeffizient anders, als wenn Knoten mit den Graden 20 und 2000 verlinkt sind. Zusätzlich, durch die baumartige Struktur bedingt, kann aufgrund von Verzweigungen ein tatsächlicher, leichter Zusammenhang zwischen den Knotengraden bestehen. Anschließend müssen noch die zahlreichen, isolierten Knotenpaaren berücksichtigt werden, die nur mit einer Ecke verbunden sind und somit denselben Grad besitzen.

Zusammenfassend, in Abhängigkeit von der Netzwerkebene können unterschiedliche Eigenschaften beobachtet werden. Die Referenzebene zeigt einige für ein zufällig erzeugtes Netz typische Merkmale - es existiert keine Knotenartkorrelation und die am häufigsten vorkommende Knoten werden auch am häufigsten mit den anderen verlinkt. Diese Proportionalität ist jedoch nicht linear, da die fast mit gleicher Wahrscheinlichkeit vorkommende *Informational*-Memoranda viel seltener referenziert werden. Diese Charakteristik unterscheidet das RFC-Normennetzwerk von dem Zufallsgraphen, wo ein ähnliches Verhältnis von gleich oft vorkommenden Knoten zu erwarten ist. Zusätzlich ist keine Knotengradkorrelation vorhanden, weil die häufig zitierten Dokumente eine gemeinsame Ecke mit vielen anderen Texten besitzen, die durch eine hohe Streuung von Graden charakterisiert werden.

Die Aktualitätsebene besitzt, im Gegenteil zur Referenzebene, eine relativ starke Korrelation von Knotentypen. Sie liegt an der Funktion des Netzwerks, die die Aktualisierungen von den RFC-Dokumenten widerspiegelt. Deshalb kann man erwarten, dass wenn ein Memorandum ersetzt wird, der neue Text in dieselbe Kategorie eingeteilt wird. Zusätzlich zeigt dieses Netzwerk eine Knotengradkorrelation, die zwar schwach ist, aber trotzdem stärker als auf der anderen Netzwerkebene. Die niedrigen Werte von dieser Korrelation sind typisch für Graphen, die keine sozialen Netzwerke darstellen. [12] In der Literatur fehlt jedoch die Erklärung von diesem Sachverhalt.

5 Zeitliche Entwicklung des RFC-Normennetzwerks

Jedes RFC-Dokument ist mit einem Veröffentlichungsdatum bezeichnet, außerdem sind die Texte unveränderbar und jede Korrektur bzw. Verbesserung setzt eine Publikation von neuem Memorandum voraus. Dadurch bleibt der Ausgrad von den jeweiligen Knoten konstant. Diese zwei Eigenschaften des Normennetzwerks ermöglichen eine Darstellung von dem Graphen zu jedem, beliebig gewähltem Zeitpunkt. Die einzige Charakteristik, die sich zeitlich verändert ist der Status eines Dokumentes, z.B. bei dem im Rahmen des *Standards Track* Anerkennungsprozesses. Er beeinflusst aber die Struktur des Netzwerks nicht und verändert keine statistische Metriken, außer denen, die eine Knotenartkorrelation wiedergeben.

5.1 Netzwerkwachstum - Pricesches Modell

Die Referenzebene von den RFC-Dokumenten ist ein skalenfreies Netz, was durch die Beobachtung der Gradverteilung nachgewiesen wurde. Derartige Netzwerke befinden sich seit 1965 im Interessenkreis von den Forschern. In diesem Jahr veröffentlichte Derek de Solla Price seine Arbeit, in der er das Zitationisnetzwerk von wissenschaftlichen Publikationen ausgewertet hat. [6] Das Ergebnis von dieser Studie zeigte, dass sowohl die In-, als auch die Ausgrade dem Potenzgesetz folgen. Um die Ursache von einer solchen Verteilung zu finden, veröffentlichte er im Jahre 1976 eine weitere Publikation, die das Modell von der Entstehung von derartigen Netzwerken beschreibt. [19] Das von ihm vorgeschlagene *Cumulative Advantage* Schema entspricht dem *Preferential Attachment* Prinzip und basiert auf den vorherigen Arbeiten von Herbert Simon, der die Reichtumsverteilung betrachtete und den Satz "*the rich gets richer*" formulierte.

Das Pricesche Modell kann mit dem Wachstum der Referenzebene überprüft werden, da dieses Netz, von dem Aufbau her, einem von Price erforschten Zitationsnetzwerk entspricht - die Dokumente dienen als Knoten, Kanten sind gerichtet und entsprechen den Referenzen. Zudem sind die Inhalte durch zeitliche Invarianz charakterisiert und das Netzwerk kann zu jedem Zeitpunkt dargestellt werden. Diese Bedingungen führen dazu, dass die Referenzebene des RFC-Normennetzwerk als guter Vergleichswert für dieses Modell dienen kann. Diese Aussage trifft jedoch nicht zu, wenn man die Aktualitätsebene betrachtet. Die dortigen Relationen zwischen den Dokumenten sind ungerichtet, was zwei Voraussetzungen von dem Priceschen Modell verletzt: erstens, Kanten sind ungerichtet

und zweitens, der Ausgrad von den jeweiligen Knoten bleibt nicht konstant mit der Ablauf der Zeit - da die Relationen nicht mit Richtung vorgesehen sind, ist der Ingrad gleich dem Ausgrad und jede Aktualisierung eines RFC-Dokumentes eine Veränderung von diesem Parameter verursacht.

