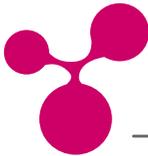


Technische Universität Dresden – Fakultät Informatik  
Professur für Multimedialechnik, Privat-Dozentur für Angewandte Informatik

Prof. Dr.-Ing. Klaus Meißner  
PD Dr.-Ing. habil. Martin Engeliem  
(Hrsg.)



# GENE '08

---

GEMEINSCHAFTEN IN NEUEN MEDIEN

an der  
Fakultät Informatik der Technischen Universität Dresden

mit Unterstützung der

GI-Regionalgruppe Dresden  
Initiative D21 e.V.  
Kontext E GmbH, Dresden  
Medienzentrum der TU Dresden  
SALT Solutions GmbH, Dresden  
SAP Research CEC Dresden  
Saxonia Systems AG, Dresden  
T-Systems Multimedia Solutions GmbH  
3m5. Media GmbH, Dresden

am 01. und 02. Oktober 2008 in Dresden  
<http://www-mmt.inf.tu-dresden.de/geneme/>  
[geneme@mail-mmt.inf.tu-dresden.de](mailto:geneme@mail-mmt.inf.tu-dresden.de)

## F.2 Zeitliche und strukturelle Untersuchung von Kommunikationsbeziehungen in der Blogosphäre

*Marius Feldmann, Oliver Gepp, Alexander Schill  
TU Dresden, Fakultät Informatik*

### 1 Motivation

„Für mich ist es eine Sucht. Ein unstillbarer Hunger nach Aufmerksamkeit. Oder, um es positiver und weniger egozentrisch zu sagen: nach Kommunikation.“ [9] Mit diesen Worten fasste Stefan Niggemeier, der Mitbegründer von bildblog.de, seine Motivation beim Bloggen zusammen. Das Interessante an seiner Aussage ist die eingeflochtene, implizite Charakterisierung von Weblogs. Niggemeier spricht nicht von seinem Verlangen, Inhalte zu veröffentlichen. Ihm geht es um die Kommunikation - die Interaktion mit anderen Menschen. Ein Weblog ist für ihn somit nicht einfach nur ein Publikations- sondern vor allem ein Kommunikationsmedium. Dass diese Eigenschaft nicht nur eine von wenigen Bloggern empfundene nebensächliche Eigenart, sondern hauptsächliche Erfolgsursache für die Beliebtheit von Weblogs ist, belegen zahlreiche Beispiele der alltäglichen Kommunikation, aber auch der Nutzung zur politischen Debatte – vor allem dann, wenn sie auf anderem Wege nicht möglich ist (siehe z.B. [4]).

Der Begriff Weblog allein reicht allerdings nicht aus, um betonen zu können, was hauptsächliche Eigenschaft dieser Plattform ist. Deutlich besser werden die Charakteristika von Weblogs durch den Begriff der Blogosphäre – von manchen Autoren auch als Blogspace benannt – beschrieben. Er bezeichnet nicht nur die Gesamtheit aller Blogs, sondern betont das Vorliegen gegenseitiger Verbindungen, die sich als Ausdruck für eine Kommunikationsbeziehung verstehen lassen. Für eine optimale Nutzung dieses Mediums ist allerdings eine tiefgreifende Kenntnis seiner besonderen Charakteristika unabdingbar. Hierzu können beispielhaft folgende Fragestellungen aufgeworfen werden:

- Wie lassen sich Kommunikationsbeziehungen zwischen Blogs erkennen?
- Zu welchen Weblogs hat ein vorgegebener Weblog intensive Verknüpfungen?
- Wie aktiv sind Kommunikationsbeziehungen in der Blogosphäre?
- Zu welchen Tageszeiten ist die Schreib- und Leseaktivität besonders hoch?

Diese und weitere Fragestellungen können größtenteils durch eine strukturelle Untersuchung der Blogosphäre beantwortet werden. Unter einer strukturellen Untersuchung wird dabei die Analyse der Linkstruktur als Ausdruck von Kommunikationsbeziehungen samt Eigenschaften von Verknüpfungen, insbesondere von zeitlichen Eigenschaften, verstanden. Nach einer kurzen Einordnung der Forschung und einer Darstellung assoziierter Grundlagen soll dazu ein mögliches Vorgehen schematisch dargestellt und zentrale Ergebnisse beschrieben und

interpretiert werden. Die aus dieser Untersuchung erhaltenen Erkenntnisse sollen zum besseren Verständnis der Kommunikation mittels Weblogs beitragen und als Grundlage konkreter Anwendungen dienen, die im letzten Abschnitt exemplarisch angesprochen werden.

