Jens Meichsner

Netzwerk und Internet

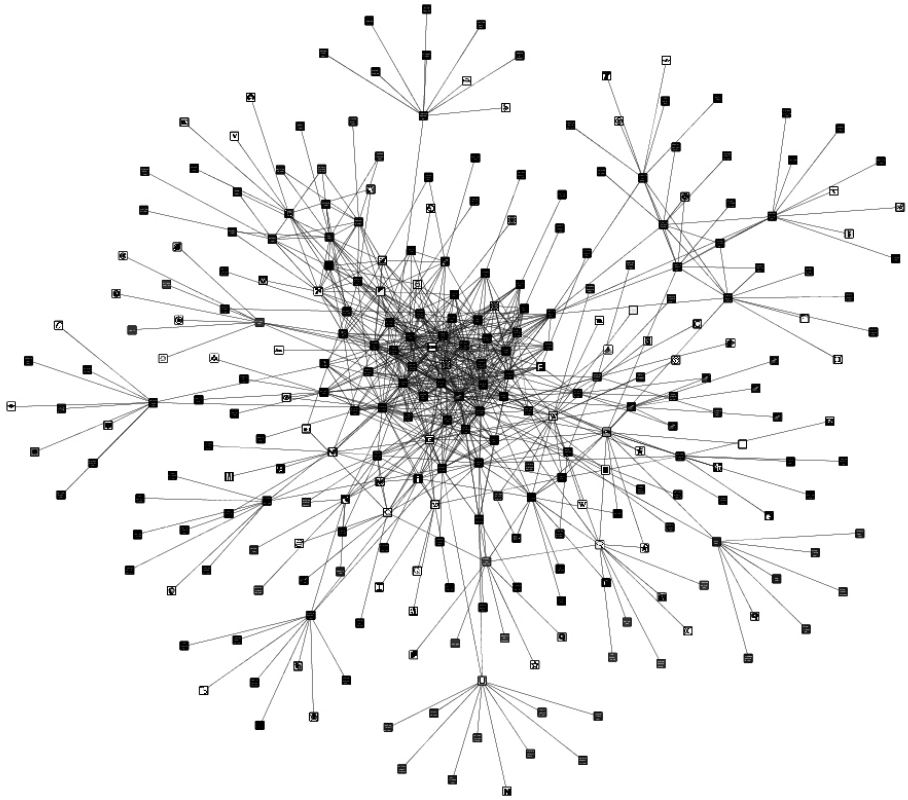

Abb. 1

Beschäftigen wir uns heutzutage mit Netzwerken im Allgemeinen, kommen wir kaum darum herum uns im Speziellen auch mit dem Internet auseinanderzusetzen. Denn das globale Computernetzwerk, so scheint es mir, ist zugleich Ausgangspunkt und Zentrum wissenschaftlicher Netzwerktheorien und Wissensmodelle. Zwar wurde ohne Frage bereits weit vor Erfindung und Etablierung des Internets in der Wissenschaft über Netzwerke geschrieben und theoretisiert, aber nie mit einem solch universellen Anspruch, wie es im beginnenden 21. Jahrhundert der Fall ist: Fast alles wird in Netzwerkstrukturen gefasst oder, wie ein Netzwerkkritiker sagen würde, gepresst. Ich möchte mich im Folgenden allerdings weniger kritisch mit universellen Netzwerktheorien allgemein auseinandersetzen. Vielmehr liegt es mir daran, das Internet als Netzwerk selbst ins Visier zu nehmen.