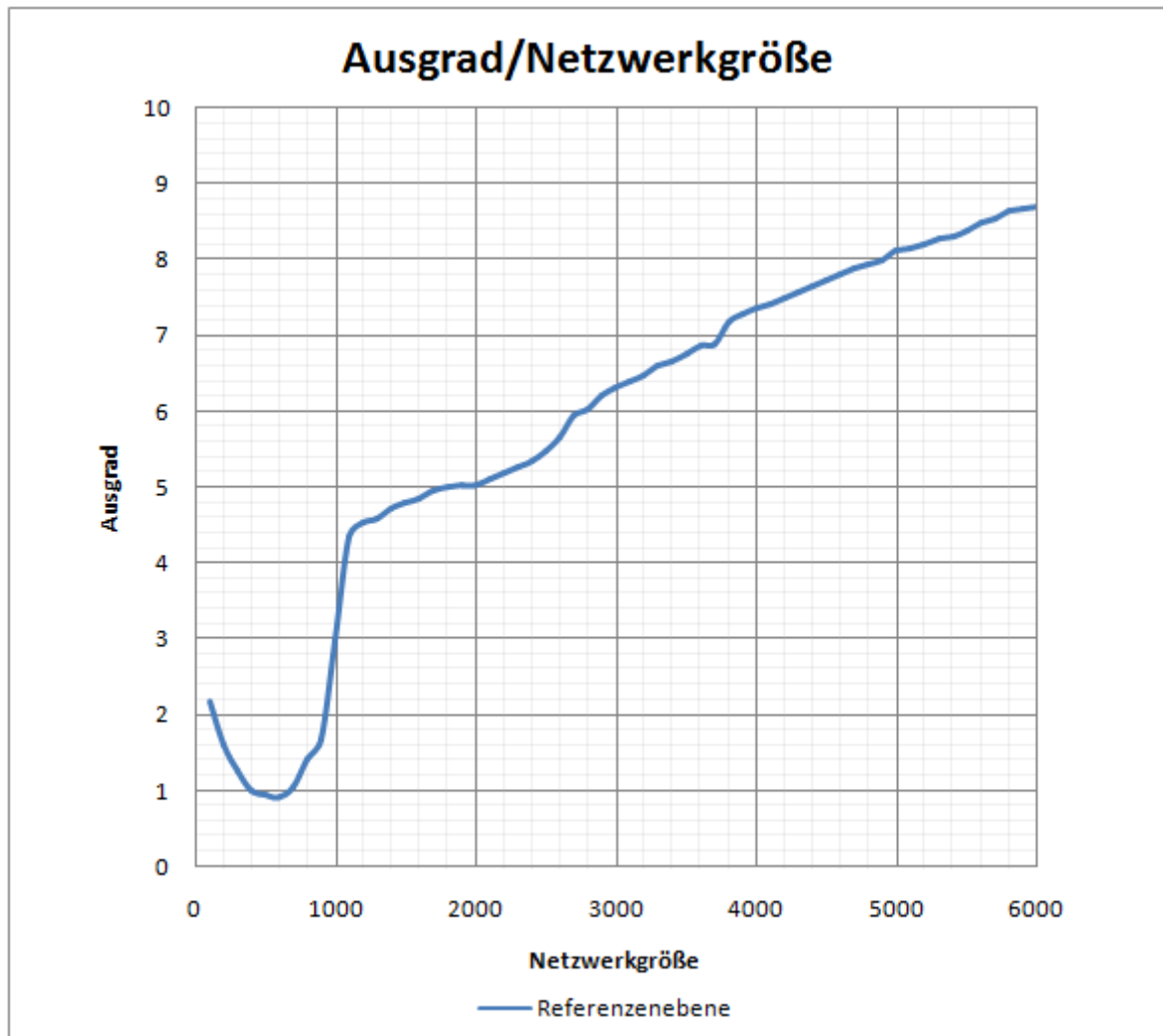


Abbildung 22: Durchschnittlicher Ausgrad in Abhängigkeit von der Netzwerkgröße

Das Pricesche Modell setzt voraus, dass der durchschnittliche Ausgrad m_a zu jedem Zeitpunkt konstant bleibt. [12] Diese Netzwerkeigenschaft kann sehr einfach überprüft werden - wenn man die Referenzen zeitlich betrachtet, ist der Anfangsknoten immer neuer, bzw. gleich alt wie der Endknoten. Es liegt daran, dass die Dokumente in bestimmter Reihenfolge veröffentlicht werden und es darf kein Memorandum zitiert werden, das neuer

als der zitierende Text ist. Es gibt nur eine Möglichkeit, dass ein Dokument ein anderes mit höherer Nummer referenziert. Diese Abweichung passiert, wenn die Texte gleichzeitig, in einem Bündel veröffentlicht werden. In dem gesamten Netzwerk gibt es lediglich 1500 solche Verbindungen, was 3% der gesamten Anzahl von Kanten entspricht und beeinflusst somit das Endergebnis nur sehr gering. Die Auswertung von $m_{a|t}$ zum bestimmten Zeitpunkt t erfolgte, indem die Kanten, die von den Dokumenten mit dem Veröffentlichungsdatum $\leq t$ herausgehen, gezählt und durch die Anzahl von diesen Dokumenten dividiert wurden.

Die Auswertung von dem durchschnittlichen Ausgrad der Referenzebene ergibt, dass er nicht konstant bleibt. Bis zur etwa 500 veröffentlichten RFCs fällt er exponentiell ab und danach steigt plötzlich. Dieses Ereignis kann mit Sicherheit dem Erscheinen von RFC 1012 zugeordnet werden, weil dieses Dokument ein Inhaltsverzeichnis ist und über 900 andere Texte referenziert. Der weitere Verlauf der Kurve ist sehr linear und wächst um etwa 1 für jede 1000 veröffentlichte Texte. Als Grund für dieses Wachstum kann die steigende Komplexität und unterschiedlicher Status von den Dokumenten angenommen werden. Im Gegensatz zu den wissenschaftlichen Arbeiten, die sich ständig nur auf kleinem Themenbereich konzentrieren, müssen die RFC-Dokumente die stets wachsende Basis beachten, wo die Aktualisierungen eine baumartige Form annehmen und dadurch eine Vervielfachung der zu zitierenden Dokumenten verursachen. Diese Tendenz kann insbesondere bei den *Informational* und *Proposed Standard* Texten beobachtet werden, die einen überdurchschnittlichen Ausgrad besitzen. Als weiterer Aspekt des Priceschen Modells wird das Netzwerk auf den *Preferential Attachment* Prozess getestet. Er besagt, dass die Knoten, die bereits viele Referenzen haben, immer öfter zitiert werden. Diese Regelmäßigkeit liegt daran, dass mit der steigenden Anzahl von zitierenden Dokumenten erhöht sich die Wahrscheinlichkeit, dass eine bestimmte Publikation von einem potenziellen Autor bemerkt wird - und das widerspiegelt sich in der Wahrscheinlichkeit, dass sie in den Arbeiten von diesem Wissenschaftler zitiert wird. Dieser Faktor lässt sich für die einzelnen Dokumente relativ leicht bestimmen, indem man für einen bestimmten, kleinen Zeitraum $t_1 - t_2$ die Anzahl von Zitierungen berechnet und diesen in dem kompletten Bereich von der Veröffentlichung bis zum Erscheinen von dem letzten RFC-Dokument wiederholt. Falls die Aussage "*richer gets richer*" wahr ist, soll die Steigung der resultierenden Kurve größer als 0 sein. Ein negatives Ergebnis würde zeigen, dass trotz steigender Anzahl von Referenzen der Text immer seltener berücksichtigt wird.

