Zuverlässige und Schnelle Erzeugung von Zufallsnetzwerken für Evaluationszwecke

Darko Obradović, Wolfgang Schlauch

AG Wissensbasierte Systeme, Fachbereich Informatik, TU Kaiserslautern
Kaiserslautern, Germany
darko.obradovic@dfki.uni-kl.de
wolfgang.schlauch@dfki.uni-kl.de

***Abstract*— Die Analyse von Netzwerken hat in der Soziologie seit dem 20. Jahrhundert einen hohen Stellenwert, ist aber auch für andere Disziplinen interessant. Zum Beispiel bei der Untersuchung von neuronale Netzen oder DNA-Transkriptionen in der Biologie. Für die Evaluation von Mustern gibt es verschiedene Methoden, wobei wir in dieser Arbeit den Fokus auf den statistischen Vergleich des realen Netzwerks mit gleichartigen Zufallsnetzwerken legen, da diese in besonderem Maße objetive Aussagen liefern kann. Für diese Methode existieren zwei Verfahren mit verschiedenen Unzulänglichkeiten, welche beide nicht in allen Fällen zuverlässig machen. Für das Markov-Ketten Monte-Carlo Verfahren versuchen wir, das offene Problem einer effizienten, aber auch zuverlässigen Schrittzahl mit einem neuen Ansatz anzugehen. Das Idee besteht darin, eine möglichst exakte Berechnung dieser Schrittzahl für bestimmte KLassen von Netzwerken über ein Modell der Markov-Kette zu erreichen. Sollte uns das in unserer zukünftigen Arbeit fundiert gelingen, wäre dies der bisherigen empirischen Methode deutlich überlegen.**

I. Einführung

In diesem Papier beschäftigen wir uns mit der Analyse von Netzwerkstrukturen, welche theoretisch stark in der Graphentheorie der Mathematik verwurzelt ist, praktisch aber vor allem durch die Untersuchung von sozialen Netzwerken vorangetrieben wird [1]. Trotz der Fokussierung dieser Disziplin auf soziale Netzwerke, woher sie auch ihren Namen *Soziale Netzwerkanalyse* (SNA) erhielt, lassen sich die Methoden auf jede andere Disziplin übertragen, welche ihre zu analysierenden Strukturen in Netzwerken repräsentieren kann. Folglich zeigten sich auch schon erste Anwendungen der Biologie in Neuronalen- oder Gen-Netzwerken.

A. Historie der Sozialen Netzwerkanalyse [2]

Ende der 1930er Jahre tauchte bei John Moreno erstmals das Konzept des *Soziogramms* auf, mit dem er soziale Beziehungen formalisierte, um sie danach methodisch zu untersuchen. Im Jahre 1954 wurde der Term „Soziales Netzwerk“von J. A. Barnes als ein feststehender Ausdruck verwendet, um sich wiederholende Strukturen in verschiedenen Gruppierungen zu beschreiben, zum Beispiel in Stämmen, Familien oder auch in sozialen Kategorien.

Seitdem ging es mit der Untersuchung von sozialen Netzwerken langsam, aber stetig, weiter. Einerseits untersuchte Harrison White mit seiner Studiengruppe an der Harvard University, Department of Social Relations, dieses Themengebiet auf einer sozialen Ebene, während Charles Tilly sich mit sozialen Bewegungen und politischen Netzwerken beschäftigte. Besondere Erwähnung verdient insbesondere Stanley Milgram, der mit der These des *Kleinen-Welt-Phänomens* die Untersuchung der sozialen Netzwerke vorantrieb und durch die These, dass in den USA jeder Mensch mit jedem anderen über durchschnittlich sechs Verbindungen bekannt sei, großes Interesse erweckte, auch wenn dieses Experiment methodisch stark umstritten ist.

Durch Duncan J. Watts und Steven H. Strogatz wurde das Kleine-Welt-Phänomen 1998 weiter untersucht [3]. Sie brachten interessante neue Ergebnisse hervor im Bezug auf bestimmte Effekte wie den Abstand zweier Knoten voneinander. Sie nahmen in der realen Welt vorkommende Netze, zum Beispiel den Graphen der Zusammenarbeit von Schauspielern, und untersuchten sie, schlugen aber auch die genauere Untersuchung von anderen sozialen Netzwerken vor.

B. Evaluation von Netzwerkeigenschaften

Bei der Analyse von realen Netzwerken sind zwei grundlegende Methoden stark verbreitet.

Zum Einen die *explorative Analyse*, bei welcher der Forscher eine graphische Darstellung des Netzwerks interaktiv durchsieht, oder spezielle Layouts betrachtet, um auffällige Muster oder Eigenschaften zu entdecken. Diese Entdeckungen sind dann das Ergebnis der Analyse. Hierfür existieren eine Vielzahl von Tools wie bspw. **GUESS**.

