reicht hat (Erreichbarkeit). Zur Erreichbarkeit sind in anderen Graphen (Bekanntschaften zwischen
Menschen[4, 5] sowie Mailbekanntschaften[6]) interessante Untersuchungen unter dem Begriff "Sm
World" veröffentlicht worden.

Um dieses Vorhaben in die Realität umzusetzen entwickelten wir einen Crawler. Im Internet s
zwar einige Crawler zu finden (Limewire[3], gprobe[7]), diese konnten wir aber wegen fehlenden Sour
oder Abhängigkeiten nicht verwenden.

Ein Crawler verbindet sich mit dem GNetz, fragt jeden erreichbaren Host nach dessen benachbar
Hosts und wiederholt dieses Verfahren für jeden gefundenen Host. Unser Crawler basiert auf dem C
von Keno Albrecht. Zu Beginn konzentrierten wir uns auf die sogenannten "browsing Pings". Hier
stellt der Crawler eine Verbindung gemäss Protokoll (0.6)[8] mit einem Host her und sendet danach
Ping mit einer TTL[1] von 2 und einem Hop-Count[2] von 0. Als Antwort (Pong) sollte er die Nachbarn
angefragten Hosts erhalten. Als grösstes Problem erwiesen sich die abgelehnten Verbindungswünsc
Die Erlösung folgte im Februar 2003, als im Gnutella Developper Forum (GDF) Vorschläge ei
speziellen Crawler-Headers erschienen (siehe Abschnitt 2.1).

1.3 Gnutella

Das GNetz ist ein sogenanntes Peer-To-Peer Netz (P2P). Dabei werden einzelne Computer, auch Ho
Clients oder Peers genannt, auf einem vorhandenen Netz (meist das Internet) virtuell zusammen
schlossen. Das GNetz ist ein dezentrales Netz, da es aus gleichberechtigten Teilnehmern besteht. Bis
wird das GNetz selten kommerziell verwendet, meist sind es normale Internetbenutzer die Dateien
tereinander austauschen. Es ist aber durchaus denkbar, dass Firmen diese Art des Dateiaustausc
in einer ihren Sicherheitsbedürfnissen angepassten Variante nutzen.

Gegenüber der ersten erfolgreichen Version (0.4) sind in der aktuellen Protokollversion (0.6
einige Anpassungen vorgenommen worden. Die im Zusammenhang mit der Struktur interessanteste
die Einführung von Ultrapeers. Damit gibt es im GNetz verschiedene Typen von Hosts, die norma
Hosts sowie die Ultrapeers. Ultrapeers haben die Aufgabe das GNetz zusammenzuhalten, sie bil
das Rückgrat des GNetzes. Voraussetzungen damit ein Host ein Ultrapeer sein kann, sind schn
Verbindungen zu anderen Hosts sowie eine lange Laufzeit.

[1] TTL: Time To Live, über wieviele Hosts die Nachricht noch gesendet werden soll.
[2] Über wieviele Hosts die Nachricht schon gesendet wurde.

Tabelle 1: Verbindungsaufbau Crawler zu Host

Nicht alle Clients halten sich strikt an diesen Ablauf. Häufig ist ein Abbruch statt einem korrek Verbindungsabbau nach dem Senden der ersten Antwort zu beobachten. Ein weiterer Umstand dass die Clients die Zeile mit den Leaves und Peers über mehrere Zeilen verteilen. Dabei kann un hoher Last nicht unterschieden werden, ob die Verbindung vom Client abgebrochen wurde oder auf die nächste Zeile gewartet wird.

Verglichen mit der Variante der Pings hat die Methode des Crawler-Headers den Vorteil, dass ni eine Verbindung hergestellt werden muss, bevor die Anfrage (Ping) gesendet werden kann. Es ka direkt die Anfrage (Crawler-Header) gesendet werden.

Fehlermeldungen wie "500 Server too busy", wie sie beim Verbindungsaufbau für eine Ping-Anfr häufig zu sehen waren, treten dank der prioritären Behandlung bei den Clients nicht mehr auf.

2.2 Crawler

Aufgrund des neu hinzugefügten Crawler-Headers im Gnutella Protokoll passten wir unseren Craw an. Die Nachteile der Variante Ping waren zu erheblich. Wenige Verbindungen waren erfolgreich. We war problematisch, dass einige Clients gar nicht, nicht korrekt oder dann nur mit niedriger Priori auf Anfragen antworten.

