Methodology for Weighted Social Networks Navigation

Fatima Mohsen Bin Alnaqeeb

Master of Science in Information Technology
Faculty of Informatics
The British University in Dubai
May, 2010

To My Family and Friends for their Unlimited Love, Support and Prayers.

Abstract

Our world is becoming an increasingly interconnected world. Connection between different people is being expanded dramatically especially after the vast use of technologies in this area. This expansion necessitates a deep analysis to capture the richness of information that these connections contain. Recently, social networks studies have attracted many researchers from different fields due to their common patterns that exist in wide range of real world networks and the exponential growth of social network sites. One of the important problems in studying social networks is network navigation: how to reach a destination node from a source node using minimum information.

In this thesis, our goal is to study the effect of weights in the network navigation and analyze the inter-play between the homophily, node degree, node strength and node continuous degree. We have identified three query routing paradigms based on defining different weights for nodes' edges to guide the navigation process through the network. We then have an extensive experimental study of the performance of incorporating weights into the network for different degrees, homophily parameters and different types of networks.

Acknowledgement

I would like to express my sincerest gratitude to my supervisor, Dr. Sherief Abdallah, for his guidance, support and inspiration during my work on this thesis. His rich knowledge and encouragement have been a great value to me throughout this thesis. I would like to thank Dr. Ibrahim Kamel for many useful discussions and also to express gratitude to my colleague Osama Al-koky, Research Assistant, for his great help and advice in the Java implementation part. Special thanks to my work supervisor Mr. Essam Hasan for his support and encouragment.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Fatima Mohsen Bin Alnaqeeb*)

Contents

List of Figures

List of Tables

Chapter 1

Overview

1.1 Introduction

The principle of living in a small world has inspired scientists to investigate many disciplines related to social, biological, mathematical and computer sciences. A considerable amount of work has been performed to develop models that can confine the properties of networks concerned with small world phenomena. Scientists have proved that small world characteristics exist in many real world paradigms. Social networks, worldwide webs and telephone communications are examples of these networks.

One of the important problems in studying social networks is network navigation: how to reach a destination node from a source node using minimum information. In distributed systems, only local knowledge is available for each node in the system to route messages. There are many practical applications that are based on a structure of distributed topology. Ad- hoc wireless networking, routing algorithms in telecommunication networks, peer-to-peer networks, systems for routing messages on World Wide Web and many other applications are examples of distributed systems. This widespread and extensive usage of such systems in different fields attracted scientists and researchers to propose algorithms and models in order to handle the challenge of routing the messages in the network using local information only.

In this thesis, we are interested in understanding simple heuristics that can be used to navigate complex networks that fall under small world networks. State of the art uses un-weighted network measures in computing heuristics. Here, we study different alternatives of incorporating weights into the network measures. This thesis stems from the interest in weighted networks and the confidence in the weights which empower the richness of information gained from the network. Hence, it eases the navigation process through the network. A state-of-the-art algorithm for addressing this problem is expected-value navigation (EVN) algorithm [15]. EVN depends on nodes degree and homophily parameter in guiding the navigation process through the network.

The main limitation of this algorithm is ignoring edge weights. In this thesis, we identify different alternatives for incorporating weights into the EVN algorithm. We study the effect of weights in the network navigation and analyze the inter-play between the homophily, node degree, node strength and node continuous degree. We then have an extensive experimental study of the performance of these alternatives for different degrees, homophily parameters and different types of networks. Our method shows an improvement in the performance of the navigation process when the nodes in the network are not highly correlated.

1.2 Problem Statement

Social networks have increasingly attracted the attention of academic and industry researchers from wide range of disciplines. The key objective of this thesis is to evaluate and extend a recent algorithm for searching social networks called expected-value navigation (EVN) [15]. It tackles the navigation process through the network based on decisions taken by the individuals in the network. Each individual has local knowledge about the network that includes information about its direct neighbors comprising their identity, certain attributes and degree but it is unaware of the rest of the network. In the network navigation stage, the aim is to achieve the requested query of reaching the destination with the minimum length of the path.
To accomplish our objectives of having new and useful methodologies in network navigation from weighted networking perspective, EVN was evaluated and compared with our proposed methods. This thesis work differs from earlier work on how it takes weights into account. EVN algorithm ignores edges weights while they can be an important source of information in the networks.

1.3 Research Questions

In this thesis, the following questions are addressed:

- How to extend EVN algorithm [15] to include weights?
  - What are the effects of node strength in network navigation process?
  - What are the effects of node continuous degree in network navigation process?
  - What are the effects of node strength and continuous degree in network navigation process jointly?
- The study of homophily effects on the network.
- How does network structure affect the navigation?

1.4 Contributions

In this thesis, we have studied weighted social network navigation, starting from network generation and working up to computational models for navigating networks.

The contributions of this thesis include:

- Identifying different alternatives for incorporating weights into the EVN algorithm [15].
- An extensive experimental study of the performance of these alternatives for different degrees, homophily parameters and different types of networks.

1.5 Scope

In this thesis, we focus on scale free networks with two types of degree distributions: Power law and Poisson. We assume network edges are weighted where each weight reflects in some way the similarity between nodes. It is our aim that the contributions made in the thesis have an impact on the researches conducted in social networks.

1.6 Organization of Thesis

The remaining chapters are arranged as follows: Chapter 2 presents a literature review about social networks, complex networks and some related works. Chapter 3 illustrates the EVN algorithm, which is designed to navigate networks by using homophily and degree. Chapter 4 demonstrates our proposed algorithm in the navigation process and investigates the influence of nodes' strength degree and continuous degree as a weight of the network edges. In chapter 5, the experimentation and evaluation of the algorithm is shown. Finally, in chapter 6, ideas for future researches and enhancements are highlighted.

Chapter 2

Literature Review

In this chapter, the basic background of networks in general, social networks, the concept of scale free networks, small world networks, and navigation through them are covered. At the end of the chapter, some related works in this field are elaborated.

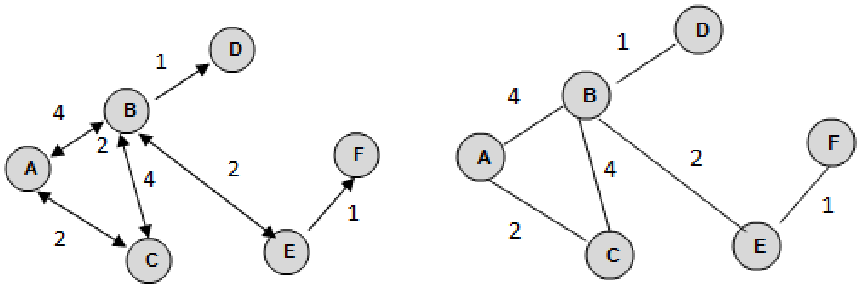
2.1 Networks

A network (*graph*) is comprised of a set of connected elements (*nodes*) and interacting entities (*edges*) in a system. Network that is represented as nodes and edges can appear in different systems. The below table shows examples of some systems and the corresponding nodes and edges for each system.

System	**Nodes**	**Edges**
World Wide Web	web pages	hyperlinks
Communication	computers	optic cable
Collaboration	researchers, co-authors	research collaboration
Call graphs	telephones	calls
Citation	papers	referring to
Social networks	people	realtionships
Semantic networks	conceptss	relations
Ideas distribution	people	ideas

Table 2.1: Network System Examples

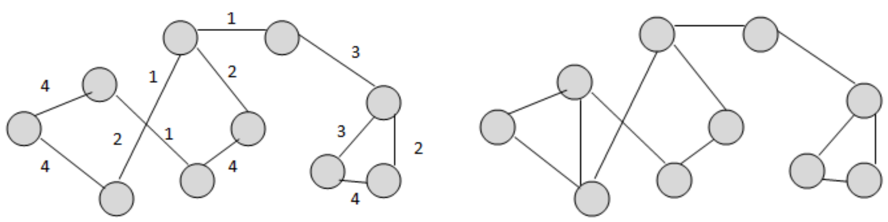
2.1.1 Undirected and Directed Networks



(a) Directed Network (b) Undirected Network

Networks can be directed or undirected. Directed networks is a one way network that means if an edge exists between node A to node B, this does not necessarily apply that an edge exists between node B to node A. Nodes in a directed network can have in-degree and out-degree. In-degree equals to the number of nodes that connect in to the node while out-degree represents the number of nodes that the node connects to. For example, An advice network where edges are created when people ask for advice from others are generally recorded as directed [11]. Figure 2.1(a) represents a directed network. Let's have node E as an example, it has an out-degree of 2, but only an in-degree of 1.
Undirected network is a two ways interaction network where if an edge exists between node A to node B, applies the existence of an edge between node B to node A. A collaboration network is an example of undirected networks where edges are created when two people collaborate on a project. Figure 2.1(b) represents an undirected network. Let's have node E as an example, it has a degree of 2 since it is connected to nodes B and F.
Our concern in this study is on directed networks in the case of synthetic networks and undirected for the real- world networks.

2.1.2 Weighted and Un-weighted Networks



(a) Weighted Network (b) Unweighted Network

It can be seen clearly that not all edges in a network have the same weight. This means that the capacity of some edges is more than others. Hence, a special treatment should be taken with these edges. To clarify this, here is an example, in a social network some contacts are friends, whereas others are simply acquaintances. Friends will be connected to a certain nodes more strongly than any acquaintances. This should be reflected in the network parameters and measures, therefore, two types of network exist: un-weighted and weighted network

- An un-weighted network is a network where all the edges among nodes are considered equivalent.
- A weighted network is a network where the edges among nodes have weights assigned to them.