Zum Anderen gibt es noch die *Metriken-basierte Analyse*, bei welcher Metriken auf dem Netzwerk (z.B. Dichte, Durchmesser, ...) oder Metriken für einzelne Knoten (z.B. Zentralität, Clustering-Koeffizient, ...) im Durchschnitt oder in ihrer Verteilung betrachtet werden. Hier können aber auch Algorithmen zur Suche bestimmter quantifizierbarer Muster zum Einsatz kommen (z.B. Clustering, Motive, ...).

Die explorative Methode ist bei vielen Forschern, die anwendungsorientiert arbeiten, sehr beliebt, da man mit Standardwerkzeugen sehr schnell und einfach zu Ergebnissen kommen kann. Diesen Ergebnissen haftet jedoch stets ein Anschein der Subjektivität an, wo man in der Wissenschaft doch aber stets um Objektivierungen bemüht ist. Folglich ist die Metriken-basierte Analyse nach allgemeiner Ansicht die aussagekräftigere und stärkere Methode. Für viele Standardmetriken gibt es heute auch schon mächtige Werkzeuge wie z.B. **Pajek**. Diese stoßen zwar an ihre Grenzen, wenn das

Repertoire der Standardmetriken zur Beantwortung der Forschungsfragen nicht ausreicht, und erfordern dann weitreichende Programmierfähigkeiten. Dennoch hat sich die Metriken-basierte Methode heute für wissenschaftliche Untersuchungen weitgehend durchgesetzt, auch in Kombination mit der explorativen Methode.

Eine dritte, recht neue Methode ist die *statistisch-stochastische Analyse*, welche die Metriken-basierte Analyse erweitert. Die Objektivität der Metriken-basierten Analyse ist nämlich an jener Stelle nicht mehr gegeben, an welcher die numerischen Ergebnisse durch den Forscher interpretiert werden. Problematisch sind hier Aussagen, welche die Besonderheit von Mustern, bzw. deren erwartetes Auftreten subjektiv bewerten. Dies wurde in einem sehr bedeutenden Artikel von Watts und Strogatz 1998 erstmals eindrücklich durch die Untersuchung von Zufallsnetzwerken demonstriert [3]. In diesem Licht erscheinen z.B. auch die „Six Degrees of Separation“ weit weniger spektakulär als ursprünglich angenommen, da das Gegenteil in einem nicht perfekt-regulären Netzwerk objektiv gesehen eine viel größere Sensation gewesen wäre.

Diese Beobachtung machten sich Forscher dann auch erstmals zu Nutze, um Interpretationen von Metriken zu objektivieren. Hierfür wird eine Anzahl vergleichbarer[1] Zufallsnetzwerke generiert, und die gewünschte Metrik für diese ermittelt. Somit erhält man einen Anhaltspunkt, welchen Wert der Metrik man statistisch in gleichartigen Netzwerken erwarten kann.

So wies beispielsweise Alon 2007 nach, dass bestimmte *Motive* in biologischen Netzen um die vielfache Standardabweichung öfter auftauchen, als im Durchschnitt von gleichartigen Zufallsnetzwerken. Erst jetzt ist also die Aussage, dass die gemessene Ausprägung einer Metrik ungewöhnlich sei, tatsächlich objektiv nachgewiesen.

C. Motivation

Die statistisch-stochastische metrik-basierte Untersuchungsmethode für reale Netzwerke ist eine recht junge, aber sehr vielversprechende Methode, wie die Rückmeldungen aus der Forschergemeinschaft zeigen. So erregte damals die Arbeit von Watts and Strogatz schon viel Aufmerksamkeit, die Ergebnisse Alons aus der Praxis wurden ebefalls renommiert publiziert. Und auch wir haben für Analysen der Blogosphäre mit Hilfe dieser Methode viel Anerkennung erhalten [4], weshalb wir diese Methode intensiv weiterverfolgen möchten.

Als kritischer Punkt hat sich hierbei jedoch die Wahl des Algorithmus zur Generierung dieser gleichartigen Zufallsnetzwerke erwiesen. Die Gleichartigkeit wird zuallererst natürlich über die Anzahl an Knoten und Kanten definiert, nach allgemeinem Konsens aber auch über die Knotengrade der einzelnen Knoten, welche eine vorgegebene Knotengradsequenz für das Netzwerk zur Folge hat. Nun gibt es zwei Algorithmen zur Generierung eines Zufallsnetzwerkes mit einer solchen vorgeschriebenen Sequenz, welche in [5] erläutert und verglichen werden. Zum Einen das *Konfigurationsmodell* als direkt generierender Algorithmus, welcher sehr einfach, und in der Praxis am weitesten verbreitet ist, jedoch die ihm gestellte Aufgabe nur bis zu einem bestimmten Grad erfüllt, und Zweifel an seiner Zuverlässigkeit offen lässt. Zum Anderen gibt es den *Markov-Ketten Monte-Carlo Algorithmus*, welcher deutlich aufwändiger ist, und sich in der Praxis noch nicht etablieren konnte. Er erfüllt zwar die Anforderungen an die Zufallsnetzwerke voll, hinterlässt beim Anwender aber, wie viele andere Markov-Ketten-basierte Algorithmen auch, die Frage nach der richtigen Schrittzahl, welche einerseits für die Geschwindigkeit und andererseits für die Uniformität des Verfahrens, und somit auch für seine stochastische Aussagekraft von entscheidender Bedeutung ist.