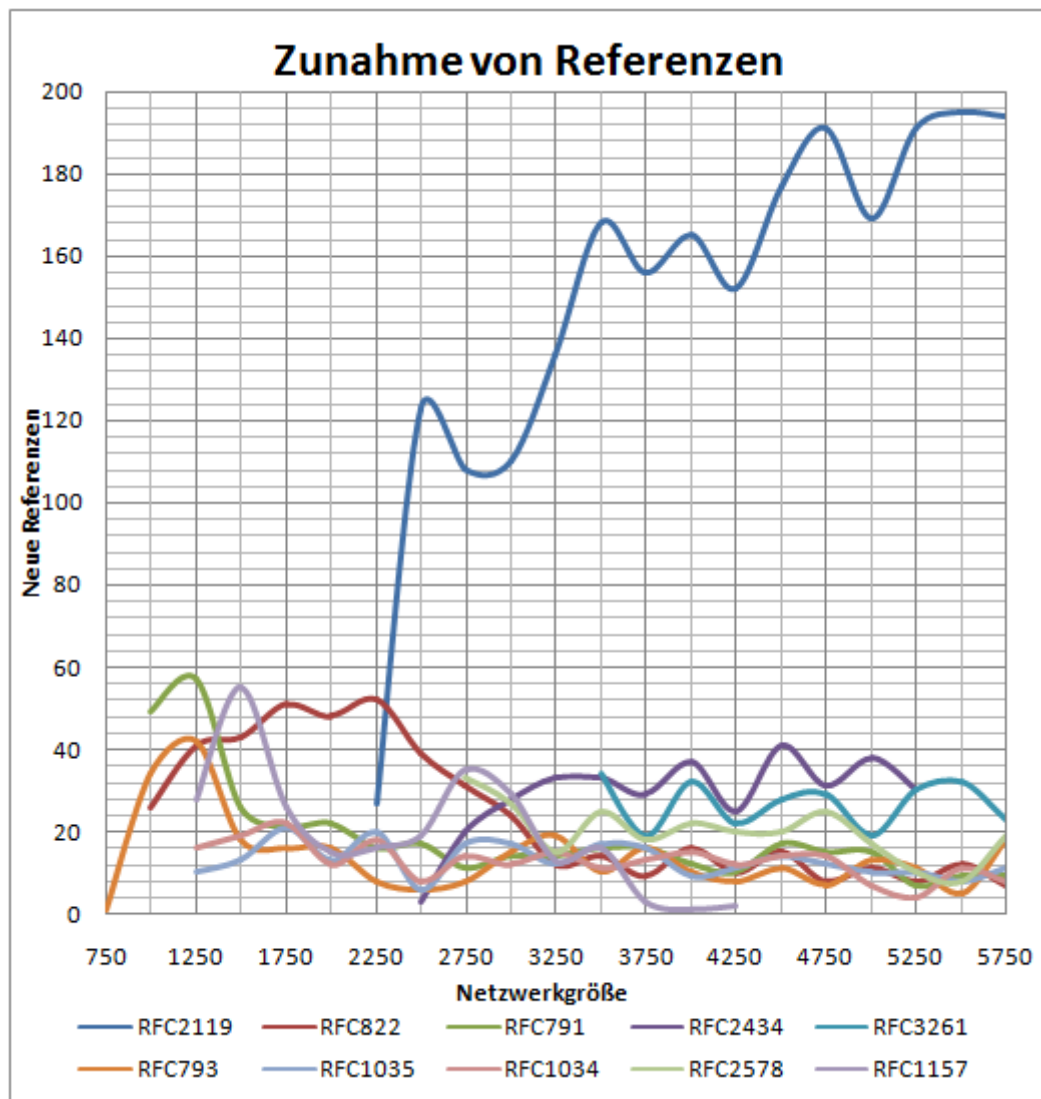


Abbildung 23: Zunahme des Ingrads für 10 am häufigsten zitierte RFCs

Die Untersuchung der Zunahme von Referenzen für die 10 am häufigsten zitierte RFC-Dokumente ergab, dass lediglich RFC2119, wo die in den Texten verwendete Begriffe spezifiziert werden, die *Preferential Attachment* Charakteristik zeigt. Zahlreiche andere Memoranda werden anfangs sehr häufig referenziert, danach wird jedoch eine sinkende Tendenz beobachtet, die nach einer bestimmten Zeit ausgebremst wird und die Anzahl von Zitierungen in den neuen Dokumenten sich stabilisiert. Derartiges Verhalten von den Werten kann dadurch erklärt werden, dass eine Veröffentlichung von den wichtigen Dokumenten einen Boom von darauf basierenden Texten verursacht, der entweder durch neue Möglichkeiten oder wegen Anpassung der alten Memoranda an die neuen Voraussetzungen verursacht wird.

Die dritte Eigenschaft, die im Priceschen Modell der Netzwerkentwicklung vorgesehen ist, ist die Abhängigkeit des Koeffizienten α von dem durchschnittlichen Ausgrad m . [12] Dieses Verhältnis ist folgendermaßen definiert:

$$\alpha = 2 + \frac{1}{m}$$

Da die Werte von den Koeffizienten für den Ingrad, Ausgrad und Durchschnittsgrad bereits berechnet wurden, können sie einfach mit dem im Modell vorgesehenen Exponenten verglichen werden. Der durchschnittliche Ausgrad beträgt $m = 8,425$. Daraus folgt, dass der erwartete Wert von $\alpha_{erw} \approx 2,12$ ist.

- **Ingrad** $\alpha_{ber} = 1,86$
- **Ausgrad** $\alpha_{ber} = 2$
- **Gesamtgrad** $\alpha_{ber} = 1,82$

Die errechneten Werte für die Knotengradverteilung liegen zwar in derselben Größenordnung, wie der erwartete Koeffizient, sie unterscheiden sich jedoch von ihm um zumindest $\frac{1}{m}$. Aus diesem Grund findet die im Priceschen Modell gegebene Gleichung keine Anwendung im RFC-Netzwerk.