Network's degree distribution is a measure of relative frequencies of nodes that have different degrees. In this thesis, we use two types of degree distributions: Power law and Poisson.

2.1.3 Power Law Networks

Many networks exhibit power law distribution in their node degree such as social networks[2]. Power law represents the mathematical relationship between two objects when the frequency of one of them varies as a power of some attribute of the other object. Power law networks represent networks where majority of node are with low degree and minority with very high degree [4].

> *'A network is named scale-free if its degree distribution, i.e., the probability that a node selected uniformly at random has a certain number of links (degree), follows a particular mathematical function called a power law. The power law implies that the degree distribution of these networks has no characteristic scale'*.[1]

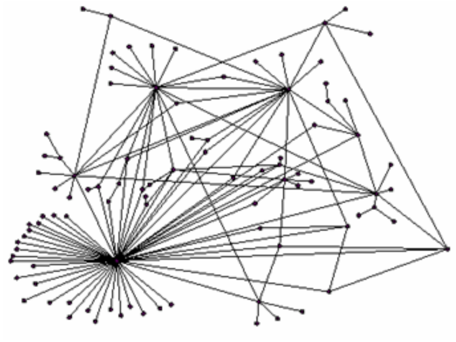


Figure 2.1: Scale-Free Network

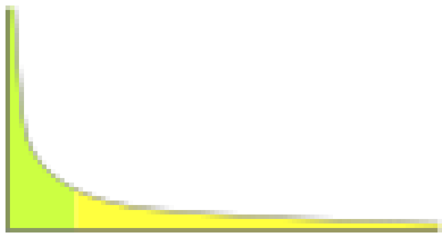


Figure 2.2: Degree Distribution in Power law Networks
[7]

2.1.4 Poisson Networks

Poisson distribution is an approximated distribution of the binomial distribution given large number of events and small probability of success.
The equation for the Poisson probability mass function is

$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}, \tag{2.1}$$

where λ is a parameter represents the average number of events in a certain period.

Figure 2.3: Poisson Network

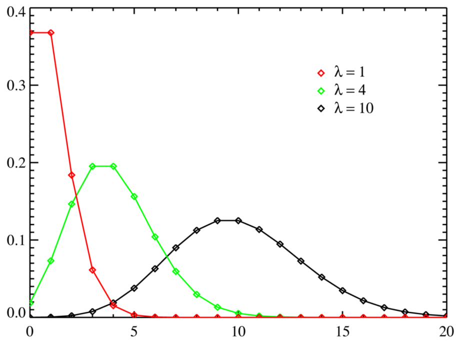


Figure 2.4: Degree Distribution of Poisson
[16]

2.2 Social Networks

In this chapter, we will review the study of social network. In this thesis, we concern about the small-world phenomenon from social network perspective that has been considered as complex network type due to its sophisticated characteristics.

2.2.1 Social Science

Recently, there has been a large expansion in social network sites from IT perspective but social networks have been deliberated comprehensively in social science for decades. It combines studies from different disciplines including anthropology, psychology and sociology.

Small World Phenomenon

We start by describing the pioneering experiment of Milgram [17] that is considered as the basis of the practical work in this field. It has been followed by many of probabilistic network models, resulting novel algorithmic and graph-theoretic studies of social networks.

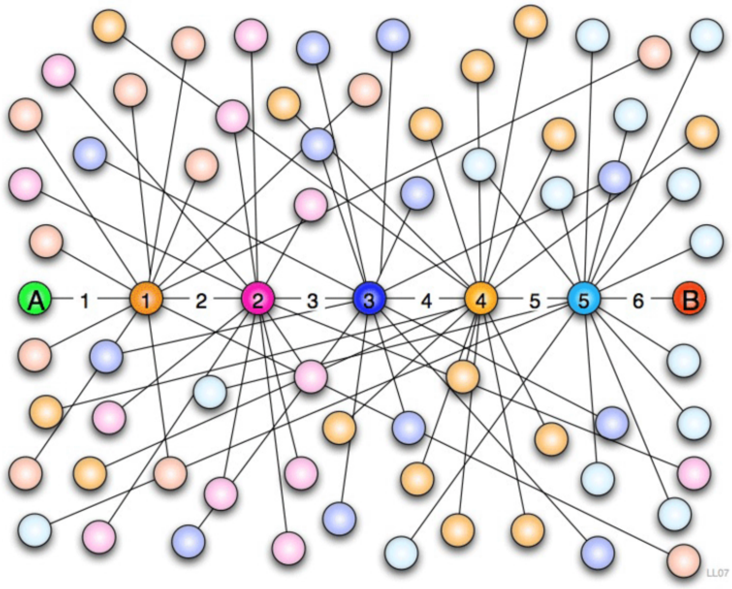


Figure 2.5: Six Degrees of Separation
[10]

The belief that we are all connected by short links, the small world phenomenon, is a foundation stone in the field of social networks. The small world phenomena analysis has its initiative steps in experiments conducted by the social psychologist Stanley Milgram in 1960s. The aim of the experiment was to find out short paths between people in the social network of United States. Milgram's experiment was conducted to estimate the average path lengths between any two individuals.
In the experiment, Milgram chose individuals in some US cities to be the starting source points, and others to be the end destination points. The reason behind selecting these cities was because of both socially and geographically distance between them in the United States. Invitations were randomly sent to individuals to participate in the experiment. The invitation described the idea behind the experiment and its purpose with basic information (address and occupation) about a target contact person. When the experiment began, the participants were forwarding the message directly to the target if known personally or to think of a friend or relative that is more likely to know the target according to their best knowledge. Knowing someone personally is based on first name. The progress of the sent messages was tracked by the researchers at Harvard University. The results of the experiment showed that the average path length between individuals in U.S is 6.

Followed by this experiment, scientists have proposed a number of network models as frameworks to study the problem of small world phenomena analytically.

Researchers who were interested in studying large scale real networks found that patterns appear in real networks [5]. Small-world networks are distinguished by their small average path length and the high clustering coefficient.

2.3 Social Network Navigation

The main purpose behind studying the structure of networks and proposing models to capture their properties is to understand how the network is generated and how the navigation through the network is performed. This section presents some related works in network navigation.

2.3.1 Kleinberg model

Small world networks have been navigated using the structure of the network and some characterstics of the users such as locatoin and job titles.

Kleinberg [14, 9] was interested in structuring the network in a way to enable a greedy search to find the best paths to be followed. Kleinberg's model structure is a lattice network with extra edges pq of probability proportional to $\frac{1}{|pq|^{\alpha}}$ where $|pq|$ is the Euclidean distance between p, q and α is a parameter. A simple greedy algorithm based on geography was able to find short paths to the destination, if $\alpha =$ d. The navigation through the network fails, if d $\neq \alpha$.

2.3.2 Watts, Dodds, and Newman Modell

The model of Watts, Dodds, and Newman [18] depends on the assumption that individuals in networks are organized based on a hierarchically structure into groups. They considered a hierarchical professional organization of individuals. The closer the nodes are in this hierarchy, the higher probability is given to their edges to indicate the higher closeness and similarity between them (higher homophily). A simple greedy algorithm that selects the next neighbor to pass the message is based on the selection of the most similar to the target along any property or dimension. It successfully sends a portion of the message before it gets stopped by the attrition probability of the nodes.
Similar results on a hierarchical network were confirmed by Kleinberg in [8]. In 2002, Dodds, Muhamad and Watts conducted an experimental study of search in global social networks in attempt to explain the small world phenomenon. More than 60,000 individuals were able to repeat Milgram's experiment using email chains. Based on the separation between the source and the target, the researchers estimated that the search process can be fulfilled in an average of five to seven steps[6].
Microsoft researchers conducted a newer research on its instant-messaging system that shows the average path length is 6.6 [12].

2.3.3 Expected- Value Navigation Model

Simsek and Jensen [15] proposed to use a simple product of the homophily and degree to navigate the network. Their algorithm tackles the navigation process through the network based on decisions taken by the individuals in the network. This paper is the key paper in this thesis; we will go through it in more detail in the following chapter.

2.4 Summary

In this chapter, we showed the dominant role of networks which spans to many disciplines that are growing exponentially. Social network is part of these growing networks where the nodes are a set of humans that are tied by one or more relationships. After that, we talked about scale free and small word phenomena. Finally, we described some experimental related works that have been performed in the area of social network modeling and searching.

Chapter 3

Background: Navigating Networks by Using Homophily and Degree

In this chapter, different parts of the experimental research of navigating networks using homophily and degree performed by O. Simsek and D. Jensen [15] are investigated in more detail. In Section 3.1, description and results of the research are discussed. In Section 3.2 and 3.3 two different heuristics which are used in their study are described. Finally, our goals for performing this project are stated in section 3.4.

3.1 Experiments on Homophily and Degree-based Navigation

The following two sections describe the main part, properties and overall results achieved by O. Simsek and D. Jensen experiment [15].

3.1.1 Description of homophily and degree-based navigation experiment

A central challenge in decentralized networks is guiding the search process and directing massages with individuals' available local knowledge.