Der Crawler besteht aus mehreren Komponenten. Sie werden anhand eines Beispiellaufes erkl Damit der Crawler einen Einstiegspunkt in das GNetz findet benutzt er GWebCaches[10]. Die von GWebCaches erhaltenen Einstiegspunkte werden in eine Priorityqueue eingefügt.

Von nun an schickt der Crawler dem Host mit der höchsten Priorität aus der Queue[3] eine Craw Anfrage. Die in der Antwort enthaltenen Hosts fügt er wiederum in die Queue ein. Ist die Queue l versucht er von Neuem bei den GWebCaches noch nicht besuchte Hosts zu finden.

Um möglichst viele neue Hosts in kurzer Zeit zu erreichen arbeiten mehrere Crawler (Threa parallel zueinander. Sie benutzen jedoch eine gemeinsame Priorityqueue.

Gestoppt wird der Crawler manuell, wenn die Queue leer ist und ebenfalls von den GWebCac keine neuen Hosts mehr gefunden werden konnten.

Als Ergebnis erhalten wir Logdaten aus denen dann die Graphen erstellt werden können. Benennung der Graphen entsprechen folgendem Muster: JJJJ/MM/TT hhmm, zum Beispiel wu der Graph 2003/06/13 1439 am 13. Juni 2003 um 14:39 Uhr gestartet.

[3] Jeder Host wird mit einem Zeitstempel (nächste Besuchszeit) eingefügt. Die kleinste Zeit hat die höchste Priori

betreffenden Hosts besteht.

- Es besteht die Problematik, dass das Netz dynamisch und ein Schnappschuss so nicht mögl ist. Die Zeit während der der Crawler läuft, ist in der Grössenordnung von 10 bis 20 Stunde
- Die Clients senden nur Informationen bezüglich Ihrer Nachbarn zurück, daher ist es nicht mögl zu eruieren ob es sich um normale Hosts oder Ultrapeers handelt.

2.5 Angaben von Limewire

Die Angaben von Limewire[3] sind als momentane Grösse des GNetzes zu verstehen. Die Werte wer mit einem periodisch laufenden, inkrementellen Crawler errechnet.

3 Auswertung

3.1 Annahme

Hosts mit einem Knotengrad grösser gleich 11 betrachten wir als Ultrapeers, alle übrigen Hosts wer als normale Hosts betrachtet.

Als Grundlage dieser Annahme dient die Tatsache, dass in den meisten Clients manuell eingest werden kann, mit wie vielen Hosts eine Verbindung aufrecht erhalten werden soll. Die Standardeins lungen sind bei ausgewählten Clients so festgesetzt:

Bearshare[11] 5 bis 10 Verbindungen, Gnucleus[12] 4 bis 6, Morpheus[13] nicht einsehbar, Swappe nicht einsehbar, Phex[15] 3, Xolox[16] 5, Limewire[1] nicht einsehbar, GTK-Gnutella[17] 4 bis Shareaza[18] 6.

[4] Die Priorität ist hier der nächste Zeitpunkt an welchem der Host angefragt werden soll. Je kleiner diese Zeit, d höher ist die Priorität.

Es ist keine Sättigung ersichtlich. Das könnte 2 Erklärungen haben. Entweder ist das Netz wesentl
grösser als die gesammelten ca. 150000 Hosts oder durch häufiges An- und Abmelden von Cients wer
ständig neue Hosts gefunden während schon erfasste Hosts schon nicht mehr Bestandteil des Net
sind. Die Angaben von Limewire drängen zur 2. Vermutung. Die erfassten Werte für eine Angabe
Anzahl der zu einem bestimmten Zeitpunkt im GNetz beteiligten Hosts sind somit tendenziell eher
zu hoch zu deuten.

3.3 Hosts

Die Anzahl der ermittelten Hosts liegt etwa im Bereich der von Limewire gemachten Angaben. Im J
2003 schwankten die bei Limewire veröffentlichten Daten im Bereich von 90000 bis 130000 Hosts
nach Tageszeit) im Netz.

Bei den längeren Läufen ist die Zahl der erfassten Hosts grösser als die Angaben von Limew
Daraus lässt sich eine Schätzung über die minimale Fluktuationen im Netz machen.[5]

$$\frac{(160000 - 130000) * \frac{11}{12}}{18 * 3600s} = 0,42\frac{1}{s} \approx \frac{2}{5}\frac{1}{s}$$

Das bedeutet, dass alle 5 Sekunden mehr als 2 Hosts sich neu zum Netz verbinden oder 2 Ho
die Verbindung zum Netz beenden.