Die Untersuchung der im Priceschen Modell für Netzwerkwachstum gegebenen Vorgaben anhand von der RFC-Referenzebene ergab, dass dieses Netzwerk nach anderen Prinzipien gewachsen ist. Sehr auffällig ist hier der linear wachsende Ausgrad, der bedeutet, dass je neuer ein RFC-Dokument ist, desto mehr Quellen referenziert er. Daraus folgt, dass die Komplexität des Netzwerks steigt und immer mehr Memoranda berücksichtigt werden müssen. Ein interessanter Verlauf charakterisiert auch die Anzahl von Zitierungen in bestimmten Zeitschlitzten. Die neuen Veröffentlichungen werden zwar am Anfang sehr häufig referenziert, nach bestimmter Zeit fällt aber dieser Wert und bewegt sich danach im relativ engen Bereich. Das ist ein Widerspruch zu dem im Modell enthaltenen *Cumulative Advantage* Prinzip, wo die Knoten immer häufiger verbunden werden sollen. Der Wert des α -Exponenten stimmt mit der Modellvorgabe nicht überein und ist niedriger als das erwartete Ergebnis. Diese drei Fakten führen zur Schlussfolgerung, dass die Referenzebene des RFC-Normennetzwerks, obwohl offensichtlich skalenfrei, nach einem eigenen Modell gewachsen ist, das nicht mit dem von Derek Price vorgeschlagenen gleichgestellt werden darf.

5.2 Aktualität von Dokumenten

Im Lauf der Zeit ändern sich die Stati von den RFC-Dokumenten, die Memoranda werden aktualisiert und ersetzt. Diese Vorgänge haben bestimmte Folgen für die Texte, die solche ersetzte Netzwerkknoten referenzieren - sie beziehen sich dann auf die nicht mehr aktuellen Bestimmungen, die möglicherweise komplett verändert wurden. Das vermindert die Qualität von den dort enthaltenen Normen und zeigt auf einen Aktualisierungsbedarf. Wegen der eindeutigen Struktur des Netzwerks und klarer Bedeutung von Kanten auf der Aktualitätsebene kann ein Maß für die Qualität des gesamten Normennetzwerks numerisch berechnet werden.

Es existieren unterschiedliche Gelegenheiten, diesen Qualitätsfaktor zu errechnen. Folgender Vorgang bietet sich als eine Möglichkeit für die Auswertung des gesamten Netzwerks an:

1. Alle Knoten wählen, die nicht ersetzt oder aktualisiert wurden
2. Anzahl von Knoten aus (1) berechnen, die einen veralteten Verweis enthalten
3. Wert aus (2) durch den Wert aus (1) dividieren

Dadurch erhält man eine dimensionslose Zahl, die die Werte von 0 bis 1, bzw. von 0% bis 100% annimmt. Je kleiner das Ergebnis, desto bessere Aktualität das Netzwerk besitzt. Außerdem, es können unterschiedliche Varianten von diesem Algorithmus durchgeführt werden, wie z.B. die Berechnung von dem Qualitätsfaktor für bestimmte Stati oder mit bzw. ohne Berücksichtigung von den *Updated by*-Referenzen.

Die Auswertung von diesem Qualitätsfaktor verdeutlicht überraschend hohe Werte für alle Typen von Knoten. Über 94% von den *Draft Standard* Dokumenten besitzt eine Referenz auf ein veraltetes Memorandum. Wenn man zusätzlich die aktualisierten RFCs betrachtet, steigt dieser Wert auf 100%. Die weiteren Typen besitzen auch einen sehr hohen Anteil von verlinkten Knoten, die bereits ersetzt wurden. Der niedrigste Wert charakterisiert die *Standard* Dokumente, es muss aber beachtet werden, dass 24,6% von ihnen überhaupt keine Referenz enthält. Bei der Betrachtung vom gesamten Netzwerk ergibt sich, dass lediglich 31,2% von allen RFCs, die selbst nicht ersetzt oder aktualisiert wurden, ausschließlich aktuelle Dokumente zitiert und nicht verbesserungsbedürftig ist.

Wenn man zusätzlich auch die Verweise auf aktualisierte Memoranda als negativ bewertet, sind nur 12,1% der Dokumente auf dem neusten Stand.

Typ	<i>Obsoleted</i>	<i>Updated</i>	<i>Obsoleted</i> o. <i>Updated</i>	Keine Re- ferenzen
<i>Best Current Practice</i>	76,4%	89,2%	97,3%	2,0%
<i>Draft Standard</i>	94,1%	77,9%	100,0%	0,0%
<i>Experimental</i>	70,4%	79,6%	90,1%	4,0%
<i>Historic</i>	71,7%	56,7%	81,9%	11,8%
<i>Informational</i>	61,5%	68,8%	80,5%	7,5%
<i>Proposed Standard</i>	67,7%	87,8%	93,1%	0,8%
<i>Standard</i>	61,4%	64,9%	73,7%	24,6%
<i>Unknown</i>	17,2%	19,1%	24,5%	55,3%
Gesamtnetz	58,9%	69,0%	77,9%	12,1%

Tabelle 11: Allgemeiner Qualitätsfaktor für das RFC-Normennetzwerk

Der bisher ausgewertete Qualitätsfaktor berücksichtigt nicht den Anteil von veralteten Referenzen. Es reichte, dass nur ein von 100 Verweisen auf ein nicht mehr aktuelles Dokument zeigt, um das ganze Memorandum negativ zu bewerten. Um einen solchen Einfluss auszuschließen, ist eine Modifizierung von dem Berechnungsprozess notwendig, so dass die Anzahl von Kanten bewertet wird.

1. Alle Referenzen-Kanten wählen, deren Anfangsknoten nicht ersetzt oder aktualisiert wurden
2. Anzahl von Kanten aus (1) berechnen, die Verweis auf ein veraltetes Dokument enthalten
3. Wert aus (2) durch den Wert aus (1) dividieren

Dadurch wird zwar der negative Einfluss von einzelnen, nicht mehr aktuellen Kanten minimiert, aber das Ergebnis sagt nicht mehr aus, wieviele Knoten derartige Referenzen enthalten, sondern wie hoch der Anteil von solchen Kanten ist.