Eine lineare Größenordnung dieser Schrittzahl, welche für die Praktikabilität des Verfahrens von entscheidender Bedeutung ist, wird bisher nur vermutet, und für den Vorfaktor gibt es nur indirekte empirische Untersuchungen. Tatsächlich bewiesen wurde bisher nur eine quadratische Größenordnung abhängig von der zu generierenden Kantenzahl. Diese Situation führt bei uns, und möglicherweise auch bei anderen Forschern, noch zu einer Zurückhaltung vor der Anwendung.

In diesem Papier präsentieren wir unsere Idee und erste Experimente, um die Schrittzahl in Einzelfällen exakt zu bestimmen. Indem wir dies für verschieden große Netzwerke einer bestimmten, sehr weit verbreiteten Klasse von Knotengradsequenzen, den sogenannten *skalenfreien Netzen* [6] ermitteln, erhoffen wir uns, allgemeine und verlässliche Rückschlüsse auf die notwendige Schrittzahl für die Uniformität in dieser Klasse zu ermitteln.

II. Zufallsnetzwerke

In diesem Kapitel gehen wir im Detail auf die bereits erwähnte statistisch-stochastische Untersuchungsmethode für Netzwerke ein, und beschreiben die daraus resultierenden Anforderungen an entsprechende Generierungsverfahren für Zufallsnetzwerke. Danach werden die bekannten Verfahren erläutert, kritisch betrachtet und verglichen.

A. Die Statistisch-Stochastische Methode im Detail

In diesem Abschnitt wollen wir genauer auf die in Kapitel I-B beschriebene statistisch-stochastische metrik-basierte Untersuchungsmethode für reale Netzwerke eingehen. Ausgangspunkt sind hierbei zum Einen das reale Netzwerk als Gegenstand der Untersuchung, und zum Anderen die vom Forscher ausgewählte Metrik, über welche er Aussagen treffen möchte.

Neben der Berechnung der Metriken auf dem realen Netzwerk, wird diese auch mit ihren erwarteten Ausprägungen bei gleichartigen Zufallsnetzwerken verglichen. Diese Gleichartigkeit definiert sich über die Anzahl der Knoten und Kanten, und zusätzlich auch über die Knotengradsequenz des Netzwerkes. D.h. dass zu jedem Knoten des realen Netzwerkes im Zufallsnetzwerk ein Pendant mit exakt der gleichen Zahl an ein- und ausgehenden Kanten existiert.

Verfügt man gedanklich über eine Urne, welche alle möglichen Netzwerke enthält, die in Größe und Gradsequenz

[1] mehr zur Vergleichbarkeit in den folgenden Abschnitten

dem realen Netzwerk entsprechen, so kann man nun eine hinreichend große Anzahl s dieser Netzwerke mit jeweils gleicher Wahrscheinlichkeit ziehen, und erhält somit eine repräsentative Teilmenge. Üblich sind hier alles zwischen 30 und 1000 Ziehungen. Für jedes gezogene Netzwerk wird nun die Metrik berechnet, und für die Menge insgesamt kann man nun die deskriptive statistische Verteilung der Ausprägung ermitteln. Dies resultiert in einem Durchschnittswert mit einer Standardabweichung.

Für das reale Netzwerk kann man nun den absoluten Abstand der Metrik zum statistischen Mittel der Zufallsnetzwerke berechnen, und gibt diesen geteilt durch die Standardabweichung als z-Wert an. Hat die Metrik auf dem realen Netzwerk ein $z < 1$ so ist die Eigenschaft im Rahmen des zu Erwartenden, weist sie allerdings ein $z > 3$ auf, so kann man davon ausgehen, eine Besonderheit des realen Netzwerks im Vergleich zu gleichartigen Zufallsnetzwerken entdeckt zu haben. Die Interpretation der Bedeutung dieser Besonderheit ist nun natürlich wieder subjektiv, stützt sich mittlerweile aber auf schon zwei objektivierte Aussagen. Im Normalfall wird damit eine vorher entsprechend aufgestellte Hypothese überprüft.

B. Anforderungen an die Verfahren

Ein konkretes Netzwerk, welches eine gegebene Knotengradsequenz realisiert, nennt man eine *Konfiguration* dieser Sequenz. Nun übersteigt schon bei kleinen Sequenzen die Anzahl der Konfigurationen aufgrund der kombinatorischen Explosion alle Speichermöglichkeiten [7], weshalb eine entsprechende Urne durch ein uniformes Generierungsverfahren für diese Konfigurationen simuliert werden muss; uniform insofern, dass jede mögliche Konfiguration exakt die gleiche Wahrscheinlichkeit haben muss, durch ein solches Verfahren generiert zu werden.

