Analyse des Normennetzwerks der Internet Requests for Comments

Maciej Wieńszczak, Robert Tolksdorf
Networked Information Systems, Freie Universität Berlin, www.ag-nbi.de
kangur@zedat.fu-berlin.de, tolk@ag-nbi.de

Abstract: In einer Netzwerkanalyse haben wir den Korpus der Internet RFC Standards analysiert. Dabei sind eine Verweis- und eine Aktualisierungsebene getrennt zu betrachten. Die Analyse ermittelt verschiedene Metriken über die jeweilige Netzwerkstruktur sowie deren temporalen Entwicklungsverlauf. Es lassen sich Aussagen über die Qualität des Normenkorpus ableiten die Rückschlüsse auf die Qualität des Standardisierungsprozess erlauben.

1 Einleitung

Die *Internet Engineering Task Force* (IETF) publiziert als *Requests for Comments* nummerierte Memoranda, die unterschiedliche Internet-bezogene Standards, Standardvorschläge, Verhalten und Forschung beschreiben. Die Dokumente sind durch eine feste, formale Struktur charakterisiert, die eine einfache, automatisierte Verarbeitung des Inhalts ermöglicht. Außerdem sind die veröffentlichten RFCs fest, 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.

Die RFCs beziehen sich unterschiedlich aufeinander. Neue Memoranda können die alten entweder komplett ersetzen oder ergänzen, es existieren auch Verweise auf andere Dokumenten direkt im Text eines Memorandums. Extrahierbar sind Referenzen/Zitierungen, Ersetzungen (*Obsoletes/Obsoletet by*) und Aktualisierungen (*Updates/Updated by*).

Die Dokumente werden als Netzwerkknoten interpretiert. Jedes Memorandum besitzt einen Status, der die Wichtigkeit wiedergibt und ein Veröffentlichungsdatum, welches es, 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 Netzwerkextraktion und -analyse

Mit einem für diesen Zweck entwickelten Werkzeug wurde das RFC-Verzeichnis im XML-Format [RFC11] verarbeitet, um die Bezüge der *Aktualitätsebene* zu extrahieren. Wir

erschienen im Tagungsband der INFORMATIK 2011
Lecture Notes in Informatics, Band P192
ISBN 978-3-88579-286-4
weitere Artikel online:
http://informatik2011.de/519.html

erhalten Informationen über die Aktualisierungen und Ersetzungen von den jeweiligen RFCs. Zusätzlich stehen die grundlegenden Informationen zu jedem Text, wie 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.

Um die Referenzen aus einem Dokument zu extrahieren, wird der komplette Text durchgearbeitet. Trotz der starken Strukturierung von RFC-Dateien, können die Referenzen im Text unterschiedliche Formen annehmen. Zusätzlich sind einige so umgebrochen, dass das Wort „RFC“ und die dazugehörige Nummer sich in unterschiedlichen Zeilen befinden.

Aufgrund des Textaufbaus wurden einige Referenzen wurden nicht extrahiert. Das sind hauptsächlich 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. Allerdings haben diese Beziehungen keinen Einfluss auf die Wichtigkeit von Netzwerkknoten, da es sich lediglich um einen Index handelt, ohne beitragenden Inhalt.

Die Modellierung und Auswertung des Graphes wurde mithilfe der Werkzeuge *Gephi* und *Network Workbench Tool* (NWB) durchgeführt. Die implementierten Metriken ermöglichen eine Berechnung von typischen Netzwerkstatistiken, wie die Verteilung des Knotengrades, der Clusterkoeffizient oder die mittlere Weglänge sowie Tests auf Skalenfreiheit.

3 Resultate

Unsere Analyse betrachtet das RFC-Normennetzwerk im Zustand von 30.7.2010 mit RFC 5942 als letztem berücksichtigem Dokument. Die Berechnung und Auswertung des Graphen wird auf Referenzenebene und Aktualitätsebene durchgeführt. Das Ergebnis wird mit einem Zufallsgraphen mit gleicher Anzahl von Knoten und Kanten verglichen, um zu untersuchen, welche Charakteristiken das Normennetzwerk von ihm unterscheiden.

3.1 Mittlere Länge des Weges

Der durchschnittliche Weg d zwischen allen Knoten in einem Graphen ist definiert als $d = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}$ wobei n die Anzahl von Knoten angibt und d_{ij} der geodesische 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 [New03]. Das 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 Netzwerkbereichen signifikant reduziert [Wat99].