Begriffliches – ein ‚visueller' Vorschlag

Was macht also das Internet, sprich das zeitgenössische Netzwerk par excellence, überhaupt als Netzwerk aus? Um diese Frage zu beantworten, muss ich der Betrachtung logischerweise einen Netzwerkbegriff zu Grunde legen. Um die Vielfältigkeit des Begriffs

Abb. 1 TouchGraph Google Browser: http://www.touchgraph.com/TGGoogleBrowser.html (30.06.07)

und seine dazugehörigen Theorien möglichst wenig zu beeinträchtigen, möchte ich versuchen, auf kleinster Ebene zu definieren. D.h. ich suche eine ‚Netzwerkdefinition', die es uns einerseits erlaubt in unserem Kontext sinnvoll mit ihr zu arbeiten, andererseits aber so offen und erweiterbar strukturiert sein muss, dass sich komplexere Netzwerktheorien darauf aufbauen lassen. Der Einfachheit halber möchte ich Ersteres als Praktikabilitätsbedingung und Zweites als Kohärenzbedingung bezeichnen. Um mein Bestreben und deren Umsetzung deutlicher zu machen, sollten wir einen Blick auf Abbildung 1 werfen.

Ich möchte mit Hilfe dieser Grafik versuchen zu zeigen, dass sich Netzwerke sinnvoll als rein *grafisch-geometrische* Konstrukte beschreiben und erklären lassen. Oder etwas vorsichtiger formuliert: Wir werden sehen, wie weit wir mit diesem Manöver kommen werden. Ein Netzwerk wäre nach diesem Gedanken also nichts weiter als eine grafische Zeichnung von verzweigten Knoten und Kanten, die eine praktikable Beschreibung eines, wie auch immer gearteten, Phänomens darstellt. Oder anders: Unter den Netzwerkbegriff fällt all das, was sich sinnvoll mit einer Grafik ‚ähnlich' zu Abbildung 1 verbildlichen lässt.[2] Der Vorteil dieses rein visuellen Definitionsvorschlags liegt in meinen Augen darin, dass er sich zu den diversen Netzwerktheorien neutral verhält. Egal wem wir die obige Abbildung 1 oder eine ihr ‚ähnliche' Grafik auch zeigen, er wird die Frage, ob damit ein Netzwerk dargestellt ist, wohl kaum verneinen. Für einige mag dies sicherlich noch zu unspezifisch sein, um es als hinreichende und vor allem als alleinige Bedingung für einen Netzwerkbegriff zu akzeptieren. Doch auf empirisch überprüfbarer Ebene scheint mir der Vorschlag der einzige und damit größtmögliche Kompromiss zu sein. Den Beweis für diese Hypothese werde ich hier allerdings nicht antreten, da der Gedanke eigentlich nur darauf abzielt, dass wir uns ohne große theoretische Manöver und Kenntnisse problemlos darauf einigen können Netzwerke auf die oben beschriebene visuelle Sichtweise als reine Knoten/Kanten-Darstellung zu fassen. Die Kohärenz-Bedingung ist für unseren Netzwerkbegriff damit, wie ich finde, leicht nachvollziehbar erfüllt. Etwas mehr Aufwand müssen wir betreiben, um zu überprüfen, ob auch die zweite Bedingung, die Praktikabilität, von der visuellen Netzwerkdefinition erfüllt wird. Schauen wir deshalb nochmals vertiefend auf unsere Grafik.

2 Mit ‚ähnlich' meine ich, dass alle Abbildungen Verzweigungen haben sollten, also keine rein, linearen Listen sind. Wobei ich durchaus einen graduellen Übergang von Liste zu Netzwerk gewähren möchte. Denn es gibt sicherlich Grenzfälle, etwa eine Darstellung in der nur zwei Verzweigungen vorhanden sind und alle restlichen Knoten in einer langen lineare Kette verbunden sind. Bei einem solchen konstruierten Fall ist es auf Anhieb schwer zu entscheiden, ob es sich dabei schon um ein Netzwerk handelt oder noch um eine Liste. Es ist wohlgemerkt damit nicht auszuschließen, dass sich ein Phänomen sowohl mittels Listengrafik als auch mittels Netzwerkgrafik praktikabel darstellen lässt (Vgl. Abschnitt zu PageRank).

