Department of Computer Science

Rochester Institute of Technology

Master's Project Proposal

Exploring the Topology of Small-World Networks

Student: Min Hu

Committee Members

Chairman: Prof. Stanisław P. Radziszowski

Reader: Prof. Roger S. Gaborski

Observer:

November 18, 2002

Exploring the Topology of Small-World Networks

Abstract

Complex networks have been studied for a long time in order to understand various real-world complex systems around us. Complex systems, such as the WWW, the movie-actor network, social networks and neural networks, are systems made of many non-identical elements connected by diverse interactions. To study the network topology is one of important issues on the way of exploring such systems, because the structure always affects the system function. Traditionally, these systems have been modeled as either completely ordered graphs or completely random graphs. Until recently, some surprising empirical results in the field of complex networks, like 19 clicks of the web's diameter and 6 degrees of separation in social networks, show us the small-world phenomena existing in some large sparse networks. This finding motivates the interest in small-world networks. The objective of the project is to study the properties of small-world networks and the network evolution over time via experiments on a movie actor collaboration network; to find their different characteristics by comparing small-world networks with random networks; and to analyze the factors that result in such differences.

1. Background

1.1 Introduction

Complex networks are ubiquitous in nature and society. The Internet is a network of routers and computers connected by physical links. The World Wide Web is a network of billions of web pages connected by hyperlinks. The nervous system is a network of nerve cells connected by axons. The cell is a network of chemical molecules linked by chemical reactions. The organization is a network of people linked by social interactions. Although we know their existence, it's so hard for people to understand the topology of these networks due to their large size and inherent complexity. Intuitively, these systems seem to be disordered. Thus, these systems were traditionally constructed as random network models. But, the recent discovery of the small-world phenomena greatly changed the way in which people looked at these complex systems. The small-world phenomenon was cited from the social category. Stanley Milgram, a sociologist at Harvard University in the US, declared the known claim "six degrees of separation"[10] in 1967. It means that one person can contact anybody else in this world by only six intermediate people on average. That is, we live in a "small world". The small-world phenomenon is not just one exception occurring in the social network. Surprisingly, it is proved that this feature also exists in several large sparse networks like movie-actor networks, www and the neural network of the nematode worm C. elegans. Is it possible that the small-world phenomenon is common in many sparse networks? Under what conditions can one network be in the small-world network set? If a system can be constructed as a small-world network, what does this imply about the dynamic behavior of this system? The project will focus on investigations of these questions. This includes doing experiments in the real movie-actor network, studying four proposed models (Watts-Strogatz Small-world model [5], Erdös Rényi Random-graph model [10], A.-L. Barabási Scale-free model [10] and Jon Kleinberg Small-world model [13]), and analyzing empirical results against theoretic predictions in models.

1.2 Basic graph concepts

A **bipartite graph** *G* is a pair *(V, E)*, where the vertex set *V* is partitioned into two non empty sets V_1 and V_2, and every edge in the edge set *E* joins a vertex in V_1 to a vertex in V_2.

A **complete graph** *G* is a pair *(V, E)*, where $|E| = \binom{|V|}{2}$; that is, any two vertices are adjacent.

A **sparse graph** *G* is a pair *(V, E)*, where $|E| \sim O(|V|)$.

The **distance** *D(i,j)* is the minimum number of edges traversed from the vertex *i* to vertex *j*.

The **characteristic path length** *L* of a graph *G(V,E)* is the mean of all shortest path lengths among all pairs of vertices. $L = \sum_{k=1}^{\binom{|V|}{2}} D_k(i,j) \Big/ \binom{|V|}{2}$

The **diameter** of a graph *G*, *Diam(G)*, is the maximal distance between any pair of vertices in *G*. $Diam(G) = max_{(i,j)} D(i,j)$

The **neighborhood** of a vertex *v*, Γ_v, is a set of those vertices that are directly connected to *v*. $\Gamma_v = \{x \mid D(v,x) = 1\}$ $(v \notin \Gamma_v)$

The **local clustering coefficient**, *Cv*, is $Cv = |E(\Gamma v)| \Big/ \binom{|\Gamma v|}{2}$, where $|E()|$ is the number of edges in the induced subgraph whose vertex set is Γ_v, and $|\Gamma v|$ is the number of elements in Γ_v.