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.

Referenzenebene Die Referenzenebene ist ein gerichteter Graph, da die zitierten Artikel keinen Verweis auf die zitierenden enthalten. Um den Einfluss dieser Eigenschaft zu untersuchen, wurde 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 der Richtungen führt zu einer noch kleineren Welt, weil dadurch mehrere mögliche Pfade gleichzeitig zur Verfügung stehen.

Knoten	Gerichtet	Zufallsgr. gerichtet	Ungerichtet	Zufallsgr. 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 (Referenzenebene)

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. Auffällig ist, dass bei einem Netzwerk mit weniger als 500 Knoten das Wachstum deutlich höher, 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 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 steigt das Verhältnis von n und d 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, 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.

ISBN 978-3-88579-286-4

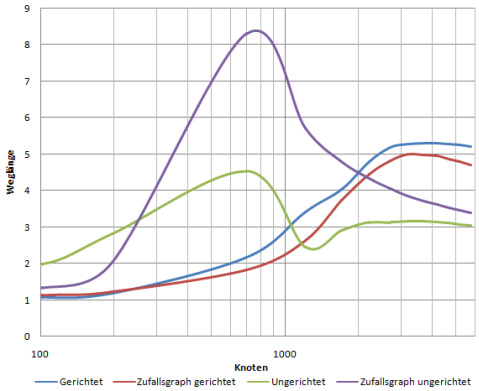

Abbildung 1: Mittlere Weglänge der Referenzebene

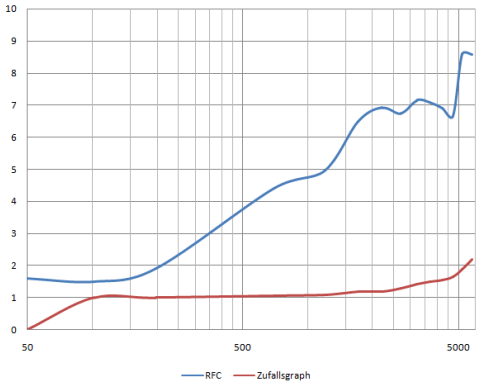

Abbildung 2: Mittlere Weglänge der Aktualitätsebene

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.

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. In strukturiert wachsenden RFC-Normennetzwerk sind dagegen mehrere solche Gruppierungen vorhanden, die durch die Entwicklung von unterschiedlichen Normen entstanden sind.

Knoten	RFC-Netzw.	Zufallsgr.
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)

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 $d \backsim \log n$ im Allgemeinen nicht verletzten. Es sind hier keine Sättigungspunkte erkennbar, ab denen der Wert von d sich senkt. Das betont die unterschiedliche Struktur beider Netzwerke. Ein weiteres solches Merkmal ist die Abwesenheit von einer nicht-logarithmischen Wachstumsphase – die Werte wachsen vom Anfang an logarithmisch.

ISBN 978-3-88579-286-4

Auffällig 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 zu, weil sie sich aufgrund des geringen Kanten/Knoten-Verhältnisses in einem eingeschränkten Bereich bewegen.

3.2 Clusterkoeffizient

Der Clusterkoeffizient C ist ein Maß für die Verlinkung eines Netzwerks.Zur Berechung werden zwei unterschiedliche Methoden angewendet. Der Globale Clusterkoeffizient gibt an, wieviele Tripel geschlossen sind und ist folgendermaßen definiert [New03]: $C =$ Anzahl von geschlossenen Tripeln/Anzahl von allen Tripeln

Das vom Duncan Watts und Steven Strogatz vorgeschlagene Verfahren zur Bestimmung des Vernetzungsgrades eines Netzwerk geht von dem lokalen Clusterkoeffizienten für die einzelnen Knoten heraus [WS98]. 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.

Der lokale Clusterkoeffizient für gerichtete Graphen ist gegeben durch:

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

Für ungerichtete Graphen nimmt die Gleichung folgende Form an:

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

Der Watts-Strogatz Clusterkoeffizient C des gesamten Netzwerks ist als ein Durchschnittswert von allen lokalen Koeffizienten definiert. $C = 1/|V| \sum_i C_i$.

