

Rheinisch Westfälische Technische Hochschule Aachen
Lehr- und Forschungsgebiet Theoretische Informatik
Prof. Peter Rossmanith

Proseminar Netzwerkanalyse im SS 2004

Globale Metriken

Andreas Feider 243675

Sascha Beckers 243616

08.07.2004

Inhaltsverzeichnis

1	Einleitung und Motivation	2
1.1	Allgemeine Einführung in Metriken	2
2	Erste einfache Metriken	3
2.1	Degree Distribution	3
2.2	Connected Components	3
2.3	Clustering Coefficient	3
2.4	Mittlere Distanz	4
2.5	Radius und Durchmesser von Graphen	4
3	Wichtige Metriken beim Internetgraphen	5
3.1	Expansion	6
3.2	Resilience	6
3.3	Distortion	6
4	Messung von Hierarchie in Graphen	7
4.1	Link Value Distribution	7
4.2	Up/Down Analysis	8
5	Laplace'sches Spektrum	8

Zusammenfassung

In diesem Proseminar wollen wir der Frage nachgehen, was Metriken sind und wofür sie benutzt werden. Wir werden zuerst das „Werkzeug Metriken“ erläutern und anschließend ein paar einfache Metriken vorstellen. Danach werden wir uns mit etwas komplexeren Metriken beschäftigen, um dann auf konkrete Anwendungsmöglichkeiten anhand des Internetgraphen und der Messung von Hierarchien in Netzwerken zu kommen.

1 Einleitung und Motivation

In der Netzwerkanalyse ist die Diskussion und Erörterung von Graphen und Netzwerken eine grundlegende Aufgabe für die man bestimmte Werkzeuge braucht. Diese variieren je nachdem was im Graph untersucht werden soll. So muss man sich erst genau darüber im Klaren sein, was für einen die entscheidenden Eigenschaften eines Netzwerkes in der jeweiligen Untersuchung sind. Deshalb wollen wir hier eine kleine Übersicht darüber geben, was grundlegende Eigenschaften von Graphen bzw. Netzwerken sind und wie diese mithilfe von Metriken gemessen werden können.

1.1 Allgemeine Einführung in Metriken

Wir geben hier jetzt zunächst eine Antwort auf die Frage was Metriken sind um dann in den nächsten Abschnitten verschiedene Arten von Metriken und deren Anwendung vorzustellen.

Die Aufgabe von Metriken ist es Graphen auf ihre topologischen Eigenschaften hin zu untersuchen und sie anhand dieser Werte vergleichbar machen zu können. Dazu bestimmt man mathematische Verfahren, die diese Eigenschaften und Attribute abbilden sollen und hat so exakte Werte, die für einen Vergleich herangezogen werden können. Außerdem ist es bei komplexen Netzwerken mit Tausenden oder Millionen von Knoten, wie sie in der Praxis häufig vorkommen, nicht möglich durch „scharfes Hinsehen“ zu entscheiden, ob zwei Graphen ähnlich sind oder nicht. Deshalb muss man auf Metriken zurückgreifen, die man durch ihre mathematische Methodik bedingt leicht in Algorithmen bestimmen kann.

Erwähnt sei an dieser Stelle jedoch, dass es für die Analyse eines Netzwerkes in der Regel nicht ausreicht nur eine Metrik zu betrachten. Gerade bei komplexeren Netzwerken ist es absolut notwendig mehrere Metriken auszuwerten bevor es möglich ist, sich ein klares Urteil zu bilden.

2 Erste einfache Metriken

In diesem Abschnitt stellen wir einige einfache Metriken vor, die leicht einsehbare Eigenschaften von Netzwerken verdeutlichen.

2.1 Degree Distribution

Die erste Metrik, die wir hier betrachten wollen ist die *Degree Distribution* (Gradverteilung).