Webcrawler & Co: die Struktur des World Wide Web

Abbildung 1 verbildlicht, ganz grob gesprochen, Linkstrukturen im Internet. Grundlage bilden dabei Suchergebnisse der Internetsuchmaschine Google. Die Grafik wird mittels eines Java-Applets dynamisch generiert. Der User gibt dazu einen beliebigen Suchbegriff ein; ich habe mich in Abbildung 1 für *insurance*, also für das Versicherungswesen, entschieden. Daraufhin werden die Suchergebnisse, die Google für diese Suchanfrage ausgibt, ausgewählt und ihre jeweiligen Verbindungen (Links) untereinander analysiert. Die Anwendung basiert dementsprechend ausschließlich auf den Daten, die Google ihm liefert. Ich möchte deshalb folgend einen kurzen Blick auf Technik und Funktionsweise der größten und bedeutendsten Suchmaschine der Welt werfen. Ziel ist es das Verständnis für Abbildung 1 und für die gesamte Struktur des Internets zu schärfen.

Google besteht im Grunde genommen aus zwei zentralen Prozessen bzw. Algorithmen, Webcrawling und PageRank. Es gibt zweifelsohne Unmengen weiterer Methoden und Techniken, die eine solch komplexe Anwendung wie Google am laufen halten. Um ein *Bild* vom Internet zu bekommen, reicht es meiner Ansicht nach aber völlig aus, sich auf Webcrawler und PageRank zu konzentrieren: Für das ‚Suchmaschinen-Rohmaterial' sind die sogenannten Webcrawler verantwortlich. Sie starten prinzipiell bei einer beliebigen Webseite und verfolgen bzw. speichern alle von ihr ausgehenden Links auf andere Webseiten. Auf allen von der ‚Urseite' aus verlinkten Webseiten wiederholt sich dieser Vorgang und auch bei den von dort aus verlinkten Internetseiten usw.. Die Webcrawler kommen erst zum ‚stehen', wenn sie an einer Webseite angelangt sind, die selbst nicht auf andere Webseiten verweist. Mehrere dieser Webcrawler, mit geschickt gewählten Anfangspunkten, ermöglichen damit einen Großteil des Internets zu indexieren. Man erhält (unter anderem) eine rein syntaktische Struktur des globalen Computernetzwerks, die prinzipiell schon ausreichen würde um eine Grafik wie Abbildung 1 zu zeichnen.

Der PageRank Algorithmus nach Larry Page, einem der Google Gründer, benannt, hat einzig und allein die Aufgabe die Ergebnisse der Webcrawler, also das Rohmaterial, zu ordnen. Ordnung ist hierbei durchaus im preußischen Sinne zu verstehen, denn PageRank bringt den für Menschen nicht sinnvoll rezipierbaren Wust der Crawlingdaten in eine uns sehr vertraute, ‚bürokratische' Form: die Liste. Welche magische Kraft steckt nun hinter Pages Ranking Verfahren, dass es Netzwerke in Listen verwandeln kann? Oder ist dabei weniger Magie, sondern vielmehr List im Spiel? Wie dem auch sei, grob gesprochen, analysiert PageRank den Linkwust der Webcrawler folgendermaßen: Je mehr Internetseiten auf eine Seite verlinken, desto höher ist ihr PageRank. Verweist eine Webseite A mit einem hohen PageRank auf eine Seite B, so ist der PageRank von B wiederum höher, als wenn sie von einer Seite C mit sehr niedrigerem PageRank aus verlinkt wäre. Larry Page simuliert damit das fiktive Verhalten eines zufällig umhersurfenden Users:

> PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. And, the *d* damping factor is the probability at each page the "random surfer" will get bored and request another random page.[3]

Der PageRank einer Webseite drückt also die Wahrscheinlichkeit aus, dass dieser wahllos umhersurfende Internetnutzer auf dieser Seite landet oder eben nicht. PageRank liefert im Endeffekt, wiederum ohne jeglichen Bezug auf den Inhalt der Webseiten, eine komplette Hierarchie des Internets.[4] Die Suchmaschine verfügt dank Pages Verfahren über ein erstes Kriterium zur Bewertung der Relevanz von Webseiten, bevor überhaupt eine semantische Analyse der Webseiten durchgeführt wird. Eine solche Analyse muss von Google natürlich trotzdem noch geleistet werden, schließlich wollen wir beim Versicherungsabschluss nicht von Paris Hilton beraten werden, nur weil ihre Webseite einen höheren PageRank hat und damit für Google relevanter ist als alle Maklerseiten des Internets.