Aufgrund der Beschaffenheit fast aller realer Netzwerke kommen noch folgende Einschränkungen an die Menge der Konfigurationen hinzu. Erstens soll die Gesamtmenge aller Konfigurationen nur einfache Graphen enthalten, d.h. sie sollen weder parallele Kanten noch Kanten zwischen ein und demselben Knoten, welche wir hier Selbst-Kanten nennen, enthalten. Zweitens sind oftmals auch nur zusammenhängende Graphen gewünscht, also Graphen ohne vollständig isolierte Subgraphen. Diese Anforderungen sollen die statistisch-stochastischen Werte auf Grundlage der real überhaupt nur in Frage kommenden Netzwerke ermitteln.

C. Das Konfigurationsmodell

Das Konfigurationsmodell ist ein sehr einfaches, direkt generierendes Verfahren (siehe [8] Abschnitt IV.B.1). Die Idee besteht darin, alle Knoten zu Beginn mit der entsprechenden Anzahl an Kantenstummeln zu versehen, und danach iterativ jeweils zwei zufällig ausgewählte Stummel mittels einer Kante miteinander zu verbinden. Dies funktioniert in linearer Zeit, generiert aber zum Einen auch nicht-einfache Graphen, und kann auch die Verbundenheit des Netzwerkes nicht garantieren.

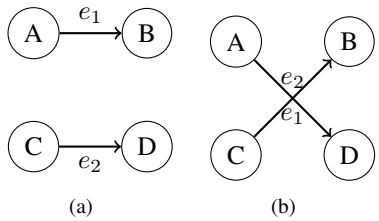

Fig. 1. gültiger Kantentausch

Um dennoch einfache Graphen zu erhalten, werden in der Praxis parallele Kanten und Selbst-Kanten einfach gelöscht, was aber sowohl die Anzahl der Kanten als auch die Gradsequenz des damit generierten Netzwerkes verfälscht. Dies wird in der Praxis bei großen Netzwerken gerne als vernachlässigbar dargestellt. Zwar mag dies oftmals stimmen, dennoch gibt es Erkenntnisse, welche Probleme bei der Uniformität des Verfahrens nachweisen [5]. Insgesmat sind wir der Meinung, dass das Verfahren unter strengsten wissenschaftlichen Kriterien als nicht zuverlässig betrachtet werden muss.

Die Verbundenheit der Graphen kann in keinster Weise garantiert werden, aber auch der Versuch, alle nicht verbundenen generierten Graphen zu verwerfen ist für fast alle real vorkommenden Gradsequenzen hoffnungslos, wie fundierte Abschätzungen über den Anteil nicht verbundener Graphen für gegebene Sequenzen zeigen [9]. Benötigt man in der Praxis dennoch diese Einschränkung, so ist es üblich, nur den größten zusammenhängenden Teilgraphen des Netzwerkes zu nehmen. Eine solche *giant component*, welche nahe an die vorgegebene Größe des Netzwerkes herankommt, wird es zwar meistens mit hoher Wahrscheinlichkeit geben [10], das Resultat verliert aber noch weiter an Zuverlässigkeit.

Von Uniformität kann man bei diesem Verfahren kaum noch sprechen, da die generierten Netzwerke gar nicht in der vorgegebenen Menge der Konfigurationen liegen. Hier werden stattdessen die statistischen Werte bestimmter Metriken als Vergleich herangezogen, die Uniformität also rein empirisch betrachtet.

D. Der MCMC-Algorithmus

Der Markov-Ketten Monte-Carlo Algorithmus ist ebenfalls ein Verfahren für die Generierung entsprechender Konfigurationen. Die Idee ist hierbei, aus einer Startkonfiguration heraus solange zwei zufällig ausgewählte Kanten miteinander zu vertauschen, bis jede Konfiguration mit der gleichen Wahrscheinlichkeit erreicht wird.

Als Startkonfiguration für das Verfahren nimmt man entweder das reale Netzwerk, oder, falls nur eine Gradsequenz gegeben ist, generiert man eine Konfiguration mit dem Havel-Hakimi-Algorithmus [11]. Beim Tausch zweier zufällig ausgewählter Kanten gibt es gültige und ungültige Folgekonfigurationen. Ungültig ist ein Kantentausch dann, wenn die Folgekonfiguration kein einfacher Graph mehr wäre, dies lässt sich jedoch sehr leicht anhand der ausgewählten Kanten vorab überprüfen, wie in den Abbildungen 1 und 2 zu sehen ist.