Kante von	<i>Obsoleted</i>	<i>Updated</i>	<i>Obsoleted</i> o. <i>Updated</i>
<i>Best Current Practice</i>	20,9%	31,3%	44,8%
<i>Draft Standard</i>	38,1%	23,4%	55,1%
<i>Experimental</i>	23,1%	28,4%	43,3%
<i>Historic</i>	35,6%	22,0%	50,6%
<i>Informational</i>	23,2%	24,9%	41,9%
<i>Proposed Standard</i>	20,1%	26,6%	40,2%
<i>Standard</i>	36,7%	25,1%	55,4%
<i>Unknown</i>	25,0%	24,8%	42,0%
Gesamt	22,5%	26,0%	42,0%

Tabelle 12: Allgemeiner Qualitätsfaktor für Kanten des RFC-Normennetzwerks

Die prozentuellen Anteile von veralteten Kanten sind zwar deutlich niedriger, als der Qualitätsfaktor für Knoten, aber trotzdem sind zahlreiche Verweise veraltet. Die meisten solchen Ecken besitzen die als *Draft Standard* bezeichneten Dokumente. Interessanterweise zeigen die *Standard*-RFCs einen zweithöchsten Wert bei der Berücksichtigung von Zitierungen von den veralteten Dokumenten und sie sind auf dem ersten Platz, wenn man sowohl *Updated* als auch *Obsoleted* in die Statistik mit einbezieht. Das bestätigt die These, dass der relativ niedrige Wert für solche RFCs nur von der großen Anzahl von überhaupt nicht verlinkten Dokumenten verursacht wurde. Insgesamt sollen 22,5% von allen Verweisen aktualisiert werden, weil sie auf die veralteten Memoranda zeigen.

Die beiden Ansätze für die Untersuchung der Qualität des RFC-Normennetzwerks haben ein überraschend schlechtes Ergebnis gezeigt. Mehrere Dokumente, die entweder bereits als Standard anerkannt wurden, bzw. sich in der Anerkennungsphase befinden, beziehen sich auf die veralteten Texte, die von den neueren Versionen abgelöst wurden. Einerseits zeigt es eine relativ schlechte Aktualität von dem RFC-Normennetzwerk, andererseits müssen folgende zwei Aspekte berücksichtigt werden. Erstens, bei der schneller Geschwindigkeit von der Entwicklung von neuen Standards wäre es enorm aufwendig, alle Änderungen zu beachten und ab sofort einzupflegen. Das könnte sogar einen lawinenartigen Prozess verursachen, der die Aktualisierung von fast allen vorhandenen RFCs auslösen würde. Zweitens, die maschinelle Auswertung von den Referenzen kann keine Aussage über die Wichtigkeit von solchen Zitierungen machen. Es ist durchaus möglich,

dass die veralteten Verweise nur am Rande des Textes erwähnt wurden und sich nur wenig oder gar nicht beteiligen, wenn es sich um tatsächliche Bedeutung des Dokumentes handelt. Außerdem, der Fakt, dass ein RFC-Memorandum bereits ersetzt wurde setzt nicht voraus, dass alle Implementierungen ab sofort abgeschafft werden oder alle dort enthaltenen Ideen vom Prinzip falsch und veraltet sind. Aus diesem Grund kann nicht eindeutig bestätigt werden, dass die Qualität des RFC-Normennetzwerks sehr schlecht und dringend verbesserungsbedürftig ist.

6 Zusammenfassung und Schlussfolgerungen

Das Ziel dieser Arbeit - die Untersuchung der Eigenschaften des RFC-Normennetzwerks wurde erfolgreich erreicht. Bedingt durch den sehr strukturierten Aufbau von den Dokumenten und die Existenz von hochqualitativer Verzeichnisdatei im XML-Format, war der Einstieg in die Extrahierung von den in den Dokumenten enthaltenen Informationen relativ reibungsfrei. Dank der Flexibilität von speziell zu diesem Zweck entwickelter Software, war die Verarbeitung von Daten nicht nur auf ein einfaches Auslesen beschränkt. Auch die Errechnung von Qualitätsfaktoren, die nicht in den anderen Anwendungen vorkommen ist dadurch unproblematisch abgelaufen. Dagegen war die Auswahl von einer geeigneten Fremdsoftware war sehr zeitaufwendig. Viele Versuche scheiterten, weil die Größe des Netzwerks die Fähigkeiten von den Programmen überschritten hat. Manchmal fehlte auch die Möglichkeit, die Daten in einem gängigen Format zu importieren. Durch dieses Problem wurde die Verwendung von SONIVIS ausgeschlossen.

Erst *Gephi* und *Network Workbench Tool* zeigten, dass die Bearbeitung eines aus fast 6000 Knoten und über 50000 Kanten bestehenden Netzwerks mit der heutigen Rechnerleistung auch möglich ist. In diesem Fall "möglich" kann leider nicht mit "bequem" und "reibungsfrei" gleichgestellt werden. Die Software stürzte oft ab, was die Berechnungen von den jeweiligen Metriken erschwerte. Außerdem war die Bedienung nicht immer ergonomisch, es fehlten auch die Automatisierungsmöglichkeiten, die die Auswertung von vielen Statistiken deutlich erleichtern würden. Obwohl beide verwendete Programme aus der Open-Source-Welt stammen, scheint die Entwicklung sehr langsam, bzw. komplett eingestellt (*Network Workbench Tool*) zu sein. Trotzdem konnten alle Aspekte der Netzwerkkuntersuchung erfolgreich durchgeführt und ausgewertet werden.

Die Auswertung erfolgte auf zwei unabhängigen Ebenen, die auf der gleichen Knotenmenge basierten. Die erste, Referenzebene, enthält ausschließlich Informationen über die in den Texten enthaltenen Referenzen. Die zweite, Aktualitätsebene, bezieht sich auf die Aktualisierungen und Ersetzungen von den RFC-Dokumenten. Diese Teilung war durch unterschiedliche erwartete Struktur bedingt wurde eingeführt, um den gegenseitigen Einfluss von beiden Ebenen auszuschließen. Die im Rahmen der Auswertung erhaltene Daten zeigten, dass diese Entscheidung richtig war.