## **2 Kommunikation in der Blogosphäre**

Bevor auf Eigenschaften von Kommunikationsaspekten eingegangen werden kann, muss zunächst festgelegt werden, welche Kommunikationsmittel in der Blogosphäre existieren. Es lassen sich zwei Kommunikationskategorien identifizieren:

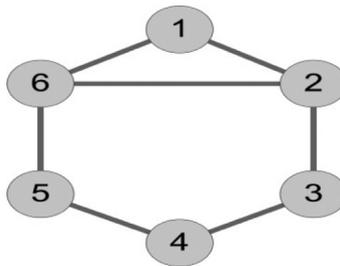
- 1) Innerhalb von Blogs durch das Schreiben von Artikeln und Kommentaren
- 2) Zwischen Blogs durch die Erzeugung von Hyperlinks

Entscheidend ist, dass beide Kommunikationsformen durch Auswertungen der Struktur von Weblogs und ihrer Linkbeziehungen automatisiert ermittelbar sind. Über einen Webcrawler, eine Software, die von Weblog zu Weblog die Blogosphäre durchwandert, können diese Kommunikationsformen ermittelt und zusätzliche Informationen zu ihnen aus den Weblogs extrahiert werden.

Die Blogosphäre ist bereits seit einigen Jahren Gegenstand wissenschaftlicher Untersuchungen, die vor allem im angloamerikanischen Raum durchgeführt werden. Der Fokus ist überwiegend auf die Detektion und Analyse Sozialer Netzwerke bzw. gesellschaftlicher Aspekte gerichtet. Überwiegend werden dabei bewährte Methoden aus dem Bereich des Information Retrieval mit Methoden der Analyse Sozialer Netzwerke kombiniert. Eine Möglichkeit der Identifikation von sozialen Beziehungen in solchen Netzwerken stellt eine inhaltliche Analyse dar, wie dies beispielsweise in [5] beschrieben wird. Einer grundlegend anderen Möglichkeit liegt eine einfache Annahme zugrunde: Eine soziale Beziehung zwischen zwei Blogs wird häufig beim Vorliegen eines Hyperlinks zwischen ihnen vermutet. Ein signifikantes Beispiel für derartige Untersuchungen wird in [1] beschrieben. Die Kernthese dieser Arbeit ist, dass sich nicht durch einheitliche Links Soziale Netzwerke in der Blogosphäre ausbilden, sondern dass zwischen verschiedenen Linkarten differenziert werden könne. Den dabei aufgeführten vier Linktypen – Links innerhalb von Beiträgen, innerhalb von Kommentaren, innerhalb der Blogroll und durch Trackback/Pingback [6] entstandenen Links – werden unterschiedliche Semantiken zugewiesen. Ein durch einen Trackback entstandener Link deute beispielsweise auf ein derartig großes Interesse eines weiteren Bloggers an einem fremden Artikel, dass er ihn in seinem eigenen Blog kommentiert hat. In [3] wird eine neue Definition für die Kommunikation in der Blogosphäre gegeben und diese zur Detektion von sozialen Beziehungen verwendet. Kommunikation sei demnach keine einseitig gerichtete Handlung. Eine Verlinkung ausgehend von einem Weblog auf einen zweiten sei nicht ausreichend, um von einer Kommunikation zu sprechen. Erst wenn auch eine rückwärtige Verknüpfung vorliegt, beruhe die Beziehung auf gegenseitiger Kenntnis

(*mutual awareness*) und damit auf einer tatsächlichen Kommunikation. Aufbauend auf dieser These wurden daraufhin *Communities* in der Blogosphäre identifiziert, die das genannte Kriterium der Kommunikation aufweisen.

Der in der Folge gebrauchte soziologische Begriff des *n*-Clans ist deutlich präziser als der in zahlreichen anderen Bereichen verwendete und dadurch unscharfe Begriff der *Community*. Unter einem *n*-Clan versteht man eine Gruppe von Personen - bzw. hier: von Weblogs – die alle untereinander über maximal *n* Verbindungen miteinander verknüpft sind. So enthält der Graph in Abbildung 1 beispielsweise den 2-Clan {2,3,4,5,6}.