Expected-value navigation (EVN) algorithm is an algorithm for searching social networks. It tackles the navigation process through the network based on decisions taken by the individuals in the network. Each individual has local knowledge about the network that includes information about its direct neighbors comprising their identity, certain attributes and degree but it is unaware of the rest of the network. In the network navigation stage, the aim is to achieve the requested query of reaching the destination with the minimum length of the path.
EVN uses a simple intuitive idea: choose the neighbor that is more likely to be on the shortest

path to the destination. To do so, the authors proposed a simple heuristic that approximates this objective (since it is impossible to know this for sure without knowing the whole network). The heuristic was to choose the neighbor with maximum product ($g(n)$) of node degree (k_s) and a homophily preference (f_{st}).

$$g(n) = f_{st}k_s \tag{3.1}$$

$$f_{st} = (max|a_s - a_t|, 0.01)^{-r} \tag{3.2}$$

where a_s and a_t are attribute values on nodes s and t, and r is a homophily parameter.
Let's have the below figure as an example by numbers on how to consider the node's strength:

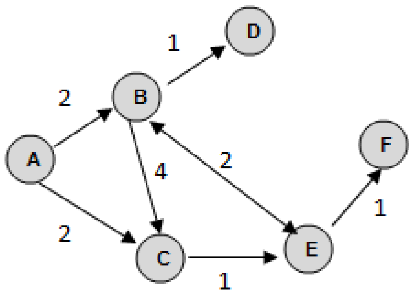


Figure 3.1: Example of using strength of a node

Let's assume that we have the following navigation request:

- Search query: Send a message in source node (A) to destination node (F)
- Considering neighbors: node A considers its neighbors (B and C) to select one of them to pass the message to it.
- Node's degree determination:
  - o B's degree: node B has a degree of 3.
  - o C's degree: node C has a degree of 1.
- Homophily preference calculation:
  - o Let's assume that the homophily attributes for B, C and F equals 0.4, 0.1, 0.5 respectively.
  - o Applying the homophily equation between node B and F(the target) gives 10
  - o Applying the homophily equation between node C and F(the target) gives 2.5
  - o This means that node B is more similar to node F than node C.
- The product of homophily and node's strength of nodes B and C:
  - o For node B, the product is 3*10=30
  - o For node C, the product is 1*2.5=2.5
- Nodes selection:
  - o The source node (A) will select the node that maximize the product of homophily and node's degree. In this step of this example, node B will be selected to continue directing the message.
- This procedure is repeated till the destination is reached or the search is stopped. The algorithm ignores visited neighbors in the presence of unvisited neighbors, and selects randomly among them otherwise.

EVN compromises the use of both network measures as follows:

- If dissimilar nodes are most likely to be connected together (no homophily in the network), the selection of the next node to direct the message will based on the degree (the highest degree will be chosen).
- If the nodes in the network tend to have equal number of degree, the homophily will take the control of guiding the navigation in the network.
- Increasing the two parameters enhances the performance of the navigation. This means that the higher the degree of the neighbor node and the more similarity to the target, the better the navigation becomes to lead directly to the target.

3.1.2 Results of homophily and degree-based navigation experiment

EVN greatly outperforms the previous algorithms that based only on homophily or only on degree. It succeeded in many cases where other algorithms failed in finding out short paths to route the message through the network. It compromises the knowledge obtained by both degree and homophily.

3.2 Experiments on degree-based navigation

This section describes the experiments on degree- based navigation along with the results of the experiments.

3.2.1 Description of degree-based navigation experiment

The navigation that depends on the degree of the nodes chooses the neighbor with the highest number of degree and passes the message to it.

3.2.2 Results of degree-based navigation experiment

Power law networks are characterized by small number of nodes with high number of degree and high number of nodes with small number of degree.
Hence, the degree based navigation in these networks will be efficient when the network shows no homophily or small values of homophily.

Increasing the degree parameter decreases the probability of having high degree nodes that results a decrement on the performance of the degree based navigation. Also, increasing the homophily parameter will allow the homophily to control over the navigation process.

In Poisson networks, networks show less variation in their degree compared with power law networks.

- Nodes do not high degree nodes as in power law.
- Increasing the degree parameter increases the performance of the navigation.

3.3 Experiments on Similarity-based (Homophily) Navigation

This section describes the experiments on similarity- based navigation along with the results of the experiments.

3.3.1 Description of Similarity-based Navigation Experiment

The similarity-based navigation selects the neighbor most similar to the target node in attribute values.

Homophily is the tendency of like to know who are like. In real world network, this similarity between the nodes can be based on many characteristic matrices such as nationality, geography, and occupation, etc.

3.3.2 Results of similarity-based navigation experiment

In power law networks, homophily-based navigation show a different tendency as follows:

- When the network shows a low homophily structure, the homophily based navigation becomes ineffective in guiding the navigation.
- A network with average homophily values is where the homophily based navigation has its best performance.
- High homophily network structure does not have many short paths.

In Poisson networks:

- Zero homophily parameter causes a reduction in the performance of the local navigation because in these networks no high degree of nodes are available. Hence, there is nothing to guide the navigation.
- Increasing the degree parameter improves the performance.

In general, using both homophily and degree was able to successfully guide the navigation. It outperforms other algorithms that depend only on one of these parameters.

3.4 Our Goal

EVN algorithm ignores edges weights while they can be an important source of information in the network. We have proposed different heuristics and performed numerous experiments to analyze the following points:

- How to extend EVN algorithm [15] to include weights?
  - What are the effects of node strength in network navigation process?
  - What are the effects of node continuous degree in network navigation process?
  - What are the effects of node strength and continuous degree in network navigation process jointly?
- The study of homophily effects on the network.
- How does network structure affect the navigation?

Chapter 4

Investigating the influence of nodes' strength and continuous degree

According to the EVN algorithm [15], authors did not consider the weights of the edges in finding the shortest path. Inspired by their algorithm, we propose the use of nodes' strength and continuous degree *'a new methodology for generalizing measures of unweighted networks through a generalization of the cardinality concept of a set of weights'* [3]. In this chapter, we describe how we apply the algorithm and present the results of the empirical study. The algorithm and its effects are discussed and compared with the results of experiment performed by O. Simsek and D. Jensen [15].

4.1 Node degree, node strength and continuous degree

In this section, the three main network measures that are used in this thesis are explained.

The degree of a node is one of the simplest network measures that have been widely used. It is equal to the number of other nodes connected to the node. It reflects how connected the node is. One of the degree measure drawbacks is that it ignores the weights of the connection between nodes. This means that an important source of information is ignored. Another problem with the degree is being discrete. A neighbor is either counted in the degree or not.

The strength of a node is equal to the sum of weights attached to the edges that connect a node to others. Node strength measures the intensity of the relationships of each node. It takes into account both the connectivity and the weights of the links.

The continuous degree (C-degree) of a node is a newly introduced measure for analyzing weighted networks [3] that we will use in this thesis. Different from all previously developed measures, C-degree succeeded in giving a proper generalization of the degree measure. It is a continuous generalization of the degree measure that captures the disparity of interaction between nodes.

C-degree was proved to be a proper generalization in [3]

The C-degree of a node in a network is

$$c(E') = \begin{cases} 0 & \text{if } E' \text{ is empty} \\ 2^{\left(\sum_{e \in E'} \frac{w(e)}{\sum_{o \in E'} w(o)} log_2 \frac{\sum_{o \in E'} w(o)}{w(e)}\right)}, & \text{otherwise} \end{cases}$$

where w(e) represents each node weight and $\sum_{o \in E'} w(o)$ represents the summation of all nodes weight (strength).

The strength of a node becomes similar to the node's degree if all weights are equal to 1. The C-degree becomes similar to the node's degree measure of the same node if every node interacts equally and uniformly with all its neighbors.

Chapter 5

Evaluation and Experimental Results

In this chapter, we discuss the detailed implementation of our method to incorporate weights in the networks. In this thesis, we are using synthetic networks and several real- world networks to represent our social networks.

The experiment goes through stages: network generation and network navigation.

5.1 Network generation

We consider directed networks with power-law degree distribution and Poisson degree distribution.

Networks with different degree parameters and homophily parameters have been used in this experiment.

Synthetic networks are generated for our experiment as follows:

- First, we generate the out-degree of each node according to the degree distribution (Power law or Poisson).
- Second, we generate the homophily attribute of each node. This value follows a uniform distribution over the range between 0 and 1.
- Then, for each node we generate the actual links and edges between nodes according to the homophily parameter equation. $f_{st} = (max|a_s - a_t|, 0.01)^{-r}$

Based on 3.2 equation, when the network shows no homophily (r=0), nodes have an equal probability to be linked to any other node in the network. Increasing the homophily parameter enables the nodes to connect to other nodes that are similar to them.

The following example explains in detail the network generation part. Suppose we want to generate a Poisson network of 5 nodes. We first generate 5 random numbers according to the Poisson distribution. These numbers are the out-degree of each of the 5 nodes; let's say 2, 3, 1, 1, 2. Then, we generate 5 real numbers representing the homophily attribute of each node, let's say 0.1, 0.5, 0.3, 0.4, 0.8. After that, we generate for each node the assigned number of links according to the node's out-degree. The first node has out-degree of 2, so we choose two random nodes of the other 4 nodes according to the homophily equation and add the corresponding links. The following table shows the steps for generating the links and edges for one node as an example (let's consider node A).