Eine wichtige Tatsache eines komplexen Netzwerkes ist es, dass nicht jeder Knoten dieselbe Anzahl an nächsten Nachbarn besitzt. Gewöhnlich variiert diese als *Grad* definierte Eigenschaft von Knoten zu Knoten. So handelt es sich bei der Gradverteilung $P(k)$ um die Wahrscheinlichkeit, dass ein zufällig ausgewählter Knoten v des Graphen den Grad k besitzt. Interessanterweise gehorcht diese Verteilung oft dem so genannten *Power Law* der Form $P(k) \sim k^{-y}$. Daraus folgt, dass das zugrunde liegende System keine charakteristische (Grad) Skala besitzt, daher der Name skalenfreies Netzwerk. Das heißt, dass die Grade der Knoten nicht in der näheren Umgebung des Mittelwertes über die Grade aller Knoten liegen. Die Konsequenz davon ist das Auftreten von *Hubs* (Knoten hohen Grades), welche das Netzwerk zusammenweben. Diese Knoten spielen eine zentrale Rolle in der Robustheit und Verwundbarkeit eines solchen Netzes.

Die Komplexität dieser Metrik ist linear in $O(|V|)$.

2.2 Connected Components

Bei den *Connected Components* (Zusammenhangskomponenten) handelt es sich um die Anzahl der voneinander unabhängigen Teilgraphen. In einer Zusammenhangskomponente gibt es zwischen je zwei beliebigen Knoten einen Weg im Graphen, der diese miteinander verbindet. In gerichteten Graphen unterscheidet man zusätzlich noch zwischen starken und schwachen Zusammenhangskomponenten. Wobei in den *Strongly Connected Components* jeder Knoten von jedem aus erreichbar ist mit Berücksichtigung der Richtung der Kanten und in den *Weakly Connected Components* zwar auch jeder Knoten von jedem anderen aus erreichbar ist, jedoch nur ohne Berücksichtigung der Richtung der Kanten.

Auch diese Metrik hat lineare Komplexität.

2.3 Clustering Coefficient

Der *Clustering Coefficient* (Clusteringkoeffizient) ist ein Maß für den Grad der Verlinkung in einem Graphen. Man unterscheidet den lokalen Clusteringkoeffizienten für einen bestimmten Knoten des Graphen und den globalen Clusteringkoeffizienten für den gesamten Graphen (auch Vernetzungsgrad).

Der *lokale Clusteringkoeffizient* $C(v)$ eines Knotens v in einem Graphen G bezeichnet in der Graphentheorie den Quotienten aus der Anzahl der Kanten die zwischen seinen Nachbarn tatsächlich verlaufen und der Anzahl Kanten, die zwischen seinen Nachbarn maximal verlaufen könnten (Wenn v k_v direkte Nachbarn hat $\frac{k_v(k_v-1)}{2}$) (aus [8]).

Der *globale Clusteringkoeffizient* gibt das Verhältnis der vorhandenen Links zu den möglichen Links an. Ein *vollständiger Graph*, in dem jeder Knoten mit jedem verbunden

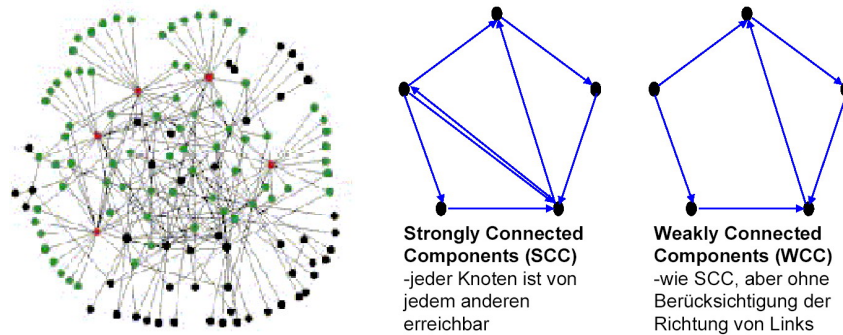


Abbildung 1: **Skalenfreies Netzwerk:** Die fünf dominierenden (roten) Knoten sind in direktem Kontakt mit 60% aller übrigen Knoten (grün).

Connected Components: Hier wird der Unterschied zwischen Strongly und Weakly Connected Components verdeutlicht.

ist, hat, wie leicht einzusehen ist, den maximal möglichen Clusteringkoeffizient 1. Der globale Clusteringkoeffizient C lässt sich auch als Mittelwert der lokalen Clusteringkoeffizienten aller Knoten berechnen. Vereinfacht gesagt misst der Clusteringkoeffizient wie ähnlich die Nachbarschaft eines jeden Knotens zu einer Clique ist.