Viele aus der echten Welt stammende Netzwerken werden als *Small-World* klassifiziert. Dieser Begriff hat sich in Folge von dem in sechsziger Jahren von Stanley Milgram durchgeführten Experimentes durchgesetzt. Der Versuch zeigte, dass das Netzwerk von Bekanntschaften in den Vereinigten Staaten durch eine mittlere Weglänge von etwa 6 Schritten charakterisiert wird. Dieser relativ kleine (wenn die Anzahl von Einwohnern berücksichtigt wird) Wert verursachte eine verstärkte Forschung von sozialen Netzwerken. Die Ergebnisse von der wissenschaftlichen Betrachtung der zwischenmenschlichen Beziehungen haben auf eine Menge von Merkmalen hingewiesen, die ein natürlich erzeugtes Netzwerk charakterisieren. Zu diesen gehören:

- Niedrige mittlere Weglänge. Der Wert für die Referenzebene beträgt $d = 5,215$ mit Berücksichtigung der Kantenrichtung und $d = 3,037$ bei Betrachtung als ungerichtetes Netzwerk. Für die Aktualitätsebene beträgt $d = 8,581$.
- Höchstens logarithmisches Wachstum der mittleren Weglänge. Für beide Netzwerkebenen lässt sich derartige Steigung von der Kurve beobachten. Bei der Referenzebene kann, ab einer bestimmten Anzahl von Knoten, eine deutliche Senkung von dieser Metrik beobachtet werden.
- Höherer Clusterkoeffizient. Dieser Wert widerspiegelt die Neigung von Knoten zur Bildung von Cliquen - völlig vernetzten Gruppen von Knoten. Der Wert vom globalen Clusterkoeffizienten beträgt $C = 0,355$ für die Referenzebene und $C = 0,076$ für die Aktualitätsebene. Watts-Strogatz-Clusterkoeffiziente betragen entsprechend $0,351$ und $0,026$. Beide Metriken sind viel höher als für einen vergleichbaren Zufallsgraphen.

Die drei oben beschriebenen Voraussetzungen sind erfüllt und das RFC-Normennetzwerk gehört zur Gruppe von *Small-World-Networks*.

Eine weitere Eigenschaft, die viele natürlich erzeugte Netze charakterisiert, ist die Skalenfreiheit von der Knotengradverteilung. Dies bedeutet, dass die Knotengrade so verteilt sind, dass der Verlauf der Kurve einem Potenzgesetz folgt. Dieser Sachverhalt lässt mehrere Aussagen über die Struktur des Netzwerks machen. In solchen Graphen bilden die Knoten mit niedrigem Grad eine deutliche Mehrheit. Es sind aber auch wenige Knoten vorhanden, die durch einen extrem hohen Grad charakterisiert werden und als "Abkürzungen" dienen, die die mittlere Weglänge im Netzwerk senken. Eine derartige Verteilung spielt auch eine entscheidende Rolle bei der Ausfallsicherheit des Netzwerks. Die Überprüfung dieser Eigenschaft erfolgte grafisch, indem die Knotengradverteilung (berechnet mit der Methode von exponentieller Einteilung) auf einer logarithmischen Skala dargestellt wurde. Die resultierende Kurve hat in weiten Bereichen einen nahezu linearen Verlauf, insbesondere im rechten Teil. Das ist ein direkter Beweis für die Skalenfreiheit der Knotengradverteilung auf beiden RFC-Netzwerkebenen. Diese Graphen werden durch folgende Funktionen approximiert:

- Referenzebene - Ingrad: $f(x) = 0,688x^{-1,86}$
- Referenzebene - Ausgrad: $f(x) = 0,872x^{-2}$
- Referenzebene - Gesamtgrad: $f(x) = 0,861x^{-1,82}$
- Aktualitätsebene - Gesamtgrad: $f(x) = 0,0891x^{-2,0635}$

Die Werte von den Exponenten liegen im Bereich, der typisch für natürlich erzeugte Netzwerke ist. Ferner, die am häufigsten verlinkten Dokumente wurden explizit identifiziert und genannt. Das sind:

- Referenzebene - Ingrad: RFC2119 (2385 eingehende Kanten)
- Referenzebene - Ausgrad: RFC1012 (920 ausgehende Kanten)
- Aktualitätsebene - Gesamtgrad: RFC1035 (24 Kanten)

Das Memorandum mit dem höchsten Ingrad beschreibt die in den Texten verwendeten Begriffe, was die häufige Verlinkung erklärt. RFC1012 ist ein Verzeichnis von allen anderen RFCs mit laufenden Nummern von 1 bis 1000. Das Top-Dokument der Aktualitätsebene beschreibt das DNS-System und wurde von 21 anderen RFCs aktualisiert (selbst hat es 3 andere erweitert).

Wie bereits erwähnt, hat die Skalenfreiheit der Knotenverteilung einen gravierenden Einfluss auf die Robustheit des Netzwerks. Solche Graphen sind zwar auf eine zufällige Entfernung von Knoten unempfindlich, aber ein gezielter Angriff auf die Knoten mit dem höchsten Grad sehr große Wirkung auf die Netzwerkmetriken hat. Beide Ebenen des RFC-Normennetzwerks wurden getestet, wie sie sich nach der Entfernung von 1% bis 15% von Knoten verhalten - zufällig und gezielt. Dabei wurde nach jedem Löschvorgang die mittlere Weglänge und der Watts-Strogatz Clusterkoeffizient bemessen. Die zufällige Entfernung hat nur einen sehr geringen Einfluss auf die Werte von d und C . Das Netzwerk verhält sich ganz anders, wenn die Knoten gezielt gelöscht werden. Auf der Referenzebene steigt die mittlere Weglänge proportional zur Anzahl von den gelöschten Knoten und nimmt am Ende plötzlich ab - das Netzwerk wurde in zwei Komponenten geteilt. Der Einfluss auf den Clusterkoeffizienten ist geringer - er fällt aber relativ langsam und nach der Entfernung von 15% der wichtigsten Knoten beträgt er trotzdem etwa 0,25. Das zeigt, dass die Vernetzung von Knoten mit kleinem Grad ziemlich stark ist.