The **clustering coefficient** of a graph *G(V,E)*, *C*, is the average of all *Cv* over all *v*.

$$C = \left(\sum_{v \in V} Cv \right) \Big/ |V|$$

The **degree** of a vertex *v, dv,* is the number of edges incident on it. $dv = |\Gamma v|$

The **degree distribution** of a graph *G(V,E)***,** is the probabilities of obtaining a given value of the discrete random variable, X_k. The probability can be denoted by $P(X_k=r)$, where X_k represents the number of nodes with degree $k \in \{1,2,...,|V|\}$, and $r \in \{0,1,2,...,|V|\}$.

The **giant component** of a graph *G(V,E)* is a component with $\succ |V|/2$ vertices.

2. Project Description

2.1 Properties of Small-world Networks

The large-scale networks, which fall in the small-world network set, have three significant properties in common: sparseness, small diameter and clustering. This project will verify these three properties by investigating the movie-actor network. Besides them, other two properties, threshold of the giant cluster appearance and degree distribution, will be studied as well. The latter two properties are thought to be related to the network dynamics. With growth of the network over time, how they change is an interesting and important question hidden in the class of small-world networks. The explanation of these properties is given as follows:

- Sparseness

  In a small-world network, the graph with *n* nodes has edges closer to *O(n)* than to $O(n^2)$. So, comparing with a complete graph with the same number of nodes, the small-world graph has much smaller number of edges.

- Small diameter

  A random graph with a large connection probability usually has a small diameter. Also, this characteristic is observed in small-world networks. Shortcuts in a small-world graph result in its small diameter. The diameters in small-world graphs increase logarithmically with the number of nodes.

- Clustering

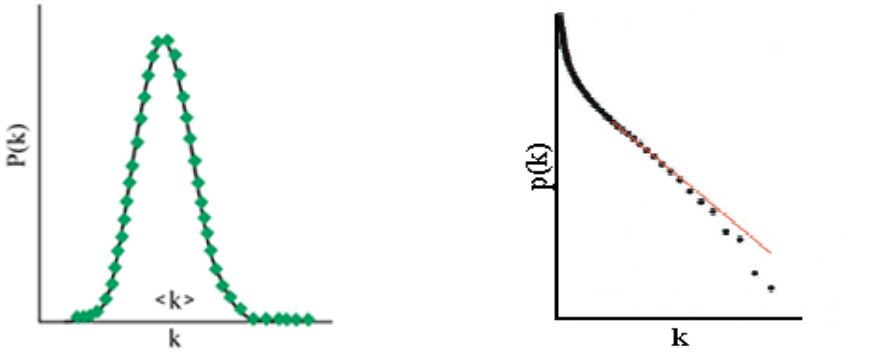
  Clustering coefficient *C* is the measurement of how much a graph is clustered. In a small-world graph, there is much higher clustering coefficient than in a random graph.

- Threshold of the giant cluster appearance

  The giant component emerges when the number of edges is greater than *(n/2)*[3] in random graphs. I would like to see if this property also appears in small-world graphs.

- Scale-free power-law degree distribution

  The degree distribution measures the network connectivity. According to the random graph theory, the degree distribution should be a Poisson distribution shown in Figure 1. left. For the Poisson distribution, *P(k)* is strongly peaked at $k = \langle k \rangle$, where $\langle k \rangle$ is a mean degree that depends on *p*; then exponentially decays as *k* tends to be very large [10]. But, an unexpected degree distribution was found in small-world graphs. It is a power-law degree distribution shown in Figure 1 right. A power-law distribution shows that most vertices have similar low degree, but a small number of vertices have much higher degree than in random graphs [2].



a. Poisson Distribution [10] **b. Power-law Distribution [2]**

Figure 1. : Two different kinds of degree distribution

2.2 Experiment environment

In this project, the study of small-world networks is closely related to understanding one real system, movie-actor collaboration network. The movie-actor network is selected as one example of complex systems based on each of the following requirements:

1. Be a large and sparse network
2. Possible to obtain the complete movie-actor database [11]
3. Construction as a bipartite graph
4. Network evolution over time

Several research groups studied the movie-actor network. They usually constructed this network as a unipartite graph in this way: there is only one type of vertex set in the graph: actors. Two corresponding actors are connected by an edge if they appear together in a movie [1]. In fact, the movie-actor network has an underlying bipartite nature. Exploring these real networks with the bipartite structure has not been done much. Such networks can be represented by bipartite graphs. Since bipartite graphs usually contain more information than unipartite graphs and both node sets of bipartite graphs will be increased over time, then, do the characteristics of small-world networks still exist in such graphs? In this project, the construction method is based on the bipartite nature of the movie-actor network. There are two distinct types of vertex sets in the graph: actors and films. Each edge always connects two nodes of different types. Any two nodes in the same set are never connected. For example, node A from the film set *m*, node B from the actor set *n*, there is an edge between A and B if actor B was in cast of the film A.