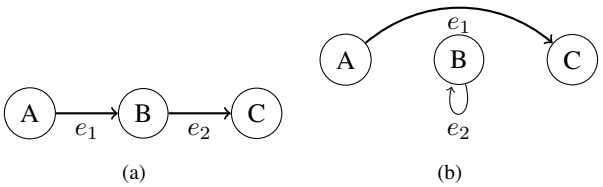

Fig. 2. ungültiger Kantentausch wegen Schleife

Aufgrund theoretischer Gegebenheiten konvergiert dieses Verfahren bei einer Schrittzahl q von erfolgreichen oder versuchten Kantentauschen gegen unendlich zu einer perfekt uniformen Generierung aller Konfigurationen. Doch genau bei der Bestimmung dieser Schrittzahl q liegt das Problem dieses Verfahrens, wie bei vielen anderen Markov-Ketten-Verfahren auch. Nachgewiesen ist, dass eine Schrittzahl in der Größenordnung von $O(m^2)$ für die Uniformität ausreichend ist, jedoch gibt es keine verlässliche Aussage über den Vorfaktor. Empirische Tests legen gar eine lineare Größenordnung nahe, und schlagen einen Vorfaktor f_q von 100 dazu vor [5], [12], also insgesamt eine Schrittzahl von $q = 100 \cdot m$. Wir werden in Kapitel III genauer auf diese Problematik und Überlegungen zu möglichen Auswegen eingehen.

Vorab wollen wir noch kurz auf die Generierung von ausschließlich verbundenen Graphen eingehen, welche mit dem MCMC-Algorithmus unter Wahrung der Uniformität möglich ist, was das Verfahren deswegen auch so attraktiv macht. Nehmen wir an, die Schrittzahl liege tatsächlich in $O(m)$. Man kann nun nach jedem Kantentausch überprüfen, ob die Nachfolgekonfiguration ein verbundener Graph ist oder nicht. Dies kann mittels einer einfachen Tiefensuche in m Schritten entschieden werden. Davon abhängig gilt der Kantentausch dann entweder als gültig, oder aber man geht wieder zur ursprünglichen Konfiguration zurück und versucht erneut einen Kantentausch, ohne den vorangegangenen mitzuzählen. Dies führt insgesamt zu einer Laufzeit des Verfahrens in $O(m^2)$. Viger und Latapy haben diese Methode weiterentwickelt, und nachweislich eine Erhöhung um nur O(log m) für die Verbundenheitsprüfung erreicht, so dass ihr optimierter MCMC-Algorithmus verbundene Netzwerke mit einer Laufzeit in $O(m \cdot log\ m)$ generiert.

Bestünde nun Gewissheit über die für die Uniformität tatsächlich notwendige Schrittzahl, so würde dieser Algorithmus zuverlässig und schnell die gewünschten verbundenen Netzwerke generieren. Tatsächlich wird dieser Algorithmus aber bis heute in keinem der Standardwerkzeuge und -bibliotheken angeboten. Stattdessen findet man stets eine Implementierung des Konfigurationsmpdells vor, trotz der oben beschriebenen Unzulänglichkeiten. Im nächsten Kapitel werden wir unsere Ideen für eine bessere Bestimmung der Schrittzahl vorstellen, um so einen Beitrag für die Etablierung der MCMC-Algorithmen zu leisten.

III. Die Schrittzahl des MCMC-Algorithmus

In diesem Kapitel stellen wir den Stand der Wissenschft bezüglich der Bestimmung der Schrittzahl für den MCMC-Algorithmus vor, so dass dieser möglichst uniforme Ergebnisse liefert, aber nicht übermäßig viel Zeit braucht. Wir hinterfragen diesen kritisch und zeigen anhand einiger Beispiele Unzulänglichkeiten und mögliche Lösungswege hierfür auf.

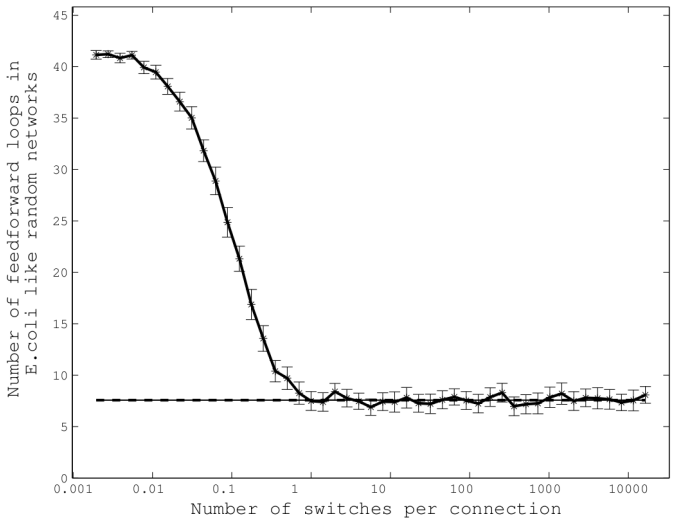

Fig. 3. Empirische Schätzung des Vorfaktors f_q aus [5]

A. Bisherige Festlegung

Dass die Schrittzahl q für den MCMC-Algorithmus linear von der Anzahl der Kanten m abhängt ist bisher wie gesagt nur eine Vermutung. Um unter dieser Voraussetzung den passenden Vorfaktor f_q zu ermitteln, wurde in vorangegangenen Arbeiten [5], [12] stets eine indirekte empirische Methode verwendet. Das Prinzip beruht darauf, sich eine Startkonfiguration mit einer sehr ungewöhnlich ausgeprägten Metrik zu nehmen, davon ausgehend wiederholt Kantentausche durchzuführen, und zu messen, ab welcher Schrittzahl die Metrik sich beim erwarteten Durchschnittswert einpendelt. Hieraus wird dann eine ausreichende Unifomität des Verfahrens geschlussfolgert.