Node Name	**Out degree number** (according to the degree distribution	**Homophily attributes**	$f_{st} = (max\|a_s - a_t\|, 0.01)^{-r}$ (let's assume r=1 in this run)
A	2	0.1	-
B	3	0.5	$f_{ab} = (max\|0.1 - 0.5\|, 0.01)^{-1} = 2.5$
C	1	0.3	$f_{ab} = (max\|0.1 - 0.3\|, 0.01)^{-1} = 5$
D	1	0.4	$f_{ab} = (max\|0.1 - 0.4\|, 0.01)^{-1} = 3.33$
E	2	0.8	$f_{ab} = (max\|0.1 - 0.8\|, 0.01)^{-1} = 1.43$

Table 5.1: Example of Homophily Preference Calculation

Let us define the algorithm as function getNeighbors(X), where X is the node we want to generate neighbors for (i.e X = A in the above example).
1. Let R be the list of neighbors we want to generate for node X, which is initially empty
2. Let L be the list of nodes to consider, which is initially all nodes except X (in the example, this should be B,C,D,E)
3. For i = 1 to out degree of X do
4. g = chooseNeighbor(X,L)
5. remove g from L
6. add g to R
7. end for
8. return R

Now, let us define the function chooseNeighbor(X, L)
1. Compute TOTAL = sum of all fst for nodes in L (initially 12.26, but may decrease as we remove nodes in step 5 above)
2. TEMP = a random number between 0 and TOTAL
3. INDEX =0, SUM = 0
4. SUM = SUM + fst (A , L[INDEX])
5. While (TEMP $\succ$ SUM) do
6. INDEX ++
7. SUM = SUM + f_{st} (A , L[INDEX])
8. end while
9. return INDEX

5.2 Network navigation

This is the core part of the experiment where the navigation process and message directing is performed. It demonstrates the steps that each node takes in order to decide to which neighbor

the message should be passed to. Within the experiment, all nodes used the same algorithm for their decision making process. In this heuristic, the simulation starts with a query routing request from the source node. The message in the request should be routed and delivered to the destination node. In each search process, the source and the destination are chosen randomly among all nodes.
Using the node strength as a weight, the decision rule of each node is to maximize:
$g_s(n) = f_{st}s(n)$ where s(n) represents the node strength.
Using the C-degree as a weight, the decision rule of each node is to maximize:
$g_c(n) = f_{st}c(n)$ where c(n) represents the node c-degree.

Beside the synthetic networks that have been used to represent the social networks, we have used several real-world networks to study and evaluate the performance of our methods. The used real- world networks consist of three different networks.

The first used network is industry-yh [13] network that represents 1798 nodes of business news stories collected from the web. The edge connection between two companies is established if they appeared together in a story and the edge weight represents the number of this appearance.
The other two networks are webkb-texas-cocite and webkb-washington-link1. They consist of web pages gathered from computer science departments in Texas and Washington universities. Six categories of: course, department, faculty, project, staff, and student is labeled for each page [13]. Co-citation represents the linking edge between two pages.

The main difference in the navigation process between the synthetic network and the real-world networks is how the homophily attribute of each node is handled. In the synthetic network, as mentioned earlier, the homophily attribute follows a uniform distribution over the range between 0 and 1 and as in equation 3.2 the difference is calculated based on these values. While in the real-world networks, if two nodes belong to the same class, the difference value is set to be 0 otherwise 1.

5.2.1 Message directing process using homophily and nodes' strength

- The process starts by search query in the source node to passed to the distention node.
- The source node starts by considering its neighbors.
- The node strengths of the neighbors are calculated. Node strength: the sum of the weights of all links attached to the node
- The homophily preference between the neighbors and the destination (target) node is calculated based on the homophily equation: $f_{st} = (max|a_s - a_t|, 0.01)^{-r}$
- The product of each neighbor homophily and its corresponding node strength is calculated.
- The source node chooses the neighbor that maximizes this product.

5.2.2 Message directing process using homophily and the node C-degree

The same steps in section 5.2.1 are repeated when using the node C-degree except node's strength calculation. Instead C-degree equation is computed, multiplied by homophily and the neighbor that maximizes the product is selected to pass the message. C-degree is used to capture the disparity of interaction between nodes.

5.2.3 Experiments Setup

We conducted several experiments where all possible combinations of degree parameter and homophily parameter were considered. Each plot in our results represents 10 to 30 randomly generated networks. Each network is of 1000 node and in each run 5000 random search queries were selected.

The main performance criteria we used, similar to EVN experiment, is the empirical CDF (cumulative distribution function) of path length in power-law and Poisson networks as a function of degree distribution and homophily parameters.

Our simulations and empirical studies show the influence of node's degree strength and continuous degree as a weight of the edges each one separately. Other experiments were held to see the influence of both node's degree strength and continuous degree jointly aiming to have a method that can combine the goodness of both measures. These measures included computing the mathematical average of node's strength and node's C-degree, computing the geometrical average of node's strength and node's C-degree and computing the simple product of the two terms. After that each term were multiplied by the homophily as in all the experiments.
In [15], they assumed that the homophily parameter in network generation is equal to homophily parameter in network navigation. This assumption is not so practical, as in network navigation nodes do not know how the network were generated or structured. Hence, in addition to regenerating all plots in EVN [15], we generated more plots to show the effect of the algorithms in a wide range.

5.2.4 Results and Analysis

In power networks, increasing degree parameter decreases the probability of having high degree nodes in the network and in Poisson networks no high degree nodes. Hence, the navigation through the network depends mainly on the homophily structure.

Node's strength degree
In this part, we are basically replacing the node degree in equation 3.1.1 with the node strength. $g_s(n) = f_{st}s(n)$ where s(n) represents the node strength. Node's strength captures the capacity and intensity of the edges that is represented by the homophily. Therefore using node's strength degree outperforms using the degree alone in some of the generated networks where the homophily parameter ranges from 0 and 1.

Node's C-degree
In this part, we are basically replacing the node degree in equation 3.1.1 with the node C-degree. $g_c(n) = f_{st}c(n)$ where c(n) represents the node c-degree.
C-degree captures more on the disparity of weights. In this particular problem, we think that it is more important to focus on the absolute value of weights, which is what the strength does. Therefore using node degree and node's strength outperforms node's C-degree.

Node's strength degree and node's C-degree jointly
An open question to be answered and investigated more on how to combine different measures of node's strength degree and node's C-degree to come up with better results. Basic operations have been done to experiment its effect. It includes the mathematical average, the geometrical average and the product of both measures. The results was varying that it is shown in the resulted figures.

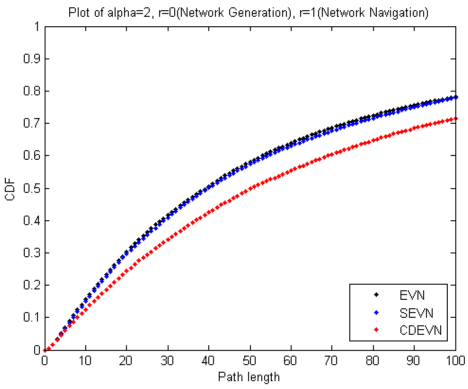
Our main aim of the conducted experiments is to study the effect of weights in the networks. We surprisingly found that weights did not improve the search process. Our definition of the weights depends mainly on the homophily of the network. The reason behind the unimproved performance in the network navigation using weights can be that homophily does not guarantee guiding the navigation through the shortest paths.

Then, we lowered the evaluated range of homophily parameter in our experiments to consider fraction values of homophily parameter. We evaluated the performance of the algorithms for the values of homophily parameter between 0 and 1. We found that the performance of using node strength outperforms node degree. An explanation of this is when the homophily parameter is 0; the network structure shows no correlation between neighbors. This implies that using the strength will not improve the performance of the search. Gradual increase of the homophily parameter shows an improvement in the performance of the search using node strength till a value of 1 where the generated networks start to show high correlation between neighbors. The effect of this high correlation results an equal effect of node degree, node strength and C-degree.

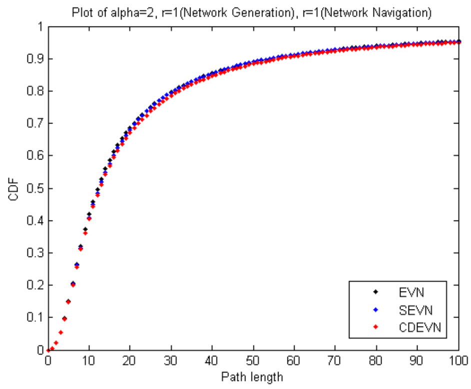
The following shows part of the plotted graphs for the different network structures.
Description of plots legend:

- EVN means implementing the algorithm using the discrete degree.
- SEVN means implementing the algorithm using node's strength.
- CDEVN means implementing the algorithm using C-degree.
- Avg-SCD means implementing the algorithm using the mathmatical average of node's strength and C-degree.
- GAvg-SCD means implementing the algorithm using the geometrical average of node's strength and C-degree.
- P-SCD means implementing the algorithm using the product of node's strength and C-degree.

