# **Kritik**: *Social Networks Spread Rumors in Sublogarithmic Time*, Benjamin Doerr, Mahmoud Fouz, Tobias Friedrich, STOC 2011

*Felix Gessert, 10.06.2012, Seminar maschinelles Lernen (Uni Hamburg)*

## Autoren

Das 2011 auf der STOC (*Symposium on Theory of Computing*) erschienene Paper nennt drei Autoren: Benjamin Doerr, Mahmoud Fouz und Tobias Friedrich. Ihr Hintergrund und Forschungsschwerpunkt, sowie eine Einschätzung ihrer thematischen Kompetenz wird im folgenden Abschnitt diskutiert.

### Benjamin Doerr

| | |
|---|---|
| Person | Professor für Informatik (Uni Saarland), Senior Researcher am MPI Saarland, Dr. der Mathematik, Habilitation an der Uni Kiel |
| Forschungs-Schwerpunkt | Ausbreitungsmuster in Graphen, probabilistische Methoden in der diskreten Mathematik, Pseudozufälligkeit, evolutionäre Algorithmen |
| Publikationen | 49 Journal Artikel, 95 Konferenz Paper, Editor bei einem Buch und 5 Proceedings |
| Zitationen | 940 (nach Google Scholar) |
| Awards | Drei Best-Paper-Awards bei der Genetic and Evolutionary Computation Conference (GECCO) |

Doerr ist ein sehr aktiver Forscher in der mathematisch orientierten, algorithmischen Informatik. Trotz seines jungen Alters hat er eine signifikante und vielzitierte Publikationsliste vorzuweisen und engagiert sich als Mitglied in Program Committees mehrerer Konferenzen und als Editor zweier Journals. Seinen als Professor erworbenen didaktischen Fähigkeiten ist sicherlich die instruktive Darstellung der Ergebnisse des vorliegenden Papers in den *Communications of the ACM* geschuldet. Die Erfahrung in zahlreichen Publikationen zu Themen der stochastischen Graphentheorie und sein offensichtlicher Erfolg auf diesem Gebiet machen Doerr nach meiner Einschätzung zu einem sehr kompetenten Erstautor.

### Mahmoud Fouz

| | |
|---|---|
| Person | Doktorand an der Uni Saarland, Master der Informatik an der Uni Saarland im Gebiet *Computational Complexity*, keine Website mit Informationen |
| Forschungs-Schwerpunkt | Stochastik und Graphentheorie |
| Publikationen | 13 (veröffentlichte) Konferenz Paper, davon 5 zum Thema der Verbreitung von Gerüchten in Netzwerken |
| Zitationen | ca. 50 (nach Google Scholar) |
| Awards | keine |

Mahmoud Fouz ist Doktorand von Doerr an der Uni Saarland. Über seine genaue Tätigkeit können nur indirekte Schlüsse gezogen werden, da er seine Forschungstätigkeit nicht dediziert im Web darlegt. Seinen Publikationen nach zu urteilen, scheint das Thema dieses Papers dem Kern seiner bisherigen Forschung zu entsprechen: der mathematischen Modellierung von Ausbreitungsprozessen in Netzwerken.

### Tobias Friedrich

| | |
|---|---|
| Person | Forschungsgruppenleiter an der Uni Saarland, Senior Researcher am MPI Saarland, Dr. der Informatik, Diplom Mathematik, Master Informatik |
| Forschungs-Schwerpunkt | Verteilte Algorithmen, Pseudozufälligkeit, diskrete Mathematik, evolutionäre Algorithmen |
| Publikationen | 44 Konferenz Paper (die meisten bei *Symposium on Discrete Algorithms* und *Genetic and Evolutionary Computation Conference*), 25 Journal Paper |
| Zitationen | 575 (nach ISI Web of Knowledge) |
| Awards | Drei Best-Paper-Awards bei der Genetic and Evolutionary Computation Conference (GECCO) |

Das Forschungsgebiet von Tobias Friedrich ähnelt dem von Doerr stark. Dies schlägt sich in zahlreichen gemeinsamen Publikationen zu verschiedenen Themen der diskreten Mathematik und Graphentheorie nieder. Wie Doerr gelingt es auch Friedrich, in einer sehr kurzen akademischen Karriere (Promotion 2007) eine große Zahl an Publikationen von hoher Bedeutung zu veröffentlichen. Der Fokus seiner Forschung ist jedoch weniger stark auf Graphentheorie und Netzwerke ausgerichtet als der von Doerr. Seine Kompetenz bezüglich des Themas ist nichtsdestotrotz als hoch einzuschätzen, nicht zuletzt aufgrund zahlreicher Arbeiten, die Stochastik mit diskreter Mathematik verknüpfen. Friedrich ist ebenfalls in Programm Comittees mehrerer Konferenzen aktiv.