Abbildung 3 aus [5] zeigt eine solche Bestimmung anhand des Transkriptionsnetzwerkes des E.Coli-Bakteriums. Als Metrik wurde die gefundene Anzahl des Motivs „Feed-Forward Loop“ gewählt. Man sieht recht gut, dass der Erwartungsswert gleichartiger Zufallsnetzwerke in diesem Beispiel schon mit $f_q = 1$ erreicht wird, dennoch ist es die Schlussfolgerung der Autoren, im allgemeinen Fall lieber $f_q = 100$ zu wählen.

Das Problem dieser Art der Bestimmung sind die Auswahl der Metrik und einer dazugehörigen außergewöhnlichen Startkonfiguration. Nun haben schon Watts und Strogatz 1998 gezeigt, dass verschiedene Metriken verschieden schnell auf Kantentausche reagieren [3], eine unter allen Umständen „trägste“ Metrik ist jedoch nicht bekannt. Dennoch erscheint es fragwürdig, eine besonders dynamische Metrik wie den Durchmesser eines Netzwerkes für solch einen Test heranzuziehen, wie in Vorarbeiten geschehen [12].

Das andere Problem betrifft die Startkonfiguration. Hat das Transkriptionsnetzwerk zwar mit 42 Motiven einen z-Wert von ca. 10 bei der Abweichung von der erwarteten Motivanzahl,

so lassen sich bei der gegebenen Sequenz dennoch auch leicht Konfigurationen mit über 100 Motiven realisieren, was das Konvergieren aus einer solchen Konfiguration heraus weiter verlangsamen würde. Auch die Größen der Netzwerke bleiben unzureichend berücksichtigt, da diese Art der Bestimmung sehr zeitaufwändig ist. Das kommunizierte Ergebnis ist hier immer, dass „für eine Vielzahl verschiedener Netzwerke und verschiedener Metriken sich $f_q = 100$ immer als ausreichend groß herausgestellt hat".

Wir sind mit diesem Ergebnis bezüglich der Zuverlässigkeit des Verfahrens nicht zufrieden, und die Zurückhaltung bei der Akzeptanz dieses Verfahrens scheint dies zu bestätigen. Wir stellen in den folgenden Abschnitten unseren alternative Ansatz für die Bestimmung der Schrittzahl q vor.

B. Die Markov-Kette

Zu Beginn unserer Überlegungen beschäftigen wir uns mit der dem MCMC-Algorithmus zu Grunde liegenden Markov-Kette. Jede mögliche Konfiguration einer Gradsequenz ist ein Knoten der Markov-Kette, den wir zur besseren Unterscheidung *Konfigurationsknoten* nennen werden. Jeder gültige Kantentausch führt zu einem anderen Konfigurationsknoten, und die Umkehroperation auch immer wieder zurück, so dass immer bidirektionale Verbindungen zwischen zwei Konfigurationsknoten bestehen. Alle ungültigen Kantentausche zählen ebenfalls als Schritt des Algorithmus, da sie aber zu keiner neuen Konfiguration führen, werden sie deshalb in einer Selbst-Kante mit entsprechender Gewichtung zusammengefasst.

Es gibt insgesamt $\frac{m^2+m}{2}$ ungeordnete Kantenpaare, so dass jeder Konfigurationsknoten diesen Wert als Ein- und Ausgangsgrad besitzt. Da auch die Verbundenheit der Markov-Kette bewiesen ist, führt dies theoretisch zu einer perfekten Uniformität. Entscheidend hierfür ist in der Praxis aber die notwendige Schrittzahl, bis zum sogenannten *Mixing* der Markov-Kette.

C. Exakte Berechnung im Einzelfall

Für gegebene Gradsequenzen und sehr kleine Netzwerke lässt sich die Markov-Kette mit normaler Hardware noch vollständig erzeugen. Verfügt man über diese Markov-Kette, so kann man die für deren Mixing notwendige Schrittzahl mit einer Variante des einfachen ungedämpften PageRank-Algorithmus [13] exakt berechnen. Hierfür wird initial nicht jedem Knoten der Wert 1 gegeben, sondern der Startkonfigurationsknoten erhält den Wert m, und alle anderen Konfigurationsknoten den Wert 0. Nun werden soviele Iterationen q des ungedämpften PageRank-Algorithmus durchgeführt, bis der Wert aller Knoten $1 \pm \delta$ beträgt.

Die Werte entsprechen der Wahrscheinlichkeit geteilt durch m, nach dem jeweiligen Schritt des MCMC-Algorithmus bei dieser Konfiguration zu sein. Der Wert δ gibt die maximal auftretende bzw. tolerierte Abweichung einer Konfiguration von der geforderten Wahrscheinlichkeit an.