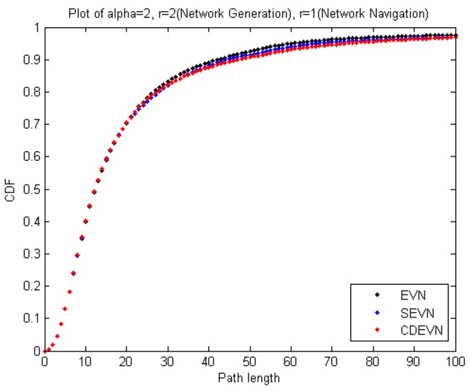
Results of Power-law networks



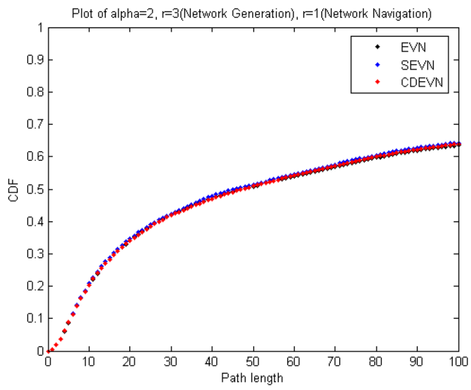
(a) Power Law-Alpha=2 r=0



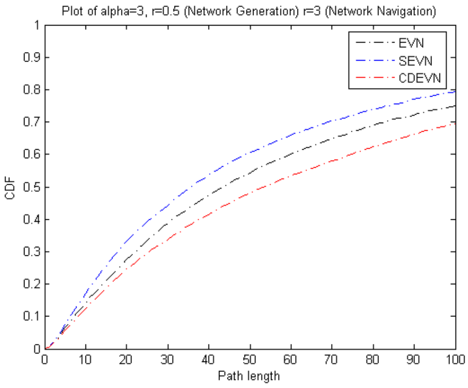
(b) Power Law-Alpha=2 r=1



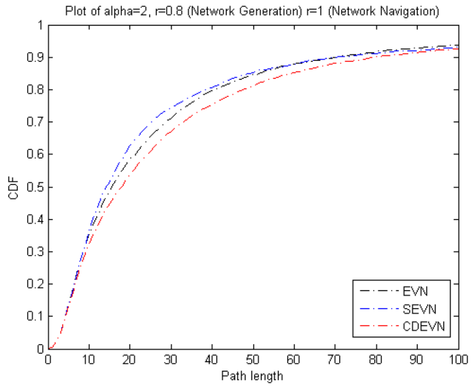
(c) Power Law-Alpha=2 r=2



(d) Power Law-Alpha=2 r=3

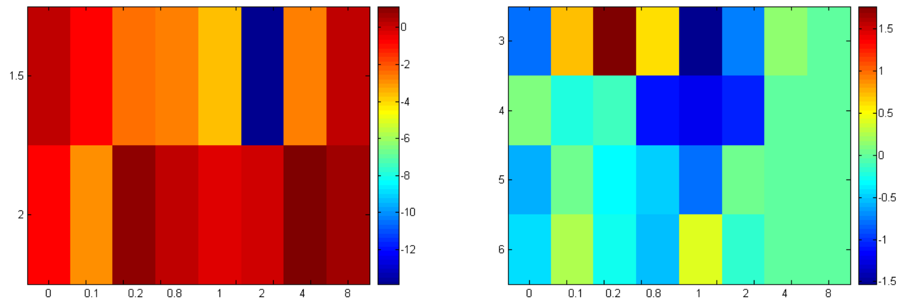


(e) Power law Network- Alpha=3, r=0.5



(f) Power law Network- Alpha=2, r=0.8

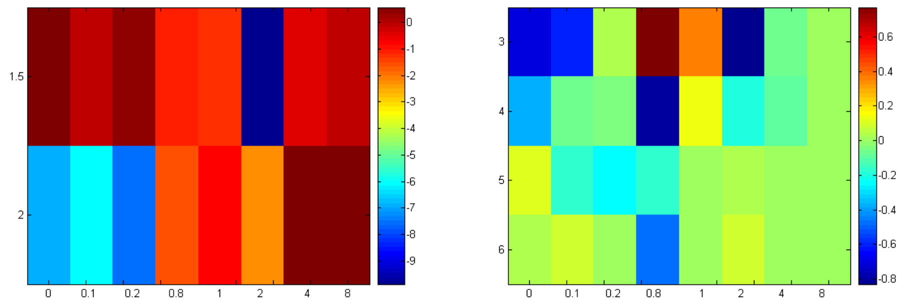
Figures 5.1(a), 5.1(b),5.1(c), 5.1(d), 5.1(e) and 5.1(f) represent of the empirical CDF of path length in the conducted experiments of power-law networks as a function of degree distribution and homophily parameter. As mentioned earlier, power-law networks have high number of nodes with low degree and low number of nodes with high degree. Increasing the degree parameter decreases the probability of having high degree nodes in the network. And the higher the homophiy parameter is, the more similar the nodes that are connected together. From 5.1(a), 5.1(b),5.1(c), 5.1(d) power-law networks, we can see how the usage of nodes' strength and C-degree did not improve the performance of the navigation process for homophily parameters of 0,1,2 and 3. When the homophily parameter equals 0, the nodes and their neighbors show no correlation. While when the homophily parameter equals 1 or 2 or 3, the nodes show high correlation. Therefore, adding the weights to the edges did not help in the navigation process. Accordingly, we have studied the effect of the weights in the range of homophily parameters that neither is uncorrelated nor highly correlated. More experiments were conducted with varying the homophily parameter between values of 0 and 1.
Figures 5.1(e) and 5.1(f) shows how the usage of the weighs by using nodes' strength outperform the results of the degree alone.



(g) Power law with alpha= 1.5 and 2

(h) Power law with alpha= 3-6

Figure 5.1: ColorMap of the Power Law network SEVN-EVN



(a) Power law with alpha= 1.5 and 2

(b) Power law with alpha= 3-6

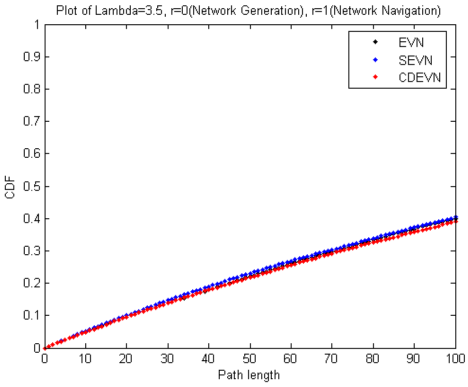
Figure 5.2: ColorMap of the Power Law network CDEVN-EVN

The colorMap figures are obtained as follows:

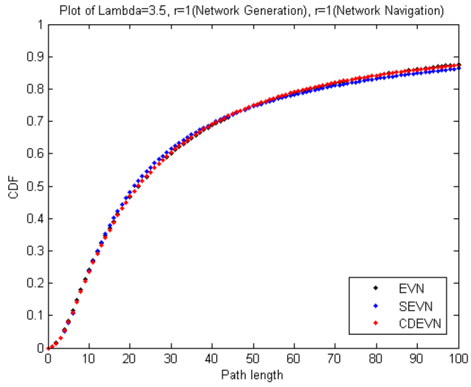
- The area under the ecdf curve for each method using EVN, SEVN and CDEVN is calculated.
- The difference between the obtained values is computed for each combination of degree and homophily parameters as represented by x-axis and y-axis.

The X axis is the homophily paramter, while the Y axis is the degree parameter. Several experiments were conducted using partial value of homophily [0, 0.1, 0.2, 0.8, 1, 2, 4, and 8]. Each color change in the x-axis represents these partial homophily values [0, 0.1, 0.2, 0.8, 1, 2, 4, and 8]. And each color change in the y-axis represents alpha of [1.5, 2, 3, 4, 5 and 6].
We conclude that, to our surprise, there is no clear advantage to taking weights into account for alpha > 2, and in fact can be harmful if alpha ≤ 2.