Der Clusteringkoeffizient entspricht somit auch der Wahrscheinlichkeit, dass zwei Knoten durch eine Linie miteinander verbunden sind, bedingt dass sie einen gemeinsamen nächsten Nachbarn besitzen. Ein niedriger Clusterkoeffizient impliziert, dass die lokale Topologie mehrheitlich baumartig ist, wohingegen hohe Werte auf eine reichhaltige Schleifenstruktur hinweisen. Interessant ist außerdem, dass so genannte *Small-World-Netzwerke* einen sehr hohen durchschnittlichen Clusteringkoeffizienten haben. In einem Zufallsgraphen ist der Clusteringkoeffizient dagegen relativ gering.

2.4 Mittlere Distanz

Eine andere interessante Frage ist die folgende: Durch wie viele Kanten sind zwei Knoten im Mittel getrennt? Die Antwort darauf liefert einen Anhaltspunkt hinsichtlich der globalen Vernetzung eines Graphen. Das entsprechende Maß ist die *mittlere Distanz*: Man zähle für alle möglichen Knotenpaare die Anzahl Kanten, welche im kürzesten Weg zwischen den dazugehörigen Knoten enthalten sind und bilde dann den Mittelwert.

Die mittlere Distanz wird auch *Characteristic Path Length* genannt.

Ein Netzwerk, in dem globale Distanzen klein sind, und das hohes Clustering aufweist, wird kleine Welt (Small-World-Netzwerk) genannt. Das gleichzeitige Auftreten von hoher lokaler und globaler Vernetzung in solchen Netzwerken bestimmt maßgeblich die Ausbreitung von Computerviren und Infektionskrankheiten, das Funktionieren unserer Gehirne, das synchrone Zirpen von Heuschrecken sowie eine Vielzahl von weiteren Problemen.

2.5 Radius und Durchmesser von Graphen

Abschließend in diesem Abschnitt beschäftigen wir uns nun noch mit dem *Radius* und dem *Durchmesser* von Graphen. Dafür benötigen wir die Eigenschaft der *Exzentrizität* eines

Knotens, die definiert ist als dessen Abstand zum von ihm aus am weitesten entfernten Knoten im Netzwerk.

Der Radius des Graphen ist dann einfach das Minimum und der Durchmesser (im engl. diameter) das Maximum über alle Knotenexzentrizitäten.

3 Wichtige Metriken beim Internetgraphen

An dieser Stelle werden wir mal eine konkrete Praxisanwendung (entnommen aus [5]) von Metriken betrachten. Dabei geht es um die Modellierung des Internetgraphs. Das betrachtete Problem ist die Frage: Mit welcher Art von Netzwerkgeneratoren kann man den Internetgraphen am besten nachbilden? Dazu wurde das Internet auf zwei verschiedene Arten gemessen (aufgrund der technischen Möglichkeiten nicht vollständig möglich) – einmal mit Autonomen Systemen (AS-level graph) und einmal mit Routern (Router-level graph) als Knoten des Graphs. Dann wurden mehrere Netzwerktopologiegeneratoren mit diesem Graphen verglichen, die sich in mehrere Klassen unterteilen lassen: die random generators, die mit gewissen Wahrscheinlichkeiten zufällig Knoten verbinden, die structural generators, die Hierarchie als Grundlage haben und die degree-based generators, die die Gradverteilung (s.o.) als Grundlage haben. Das Ergebnis der Untersuchung war, dass die degree-based generators (im Besonderen der Power Law Random Generator, der dem Power Law folgt – wie auch das Internet selbst) den Internetgraph am besten modellieren konnten, auch wenn sie Hierarchien nicht explizit aufbauten (das Internet ist zweifellos hierarchisch aufgebaut).

Die Beschreibung dieser Analyse halten wir knapp, da große Teile davon, insbesondere die Details der Netzwerkgeneratoren, der gemessenen Internetgraphen und die genauen Ergebnisse samt zugehöriger Diagramme und Tabellen bereits im Vortrag zum Thema *Internet* vorgestellt worden sind.