**Abbildung 1: Beispiel n-Clan**

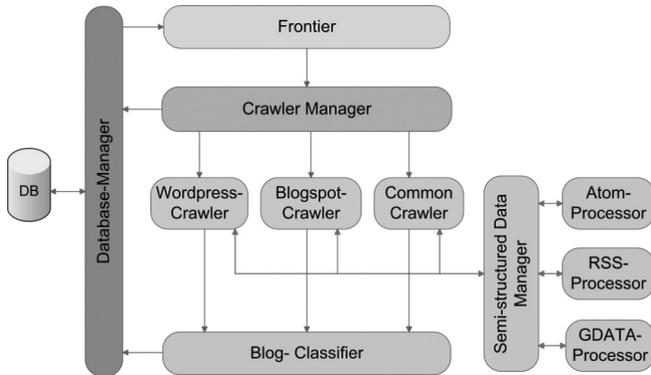
Aufgrund der präziseren Definition und seiner besseren Eignung für die Beschreibung struktureller Eigenschaften soll der Begriff des *n*-Clans auch in der Folge als Ersatz für den häufig verwendeten Begriff *Community* verwendet werden. Laut soziologischer Forschung ist  $n \leq 3$  vertretbar, um ein Aufweichen der Beziehung zwischen den Akteuren zu vermeiden (siehe dazu [7]). Interessant ist hierbei die noch offene und in anderen Untersuchungen nicht behandelte Frage, ob das Konzept der *n*-Clans verwendet werden kann, um Gemeinschaften in der Blogosphäre erkennen zu können. Die gegenseitige Kenntnis soll dabei ausgeweitet werden auf eine transitive Beziehung, um auch indirekte gegenseitige Bekanntschaften berücksichtigen zu können. Ob diese Kombination überhaupt ausreichend große zusammenhängende Bereiche erzeugt, ist eine der Fragestellungen, die in der Folge beantwortet werden sollen. Ist dies der Fall, so soll die Qualität der erhaltenen Gemeinschaften ermittelt werden. Zentrale Fragestellung an dieser Stelle ist die Messbarkeit von Qualität. Eine Möglichkeit der Qualitätsbewertung ist die Durchführung von inhaltlichen Analysen. Doch was sollen dabei die Kriterien sein, gegen die etwa erstellte Inhaltsvektoren abgeglichen werden? Inhaltliche Qualität ist immer abhängig vom Auge des Betrachters. Aus diesem Grunde soll ein anderes Kriterium zur Qualitätseinschätzung eingeführt werden: Da - wie einleitend dargestellt - die Hauptmotivation für das Bloggen die Kommunikation mit anderen Menschen darstellt, wird genau aus dieser Motivation heraus der Begriff

Qualität definiert: Die Qualität einer Gemeinschaft und eines einzelnen Weblogs wird durch die Intensität der Kommunikation bestimmt.

Neben den erwähnten sehr tief greifenden Analysen sollen zudem allgemeine strukturelle Aspekte von Kommunikation untersucht werden. Dazu gehört die Ermittlung des durchschnittlichen Verlinkungsgrades differenziert nach Linktypen, die durchschnittlichen Updateintervalle und Reaktionszeiten durch Kommentare. Darüber hinaus sollen mögliche Zusammenhänge zwischen Updateintervallen und Reaktionen durch Leser des Weblogs untersucht werden. Elemente dieser Analysen dienen der Schaffung von Vergleichswerten für die Qualitätsbewertung nach obigem Qualitätsbegriff der 2- bzw. 3-Clans mit transitiver gegenseitiger Kenntnis.

### **3 Vorgehen**

Eine große Herausforderung bei der Untersuchung der Blogosphäre stellte die Sammlung von geeignetem Datenmaterial dar. Wie oben erwähnt, soll dieses ausschließlich durch ein automatisiertes Durchwandern mittels eines Webcrawlers geschehen. Die Architektur des entwickelten Crawlers wurde an gängige Crawlerarchitekturen ([8], Kapitel 8.1) angelehnt und ist in Abbildung 2 dargestellt. Zur Initialisierung des Ablaufes muss eine Menge von Ausgangsblogs über ihre Uniform Resource Identifier (URI) angegeben werden, die in der so genannten Frontier gespeichert werden. Dieser Datenstruktur werden nach und nach Einträge entnommen und die jeweiligen Seiten werden aus dem Web zur Weiterverarbeitung herunter geladen. Liegt eine Seite vor, wird zunächst durch den Crawler-Manager versucht, zu detektieren, ob es sich dabei überhaupt um einen Weblog handelt. Dazu werden verschiedene Kriterien ausgewertet, wie zum Beispiel die Prüfung, ob in der URI die Domäne eines bekannten Weblog-Systems auftaucht oder ob bestimmte Header-Informationen auf ein bestimmtes Weblog-System schließen lassen. In Abbildung 2 sind exemplarisch für Wordpress und Blogspot spezifische Crawler neben dem allgemeingültigen Common Crawler eingezeichnet.



**Abbildung 2: Architektur Web-Crawler**

Diese Erweiterung klassischer Crawler-Architekturen ermöglicht ein effizienteres und effektiveres Durchwandern der Blogosphäre. Es wurde versucht, möglichst viele Informationen aus semistrukturierten Daten, insbesondere aus Web Feeds zu ermitteln. Stichprobenartige Prüfungen des gesamten Crawling-Vorgangs haben eine hohe Eignung des Algorithmus für die automatische Erkennung von Weblogs und der abgegriffenen Informationen bestätigt. Die ermittelten Weblogs inklusive Daten zu Verlinkungen, zeitlichen Informationen usw. werden in einer Datenbank abgelegt, von wo sie zur Weiterverarbeitung extrahiert werden. Ziel der Weiterverarbeitung ist zunächst die Ermittlung aller strukturellen Informationen, die eine der oben erwähnten Kommunikationsformen bedeuten. Zusätzlich sollen zeitliche Informationen mit diesen verknüpft werden. In einem umfangreicheren Berechnungsschritt sollen danach 2- und 3-Clans ermittelt werden. Da die Ermittlung von Clans potentiell aufwendig ist, wird sie über eine Vorverarbeitung effizienter gestaltet. Dazu wird zunächst die Menge von Blogs mit dem Kriterium der gegenseitigen Kenntnis für alle Blogs in der Datenbank erstellt. Dies hat den positiven Nebeneffekt, dass die Wahrscheinlichkeit von enthaltenen Spamblogs verringert wird, da es unwahrscheinlich ist, dass diese in größeren Gruppen gegenseitiger Kenntnis eingeordnet sind.