## Der Inhalt des Papers

Das Paper unternimmt den Versuch, den in vielen Realdaten sozialer Netzwerke beobachtbaren Mechanismus extrem schneller Informationsverbreitung mit einer mathematischen Erklärung zu untermauern. Konkret werden dazu Netzwerke betrachtet, die dem sehr populären *Barabási–Albert* Modell genügen und untersuchen die Dynamik einer *Push-Pull-Ausbreitungsstrategie* für Gerüchte (bzw. beliebiger „viraler“ Informationen). Die Autoren beweisen für dieses Szenario eine obere und untere Schranke $\Theta(\log(n))$, der Anzahl diskreter Zeitschritte bis zur vollständigen Ausbreitung des Gerüchts. Die Schranke ist der bisher besten bekannten Schranke quadratisch überlegen. Für eine Anpassung des Push-Pull-Modells um die realistische Annahme, dass kein Knoten (Person eines sozialen Netzwerks) sich zweimal mit einem anderen Knoten über ein Gerücht austauscht, können die Autoren sogar eine sublogarithmische Schranke $\Theta(\frac{\log(n)}{\log(\log(n))})$ beweisen.

Das Barabási–Albert Modell ist eine generatives Modell der Klasse „Preferential Attachment“ für Graphen, die Eigenschaften mit realen sozialen Netzwerken teilen. Ausgehend von einem Dichteparameter $m$ wird die Entstehung des Netzes so charakterisiert, dass jeder in sequentieller Reihenfolge hinzukommende Knoten sich mit $m$ anderen Knoten verbindet. Die

Präferenz mit der ein bestimmter Knoten ausgewählt wird, ist mit $P(node) \sim degree(node)$ eine stochastische Größe, wobei $degree$ den Knotengrad angibt. Dieses Prinzip ist als Mathew-Effekt bekannt („the rich get richer“). Der entstehende Graph hat u.a. die Eigenschaft, dass wie in realen Netzen das Histogramm der Knotengrade einem Potenzgesetz genügt.

Die Push-Pull Strategie geht davon aus, dass die Zeit in diskrete, synchrone Runden zerlegt werden kann und jeder Knoten einmal pro Schritt einen Nachbarn zufällig auswählt und dabei entweder das Gerücht von diesem erhält (*Pull*) oder überträgt (*Push*). Als Erweiterung wird ein „Knoten-Gedächtnis“ eingeführt, dass die letzten $M$ Interaktionen berücksichtigt, um einen erneuten Austausch zu verhindern.