Aufgrund von kleinerer Anzahl der Kanten ist die Aktualitätsebene auf das gezielte Löschen empfindlicher. Hier wird die mittlere Weglänge immer kleiner, jedoch nicht aufgrund von neuen Wegen, sondern wegen Zerlegung des Netzwerks in immer mehr unabhängige Komponenten. Im Endeffekt werden alle Kanten vernichtet, was zur mittleren Weglänge von 1 führt. Der Clusterkoeffizient nimmt auch sehr stark ab und bereits nach der Entfernung von 7% der wichtigsten Kanten beträgt er 0.

Das RFC-Normennetzwerk wurde auf zwei Korrelationen untersucht: Knotenartkorrelation und Knotengradkorrelation. In dem ersten Fall wurde überprüft, ob die Knoten mit einem bestimmten Status sich häufiger mit den anderen mit gleichem Knotentyp verbinden. Diese Auswertung erfolgte mit dem Assortativitätskoeffizienten r , der wiedergibt, wie stark diese Tendenz ist. Für die Referenzebene beträgt er $r = 0,15$, was auf keinen Zusammenhang hinweist. Auch die manuelle Beobachtung von der Ergebnistabelle ergab, dass die Knotentypen, die einfach am häufigsten vorkommen auch am häufigsten verlinkt werden. Die Aktualitätsebene zeigt im Gegensatz eine bemerkbare Korrelation zwischen Knotentypem - der Wert des Koeffizienten beträgt $r = 0,509$, was einen relativ starken Zusammenhang bedeutet. Das entspricht auch den Erwartungen, da wenn ein Dokument ersetzt wird, erwartet man, dass der neue Text vom gleichen Typ ist.

Als zweite Korrelation wurde die Abhängigkeit von dem Knotengrad untersucht. Diese Auswertung erfolgte mit dem Pearson-Korrelationskoeffizienten und wurde in folgenden Konstellationen durchgeführt:

- Ingrad $\rightarrow$ Ingrad (Referenzebene): $\varrho = -0,0205$
- Ingrad $\rightarrow$ Ausgrad (Referenzebene): $\varrho = -0,0085$
- Ausgrad $\rightarrow$ Ingrad (Referenzebene): $\varrho = -0,0773$
- Ausgrad $\rightarrow$ Ausgrad (Referenzebene): $\varrho = -0,0811$
- Gesamtgrad $\rightarrow$ Gesamtgrad (Aktualitätsebene): $\varrho = 0,301$

Wie bei der Knotenartkorrelation, zeigt die Referenzebene keinen Zusammenhang zwischen dem Grad von verbundenen Knoten. Bei der Aktualitätsebene ist der Wert des Koeffizienten zwar höher, bedeutet aber auch keine starke Korrelation und ist durch den geringeren Wertebereich und baumförmige Struktur des Netzwerks bedingt.

Derek de Solla Price erstellte ein mathematisches Modell für das von ihm untersuchte Zitationsnetzwerk. Da die Referenzebene des RFC-Normennetzwerks im Prinzip auch ein Zitationsnetzwerk ist, wurde überprüft, ob ihre Entwicklung den Grundlagen des Priceschen Modells folgte. Diese Auswertung war dank der zeitlichen Invarianz von RFC-Texten und vorhandenen Veröffentlichungsdaten möglich. Es wurden drei Aspekte des Netzwerkwachstums überprüft:

- **Mittlerer Ausgrad** m bleibt im Priceschen Modell konstant. Das RFC-Normennetzwerk zeigte aber ein lineares Wachstum von dieser Metrik.
- **Cumulative Advantage** - die am stärksten zitierten Dokumente sollen immer häufiger referenziert werden. Die Untersuchung von 10 wichtigsten RFCs ergab, dass lediglich das Dokument mit den Begriffsdefinitionen diesem Prinzip folgt. Die anderen erreichen entweder ab sofort einen Sättigungszustand oder werden am Anfang sehr oft zitiert, danach fällt aber der Wert sehr schnell und stabilisiert sich auf einem viel niedrigerem Niveau.
- **α -Koeffizient** im Priceschen Modell liegt zwischen 2 und 3 und ist durch die Gleichung $\alpha = 2 + \frac{1}{m}$ gegeben. Das trifft auch nicht zu, da alle Werte von α in dem Normennetzwerk kleiner als 2 sind.

Aus den oben genannten Gründen folgt, dass das RFC-Normennetzwerk nicht wie im Priceschen Modell wächst. Die Aktualitätsebene wurde gar nicht untersucht, weil das Modell nur für gerichtete Netzwerke anwendbar ist.

Als letztes wurde die Aktualität der Dokumenten qualitativ ausgewertet. Es wurde ein Qualitätsfaktor berechnet, der angibt, wie viele von den aktuellen Dokumenten (d.h. diesen, die keinen *Updated by* bzw. *Obsoleted by* Vermerk haben) Verweise enthalten, die auf die bereits aktualisierten oder ersetzen Dokumenten zeigen. Das Ergebnis war sehr hoch, im gesamten Referenzennetzwerk beinhalten fast 58,9% aller zu diesem Zeitpunkt nicht ersetzten Dokumente eine Referenz auf einen nicht mehr aktuellen Text. Diese überraschend große Werte haben zur Berechnung vom weiteren Qualitätsfaktor geführt, der prozentuell angibt, wieviele Kanten auf veraltete Dokumente zeigen. Dieser Wert beträgt für das gesamte Netzwerk 22,5%, was zwar niedriger, aber trotzdem merkwürdig ist. Aufgrund der Komplexität der Referenzebene und unterschiedlicher Wichtigkeit von Verweisen darf jedoch nicht eindeutig gesagt werden, dass das RFC-Normennetzwerk von sehr schlechter Qualität ist.

Die in dieser Arbeit verifizierten Merkmale weisen eindeutig darauf hin, dass die Eigenschaften von dem RFC-Normennetzwerk typisch für einen natürlich gewachsenen Graphen sind. Die Entwicklung erfolgte zwar auf Basis von einem eigenen Modell, es wurden aber keine unerklärbare Werte errechnet. Das Verhalten des Netzwerks unter unterschiedlichen Bedingungen konnte auch bestimmten Prozessen zugewiesen werden und hat die entsprechenden Erwartungen erfüllt.