Ein *Bild* vom Internet

Googles strukturell-syntaktische Funktionsweise hilft uns bei unserem Vorhaben, dass Internet in eine Knoten/Kanten-Grafik zu stecken ungemein. Denn eigentlich visualisiert Abbildung 1 nichts anderes als einen Webcrawler, der auf einer bestimmten Versicherungsseite gestartet ist. Es ist wohlgemerkt mehr oder minder Zufall, dass wir in unserem Beispiel mit genau einem Webcrawler auskommen, da alle relevanten Suchergebnisse untereinander verknüpft sind. Aber wer überlässt bei Versicherungen schon etwas dem Zufall?[5] Der einzige Unterschied zum ‚eigentlichen' Webcrawler besteht darin, dass es eben notwendigerweise einen semantischen Filter, nämlich den Begriff *insurance* gibt und demzufolge auch nur die Seiten, die Google, wie auch immer, damit assoziiert, dargestellt sind. Wir sind damit schon bei der Fragestellung angekommen, was genau die Kanten und Knoten in Abbildung 1 sind bzw. was sie symbolisieren sollen.

Im Fall der Knoten haben wir die Frage eigentlich bereits beantwortet: Ein Knoten in Abbildung 1 steht für eine Webseite, die in Googles Webseitenindex enthalten ist.[6] Die Kanten wiederum stehen für einen Hyperlink (Querverweis) zwischen zwei Webseiten.

---

3 Brin, Sergey; Page, Lawrence: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Kapitel 2.1.2, http://infolab.stanford.edu/~backrub/google.html (23.06.2007)

4 Natürlich abzüglich der Seiten, die nicht mittels Webcrawling indexiert werden können. Dazu später mehr.

5 Im schlimmsten Fall ist keines der Suchergebnisse unter einander verlinkt. Da die Ergebnisse trotzdem in der Suchmaschine erscheinen, müsste es dann für jedes Ergebnis einen einzelnen Crawler geben. Wir hätten damit für unsere Betrachtung also kein Problem mit einem derartigen Fall.

6 In Googles Datenbank sind wohlgemerkt unter anderem auch Bilder oder PDF-Dateien gespeichert, die in der ‚realen' Google-Anwendung auch als Knoten fungieren können, da sie ja genauso verlinkt sind wie Webseiten. Der Einfachheit halber beschränke ich mich aber nur auf Webseiten als Knotenpunkte.

Relevant sind dabei natürlich beide Richtungen, also sowohl der Link von Seite A zu Seite B als auch, wenn vorhanden, der Link von Seite B zu Seite A. Denn schließlich verfolgt der Crawler alle Links der gerade indexierten Webseite. Diese Einteilung bzw. Zuordnung der Knoten und Kanten ist, denke ich, unstrittig.

Deutlich strittiger wird es allerdings bei der Frage, inwiefern eine Grafik wie Abbildung 1 das Internet allgemein repräsentieren kann oder ob sie nicht entscheidende Eigenschaften des globalen Computernetzwerks unbeachtet lässt. Denn hier setzen zwei durchaus relevante Einwände an. Der erste Einwand bezieht sich auf die Knoten, sprich die Webseiten, und kritisiert zu Recht, dass Google und damit auch unsere Darstellung bei weitem nicht alle Internetinhalte indexiert hat. Diese Inhalte werden im Gegensatz zum *Surface Web* (dem Suchmaschinen-Web) als *Deep Web* oder auch *Hidden Web* bezeichnet. Darunter fallen z.B. Webseiten, die nur über ein Passwortformular zugänglich sind, d.h. einen personalisierten Mitgliedsbereich oder ähnliches haben. Zentrales technisches Problem ist dabei, dass die Webcrawler rein passiv arbeiten, also eben nicht in der Lage sind Formulare aktiv ‚auszufüllen'. Selbst wenn sie rein technisch dazu in der Lage wären, wäre sicher niemand bereit seine Onlinebanking Zugangsdaten Google zur Verfügung zu stellen, nur damit die Suchmaschine mehr Webseiten indexiert. Denn schließlich handelt sich dabei um persönliche Daten, die aus nachvollziehbaren Gründen nicht öffentlich zugänglich gemacht werden sollten. Wir dürfen daraus wohlgemerkt im Umkehrschluss nicht schließen, dass alles was nicht in Google recherchierbar ist auch keine allgemeine, öffentliche Relevanz besitzt.