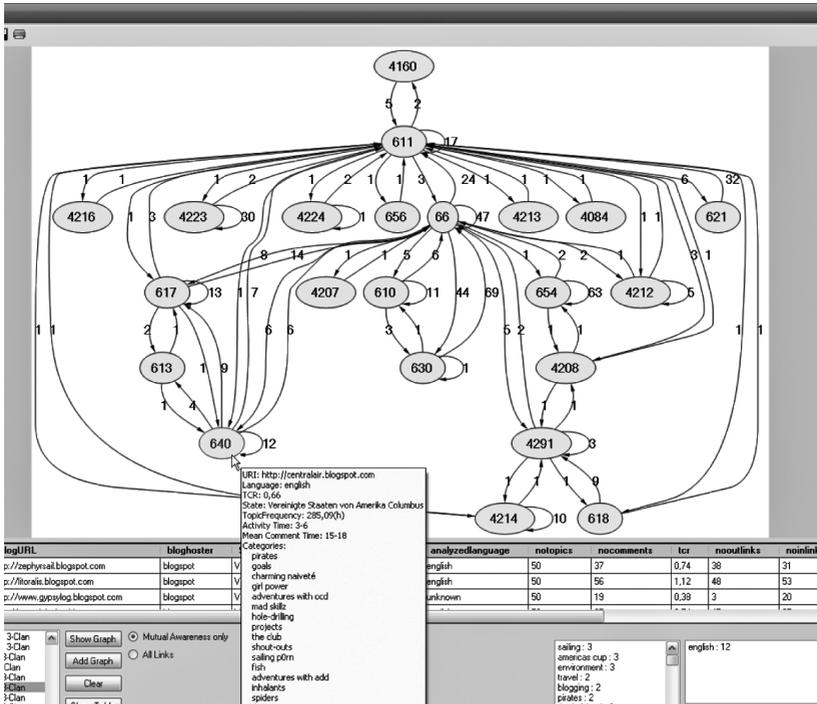


Abbildung 3: Screenshot Visualisierungs-Komponente

Für die erhaltenen Strukturen wird – um diese gegenseitige Kenntnis transitiv (und reflexiv) gestalten zu können - die reflexiv transitive Hülle gebildet. Innerhalb jedes identifizierten Bereichs ist somit sichergestellt, dass alle enthaltenen Blogs direkt oder indirekt verbunden sind. Um gefundene Gemeinschaften auch visualisieren zu können und sich zu diesen bestimmte Eigenschaften anzeigen zu lassen, wurde ein entsprechendes Werkzeug entwickelt (siehe Abbildung 3). Neben strukturellen Aspekten werden durch dieses zusätzlich Informationen über geographische Verortung oder die verwendeten Sprachen - nach entsprechender Auswertung - dargestellt.

#### 4 Beschreibung und Analyse des Datenmaterials

Nach einer kurzen Beschreibung allgemeiner Informationen soll in der Folge speziell auf Fakten zur Verlinkung, auf zeitliche Faktoren und auf die Detektion von n-Clans eingegangen werden.

#### 4.1 Allgemeine Zahlen

Für die Analyse von Kommunikationsbeziehungen wurde im Mai 2008 eine Stichprobe von mehr als 55.000 Weblogs mit insgesamt über 1.5 Millionen Artikeln abgegriffen, von denen verschiedene Teilmengen betrachtet und für die Weiterverarbeitung und die Analyse aufbereitet wurden. Folgende Tabelle 1 beschreibt einige allgemeine Kennzahlen, die aus dem Datenbestand ermittelt wurden.

<i>Durchschnittliche Anzahl von Artikeln pro Blog</i>	27,3
<i>Maximale Anzahl von Artikeln</i>	200
<i>Durchschnittliche Anzahl von Kommentaren pro Blog</i>	147,42
<i>Maximale Anzahl von Kommentaren pro Blog</i>	8601
<i>Durchschnittliche Anzahl von Kommentaren pro Artikel</i>	5,4
<i>Gefundene unterschiedliche Kategoriebezeichner</i>	39672

**Tabelle 1: allgemeine Zahlen**

Interessant daran ist zum einen die recht hohe durchschnittliche Anzahl von Kommentaren pro Blogseintrag, was die besondere Bedeutung des Kommentarmechanismus für die Kommunikation in der Blogosphäre unterstreicht. Zum anderen ist die sehr vielfältige Menge an Kategoriebezeichnern bemerkenswert. Nach dem Kategoriebezeichner „uncategorized“ (der in dieser Form tatsächlich verwendet wird) sind die fünf häufigsten Kategorien „blogging“, „family“, „music“, „books“ und „food“, was auf überwiegend private Weblogs schließen lässt.