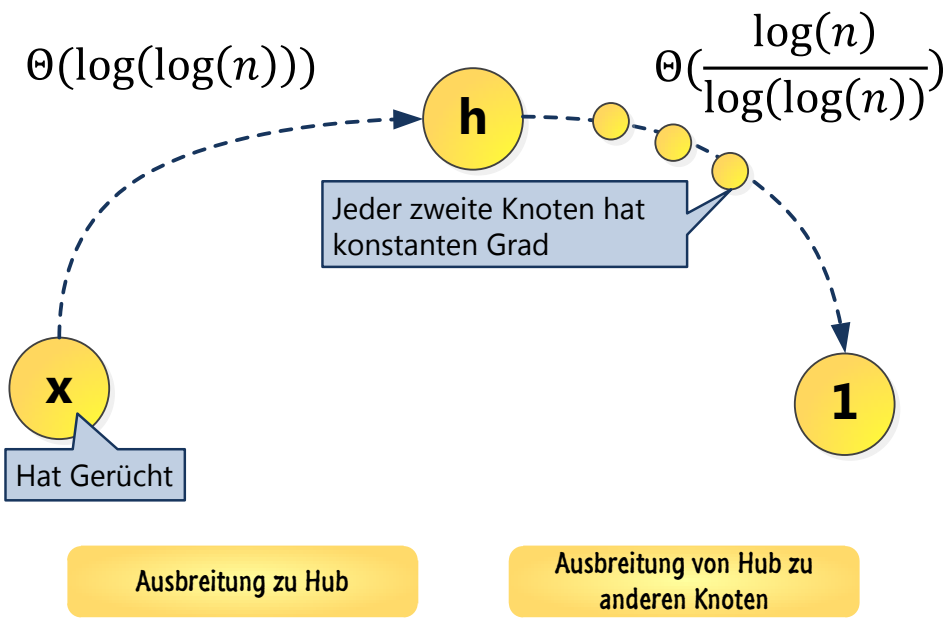


**Abbildung 1 Beweisidee für die Schranken**

Den Beweis für die Schranken erbringen die Autoren in drei Schritten (vgl. Abb. 1):

1. Ein Gerücht, das bei Knoten x startet braucht $\Theta\log(\log(n)))$ Schritte um einen Knoten mit besonders hohem Knotengrad zu erreichen („Hub“)
2. Ein Gerücht braucht $\Theta(\frac{\log(n)}{\log(\log(n))})$ Schritte um von einem Hub ausgehend Knoten 1 zu erreichen. Der Grund dafür liegt in der hohen Anzahl von Knoten mit konstantem Grad. Diese dienen als eine Abkürzung für die Verbindung zweier Hubs: sie besitzen wenige Nachbarn und können deshalb stets in kurzer Zeit ($m$ Runden) ein Gerücht von oder zu ein Nachbarn übertragen. Hubs hingegen brauchen sehr lange (polylogarithmisch von Graphengröße abhängig) bis alle Nachbarn kontaktiert sind.
3. Aus Symmetrieüberlegungen kann gefolgert werden, dass ein Gerücht in $\Theta(\frac{\log(n)}{\log(\log(n))})$ Schritten ausgehend von Knoten 1 jeden anderen Knoten erreichen kann.

Nach den anspruchsvollen Beweisen schließen die Autoren das Paper mit einem empirischem Vergleich der Push-Pull-Strategie mit und ohne Gedächtnis. Es bestätigt sich dabei, dass aus dem Gedächtnis eine asymptotisch niedrigere Laufzeit resultiert.

## Einordnung nach Qualität, Klarheit, Originalität und Signifikanz

### Qualität

Der mathematische Hintergrund der Autoren und die theoretische Natur der STOC sind Gründe für eine sehr technische und präzise Formulierung der mathematischen Schritte im Paper. Die Autoren nehmen an vielen Stellen Bezug auf vorangehende Arbeiten andere Autoren und vergessen es dabei nicht dies knapp zu erläutern. Insgesamt sind die Beweisschritte jedoch so anspruchsvoll, dass sie außerhalb der Community wohl ausgesprochen schwer nachvollziehbar sind. Die Ergebnisse des Papers sind jedoch sehr klar fassbar und präzise formuliert. Auch die Voraussetzungen des zugrundegelegten Modells werden deutlich gemacht. Dabei wir allerdings versäumt, die Implikationen des Barabási–Albert und Push-Pull Modells zu erläutern - es fehlt eine Beschreibung der Abstraktionsschwächen und Abweichungen von realen Netzwerken. Für die Qualität und Bedeutung des Papers spricht die Tatsache, dass es bereits jetzt – einem Jahr nach seiner Veröffentlichung – 11 mal zitiert wurde.

### Klarheit

Trotz der hohen mathematischen Komplexität kommunizieren die Autoren ihre Ergebnisse sehr klar und erläutern die Motivation und Relevanz des Themas durch einen ausführlichen einführenden Teil. Die Beweise, obwohl schwierig nachvollziehbar, werden sauber untergliedert und durch einen Überblick der Beweisstrategie eingeläutet. Die Einbeziehung von Daten realer sozialer Netzwerke in den abschließenden empirischen Teil hätte dem Paper geholfen. Dies haben die Autoren jedoch in der simplifizierten Darstellung ihrer Arbeit in dem *Communications of the ACM* nachgeholt.