Beide Ebenen des Normennetzwerkes weisen Clustering auf, sowohl beim globalen, als auch beim Watts-Strogatz Clusterkoeffizienten. Aufgrund von der viel höheren Anzahl von Kanten auf der Referenzenebene, 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. Auffällig ist hier auch die Diskrepanz zwischen beiden Werten bei der Aktualitätsebene, wo der globale Koeffizient ein Dreifaches von

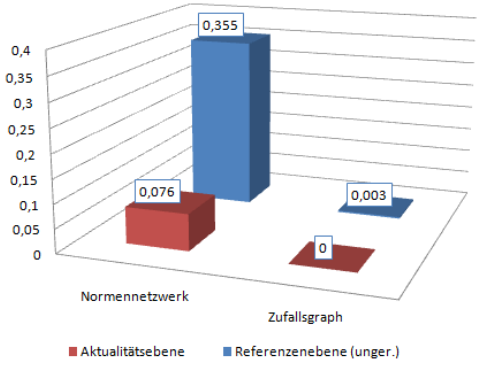

Abbildung 3: Globaler Clusterkoeffizient

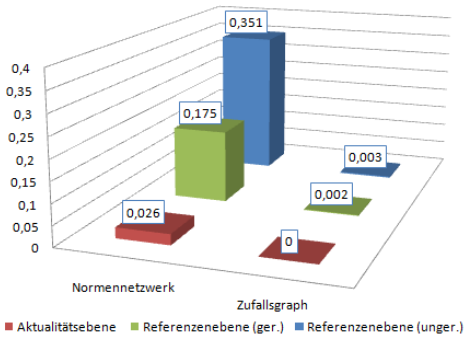

Abbildung 4: Watts-Strogatz Clusterkoeffizient

$C_{(5)}$ 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 aller Tripeln signifikant verringert.

Die beiden Teile des Normenetzwerks sind also durch eine niedrige mittlere Entfernung von Knoten charakterisiert werden. Die beiden Werte liegen im Bereich, der für die thematisierten Dokumentennetzwerke, wie z.B. wissenschaftliche Publikationen charakteristisch ist [New03]. Die Beobachtung der Abhängigkeit der mittleren Weglänge von der Netzwerkgröße führt 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. Im Vergleich zu den zufälligen Graphen derselben Größe, weichen die Werte stark ab und zeigen, dass die Entstehung von Kanten kein zufälliger Prozess war.

3.3 Verteilung des Knotengrades

Der Durchschnittsgrad macht eine 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 erhöht werden kann. Der Durchschnittsgrad $\overline{d}$ wird folgendermaßen berechnet:

$$\overline{d} = \frac{1}{|V|} \sum_i d_i, v_i \in V \tag{3}$$

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: $|\{v_i\}|/|V|$ mit $v_i \in V \wedge \deg(v_i) = k$.

ISBN 978-3-88579-286-4

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 der exponentiellen Einteilung (*Exponential Binning*) besser geeignet und wurde in dieser Arbeit verwendet.

Referenzenebene Die Berechnung der Metriken erfolgte mit dem *Network Workbench Tool*. Die dort eingebauten Algorithmen ermöglichen nicht nur die Berechung 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 10 Dokumente mit dem maximalen Ingrad und 10 mit dem maximalen Ausgrad extrahiert. Diese können den Tabellen 3 und 4 entnommen werden.

Ingrad	RFC	Titel	Status
2385	2119	Key words for use in RFCs to Indicate Requirement Levels	BCP
468	822	Standard for the format of ARPA internet text messages	BCP
372	791	Internet Protocol	Standard
347	2434	RTP Payload Format for Bundled MPEG	Experimental
295	793	Transmission Control Protocol	Standard
290	3261	SIP: Session Initiation Protocol	Prop. Standard
256	2578	Structure of Management Information Version 2 (SMIv2)	Standard
254	1157	Simple Network Management Protocol (SNMP)	Historic
248	1034	Domain names - concepts and facilities	Standard
247	1035	Domain names - implementation and specification	Standard

Tabelle 3: 20 Dokumente mit höchstem Ingrad

Wenn die Verteilung des Knotengrades einem Potenzgesetz folgt, wird ein Netzwerk mit dieser Eigenschaft skalenfrei genannt [BA99]. Diese Einteilung setzt jedoch nicht voraus, dass weitere Eigenschaften und Metriken, wie z.B. mittlere Weglänge in Abhängigkeit von der Netzwerkgröße auch skalenfrei sind.