#### 4.2 Statistiken zu Links

Die Vielzahl von privaten Weblogs im Datenbestand, welche alltägliche Meinungen, Ansichten und Geschehnisse präsentieren, mag auf eine geringe Verlinkung, sprich eine geringe Kommunikationsintensität schließen lassen. Eine nähere Untersuchung ergibt allerdings ein gegenteiliges Bild. Durchschnittlich liegen pro Weblog 60.8 ausgehende und 77.1 eingehende Links vor. Für die 310.581 in den ausgewerteten Blogs gefundenen Links ergibt sich die in Tabelle 2 dargestellte Verteilung auf die drei Linklokationen „Artikeltext“, „Kommentar“ und „Blogroll“.

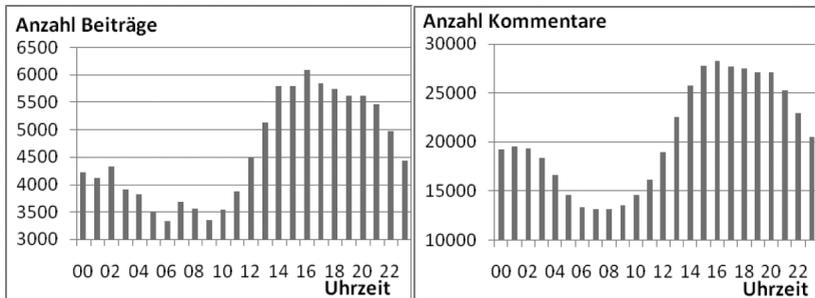
Gesamtanzahl Links innerhalb Artikeltext	203955
Gesamtanzahl Links innerhalb von Kommentare	74442
Gesamtanzahl Links innerhalb der Blogroll	32184

**Tabelle 2: Allgemeine Statistiken zu Links**

Immerhin fast zwei Drittel der Links stammen also aus den Artikeln selbst, was auf ein großes Interesse der Blogger schließen lässt, Informationen auf anderen Weblogs zu kommentieren. Ebenfalls hervorzuheben ist das geringe Verhältnis von Links innerhalb der Blogroll. Nur jeder zehnte Link entstammt aus dieser Linksammlung der Weblogs. Der Großteil der Links entsteht also dynamisch als Folge regulärer Kommunikation und wird nicht wie im Falle der Blogroll-Links vordefiniert und nur in größeren Abständen aktualisiert.

### 4.3 Zeitliche Aspekte

Durch das Vorhandensein eines klaren Aufbaus der Weblog-Seiten und durch das Vorliegen von semistrukturierten Daten kann für die ermittelten Artikel und Kommentare in den ausgewerteten Blogs die zeitliche Dimension gut dokumentiert werden. Dabei ist für eine optimale Nutzung des Kommunikationsmediums zunächst die Frage interessant, wann im Tagesverlauf eine besonders hohe Anzahl von

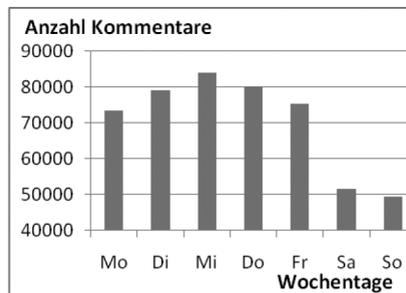


**Abbildung 4: Vergleich zwischen Aktivitätszeiten**

kommunikativen Aktivitäten gemessen werden kann. Die gegenüberstellende Abbildung 4 zeigt die Menge an erstellten Beiträgen und Kommentaren in Abhängigkeit der Uhrzeit des Blogsystems mit einer stundengenauen Granularität.

Es ist deutlich zu erkennen, dass beide Kommunikationsformen eine äußerst ähnliche zeitliche Verteilung aufweisen. Nach einem Aktivitätstief zwischen 06:00 und 09:00 Uhr erfolgt eine signifikante Zunahme der Kommunikationshäufigkeit bis zum Erreichen des Maximums, das jeweils im Zeitraum von 16:00 bis 17:00 Uhr liegt. Die wesentlich höhere Anzahl von Kommentaren zu jeder Stunde belegt noch einmal die herausragende Bedeutung des Kommentierens für die Kommunikation in der Blogosphäre.