In Tabelle 3 ist zu sehen, dass sich das Verhältnis normaler Hosts zu Ultrapeers bei allen Grap
um 14 bewegt. Auf einen Ultrapeer kommen also 14 normale Hosts.

	normale Hosts	Ultrapeer	Verhältnis
2003/06/13 1439	142280	12035	11,82
2003/06/14 2354	94686	6552	14,45
2003/06/15 1753	121481	8564	14,19
2003/06/16 0922	116175	8779	13,23
2003/06/17 1453	147130	11916	12,35
2003/06/18 1059	95571	6397	14,94

Tabelle 3: Verhältnis normale Host zu Ultrapeer

[5] Der Faktor $\frac{11}{12}$ entspricht dem tiefsten gemessenen Verhältnis von normalen Hosts zu Ultrapeers (siehe Tabelle
Die Annahme ist, dass die Ultrapeers sehr lange im Netz bleiben und die Änderungen nur die normalen Hosts betref

Abbildung 1: Knotengrad

Abbildung 1 zeigt die Verteilung der Knotengrade für den Graph 2003/06/13 1439. Es ist erkennen, dass im GNetz deutlich mehr (etwa das 12fache) Hosts mit Knotengrad kleiner gleich (normale Hosts) existieren, als solche über 10 (Ultrapeers). Das Verhältnis schwankt zwischen 1 11,82 (2003/06/13 1439) und 1 zu 14,45 (2003/06/13 2354) wie in Tabelle 3 zu sehen ist.

In Abbildung 1 sind typische Verteilungen der Hosts zu sehen. Im linken Bild sind die norma Hosts, im rechten Bild die Ultrapeers gezeigt. Die Mittelwerte der Knotengrade bewegen sich bei Ultrapeers um 42 bis 45. Allerdings mit grosser Standardabweichung (siehe Tabelle 4).

Bei allen Graphen ist eine Gemeinsamkeit zu erkennen. Immer gibt es viele Hosts, welche erst relativ kurzer Zeit neu gestartet wurden. Diese drücken den Mittelwert nach unten und erhöhen Standardabweichung. Weiter gibt es jeweils eine Häufung von Ultrapeers, welche einen Knoteng von 80 bis 85 haben. Es scheint, dass dies bei den Ultrapeers der Standardeinstellung für die Anz der akzeptierten Nachbarn entspricht.

	gemeinsam		nur normale Hosts		nur Ultrapeer	
Graph	Mw	Std	Mw	Std	Mw	Std
200306131439	7,11	13,88	4,06	2,09	43,66	31,79
200306142354	6,26	12,43	3,76	1,89	42,79	30,76
200306151753	6,61	13,37	3,90	2,02	45,68	32,61
200306160922	6,83	13,54	3,96	2,05	44,74	31,77
200306171453	7,07	14,07	4,05	2,12	44,09	32,87
200306181059	6,25	12,56	3,75	1,90	44,16	31,21

Tabelle 4: Mittelwerte / Standardabweichung $s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{k} n_j (x_j - \bar{x})^2}$

3.4.1 Erreichbarkeit aller Hosts

Um alle Hosts des GNetzes zu erreichen, würden im gleichen Verhältnis wie Hosts zu Ultrape vorhanden sind, Verbindungen zu anderen Hosts aufgebaut, wären mindestens 7 Stufen nötig.[6]

[6] Hier wird mit einem Verhältnis von $\frac{1}{14}$ etwa das Mittel aller gemessenen Verhältnisse in Tabelle 3 gewählt.

dabei möglichst nur untereinander verbunden sein. Dann wären noch 4 Stufen nötig um alle Hosts
GNetz zu erreichen.

$$4 * 44^{x-1} > 130000$$

$$x > 3,75$$

Dabei wird allerdings vernachlässigt, dass die Ultrapeers viele Verbindungen zu normalen Ho
haben müssen. Es gibt in der Grössenordnung von 10 mal mehr normale Hosts als Ultrapeers.
normalen Hosts sollen sich möglichst nur mit Ultrapeers verbinden, was bedeutet, dass der grösste
der Verbindungen eines Ultrapeers nicht zu Ultrapeers gehen, sondern zu normalen Hosts. Das ist a
die Erklärung warum dieser Ansatz zu gute Ergebnisse liefert und die realen Werte tiefer liegen.

Obige Annahme bestätigen die Werte in Tabelle 5. Das Ziel eines jeden Hosts ist es folgl
möglichst nur Verbindungen zu Ultrapeers aufzubauen und andere Verbindungen wahrscheinlich so
abzulehnen.