$$p_k \sim \frac{1}{k^\alpha} \Leftrightarrow p_k \sim k^{-\alpha} \tag{4}$$

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

ISBN 978-3-88579-286-4

Ausgrad	RFC	Titel	Status
920	1012	Bibliography of Request For Comments 1 through 999	Informat.
413	2626	The Internet and the Millennium Problem (Year 2000)	Informat.
267	3795	Survey of IPv4 Addresses in Currently Deployed IETF Application Area Standards Track and Experimental Documents	Informat.
201	3790	Survey of IPv4 Addresses in Currently Deployed IETF Internet Area Standards Track and Experimental Documents	Informat.
198	1000	Request For Comments reference guide	Unknown
184	1700	Assigned Numbers	Historic
162	3796	Survey of IPv4 Addresses in Currently Deployed IETF Operations & Management Area Standards Track and Experimental Documents	Informat.
144	1340	Assigned Numbers	Historic
143	2896	Remote Network Monitoring MIB Protocol Identifier Macros	Informat.
133	1011	Official Internet protocols	Unknown

Tabelle 4: 10 Dokumente mit höchstem Ausgrad

Verlauf bedeutet, dass sehr viele Knoten mit einem geringen Grad existieren. Zusätzlich sind auch wenige Knoten mit sehr höhem Grad vorhanden, die als lokale Zentren, mit denen die anderen Knoten vebunden sind, dienen.

Viele Netzwerke weisen diese Eigenschaft auf – darunter das Internet, Netzwerke von Schauspielern und sogar das Netzwerk von sexuellen Kontakten [New03]. Für die Untersuchung des RFC-Normennetzwerks sind als Vergleichswerte die von Price angegebenen Koeffizienten für das wissenschaftliche Zitationsnetzwerk, das zur RFC-Referenzenebene vergleichbar ist, geeignet [Pri76].

Mit derartiger Verteilung ist der *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 lässt sich beispielsweise bei den wissenschaftlichen Arbeiten beobachten. Die, die oft zitiert wurden, gewinnen an Bedeutung und werden Grundlagen vom bestimmten Themenbereich – was wiederrum zur Entstehung von mehreren Referenzen führt.

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 höhrere 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 Referenzenebene beträgt also 2.

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. Der α-Exponent beträgt $1,86$ und die Ausgleichsgerade ist durch folgende Funktion gegeben: $f(x) = 0,688x^{-1,86}$.

ISBN 978-3-88579-286-4

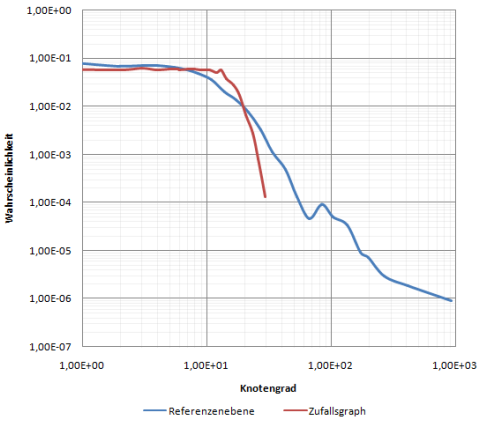

Abbildung 5: Ausgradverteilung (Ref.ebene)

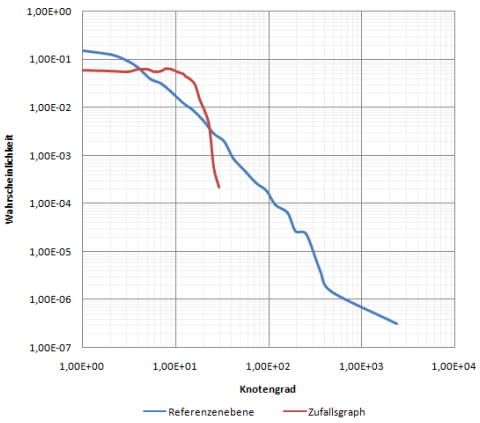

Abbildung 6: Ingradverteilung (Referenzenebene)

Bei der Betrachtung der Referenzenebene 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 [Red98]. 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: $f(x) = 0,861x^{-1,82}$.