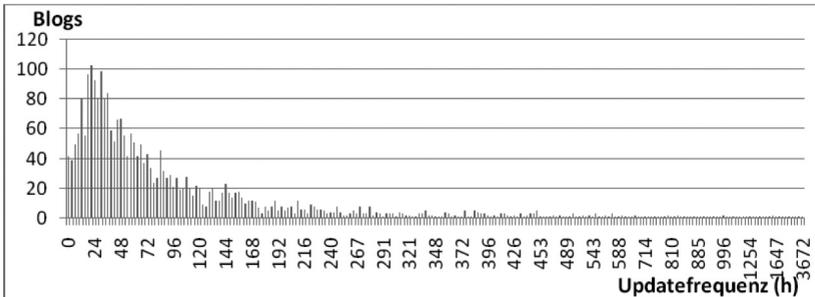
Neben der stundenfeinen Auflösung sind auch Aussagen zur Nutzung innerhalb von Wochentagen möglich, die in Abbildung 5 gezeigt werden. Auffällig sind die Konzentration der Kommentare auf den Mittwoch und der relativ symmetrische Verlauf von Montag bis Freitag. Sehr gravierend ist jedoch der starke Einbruch von Kommentaren am Wochenende um ein Drittel gegenüber dem Mittel an Werktagen, wobei die Anzahl der Beiträge der Anzahl der Kommentare entspricht. In Zusammenhang mit Abbildung 4 deutet das Ergebnis darauf hin, dass die Nutzung von Blogs eine Verbindung zur gewöhnlichen Arbeitszeit aufweist und die freizeithliche Nutzung aufgrund des Einschnitts am Wochenende relativ eingeschränkt zu sein scheint. Gründe für das gemessene Nutzungsverhalten sollten durch soziologische Studien untersucht werden.



**Abbildung 5: Aktivität nach Wochentagen**

Da nicht nur die Frage, wann kommuniziert wird, interessant ist, sondern auch, wie häufig dies geschieht, wurde ebenfalls die Updatehäufigkeit von Weblogs ausgewertet. Dazu ist die Anzahl der Blogs, die im Durchschnitt nach einer bestimmten Stundenzahl aktualisiert wurden, ermittelt worden. Das Resultat dieser Untersuchung ist in Abbildung 6 dargestellt. Die meisten Blogger veröffentlichen im Abstand von mindestens vier Tagen einen Artikel. Besonders häufig findet eine Aktualisierung im Zeitraum zwischen kurz unter einem Tag (~15 h) und etwa zweieinviertel Tagen (~54h) statt (durchschn. Updateintervall: 137,67 h). Seitens der Blogschreiber kann somit durchaus von einer hohen Kommunikationshäufigkeit (zahlreiche Blogs

mit täglichem Update) gesprochen werden. Wie zu sehen ist, wurden nur in sehr wenigen Weblogs äußerst hohe Publikationsintervalle ( $> 10$  Tage) gemessen. Neben der Darstellung der reinen Aktualisierungs- bzw. Aktivitätszeiten wurde eine Analyse von Abhängigkeiten zwischen dem Publizieren von Artikeln und der rückwärtigen Kommunikation durchgeführt.



**Abbildung 6: Updatefrequenzen von Weblogs**

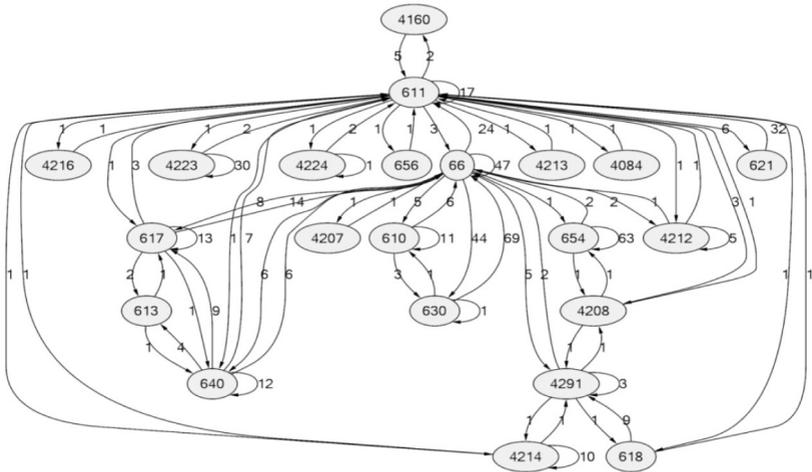
Zur Darstellung der Reaktionsgeschwindigkeit wurde die zeitliche Differenz zwischen der Veröffentlichung eines Artikels und des ersten Kommentars ermittelt. Die Ergebnisse dieser Messung sind in Tabelle 3 zusammengetragen. Geht man vom Durchschnittswert der Reaktionsgeschwindigkeit aus, werden erst nach fast zweieinhalb Tagen Kommentare für einen Beitrag hinterlassen. Ein Blick auf die Extremwerte lässt allerdings vermuten, dass dieser Wert durch starke Abweichung im Maximumbereich nicht aussagekräftig ist. Eine Segmentierung der Reaktionszeiten bestätigt diesen Verdacht. Ordnet man die Reaktionszeiten aufsteigend und bildet den Mittelwert für die schnellsten 25% der Kommentare (25% Quantil), so ergibt sich eine Reaktionszeit von weniger als zwei Stunden. Auch im Falle der schnellsten 50% der Kommentare (50% Quantil) liegt die durchschnittliche Zeitdifferenz zur rückwärtigen Kommunikation bei unter sechs Stunden. Erst für die 25% der langsamsten Reaktionen liegt die Zeit für den ersten Kommentar bei über einem Tag.