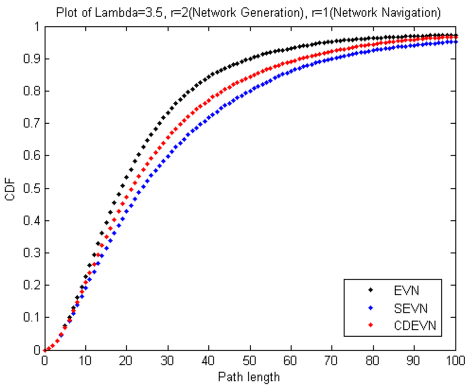
Results of Poisson networks



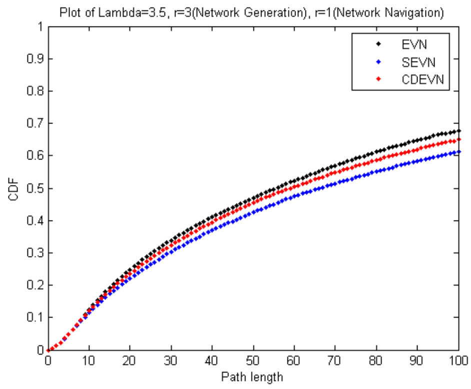
(a) Poisson Network- Lambda=3.5, r=0



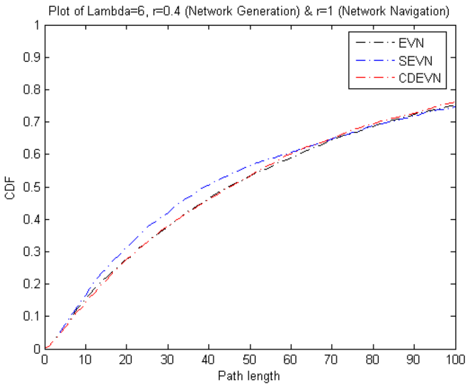
(b) Poisson Network- Lambda=3.5, r=1



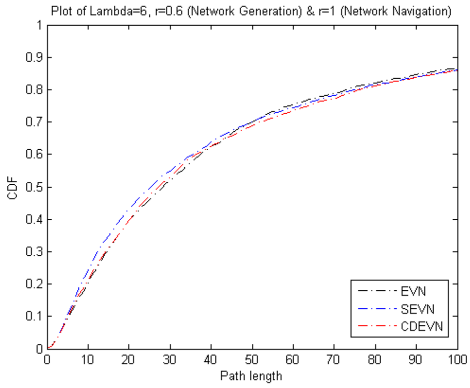
(c) Poisson Network- Lambda=3.5, r=2



(d) Poisson Network- Lambda=3.5, r=3

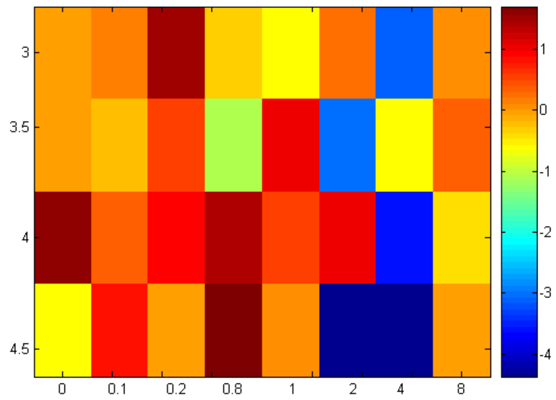


(e) Poisson Network- Lambda=6, r=0.4

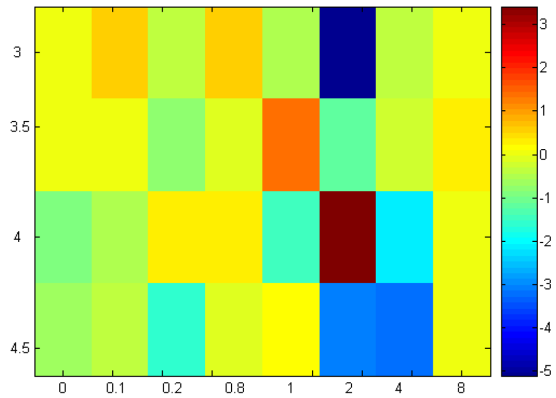


(f) Poisson Network- Lambda=6, r=0.6

Figures 5.3(a), 5.3(b),5.3(c), 5.3(d), 5.3(e) and 5.3(f) represent of the empirical CDF of path length in the conducted experiments of Poisson networks as a function of degree distribution and homophily parameter. As mentioned earlier, Poisson networks have no high degree of nodes and the navigation through the networks depends mainly on the homophily structure. Similar to our observations in the power law networks, using nodes' strength and C-degreen did not improve the performance of the navigation process in Poisson networks for homophily parameter of 0,1,2,3. When the homophily parameter equals 0, the nodes and their neighbors show no correlation. While when the homophily parameter equals 1 or 2 or 3, the nodes show high correlation. Therefore, adding the weights to the edges do not help in the navigation process. Accordingly, we have studied the effect of the weights in the range of homophily parameters that neither are uncorrelated nor highly correlated. More experiments were conducted with varying the homophily parameter between values of 0 and 1.
Figures 5.3(e) and 5.3(f) shows how the usage of the weighs by using nodes' strength outperform the results of the degree alone.



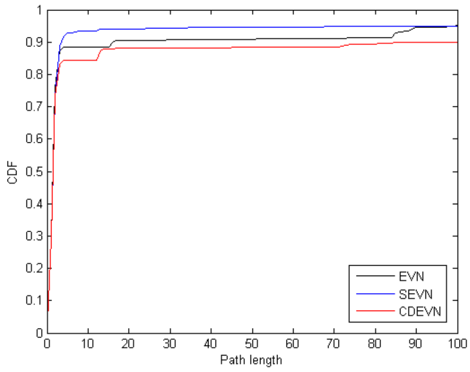
(g) ColorMap of the Poisson network SEVN-EVN



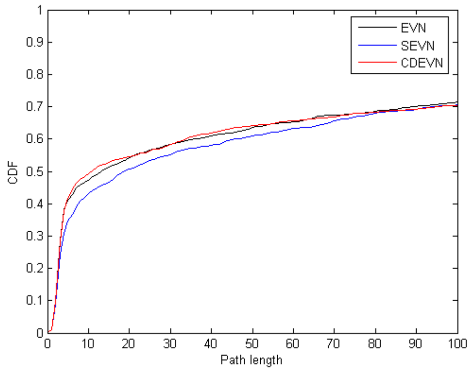
(h) ColorMap of the Poisson network CDEVN-EVN

The X axis is the homophily paramter, while the Y axis is the degree parameter. Several experiments were conducted using partial value of homophily [0, 0.1, 0.2, 0.8, 1, 2, 4, and 8]. Each color change in the x-axis represents these partial homophily values [0, 0.1, 0.2, 0.8, 1, 2, 4, and 8]. And each color change in the y-axis represents alpha of [3, 3.5, 4 and 4.5].
In Poisson networks, SEVN outperforms EVN when the correlation between edges and homophily is not large.

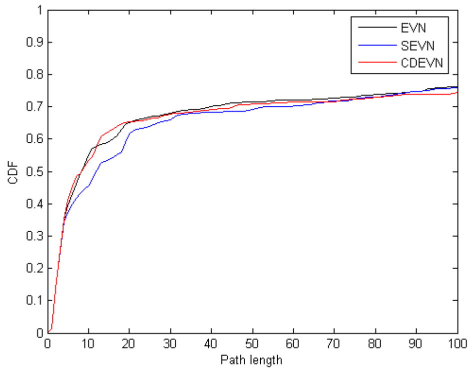
Results of Real-world networks



(i) Webkb-texas-cocite



(j) Industry-yh Network



(k) Webkb-washington-link1 Network

As mentioned earlier, we have used several real-world networks to study and evaluate the performance of our methods. The used real- world networks consist of three different networks. The first used network is industry-yh [13] network that represents 1798 nodes of business news stories collected from the web. The edge connection between two companies is established if they appeared together in a story and the edge weight represent the number of this appearance.
The other two networks are webkb-texas-cocite and webkb-washington-link1. They consist of web pages gathered from computer science departments in Texas and Washington universities. Six categories of: course, department, faculty, project, staff, and student is labeled for each page [13]. Co-citation represents the linking edge between two pages.

It can be observed from figures 5.3(i), 5.3(j) and 5.3(k) that the real- world networks show similar results to the one obtained by the synthetic networks.

Chapter 6

Discussion and Conclusion

In this thesis, we have identified different alternatives for incorporating weights into the EVN algorithm, the state-of-the-art algorithm for (social) network navigation. We have studied the effect of weights in the network navigation and analyzed the inter-play between the homophily, node degree, node strength and node continuous degree. We then have an extensive experimental study of the performance of these alternatives for different degrees, homophily parameters and different types of networks.

The experimental results show that using weights in the generated networks did not improve the navigation process for networks that shows no correlation between neighbors and others that show high correlation. Using a simple product of node's strength/C-degree and homophily measures can effectively guide the local search that outperforms using the node degree in ranges of homophily between 0 and 1 that represents a network with correlated and uncorrelated neighbors. This suggests that more studies should be performed to deeply analyze the effect of homophily in different network structures. The joint use of the node strength and C-degree shows a varying performance. However, whether the joint use of the node strength and C-degree is an important measure and to which level it can increase the performance of navigation, remains an open question.

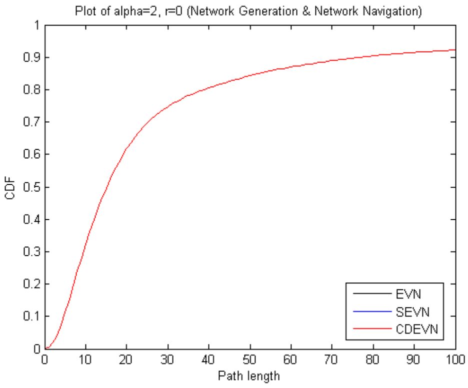
As a future work, we are interested in extending our research to define new weights measures and try other different small world networks to have a wide analysis for the effect of different network measures in the navigation process. We can define new weight measures that take into consideration the distance to the target. We can also test our methods of navigation using more real data sets that can have an effect in increasing the performance of navigation using the weights. We have a great belief that C-degree can perform better in other network structures where nodes show disparity. Therefore, analyzing such networks using C-degree will be an interesting area to be investigated and analyzed.