### Originalität

Die Autoren geben einen klaren Überblick über die verwandten Arbeiten und grenzen ihren Ansatz klar davon ab: erst ihnen ist es gelungen eine starke logarithmische und für veränderte Modellannahmen sogar eine sublogarithmische Schranke zu beweisen. Sowohl die Hauptideen der Beweise als auch die Modelle sind dabei früherer Literatur entnommen. Dies ist jedoch nicht als Einschränkung der Originalität zu werten, sondern ist vielmehr ein Beispiel für wissenschaftlichen Fortschritt durch fortwährende Weiterentwicklung und sorgfältige Quellenauswertung.

### Signifikanz

Ich schätze die Signifikanz dieses Papers als hoch ein. Es wirft jedoch auch neue Fragen auf. So bleibt z.B. ungeklärt, wie weitere Aspekte sozialer Netzwerke modelliert werden können, u.a. wie die Broadcast Natur von Posts in realen sozialen Netzwerken modelliert und ausgewertet werden kann. Denn das Punkt-zu-Punk Kommunikationsmodell das implizit in der Push-Pull Strategie verankert ist, erscheint nicht realistisch für reale Netzwerke wie Facebook oder Twitter. Die Signifikanz geht des Papers geht nach meiner Einschätzung über den Kreis der theoretischen Informatik hinaus. Die dezentrale Kommunikation von Informationen, die in einem Netzwerk verteilt werden müssen, spielt beispielweise auch für verteilte Datenbanken eine große Rolle. Als Beispiel sei das sehr einflussreiche Modell des Key-Value-Stores Dynamo genannt [1]. Dort tauschen sich Knoten in einem „Gossip-Protokoll“ pro Zeitschritt mit einem zufälligen ausgewählten Kommunikationspartner aus, um Topologieänderungen des

Netzwerkes zu verbreiten. Die Ergebnisse dieses Papers implizieren, dass es keineswegs eine gute Wahl ist den Gossip-Partner zufällig aus allen bekannten Knoten zu wählen, sondern eine Barabási–Albert Topologie für eine asymptotisch schnellere Verbreitung des Zustands sorgen kann. Dies hat direkte und praktische Auswirkungen auf reale Systeme wie DynamoDB, Riak, Cassandra, Voldemort.

| Modellannahme | Schwäche |
|---|---|
| **Barabási–Albert Modell** | Kann u.a. die stärkere Clusterung realer Netze nicht erklären. |
| **Dichteparameter m** | Nicht jeder Knoten hat in sozialen Netzwerken eine Mindestanzahl an Freunden. |
| **Push-Pull Strategie** | Punkt-zu-Punkt Modell ist unrealistisch: reale Informationen verbreiten sich zumeist durch 1:n Broadcasts (z.B. Post auf Statusseite). |
| **Synchronizität der Runden** | Nicht alle Benutzer sind gleich schnell und vor allem sind nicht alle Benutzer zu jedem Zeitpunkt aktiv. |
| **Gleichverteilung bei Wahl eines Nachbarn** | Reale Benutzer kommunizieren bevorzugt mit bestimmten Nachbarn. |
| **Ungerichteter Graph** | Reale Kommunikation ist in vielen Fällen unidirektional (z.B. Post auf Barack-Obama-Fanpage in Facebook). |

**Tabelle 1 Punkte in denen das Modell unzureichend ist**

## Zusammenfassung

Es werden nun abschließend jeweils drei Stärken und Schwächen des Papers genannt.

### Stärken des Papers

- Ein aktuelles und interessantes Thema wir bearbeitet und die Behandlung überzeugend motiviert.
- Die Ergebnisse des Papers sind stark und eine deutliche Verbesserung der bisher erzielten Schranken.
- Die Modellannahmen sind schwach genug, um praktisch relevant zu sein.

### Schwächen des Papers

- Die Kombination aus Barabási–Albert und Push-Pull Modell ist in mehreren Punkten unrealistisch (siehe Tabelle 1). Dies wird nicht ausreichend diskutiert.
- Im empirischen Teil fehlt die Einbeziehung realer Daten.
- Die Beweisidee hätte deutlicher formuliert werden können.

### Referenzen

[1] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P. und Vogels, W. 2007. Dynamo: amazon's highly available key-value store. ACM SIGOPS Operating Systems Review (2007), 205–220.
http://www.allthingsdistributed.com/files/amazon-dynamo-sosp2007.pdf