<i>Durchschnitt</i>	<i>3339 Minuten</i>
<i>Minimum</i>	<i>0 Minuten</i>
<i>Maximum</i>	<i>191410 Minuten</i>
<i>25% Quantil</i>	<i>104,75 Minuten</i>
<i>50% Quantil</i>	<i>344 Minuten</i>
<i>75% Quantil</i>	<i>1374,75 Minuten</i>

**Tabelle 3: Reaktionsgeschwindigkeit beim Kommentieren**

#### 4.4 Detektion von n-Clans

Nachdem nun auf grundlegende strukturelle und temporale Aspekte eingegangen wurde, soll als letzter Gesichtspunkt auf die oben beschriebene Detektion von n-Clans mit der gegenseitigen Kenntnis der Akteure zur Ermittlung von Gemeinschaften eingegangen werden. Wie erwähnt, war eine zentrale Fragestellung, ob durch diese Kriterien überhaupt zusammenhängende Bereiche in der Blogosphäre erkennbar sind oder ob die Kriterien derart hart sind, dass sie – wenn überhaupt – von nur sehr wenigen Akteuren erfüllt werden können – sich also nur sehr wenige, sehr kleine Gemeinschaften identifizieren lassen. Von den 29804 Blogs, deren ein- und ausgehende Links aus dem gesamten Datenbestand untersucht wurden, konnten 7315 (24,5%) nach der Ermittlung der gegenseitigen Kenntnis in eine reflexiv transitive Hülle eingeordnet werden. Ausgehend von diesen Gemeinschaften wurden sowohl 2- als auch 3-Clans bestimmt – das bedeutet, Bereiche identifiziert, in denen sich alle Akteure über maximal zwei bzw. drei Kanten (Hyperlinks) gegenseitig erreichen können. 5391 Blogs (18%) können dabei in 2-Clans mit einer durchschnittlichen Größe von 11,8 Mitgliedern eingeordnet werden. Bei der Zuordnung zu 3-Clans sind dies 2750 (9,2%) mit einer durchschnittlichen Größe von 26,6 Mitgliedern. Zwar mag man die Gesamtzahlen in 2- bzw. 3-Clans einordbare Weblogs als sehr gering interpretieren, sollte aber diese Form der Identifikation von Gemeinschaften zu hochqualitativen Ergebnissen führen, ist der Grund für wenige derart gruppierbare Weblogs evident. Ein Beispiel für einen gefundenen 3-Clan ist in Abbildung 8 dargestellt.



**Abbildung 7: Beispiel eines 3-Clans**

Die numerischen Annotationen an den Links geben die Anzahl an Verknüpfungen zwischen Quelle und Ziel wieder. Es gilt also noch die Frage nach der Qualität der Gemeinschaftsdefinition zu beantworten. Wie einführend beschrieben, soll der Qualitätsbegriff an dieser Stelle nicht über qualitative Inhalte der Artikel definiert werden, sondern über die Erfüllung des Kriteriums eines hohen Grades an Kommunikation. Dazu wurden einige der bereits für den gesamten Datensatz untersuchten Faktoren für den Kommunikationsgrad ermittelt. Die Ergebnisse sind in Tabelle 4 zusammengetragen. Die Zahlen verdeutlichen, dass für beide n-Clans eine hohe Kommunikationsintensität vorliegt. Alle fünf berechneten Durchschnittswerte sind deutlich höherwertiger (im Sinne des hohen Grades an Kommunikation) als die vergleichbaren Werte für den gesamten Datenbestand. Berechnet man noch den Mittelwert für die Kommentare pro Eintrag, so erhält man für die 2-Clans einen Wert von 11,4 (zum Vergleich der Wert für alle Weblogs: 5,4). Für die 3-Clans liegt dieser Wert immerhin noch bei 9,5. Die Anzahl der ein- und ausgehenden Links liegt deutlich über der Anzahl von Links, die die Blogs in den Clans selbst untereinander verknüpfen. Damit sind diese Bereiche einer sehr hohen Kommunikationsdichte durchaus nicht auf sich selbst fixiert, sondern stehen in engem und intensivem Kontakt zu den umgebenen Bereichen der Blogosphäre. Tabelle 4 verdeutlicht ebenfalls die leichte Aufweichung der Qualität durch die Erweiterung von der 2-Clan- auf die 3-Clan-Bildung.