Aktualitätsbene Bei der Auswertung der Knotengradverteilung 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. Trotzem 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 Referenzenebene und beträgt 24. Der durchschnittliche Grad beträgt lediglich 0,596. Auffällig ist die höhe Anzahl von isolierten Knoten, die überhaupt keine Ecken besitzen von 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. Die zahlreichen Veränderungen innerhalb vom DNS weisen darauf hin, dass dieses System eine wesentliche Erweiterung erlebt hat. Auffällig ist, dass alle bis auf ein Dokument aus dem *Standards Track* stammen.

Daraus folgt, dass sich möglicherweise die von der IETF bestimmte Bedeutung der unter-

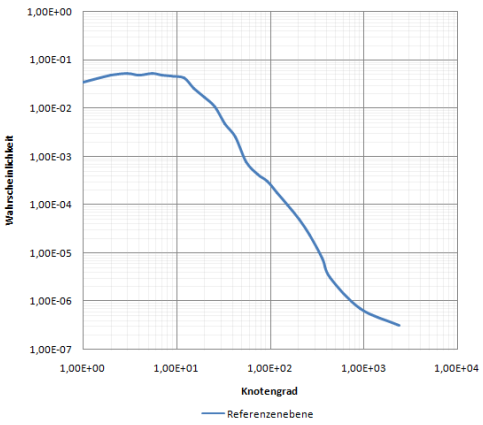

Abbildung 7: Gradvert. (Ref.ebene ungerichtet)

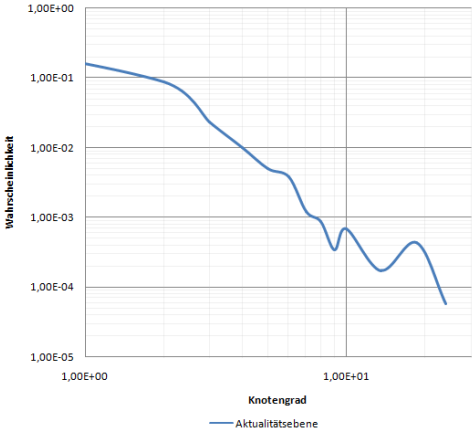

Abbildung 8: Gradverteilung (Aktualitätsebene)

schiedlichen RFC-Kategorien in der Knotengradverteilung der Aktualitätsebenewiderspiegelt. Zusätzlich haben die als *Proposed Standard* kategorisierten Dokumente am häufigsten höhere Anzahl der internen Kanten. Dagegen haben die *Standard*-RFCs mehr externe Kanten. Das entspricht dem *Standards Track*-Anerkennungsprozess. Die alten Standards werden von den neueren, noch mit dem Status *Proposed Standard* RFCs abgelöst.

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. Der α-Exponent beträgt $2,0635$ und die Ausgleichsgerade wird durch folgende Gleichung definiert: $f(x) = 0,0891x^{-2,0635}$.

Zusammenfassend ergab die Untersuchung der Gradverteilung von beiden Ebenen des Normennetzwerks, dass sie einem Potenzgesetz folgt und skalenfrei ist

3.4 Mixing Patterns

Der Begriff *Mixing Patterns* stammt aus der Analyse von sozialen Netzwerken 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 der Entstehung einer Kante zwischen einem Knotentyp A und B folgendem Wert entspricht,

Grad	RFC	Titel	Status	Ext.	Int.
24	1035	Domain names - implementation and specification	Standard	21	3
19	4035	Protocol Modifications for the DNS Security Extensions	Prop. Stand.	1	18
19	4034	Resource Records for the DNS Security Extensions	Prop. Stand.	1	18
13	5462	Multiprotocol Label Switching (MPLS) Label Stack Entry: „EXP“ Field Renamed to „Traffic Class“ Field	Prop. Stand.	0	13
10	2026	The Internet Standards Process – Revision 3	BCP	8	2
10	3473	Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions	Prop. Stand.	10	0
10	4510	Lightweight Directory Access Protocol (LDAP): Technical Specification Road Map	Prop. Stand.	0	10
9	3377	Lightweight Directory Access Protocol (v3): Technical Specification	Prop. Stand.	1	8
9	2181	Clarifications to the DNS Specification	Prop. Stand.	6	3
8	1350	The TFTP Protocol (Revision 2)	Standard	7	1
8	1349	Type of Service in the Internet Protocol Suite	Prop. Stand.	1	7
8	3658	Delegation Signer (DS) Resource Record (RR)	Prop. Stand.	4	4
8	123	Proffered Official ICP	Unknown	4	4

Tabelle 5: Dokumente mit höchstem Grad (Aktualitätsebene)

$$P(A \to B) = \frac{|A|}{n} \cdot \frac{|B|}{n} \tag{5}$$

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.