Die Debatte um Größe und Erschließbarkeit des Deep Web wird sicher eine zentrale Frage im Umgang mit dem Internet bleiben. Für unser Vorhaben ist sie allerdings kein Todesstoss. Denn egal wie viele zusätzliche Webseiten ein Webcrawler indexiert, die grundlegende Struktur des Internets verändert sich dadurch nicht. Denn auch Mitgliederbereiche basieren auf dem selben Hyperlinksystem wie ‚normal' zugängliche Webseiten. Zudem können wir sowieso nur Ausschnitte des Internets betrachten.

Der zweite Einwand nimmt wiederum genau diesen Sachverhalt ins Visier: Wie können wir etwas über das gesamte Internet aussagen wollen, wenn wir uns dabei ausschließlich auf Teilgebiete und Ausschnitte beziehen? Nun, wir könnten uns, denke ich, ohne Probleme ein Mammutprojekt vorstellen, das alle bei Google gespeicherten Webseiten in der Art und Weise von Abbildung 1 visualisieren soll. Dieses Vorhaben würde sicherlich eine Menge Zeit und Geld beanspruchen. Da wir es mit einem fiktiven Mammutprojekt zu tun haben, sollte uns das aber nicht weiter stören – schließlich können wir ja zumindest die Hoffnung hegen, dass Exzellenz nicht das Ende der wissenschaftlichen Fahnenstange ist. Wir könnten zudem auch alle passwortgeschützten Webseiten ‚crawlen'. Wir würden den Nutzern dann, vielleicht unter der Bedingung einer Entschädigung und der sofortigen

Zerstörung der Passwortdaten – nachdem wir die ausschließlich maschinell erstellten Crawlerdaten anonym gespeichert haben – bitten uns seine Daten temporär für ein wissenschaftlich extrem wichtiges Projekt zur Verfügung zu stellen. Wenn diesem ‚verlockenden' Angebot auch nur einen Bruchteil der Nutzer folgen würde, erhielten wir im Endergebnis eine Grafik, die dem Verfechter des zweiten Einwands zufrieden stellen sollte. Schließlich hätten wir damit nicht nur das Surface Web, sondern auch einen Großteil des Deep Webs erfasst. Die Darstellung wäre allerdings so differenziert, dass sie im Ganzen für uns wohl kaum überschaubar wäre. Skaliert auf die Größe von Abbildung 1 würden wir wahrscheinlich nur eine komplett schwarze Fläche erkennen können. D.h. wir müssten das Bild soweit vergrößern, dass wir doch wieder nur Ausschnitte betrachten müssten, um sinnvolle Aussagen über das Internet zu machen.

Ein Bild vom Internet zu haben, ist also nicht nur rein metaphorisch, sondern auch im expliziten Wortsinn sehr nützlich, wenn nicht sogar unabdingbar, um das Internet als Netzwerk zu charakterisieren und zu verstehen. Die rein visuell bzw. rein grafisch fokussierte Methode bietet zudem meiner Meinung nach den großen Vorteil Netzwerke auf einer festen, einheitlichen Grundlage bzw. Ebene vergleichen zu können. Sie liefert uns ein Handwerk zur Überprüfung einer wichtigen Frage für unsere heutige ‚Netzwerkgesellschaft': Wie maßgeblich und vorbildhaft ist die Netzwerkstruktur des Internets wirklich? Oder gibt es eventuell ganz andere Netzwerktypen, die sich radikal vom Internet unterscheiden lassen?