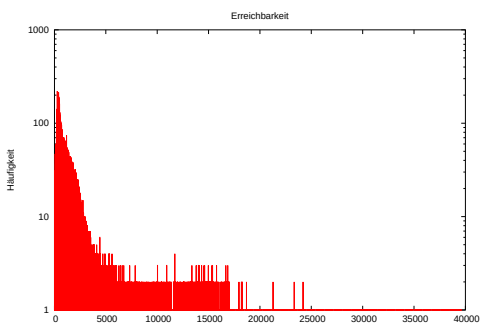

Abbildung 2: Verteilung der in 3 Schritten erreichbaren Hosts

3.5 Zyklen

Bei den Zyklen interessierten hauptsächlich solche der Grösse 3 (kleinste Kreise). Diese sorgen für
Stabilität des Netzes. Fällt ein Host aus, ist der Verlust minimal. Alle Hosts, die über den ausgefalle
Hosts erreicht werden konnten, sind dann über 1 Stufe mehr immer noch zu erreichen.

In den Graphen können etwa 1000 Zyklen der Grösse 3 gezählt werden. Auffallend ist dabei
hohe Anzahl von Zyklen, bei denen 2 Ultrapeers beteiligt sind, siehe Abbildung 3.

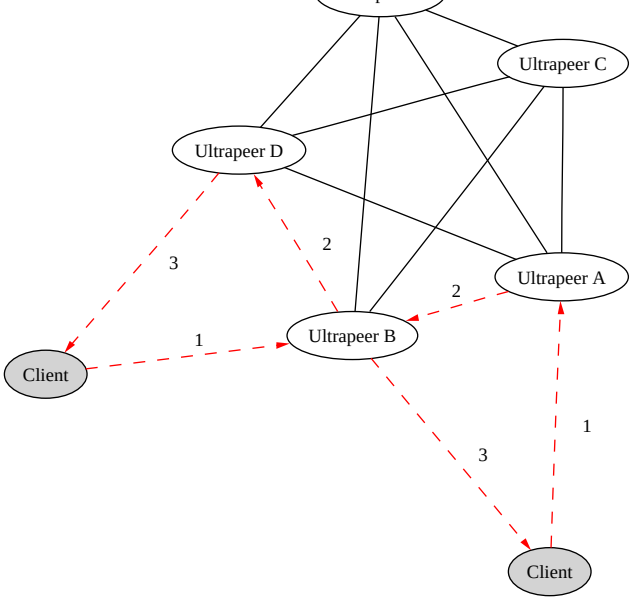

Abbildung 4: Entstehung kleinster Kreise

4 Ähnliche Projekte

4.1 Limewire Crawler

Der Limewire[3] Crawler sucht das Gnutella Netz nach ebenfalls mittels der Crawler-Header ab. wir mit unserem Crawler starteten, benutzte der Limewire Crawler noch die Ping Nachrichten.

4.2 Modeling Large-scale Peer-to-Peer Networks and a Case Study of Gn tella, Mihajlo A. Jovanovic

Mihajlo A. Jovanovic[2] entwickelte einen parallelen Crawler um das GNetz zu untersuchen. Er tersuchte die Entstehung des GNetz in Bezug auf kleine Durchmesser, Klumpenbildung (Clusteri sowie 4 verschiedenen Potenzgesetzen (Power Laws).

Um einen genaueren Schnappschuss des Netzes zu erhalten, müsste die Laufzeit reduziert werden. Ansatz dazu könnte sein, einen verteilten Crawler auf mehreren Rechnern zu verwenden.

[7] Ein Ultrapeer kann bei einer Anfrage nur einen ihm bekannten Host zurückgeben. Sobald der anfragende sich diesem verbindet, ist ein Zyklus vorhanden.

[10] http://www.gnucleus.com/gwebcache/

[11] http://www.bearshare.com/

[12] http://www.gnucleus.com/

[13] http://www.morpheus.com/

[14] http://mywebpages.comcast.net/jthomas497/swapper/swapper.html

[15] http://phex.sourceforge.net/

[16] http://www.xolox.nl/

[17] http://gtk-gnutella.sourceforge.net/

[18] http://www.shareaza.com/

[19] http://cosmos.kaist.ac.kr/cs441/project/mid-term/presentation/11-417.ppt

[20] http://disl.cc.gatech.edu/bg/homepage/

[21] Determining Characteristics of the Gnutella Network, Amogh Dhamdh
http://www.cc.gatech.edu/classes/AY2001/cs7001_fall/projects/zegura02.html