Allerdings wollen wir nochmal genauer auf die zum Vergleich der modellierten mit den gemessenen Graphen herangezogenen Metriken eingehen. Bei diesem Vergleich wurden hauptsächlich drei Metriken benutzt (andere nur zur Bestätigung der Ergebnisse), die sorgfältig aus allen möglichen ausgewählt worden nach folgenden Kriterien:

- Messungen der Großstruktur des Internets sind wichtiger als allein lokale Werte
- Die Metriken sollen oberflächliche Unterschiede wie etwa die Größe des Graphen ignorieren
- Die Metriken sollen in der Lage sein bekannte „kanonische“ Graphen voneinander zu unterscheiden. In diesem Fall wurden dafür die drei kanonischen Graphen Gitter, Baum und Zufallsgraph ausgewählt, da diese – schon rein intuitiv – sehr unterschiedlich zueinander sind und die Art dieser Unterschiede auch wichtig für Netzwerke ist und deshalb sollten die ausgewählten Metriken auch zumindest diese drei Graphen deutlich unterscheiden können.

Drei Metriken, die diese Kriterien erfüllen und so für den Vergleich benutzt worden, werden wir nun im folgenden genauer vorstellen.

3.1 Expansion

Als erstes stellen wir die *Expansion* vor, ein Maß für die Verteilung von Knoten und deren Erreichbarkeit. Man zeichne einen Ball mit Radius h um einen beliebigen Knoten des Graphen. Der Radius h gibt an, wie viele Hops – also welche Pfadlänge – zurückgelegt werden dürfen. Die Zahl der erreichbaren Knoten steigt exponentiell mit wachsendem Radius h . Um das ganze nicht nur auf einen Knoten des Graphen zu beziehen, bezeichnet $E(h)$ nun die Anzahl von Knoten, die in einem Ball mit gleichem Radius h im Mittel um einen beliebigen Knoten des Netzwerkes zu finden sind. Man berechnet also die durchschnittliche Kardinalität der Erreichbarkeitsmenge. Durch diese Technik ist es möglich Graphen unabhängig von ihrer Größe zu vergleichen, indem man den gleichen Ballradius h wählt. Um festzustellen, ob ein Knoten v innerhalb des Balls mit dem Radius h liegt, der um den Knoten v_c gezogen ist, darf der kürzeste Weg zwischen den Knoten v und v_c nicht länger als h sein.

3.2 Resilience

Die *Resilience* misst die Existenz von alternativen Pfaden. Eliminiert man in einem Baum eine einzige Kante, so ist der Graph nicht länger zusammenhängend, also man kann nicht mehr von jedem Knoten einen Pfad zu allen anderen Knoten finden. Um diese Eigenschaft nun in Zahlen ausdrücken zu können, bestimmt man wieder einen Ball um einen Knoten, diesmal ist aber nicht der Radius h , sondern die enthaltenen Knoten n relevant. $R(n)$ ist nun definiert als die durchschnittliche, minimale Anzahl an Kanten, die man durchschneiden muss, damit Graph innerhalb dieses Balls mit n enthaltenen Knoten nicht mehr zusammenhängend ist. Die Berechnung dieser Metrik ist NP-schwer. Ein Baum besitzt somit eine Resilience von $R(n) = 1$, unabhängig der Größe des Baumes.

3.3 Distortion

Die dritte Metrik ist die *Distortion*. Man betrachtet jeden Spannbaum des Graphen und berechnet die durchschnittliche Distanz zwischen allen Knoten des Spannbaums, die sich

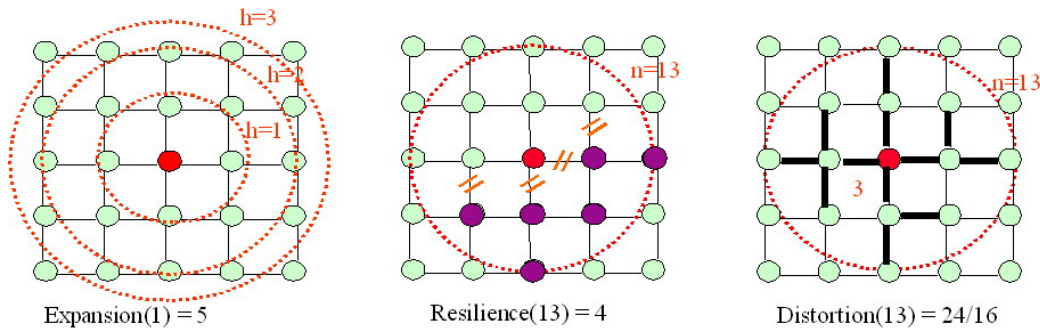


Abbildung 2:

Expansion: Streuungsrate **Resilience:** Existenz alternativer Pfade

Distortion: Baumähnlichkeit

eine Kante im Graphen teilen. Dies misst nun, wie deformiert der Graph ist, also wie viele zusätzliche Schritte nötig sind, um vom einen Knoten zum anderen Knoten zu gelangen, wobei nur Kanten des Spannbaums benutzt werden dürfen. Distortion ist nun das Minimum aller Durchschnitte über alle Spann bäume. Man könnte sagen, diese Zahl gibt an, wie baumähnlich der betrachtete Graph ist. Bestimmt wird nun die Distortion $D(n)$ – wie bei der Resilience – wieder abhängig von der im Ball enthaltenen Knotenanzahl n . Die Berechnung dieses Wertes kann wieder NP-schwer sein. Ein Baum hat unabhängig der Größe $D(n) = 1$.

4 Messung von Hierarchie in Graphen

Ein weiterer wichtiger Aspekt der Analyse in [5] ist die Messung von *Hierarchie* in Graphen. Wie zuvor bereits gesagt ist das Internet hierarchisch gegliedert – doch was heißt das eigentlich genau? Intuitiv ist damit die Vorstellung gemeint, dass es eine Menge *backbone links* gibt, die sehr viel *Traffic* befördern, d.h. der Traffic ist nicht gleichmäßig über alle Links im Netzwerk verteilt, sondern konzentriert sich auf diese zentralen backbone links. Eine genauere Betrachtung der Hierarchie folgt nun in den nächsten zwei Metriken.

4.1 Link Value Distribution

Diese Metrik basiert auf dem Gedanken, dass es in Hierarchien Links gibt, die häufiger genutzt werden als andere. Bei dieser allgemeinen Betrachtung wird die Nutzungshäufigkeit aber nicht mit dem Traffic an Datenpaketen gemessen, sondern mit der Menge an Knotenpaaren, deren „Traffic“ zueinander den Link bei Benutzung des kürzesten Weges durchquert. Dies ist die *Traversalmenge* des Links. Wenn es mehrere kürzeste Wege zwischen einem Knotenpaar gibt, dann taucht dieses Knotenpaar in den Traversalmengen eines jeden Links in jedem kürzesten Weg auf.

Nun wird der *Link Value* als die kleinste Anzahl an Knoten definiert, die die Traversalmenge „abdeckt“. Wobei mit „abdecken“ die kleinste Anzahl von Knoten gemeint ist, die entfernt werden muss, so dass bei allen Knotenpaare in der Traversalmenge zumindest ein Knoten entfernt wurde. Diese Definition erscheint zwar unnötig kompliziert, aber vereinfachte Varianten haben sich im Test als unbrauchbar erwiesen. Das „abdecken“ hier entspricht dem *Vertex Cover* auf einem bipartiten Graph von Knoten in der Traversalmenge. Zur Berechnung des Vertex Covers werden in der Praxis bekannte Approximationsalgorithmen benutzt.

In dieser Metrik werden Backbone Links also höhere Link Values haben als periphere Links. Die Verteilung dieser Link Values ist das erste Maß von Hierarchie, dass wir ansprechen wollen. Wenn in einem Netzwerk alle Links ähnliche Link Values haben, dann existiert in diesem Netzwerk keine Hierarchie, da die Nutzungshäufigkeit gleichmäßig über den Graph verteilt ist. Wenn es andererseits nur wenige Links gibt, die einen hohen Link Value besitzen, dann gibt es ein kleines und gutdefiniertes *Backbone* im Netzwerk, auf dem sich die Nutzung konzentriert. Abschließend kann man zusammenfassend sagen, dass die *Link Value Distribution* das Ausmaß aufdeckt, zu dem die Nutzung im Netzwerk sich auf backbone links konzentriert.

4.2 Up/Down Analysis

Die zweite für eine Hierarchie charakteristische Eigenschaft, die wir hier betrachten wollen, ist die, dass Wege im Netzwerk dazu tendieren erst die Ebenen der Hierarchie aufzusteigen und anschließend wieder abzustiegen. Das heißt ein Weg zwischen zwei Knoten arbeitet sich erst die Hierarchie hoch bis zu einem Backbone und fällt anschließend wieder in der Hierarchie bis der Zielknoten erreicht ist. Die zweite Methode die Hierarchie zu messen funktioniert also so, dass die Folge von Link Values entlang eines Weges betrachtet wird in Hinsicht auf das Finden von *Up-Down-Mustern*.