2.3 Applied Models

A model, which takes the form of an algorithm to generate new graphs reflecting some properties of a real-world network, is an abstraction of a real network. So, studying the models proposed by other people is a good way to understand those properties of networks. In this project, four models will be studied as follows:

- Watts-Strogatz Small-world model [8]

The generated graph of this model is shown in Figure 2 [5]. Figure 3 [5] shows two important statistics in the small-world graph.

Algorithm for generating sample networks:

Input: *(n, k, p)*

where *n* is the number of vertices, *k* is the distance in which each vertex is connected initially to its neighbors by undirected edges, and *p* ($0 \le p \le 1$) is the probability of rewiring each edge.

a) Start with a ring lattice with *n* nodes, each has the *kth* nearest neighbors; that is, $V = \{0,1,...,n\text{-}1\}$, $(i,j) \in E$ *iff* $i \neq j \cap (((i-j) \le k \cap i > j) \cup ((j-i) \le k \cap j > i))$. Thus, the degree of each vertex is *2k* and the ring has *(nk)* edges.

b) Replace original edges by random ones with the probability *p.* The following process is repeated at *k* times until all original edges are scanned: Pick up each vertex *v* from the ring in the clockwise direction; determine whether its edge connected to the *1st, 2nd, ..., or kth* nearest neighbor in the clockwise direction needs to be redrawn; if so, detach this edge and make a new edge from vertex *v* to vertex *w* that is not directly connected to *v* before.

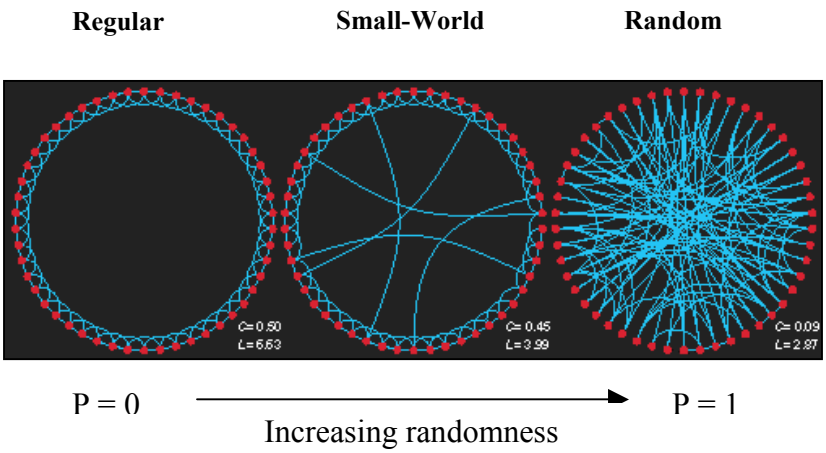


Figure 2([5]): Progression from regularity to randomness

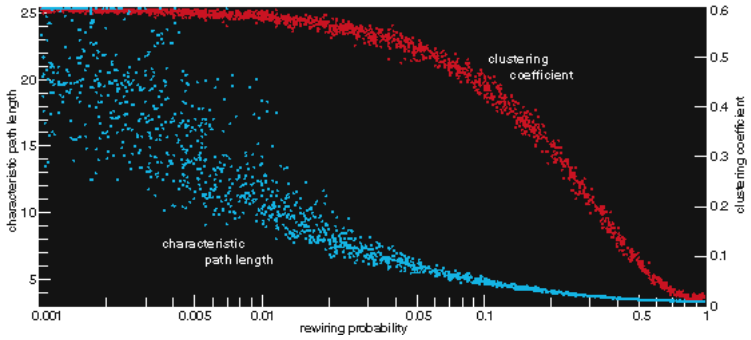


Figure 3. ([5]): The clustering coefficient and characteristic path length decline with p increasing.

- Erdös-Rényi random graph model [10]

  A sample graph of the model is shown in Figure 4 [10].

  Algorithm for generating sample networks:

  Input: *(n, p)*

  where n is the number of vertices, and p ($0 \leq p \leq 1$) is the edge probability.

  a) Start with n isolated nodes

  b) Connect each pair of nodes with the same probability p.