Appendix A

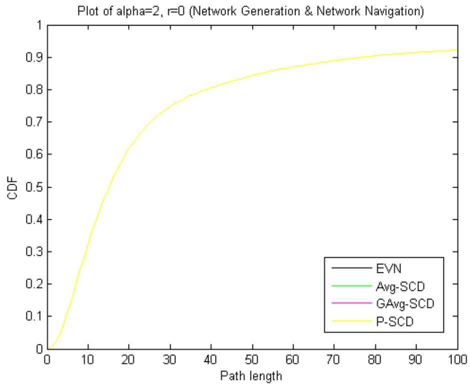
Resulted Figures of the Experiments

In this chapter, all the generated figures for the different generated networks is included. This includes the results power law networks with homophily parameter of 0, 1, 2 and 3 and degree parameter of 2 and 2.5 and Poisson networks with homophily parameter of 0, 1, 2 and 3 and degree parameter of 3, 3.5,4 and 4.5 Description of plots legend:

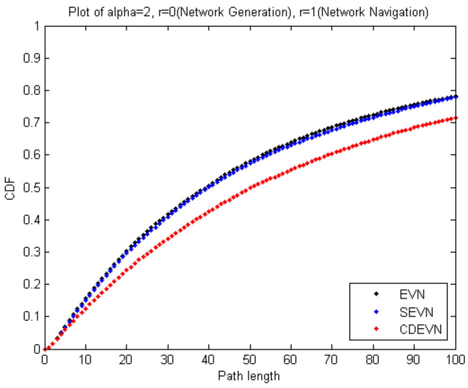
- EVN means implementing the algorithm using the discrete degree.
- SEVN means implementing the algorithm using node's strength.
- CDEVN means implementing the algorithm using C-degree.
- Avg-SCD means implementing the algorithm using the mathmatical average of node's strength and C-degree.
- GAvg-SCD means implementing the algorithm using the geometric average of node's strength and C-degree.
- P-SCD means implementing the algorithm using the product of node's strength and C-degree.