Wir untersuchen dies für vier Gradsequenzen mit 12 Kanten. Zum Einen die des zur Messung der Uniformität in [5] benutzten „toy network", welches über eine sehr spezielle, für die Realität untypische Sequenz verfügt, aber nur 91 Konfigurationen erlaubt. Zum anderen haben wir uns, unserem Interesse an real vorkommenden Sequenzen folgend, eine kleine heterogene Powerlaw-Verteilung (siehe [8] Abschnitt III.C.1) mit 12 Kanten generiert, wie sie in einem skalenfreien Netzwerk vorkommen würde. Wir haben $\alpha = 1,7$ gewählt und damit die Knotengrade $(3, 2, 2, 1, 1, 1, 1, 1)$ für 8 Knoten erhalten. Um eine Gradsequenz für einen gerichteten Graphen mit Ein- und Ausgangsgraden zu erhalten, haben wir diese einmal positiv, einmal negativ, und einmal nicht korreliert, also zufällig, auf die acht Knoten verteilt.

TABLE I
EXAKT BERECHNETE SCHRITTZAHLEN

Netzwerk	q	f_q	$\pm\ \delta$
negativ korreliert	61	~ 5	0.009
positiv korreliert	89	~ 7	0.02
zufällig	78	~ 7	0.15
Toy-Netzwerk	21	~ 2	0

Diese vier Sequenzen haben wir mit dem Havel-Hakimi-Algorithmus in einem Netzwerk realisiert, von diesen Konfigurationen aus jeweils den Zustandsraum der Markov-Kette durch Verfolgung aller jeweils möglichen 66 Kantenpaare vollständig generiert und als Netzwerk gespeichert. Die Markov-Kette des Toy-Netzwerks verfügt wie bekannt über 91 Konfigurationsknoten, die Markov-Ketten der drei skalenfreien Netzwerke über 90.000 bis 120.000 Konfigurationsknoten.

Tabelle I listet die errechneten Schrittzahlen q für die vier Beispielnetzwerke auf. Abbruchbedingung war hier, dass sich die Werte nicht mehr ändern, was bei der 32-Bit-Darstellung schon vor Erreichen der Gleichverteilung eintreten kann. Der δ-Wert gibt hierzu die maximale Abweichung eines Konfigurationsknotens vom Wert 1 an. Unter Annahme einer von m linear abhängigen Schrittzahl des MCMC-Algorithmus würde der entsprechend angegebene Vorfaktor f_q gelten.

In allen Fällen liegt f_q weit unter 100, aber auch höher als 1, wie es in Abbildung 3 möglicherweise suggeriert wurde. Dies sind aber Einzelfallbetrachtungen, aus denen man noch keine Verallgemeinerungen folgern kann.

D. Wege zur Verallgemeinerung

Die exakte Berechnung der Schrittweite ist schon bei kleinen Werten von m aufgrund der exponentiell steigenden Anzahl an Konfigurationen unpraktikabel. Um mit dieser Methode dennoch verallgemeinerbare Aussagen, zumindest für bestimmte Klassen von Netzwerken, wie z.B. die skalenfreien Netze, treffen zu können, ist noch ein weiterer Schritt notwendig. Mit einem Verständnis für die Struktur der zugrundeliegenden Markov-Ketten ließe sich die Schrittzahl für das Mixing dieser Ketten über das modifizierte PageRank-Verfahren direkt abschätzen und simulieren, anstatt sie indirekt über das beschriebene empirische Verfahren zu bestimmen.

Hierbei indentifizieren wir 3 wichtige Punkte, welche für die Geschwindigkeit des Kovergieren des PageRank-Verfahrens

von entscheidender Bedeutung sind.

1) die Anzahl der externen Verbindungen eines Konfigurationsknotens, d.h. die Anzahl der Nachbarkonfigurationen und deren Verhältnis zum Gewicht der Selbst-Kante
2) den Clustering-Koeffizienten der Markov-Kette
3) den Durchmesser der Markov-Kette

Wären diese drei Ausprägungen abhängig von der zu Grunde liegenden Gradsequenz abschätzbar, so ließe sich für eine Klasse von Gradsequenzen ein Modell der Markov-Kette aufstellen, und anhand von Simulationen auf verschieden großen Instanzen dieses Modells die notwendige Schrittzahl abschätzen oder gar ohne Simulation direkt berechnen.

Aus unseren bisherigen Erkenntnissen können wir bereits erste Hypothesen aufstellen, deren genauere Untersuchung unsere zukünftige Arbeit sein wird.

Die Topologie der in Abschnitt III-C untersuchten Markov-Ketten der skalenfreien Netze bilden ein mehrdimensionales relativ regelmäßiges Gitter, welches entlang der Dimensionen rundum verbunden ist, ähnlich wie die Oberfläche einer Kugel in zweidimensionaler Sicht. Dies ist offensichlich im allgemeinen Fall so, bedarf jedoch noch eines Nachweises. Definiert wird diese regelmäßige Struktur durch die oben genannten 3 Eigenschaften.