In den Up-Down-Wege nehmen also die Link Values entlang des Weges erst zu und dann wieder ab. Neben diesen Up-Down-Wege, die in Hierarchien überwiegen, gibt es aber auch natürlich Knotenpaare, deren Weg zueinander nur runter oder nur rauf geht oder eben verläuft. Dieses sind aber auch *gültige* Pfade einer hierarchischen Struktur; im Gegensatz dazu sind *ungültige* Pfade solche, welche ein lokales Minimum der Link Values im inneren des Weges haben. In der Praxis ist aber eine nicht ganz strenge Auslegung dieser Definition sinnvoll, da es durch Ungenauigkeiten mal kleine Abweichungen nach unten gibt (z.B. die Folge (1, 2, 3, 2.9, 3, 2, 1)). Die Messungen in den kanonischen Graphen ergeben dann auch gewünschte Ergebnisse: Im Baum sind mehr als 95% der Wege Up-Down und in einem Zufallsgraphen weniger als 60%. Im übrigen ergab die Messung des Internetgraphen bzw. des durch den PLRG erzeugten Graphen durch beide Metriken, dass ihr Hierarchiegrad zwischen dem vom Baum und dem vom Zufallsgraphen liegt.

5 Laplace'sches Spektrum

Frühere Studien zeigten, dass man auch die größten Eigenwerte der Adjazenzmatrizen zur Vergleichbarkeit von Graphen einsetzen kann. Wir wollen diese Betrachtung erweitern und uns auf die Multimenge der Eigenvektoren beziehen, dem so genannten Spektrum. Um auch Graphen verschiedenster Größe miteinander in Relation setzen zu können, verwenden wir anstelle der normalen Adjazenzmatrix, die normalisierte Laplace-Variante und erhalten somit das *Laplace'sche Spektrum* (aus [6]).

$$L = I - D \cdot A$$

I ist in unserem Falle die Einheitsmatrix entsprechender Dimension, A die Adjazenzmatrix des Graphen und D die Diagonalmatrix, die auf ihrer Diagonalen den Kehrwert des Grades des betrachteten Knotens v_i als Eintrag enthält; außerhalb dieser Diagonalen findet man nur Nullen. Dies führt dazu, dass unsere Eigenwerte λ_i im Intervall $[0, 2]$ liegen. Diese Multimenge an Eigenwerten bildet einen spezifischen Fingerabdruck für unser Netzwerk. Graphen mit ähnlichen Eigenschaften besitzen auch ein ähnliches Laplace'sches Spektrum, wobei die Größe dabei nur eine untergeordnete Rolle spielt.

Literatur

- [1] *A.-L. Barabasi.*
Linked: the new science of networks – how everything is connected to everything else and what it means for science, business and everyday life,
Perseus, Cambridge, 2002
- [2] *M. Faloutsos, P. Faloutsos, C. Faloutsos.*
On Power-Law Relationships of the Internet Topology,
In Proceedings of the ACM SIGCOMM, 1999
- [3] *M. E. J. Newman, S. H. Strogatz, D. J. Watts.*
Random graphs with arbitrary degree distribution and their applications,
Proceedings of the National Academy of Science of the USA, 2002
- [4] *P. Radoslavov, H. Tangmunarunkit, H. Yu, R. Govindan, S. Shenker, D. Estrin.*
On characterizing network topologies and analyzing their impact on protocol design,
Computer Science Department, USC, 2000
- [5] *H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, W. Willinger.*
Network topologies, power laws and hierarchy,
ACM SIGCOMM Computer Communication Review, 2001
- [6] *Danica Vukadinović , Polly Huang, Thomas Erlebach.*
On the Spectrum and Structure of Internet Topology Graphs,
In proceedings of I2CS, 2002
- [7] *Danica Vukadinović , Polly Huang, Thomas Erlebach.*
A Spectral Analysis of the Internet Topology,
ETH TIK-Nr. 118, 2001
- [8] *Duncan J. Watts, Steven H. Strogatz.*
Collective dynamics of “small-world” networks,
Nature, 1998