	Durchschn. Beiträge	Durchschn. Kommentare	Durschn. eingehende Links	Durschn. ausgehende Links	Durchschn. Frequenz von Beiträgen(h)
2-Clan	42,8629	488,4222	261,6574	279,2397	72,969718
3-Clan	40,8695	388,8207	216,5283	409,8478	127,186523

**Tabelle 4: Qualitätsbewertung von 2- bzw. 3-Clans**

Um die dargestellten Zusammenhänge aus einer anderen Perspektive zu untermauern, wurden die Kategoriebezeichner einiger 2- bzw. 3-Clans manuell ausgewertet. Dabei stellte sich heraus, dass die Titel der Kategorien auf äußerst ähnliche Themenschwerpunkte der Weblogs in einem Clan schließen lassen, selbst wenn nicht dieselben Bezeichner verwendet wurden. Beispielsweise enthalten die Weblogs eines 3-Clans unter anderem die Kategorien „poetry“, „haiku“ (jap. Gedichtform), „poem“, „writers island“, „poefusion“ und „poetic form“, die auf lyrische Inhalte schließen lassen. Mehrere dieser manuellen Überprüfungen bestätigten dabei auch aus einer inhaltlichen Perspektive eine annehmbare Definition für hoch qualitative Gemeinschaften über die Clan-Bildung.

## 5 Zusammenfassung und Ausblick

Ausgangspunkt der Überlegungen war die Feststellung, dass die Blogosphäre maßgeblich kein Publikations- sondern ein Kommunikationsmedium ist. Dabei wird eine Kommunikation durch eine wechselseitige Bezugnahme im Sinne einer Aktion wie dem Schreiben eines Artikels, eines Kommentars oder dem Setzen eines Hyperlinks verstanden. Verschiedene strukturelle und zeitliche Kriterien wurden anhand eines eigens aus der Blogosphäre extrahierten Datensatzes auf das Wesen dieser Kommunikationen hin untersucht. Da es sich bei der Blogosphäre zudem um ein Soziales Netzwerk handelt, das sich wiederum in feingranulare Gemeinschaften aufteilt, wurde dieser Grundgedanke verwendet, um eine Qualitätsbeurteilung solcher Gemeinschaften durchzuführen. Dazu wurde das Konzept der 2- bzw. 3-Clans mit einer transitiven gegenseitigen Kenntnis von Weblogsystemen kombiniert. Die Qualitätsbewertung dieses Vorgehens lieferte ein sehr positives Ergebnis. Der vorgeschlagene Mechanismus zur Identifikation Sozialer Gemeinschaften kann durch die Berücksichtigung unterschiedlicher Linktypen und Linkhäufigkeiten verbessert werden. Neben dieser Untersuchung ist eine tiefgreifende semantische Qualitätsbewertung in Kombination mit dem aufgeführten Vorgehen geplant. Für die Ergebnisse lassen sich zahlreiche Anwendungen erdenken. Dazu gehören

vor allem Suchmaschinen, die an die besonderen Eigenschaften der Blogosphäre adaptiert sind und über die Detektion von Gemeinschaften und deren Qualität eine Relevanzbewertung durchführt bzw. ähnliche Weblogs erkennt. Ebenfalls denkbar ist ein Web-basiertes Werkzeug, dass für einen vom Nutzer vorgegeben – evtl. den eigenen - Weblog 2- oder 3-Clans ermittelt, in denen dieser Weblog partizipiert, um Vorschläge für ähnlich ausgerichtete, sehr aktive Blogs zu erhalten. Dies kann auch für die gezielte Anbringung von Werbebotschaften in hochfrequentierten Bereichen der Blogosphäre ein interessantes Werkzeug sein. Daneben wird einem Blogger die Möglichkeit geboten, sich in eine vorhandene Gemeinschaft seines Interessenfokus einzubinden, um so seinen eingangs angesprochenen Hunger nach Kommunikation zu stillen.

### **Literatur**

- [1] Furukawa, T. et al.: Social Networks and Reading Behavior in the Blogosphere, ICWSM'2007 Boulder, 2007
- [2] Herring, S. et al.: Conversations in the Blogosphere: An Analysis "From the Bottom Up", Annual Hawaii International Conference, 2005
- [3] Lin, Yu-Ru et al.: Discovery of Blog Communities based on Mutual Awareness, IEEE/WIC/ACM International Conference on Web Intelligence, 2007
- [4] Adamic, L.; Glance, N.: The Political Blogosphere and the 2004 U.S. Election: Divided They Blog, International Workshop on Link Discovery, 2005
- [5] Chin A.; Chignel, M.: A Social Hypertext Model for Finding Community in Blogs, Conference on Hypertext and Hypermedia, 2006
- [6] Spezifikation von Trackback: [http://www.sixapart.com/pronet/docs/trackback\\_specPingback](http://www.sixapart.com/pronet/docs/trackback_specPingback): <http://hixie.ch/specs/pingback/pingback>
- [7] Wassermann, S.; Faust, K.: Social Network Analysis. Cambridge U. Press, 1994
- [8] Liu, B.: Web Data Mining – Exploring Hyperlinks, Contents, and Usage Data; Springer Verlag, 1. Auflage, 2007
- [9] Niggemeier, S.: „Wer bin ich? Warum das Schreiben eines Blogs so befriedigend ist“, FAZ, Ausgabe 6. Mai 2007, Seite 33