3.4.1 Knotenartkorrelation

Als Maß für die Korrelationen ohne numerische Bezeichnung des Typs, wird der Assortativitätskoeffizient r verwendet [New03]. Der Wert von r ist folgendermaßen definiert:

$$r = \frac{\mathrm{spur}(e) - \|e^2\|}{1 - \|e^2\|} \tag{6}$$

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: $e = E/\|E\|$.

Informational- und *Proposed Standard*- Memoranda bilden über 67% des gesamten Graphen. Auffällig ist die hohe Diskrepanz innerhalb vom *Standard Track*. Obwohl die Standardvorschläge selbst 36,1% des Normennetzwerks bilden, lediglich 2,4% sind in einen weitere Anerkennungsphase übergangen und nur 1,6% zum offizieller Standard wurden. Das zeigt wie häufig sich die Standardvorschläge nicht durchsetzen können.

ISBN 978-3-88579-286-4

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

Tabelle 7: Normierte Mischmatrix (Referenzenebene)

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

Tabelle 8: Normierte Mischmatrix (Aktualitätsebene)

erschienen im Tagungsband der INFORMATIK 2011
Lecture Notes in Informatics, Band P192
ISBN 978-3-88579-286-4
weitere Artikel online:
http://informatik2011.de/519.html

Aus der normierten Mischmatrix für die Referenzenebene ergibt sich, dass die am häufigsten im gesamten Netzwerk vorkommenden *Proposed Standard*-Dokumente auch sehr oft zitiert werden. Die historischen Memoranda beziehen sich häufig auf die Standards, was durch ihr Alter erklärt werden kann. Es ist eine Diskrepanz zwischen *Proposed Standard* und *Informational* Dokumenten sichtbar. Die zweite Dokumentenart wird um Faktor 2 bis 4 seltener zitiert. Daraus folgt, dass die informellen Memoranda, trotz ihrer zahlreichen Vorkommen weniger Bedeutung für das gesamte Netz haben. Der Assortativitätskoeffizient beträgt für diese Netzwerkebene $r = 0,15$.

Auf der Aktualitätsebene 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 der anderen Typen. In diesem Netzwerk tendieren die unbekannten Dokumente auch dazu, sich mit gleichartigen zu verbinden, die Anzahl von Kanten zu anderen Typen ist marginal. Der Assortativitätskoeffizient beträgt $r = 0,509$. Dieser Wert weist eindeutig auf eine Korrelation zwischen Knotenarten auf dieser Netzwerkebene hin.

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

Tabelle 6: Verteilung von Knotenarten

4 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. beim *Standards Track* Anerkennungsprozesses. Er beeinflusst aber die Struktur des Netzwerks nicht und verändert keine statistische Metriken, außer denen, die eine Knotenartkorrelation wiedergeben.

Im Lauf der Zeit ändern sich die Stati von den RFC-Dokumenten, die referenzierten Memoranda werden aktualisiert und ersetzt. Das vermindert die Qualität von den dort enthaltenen Normen und deutet auf einen Aktualisierungsbedarf hin. 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. Folgender Vorgang bietet sich als eine Möglichkeit für die Auswertung des gesamten Netzwerks an: (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 und (3) Wert aus (2) durch den Wert aus (1) dividieren.

Es existieren zahlreiche veraltete Verweise. 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 aud dem ersten Platz, wenn man sowohl *Updated* als auch *Ob-*

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

Tabelle 9: Allgemeiner Qualitätsfaktor für Kanten des RFC-Normennetzwerks

soleted in die Statistik mit einbezieht. Der relativ niedrige Wert für solche RFCs wyrde von der großen Anzahl von überhaupt nicht verlinkten Dokumenten verursacht.

Die Auswertung des Qualitätsfaktors zeigt ein überraschend schlechtes Ergebnis. 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. Jedoch, bei der schneller Geschwindigkeit von der Entwicklung von neuen Standards wäre es enorm aufwendig, alle Änderungen zu beachten und ab sofort einzupflegen. Allerdings kann die maschinelle Auswertung von den Referenzen keine Aussagen über die Wichtigkeit von solchen Zitierungen machen.