7 Literatur- und Abbildungsverzeichnis

Literatur

- [1] *RFC-Index*, <ftp://ftp.rfc-editor.org/in-notes/rfc-index.xml>, Zugriff am 27.07.2010
- [2] Bradner, S., *The Internet Standards Process – Revision 3*, 1996
- [3] Crocker, S., *RFC 1: Host Software*, 1969
- [4] Milgram, S., *The small world problem*, Psychology Today 2, 1967, S. 60–67
- [5] Bork, P., Jensen, L., von Mering, C., Ramani, A., Lee, I., Marcotte, E., *Protein interaction networks from yeast to human*.
- [6] Price, D. J. de S., *Networks of scientific papers*, Science 149, 1965, S. 510–515
- [7] *GEXF File Format*, <http://gexf.net/format/>, Zugriff am 27.07.2010
- [8] *GEXF File Format - Dynamics*, <http://gexf.net/format/dynamics.html>, Zugriff am 27.07.2010
- [9] Travers, J., Milgram, S., *An Experimental Study of the Small World Problem*, Sociometry, Vol. 32, Nr. 4, 1969, S. 425–443
- [10] *Erdős Number Project*, <http://www.oakland.edu/enp/>, Zugriff am 27.07.2010
- [11] *The Oracle of Bacon*, <http://oracleofbacon.org/cgi-bin/center-cgi?who=Kevin+Bacon>, Zugriff am 27.07.2010
- [12] Newman, M. E. J., *The Structure and Fuction of complex networks*, SIAM Review 45, 2003, S. 167–256
- [13] Watts, D. J., *Small Worlds: The Dynamics of Networks between Order and Randomness*, Priceton University Press, 1999
- [14] Holland, P. W., Leinhardt, S., *Transitivity in structural models of small groups*, Comparative Group Studies 2, 1971, S. 107–124
- [15] Watts, D. J. Watts, Strogatz, S. *Collective dynamics of "small-world" networks*, Nature 393, 1998 S. 440–442

-
- [16] Erdős, P., Rényi, A., *On random graphs*, Publicationes Mathematicae 6, 1959, S. 290-297
- [17] Barabási, A., *Scale-Free Networks*, Scientific American, 2003, S. 60-69
- [18] Barabási, A., Réka, A., *Emergence of Scaling in Random Networks*, Science 286, 1999, S. 509-512
- [19] Price, D. J. de S., *A general theory of bibliometric and other cumulative advantage processes*, Journal of the American Society for Information Science, Vol. 27, No. 5, 1976, S. 292-306
- [20] Yule, G. U., *A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis*, Philosophical Transactions of the Royal Society of London, 1925
- [21] Lotka, A. J., *The frequency distribution of scientific productivity*, Journal of the Washington Academy of Sciences, 1926, S. 317-323
- [22] Umstätter, W., *Zum Thema Lotka's law*, <http://www.ib.hu-berlin.de/wumsta/price52.html>, Zugriff am 27.07.2010
- [23] Redner, S., *How Popular is Your Paper? An Empirical Study of the Citation Distribution*, European Physics Journal B, Vol. 4, Nr. 2, 1998, S. 131-134
- [24] Newman, M. E. J., *Mixing patterns in networks*, Physical Review, 2003
- [25] Morris, M., *Data driven network models for the spread of infectious disease*, Epidemic Models: Their Structure and Relation to Data, Cambridge University Press, 1995, S. 302-322

Tabellenverzeichnis

1	Mittlere Weglänge in Abhängigkeit von Netzwerkgröße (Referenzebene) . . .	16
2	Mittlere Weglänge in Abhängigkeit von der Netzwerkgröße (Aktualitätsebene)	18
3	20 Dokumente mit höchstem Ingrad	27
4	10 Dokumente mit höchstem Ausgrad	29
5	Dokumente mit höchstem Grad (Aktualitätsebene)	35
6	Verteilung von Knotenarten	50
7	Normierte Mischmatrix (Referenzebene)	51
8	Normierte Mischmatrix (Aktualitätsebene)	51
9	Beispiel der Ergebnistabelle	54
10	Pearson-Koeffizienten des Normennetzwerks	54
11	Allgemeiner Qualitätsfaktor für das RFC-Normennetzwerk	62
12	Allgemeiner Qualitätsfaktor für Kanten des RFC-Normennetzwerks	63

Abbildungsverzeichnis

1	Kleines Netzwerk mit gerichteten Kanten	3
2	Oberfläche von Gephi	9
3	Ausschnitt aus der Graphenvisualisierung	10
4	Auswahl der Zeitspanne	10
5	Oberfläche vom Network Workbench Tool	11
6	Verteilung der Erdős-Zahl	14
7	Mittlere Weglänge der Referenzebene	17
8	Mittlere Weglänge der Aktualitätsebene	19
9	Geschlossenes und offenes Tripel	20
10	Lokaler Clusterkoeffizient: $C_A = 1$, $C_E = \frac{2}{3}$	21
11	Globaler Clusterkoeffizient	22
12	Watts-Strogatz Clusterkoeffizient	23
13	Ausgradverteilung (Referenzebene)	31
14	Ingradverteilung (Referenzebene)	32
15	Gradverteilung (Referenzebene ungerichtet)	33
16	Gradverteilung (Aktualitätsebene)	36
17	Mittlere Weglänge (Referenzebene gerichtet)	39
18	Mittlere Weglänge (Referenzebene ungerichtet)	41
19	Clusterkoeffizient (Referenzebene)	44
20	Mittlere Weglänge (Aktualitätsebene)	45
21	Clusterkoeffizient (Aktualitätsebene)	46
22	Durchschnittlicher Ausgrad in Abhängigkeit von der Netzwerkgröße	57
23	Zunahme des Ingrads für 10 am häufigsten zitierte RFCs	59

Inhaltsverzeichnis

1	Einleitung	1
2	Grundlagen	3
3	Werkzeuge	6
3.1	Extrahierung von Daten	6
3.2	Gephi	9
3.3	Network Workbench Tool	11
4	Auswertung	13
4.1	Small-World-Netzwerk	13
4.1.1	Mittlere Länge des Weges	15
4.1.2	Clusterkoeffizient	20
4.2	Knotengrad	24
4.2.1	Referenzebene	26
4.2.2	Aktualitätsbene	34
4.3	Robustheit	38
4.4	Mixing Patterns	48
4.4.1	Knotenartkorrelation	48
4.4.2	Knotengradkorrelation	52
5	Entwicklung	56
5.1	Netzwerkwachstum	56
5.2	Aktualität	61
6	Zusammenfassung	64
7	Quellen	70