Beispielsweise zeigt Abbildung 4 die Verteilung der externen Verbindungen der Konfigurationsknoten der Markov-Kette der nicht-korrelierten skalenfreien Gradsequenz, welche eine nur geringe Schwankung um den Wert 32 aufweist. Die Anzahl ungültiger Kantentausche ließe sich allgemein gut aus der Sequenz abschätzen, da Kantenpaare, welche nur 3 Knoten involvieren, nie getauscht werden können. Der theoretisch maximal mögliche Durchmesser von $m - 1$ scheint von den skalenfreien Netzen voll ausgereizt zu werden, unsere Beispielsequenzen hatten hier alle den Wert 10. Der Clustering-Koeffizient ist mit ca. $0,1$ vergleichsweise hoch, was durch den hohen Durchmesser bedingt ist, da benachbarte Konfigurationen viele Nachbarn gemeinsam haben.

Diese beiden Metriken lassen sich, die gewisse Regelmäßigkeit vorausgesetzt, durch Stichproben-Generierunge einzelner Konfigurationen und deren Nachbarschaft ermitteln und statistisch abschätzen. Dies wäre in vertretbarer Zeit berechenbar, und würde ebenfalls eine Abschätzung erlauben, und somit eine empirische Ermittlung der notwendigen Schrittzahl auf dem Modell erlauben.

IV. Zusammenfassung

In diesem Papier haben wir den Stand der Wissenschaft bezüglich der Evaluationsmethoden von konkreten realen Netzwerken zusammengefasst, und die unserer Meinung nach vorzuziehende Methode, der Evaluation mittels gleichartiger Zufallsnetzwerke kritisch analysiert. Die aufgezeigten Probleme bei der Zuverlässigkeit der gängigen Methoden wurden bewertet, und für das unserer Meinung nach überlegene MCMC-Verfahren haben wir einen Ansatz vorgestellt, mit welchem das letzte offene Problem, der zu wählenden Schrittzahl q zumindest für bestimmte Klassen von Netzwerken deutlich reduziert werden könnte. Wir werden versuchen, die dafür notwendigen Kenngrößen in Zukunft in ausgiebigen Simulationen abschätzen zu können, und damit eine für jede Netzwerkgröße dieser Klasse sichere und vor allem effiziente Wahl von q ermöglichen zu können.

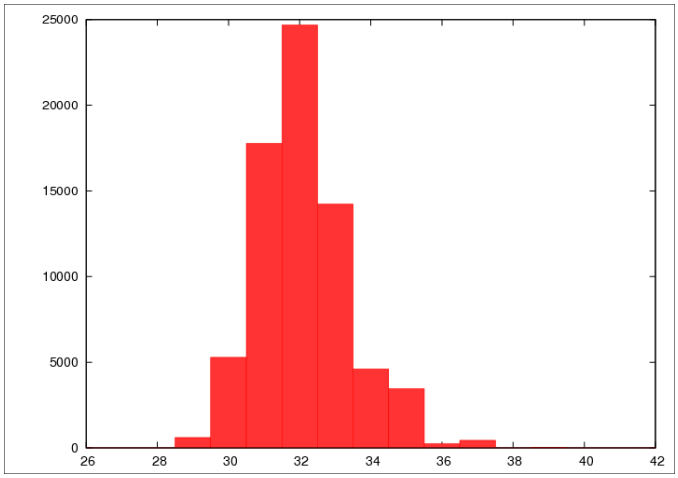

Fig. 4. Gradverteilung in der Markov-Kette des nicht-korrelierten skalenfreien Gradsequenz

References

[1] S. Wasserman, K. Faust, and D. Iacobucci, *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, 1994.

[2] L. C. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004.

[3] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, no. 393, pp. 440–442, 1998.

[4] D. Obradovic and S. Baumann, "A journey to the core of the blogosphere," in *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2009)*. IEEE, 2009, pp. 1–6, best paper award.

[5] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon, "On the uniform generation of random graphs with prescribed degree sequences," *Arxiv preprint cond-mat/0312028*, 2003.

[6] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[7] T. Snijders, "Enumeration and simulation methods for 0-1 matrices with given marginals," *Psychometrika*, vol. 56, no. 3, pp. 397–417, September 1991.

[8] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[9] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random Structures and Algorithms*, vol. 6, pp. 161–179, 1995.

[10] ——, "The size of the giant component of a random graph with a given degree sequence," vol. 7, p. 295–305, November 1998.

[11] P. L. Erdős, I. Miklós, and Z. Toroczkai, "A simple Havel-Hakimi type algorithm to realize graphical degree sequences of directed graphs," *Electronic Journal of Combinatorics*, vol. 17, no. 1, p. R66, 2010.

[12] F. Viger and M. Latapy, "Efficient and simple generation of random simple connected graphs with prescribed degree sequence." in *Proceedings of the 11th international conference on Computing and Combinatorics*, ser. Lecture Notes in Computer Science, L. Wang, Ed., vol. 3595. Springer, 2005, pp. 440–449.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford University," Technical Report, 1998.