5 Zusammenfassung und Schlussfolgerungen

Die Auswertung erfolgte auf zwei unabhängigen Ebenen, die auf der gleichen Knotenmenge basierten. Die Referenzenebene enthält ausschließlich Informationen über die in den Texten enthaltenen Referenzen. Die Aktualitätsebene bezieht sich auf die Aktualisierungen und Ersetzungen von den RFC-Dokumenten.

Die Ergebnisse der Untersuchung von sozialen Beziehungen haben auf Merkmale hingewiesen, die ein natürlich erzeugtes *Small-World-Network* charakterisieren:

- Niedrige mittlere Weglänge. Der Wert für die Referenzenebene 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 logarithmischer Wachstum der mittleren Weglänge. Für beide Netzwerkebenen lässt sich derartige Steigung der Kurve beobachten.
- Höher Clusterkoeffizient. Der globale Clusterkoeffizienten beträgt $C = 0,355$ für die Referenzenebene und $C = 0,076$ für die Aktualitätsebene. Watts-Strogatz-Clusterkoeffiziente betragen entsprechend $0,351$ und $0,026$.

Eine weitere Eigenschaft vieler natürlich erzeugter Netze ist die Skalenfreiheit der Knotengradverteilung. Die Überprüfung dieser Eigenschaft erfolgte grafisch, indem die Kno-

ISBN 978-3-88579-286-4

tengradverteilung (berechnet mit der Methode von exponentieller Einteilung) auf einer logarithmischen Skala dargestellt wurde. Die resultierende Kurve hat in weiten Bereichen einen nähezu linearen Verlauf, insbesondere im rechten Teil. Das ist ein direkter Beweis für die Skalenfreiheit der Knotengradverteilung auf beiden RFC-Netzwerkebenen.

Das RFC-Normennetzwerk wurde auf Korrelationen der Knotenarten und der Knotengrade untersucht. In dem ersten Fall erfolgte die Auswertung mit dem Assortativitätskoeffizienten r. Für die Referenzenebene beträgt er $0,15$, was auf keinen Zusammenhang hinweist. Auch die direkte Analyse der Ergebnissentabelle ergab, dass die Knotentypen, die einfach am häufigsten vorkommen auch am häufigsten verlinkt werden. Die Aktualitätsebene zeigt im Gegensatz eine Korrelation zwischen Knotentypen von $r = 0,509$.

Als zweite Korrelation wurde die Abhängigkeit von dem Knotengrad untersucht. Diese Auswertung erfolgte mit dem Pearson-Korrelationskoeffizienten und hat bewiesen, dass die Knotengrade nicht korelliert sind.

Price erstellte ein mathematisches Modell für das von ihm untersuchte Zitationsnetzwerk. Da die Referenzenebene des RFC-Normennetzwerks im Prinzip auch ein Zitationsnetzwerk ist, wurde überprüft, ob ihre Entwicklung den Grundlagen des Priceschen Modells folgte. Die Auswertung¡ von Ergebnissen zeigte, dass das RFC-Netzwerk nach anderen Prinzipien gewachsen ist und die Voraussetzungen dieses Modells nicht erfüllt sind.

Als letztes wurde die Aktualität der Dokumenten qualitativ ausgewertet. Es wurde untersucht, wieviele Verweise auf die veralteten Dokumente zeigen. Dieser Wert beträgt für das gesamte Netzwerk 22,5%. Aufgrund der Komplexität der Referenzenebene und unteschiedlicher Wichtigkeit von Verweisen darf jedoch nicht eindeutig gesagt werden, dass das RFC-Normennetzwerk von schlechter Qualität sei.

Literatur

[BA99] A. L. Barabasi und R. Albert. Emergence of scaling in random networks. *Science (New York, N.Y.)*, 286(5439):509–512, Oktober 1999.

[New03] Newman. The Structure and Function of Complex Networks. *SIREV: SIAM Review*, 45:167–256, 2003.

[Pri76] Derek D. Price. A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. Sci.*, 27(5):292–306, 1976.

[Red98] S. Redner. How Popular is Your Paper? An Empirical Study of the Citation Distribution. *European Physics Journal*, B(4):131–138, April 1998.

[RFC11] RFC Index. Bericht, Internet Engineering Task Force, http://www.rfc-editor.org/rfc-index.html, 2011.

[Wat99] D. J. Watts. *Small worlds : the dynamics of networks between order and randomness*. Princeton University Press, 1999.

[WS98] Duncan J. Watts und Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Juni 1998.