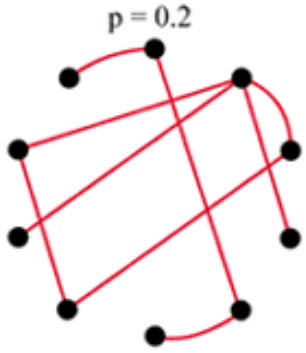


Figure 4. ([10]): This network is modeled as one Erdös-Rényi random-graph. It has $N = 10$ and $p = 0.2$. With these parameters, it is expected to have (and it has) 9 edges.

- A.-L. Barabási Scale-free model [1]

  This model (shown in Figure 5 [10]) modified two attributes of the above two models: a) with the graph evolution, the number of vertices is fixed. b) The same

probability p is applied to make new edges.

This scale-free model assumes that the network grows over time by addition of new nodes and new edges. When a new node is added to a scale-free graph, new edges are created from this node to some existing nodes. The probability that an existing vertex will receive one of such new edges is proportional to the current degree of the vertex. Hence, the more connectivity that a vertex already has, the higher probability that a vertex gains even more new edges, a phenomenon called preferential attachment.

Algorithm for generating sample networks:

Input: (n_0, m, t)

where n_0 is the initial number of vertices, m ($m \le n_0$) is the number of added edges every time one new vertex is added to the graph, and t is the repeated times.

a) Starting with n_0 nodes.
b) Every time we add one new node v, m edges will be linked to the existing nodes from v. Each existing node has a different probability $p(d_i)$ to receive the connection from v, such that $p(d_i) = d_i / \sum_j d_j$ [1]. Eventually, the graph has (n_0+t) nodes and (mt) edges.

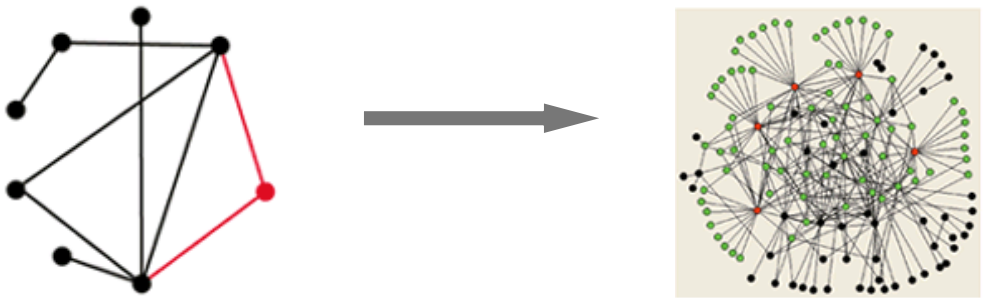


Figure 5. ([10]): Scale-free graphs

- Jon Kleinberg Small-world model [13]

As we know, there exist short paths linking together arbitrary pairs of nodes in the small-world networks. But, one question comes up: why can arbitrary pairs of nodes find these existing short paths. This model tries to figure out this question. It constructs graphs from a geographical intuition. The framework of this model can be put into *k*-dimension space.

Algorithm for generating sample networks:

Input: (k, n, p, q, r)

where k is the number of dimensions; n is the number of nodes in one dimension; p is lattice distance for constructing local connections; q ($q \geq 0$) is the number of long-range contacts for each node; and r ($r \geq 0$) is the clustering exponent that measures how wide the connections between nodes are.

a) Begin with a set of nodes which are identified by geographical coordinates in the *k*-dimension space, $\{(x_1, x_2, x_3, \ldots x_k): x_1 \in \{1, 2, \ldots, n\}, x_2 \in \{1, 2, \ldots, n\}, \ldots, x_k \in \{1, 2, \ldots, n\}\}$

b) Create local-range edges based on a constant $p \geq 1$. The node u has a directed edge to every other node within lattice distance p. The lattice distance is defined as the steps separating two nodes: $d(u, v) = d((x_1, x_2, x_3, \ldots x_k), (y_1, y_2, y_3, \ldots y_k)) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_k - y_k|$

c) Create long-range edges based on $q \geq 0$ and $r \geq 0$. Each node can choose q other nodes as the long-range contacts according to the probability proportional to $[d(u, v)]^{-r}$.

2.4 The Evolution of Small-world Networks

It is known that most of large-scale real systems, which belong to the small-world network set, dynamically grow up over time. However, it is only hypothesized whether these underlying characteristics, which determine the topologies of networks, have

impact on the dynamics of these networks. If so, how does it act? Does the dynamical behavior of such a system adversely affect its properties? What's the relationship between the topology and dynamics of these systems? In this project, I will try to investigate these questions. Since the study of the evolution, which includes variety of mechanisms like preferential attachment, local events, growth constraints [1] and so on, seems to be very complicated, it's almost impossible to cover all of them in this project. This project will concentrate only on some of them: In particular, I will extract the sub-networks from the entire movie-actor network according to consecutive years (or decades) of movies, compare the empirical data of network properties against the models mentioned above, and try to suggest a model most appropriate for this case.

2.5 Analysis of differences between empirical results and theoretic predictions

Analysis and hypothesis are important parts in the process of this project. After the movie-actor collaboration network is set up; empirical data of network properties is collected; and proposed models are built, I will collect and analyze empirical results against these models. Do empirical results match theoretic predictions very well? If there are some gaps between them, what differentiate small-world set systems from random set systems? The study of these will be the part of this project.

2.6 Software tools

The following software tools will be implemented in this project:

1) A program to extract data from the movie-actor database to construct a compact network structure.

This program reads several files from [11] about movies and actors, encode each actor and each movie in a unique number, generate a compact graph format and write it into a new file. The format is like this: each line represents the connectivity of one actor. The first column is an actor ID; the rest of columns are movie IDs, to which the actor

connects. This program can generate different sub-graphs by parameter controls on released movie year or issued movie countries. This program is written in Java or in C.

2) A simple topology generator to create topologies of the models proposed in section 2.3.

This generator will be flexible enough to propose and experiment with different models. These four models will be implemented in this generator. Each model can generate different graphs by construction parameter controls including the node number, the edge number and probabilities. The generator will output each generated graph into a file in the similar format as the program 1).

3) An analyzer program to analyze properties mentioned above based on the modeling topologies and the actual network structure.

This analyzer program reads the specific files created by the program 1) and 2). It will calculate the statistics of each given graph, such as the number of edges, a graph diameter, clustering coefficient, degree distribution and the size of a giant component. It will also provide a friendly GUI to show these statistic evidences.

3. Deliverables of the project:

- Technical report
- Code
- User manual

4. Schedule

09/03/2002 – 10/30/2002 Read references.

11/01/2002 – 11/30/2002 Project design and submission of the project proposal.

12/01/2002 – 03/10/2003 Experimental design and software implementation.

03/11/2003 – 04/30/2003 Write the document and prepare for defense.

05/01/2003 – 05/20/2003 Defense.

5. References

[1] Reka Albert, and Albert-László Barabási, Statistical mechanics of complex networks, Reviews of modern physics 74, http://www.nd.edu/~networks/PDF/rmp.pdf.

[2] L.A.N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, 2000, Classes of Small-world networks, http://polymer.bu.edu/~amaral/Papers/pnas00a.pdf.

[3] Bela Bollobás, 1985, Random Graphs, Academic press.

[4] James Case, 2001, The Continuing Appeal of Small-world Networks, SIAM News, Volume 34, Number 9.

[5] Brian Hayes, 2000, Graph theory in practice Part I/ Part II, American Scientist.

[6] Petter Holme, 2001, Characteristics of Small World NetWorks, http://www.tp.umu.se/~holme/seminars/swn.pdf.

[7] Christopher P. Mawata, 1997, Graph Theory Lessons, http://www.utc.edu/~cpmawata/petersen/index.htm.

[8] Steven H, Strogatz, 2001, Exploring complex networks, Nature 410:268-276.

[9] Ducan J, Watts, and Steven H, Strogatz, 1998, Collective dynamics of 'Small-world' networks, Nature 393:440-442.

[10] Albert-László Barabási, July 2001, The physics of the Web, Physics Web http://www.physicsweb.org/article/world/14/7/09.

[11] Movie Actor database: ftp://ftp.imdb.com/pub/interfaces/.

[12] Study of Self-Organized Networks at Notre Dame, http://www.nd.edu/~networks/.

[13] Jon Kleinberg, 1999, The Small-World Phenomenon: An Algorithmic Perspective, Cornell Computer Science Technical Report 99-1776.

[14] Jon Kleinberg, 2001, Small-World Phenomena and the Dynamics of Information, Advances in Neural Information Processing Systems (NIPS) 14, 2001.