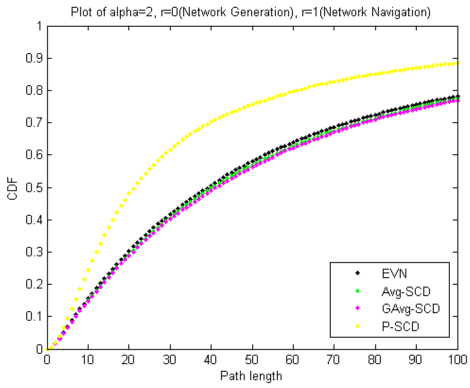
(a) Alpha=2 r=0



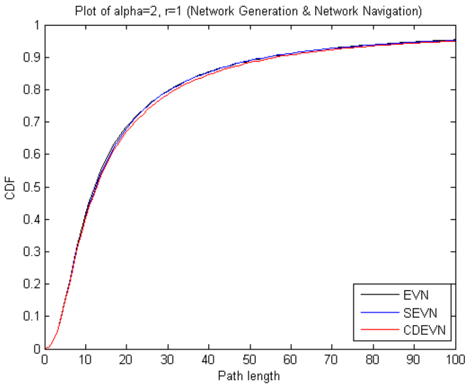
(b) Alpha=2 r=0



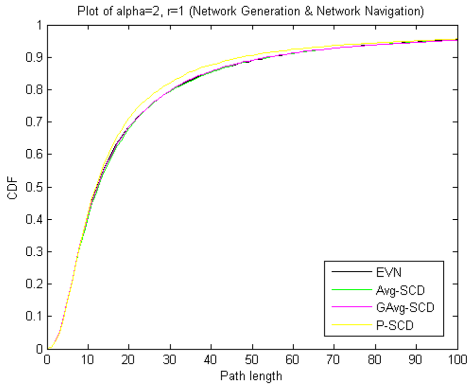
(c) Alpha=2 r=0



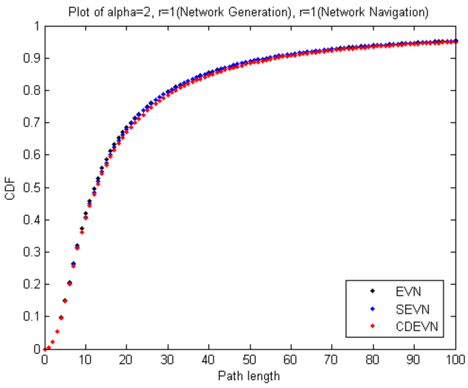
(d) Alpha=2 r=0



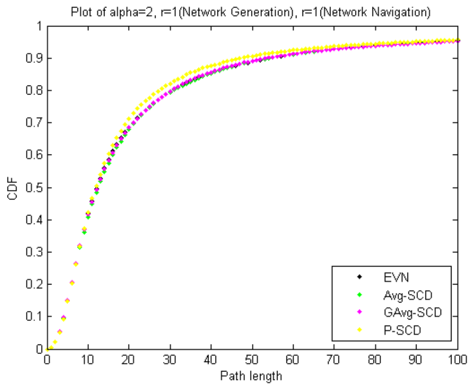
(a) Alpha=2 r=1



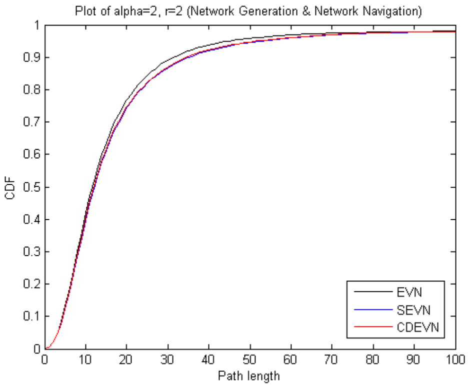
(b) Alpha=2 r=1



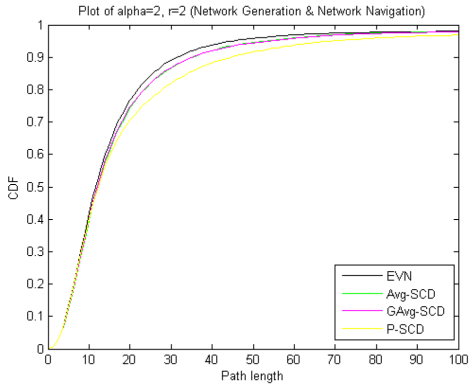
(c) Alpha=2 r=1



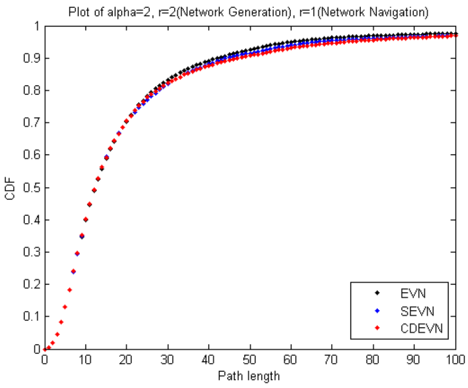
(d) Alpha=2 r=1



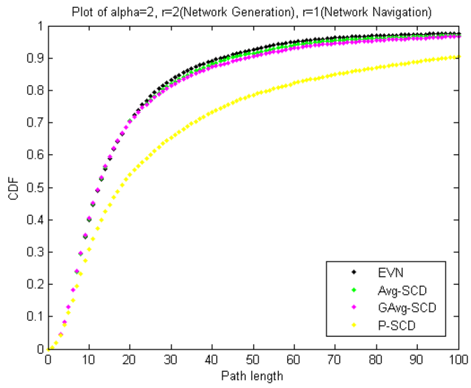
(a) Alpha=2 r=2



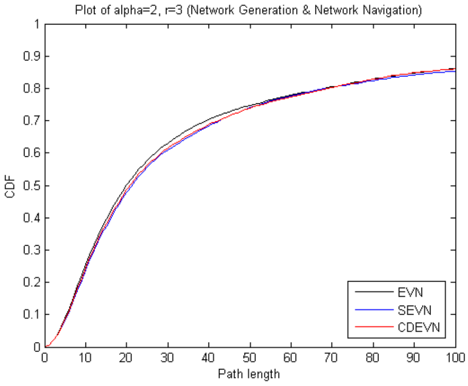
(b) Alpha=2 r=2



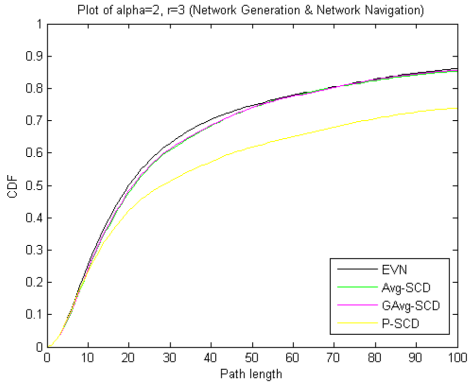
(c) Alpha=2 r=2



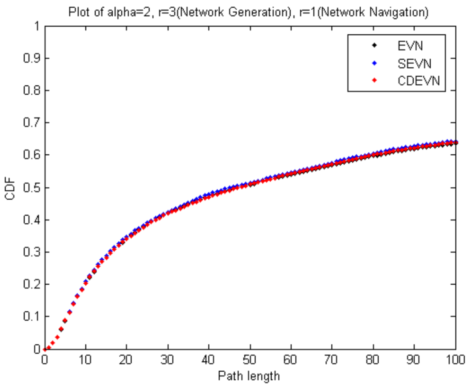
(d) Alpha=2 r=2



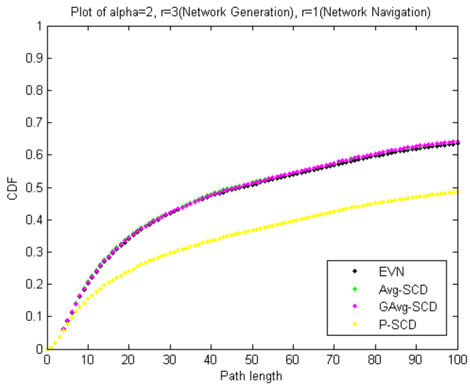
(a) Alpha=2 r=3



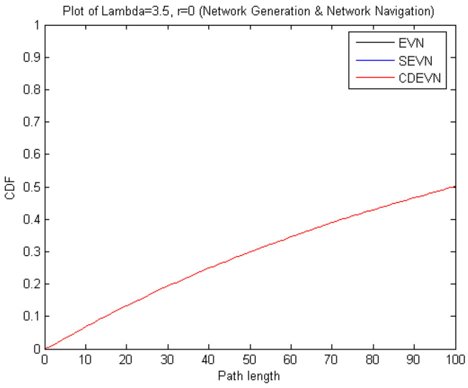
(b) Alpha=2 r=3



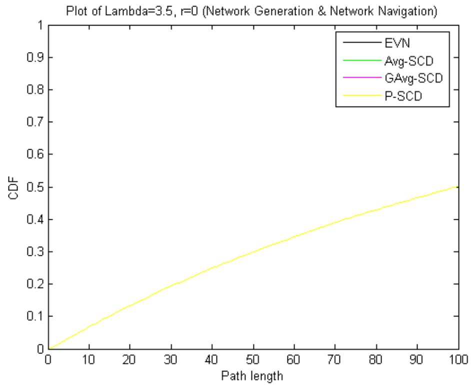
(c) Alpha=2 r=3



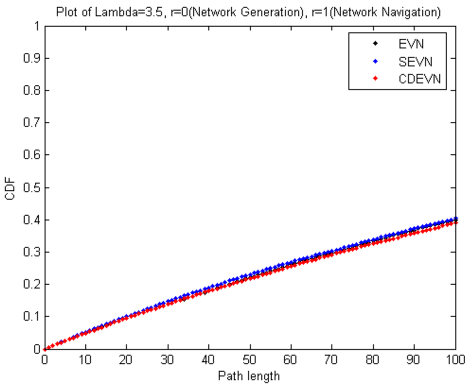
(d) Alpha=2 r=3



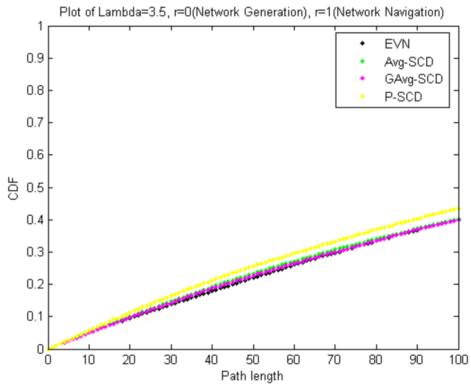
(a) Poisson Network- Lambda=3.5, r=0



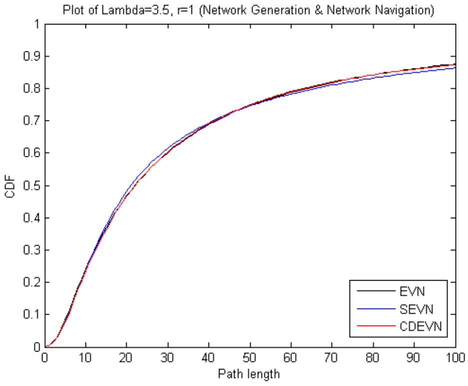
(b) Poisson Network- Lambda=3.5, r=0



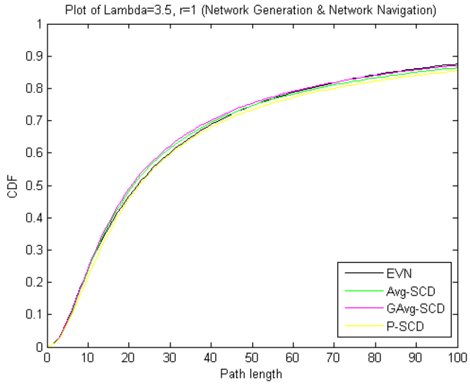
(c) Poisson Network- Lambda=3.5, r=0



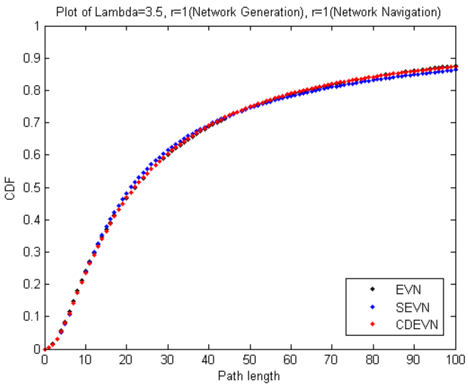
(d) Poisson Network- Lambda=3.5, r=0



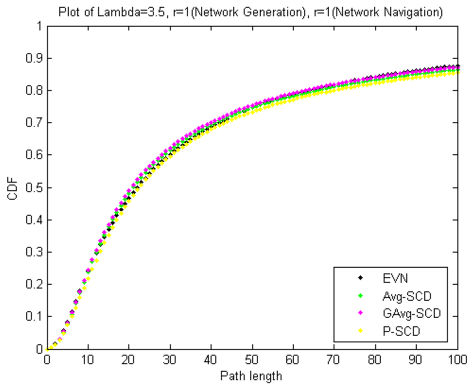
(a) Poisson Network- Lambda=3.5, r=1



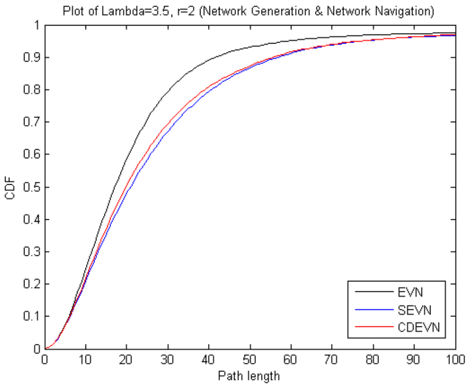
(b) Poisson Network- Lambda=3.5, r=1



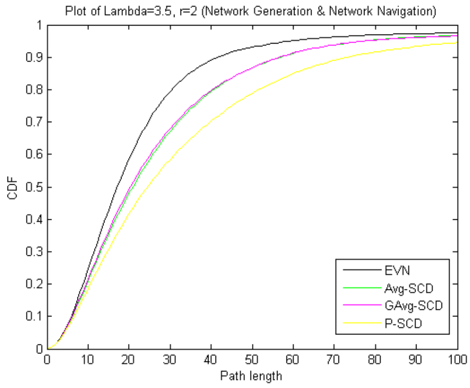
(c) Poisson Network- Lambda=3.5, r=1



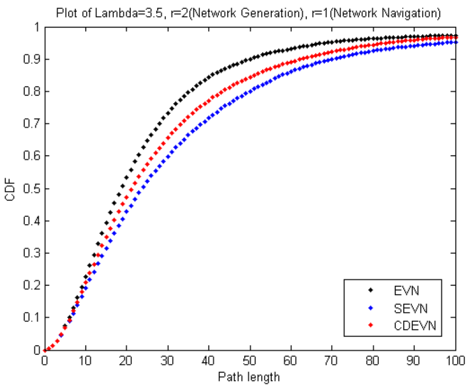
(d) Poisson Network- Lambda=3.5, r=1



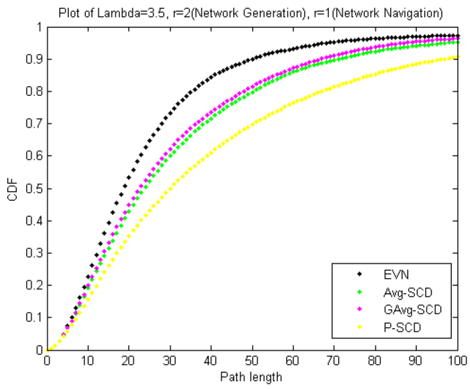
(a) Poisson Network- Lambda=3.5, r=2



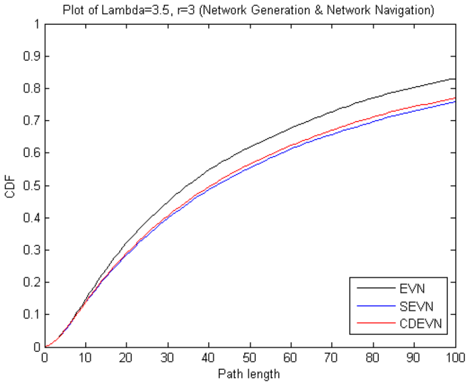
(b) Poisson Network- Lambda=3.5, r=2



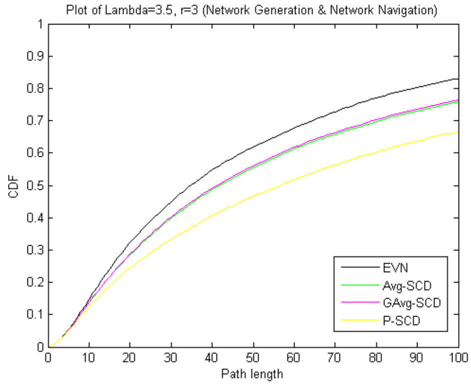
(c) Poisson Network- Lambda=3.5, r=2



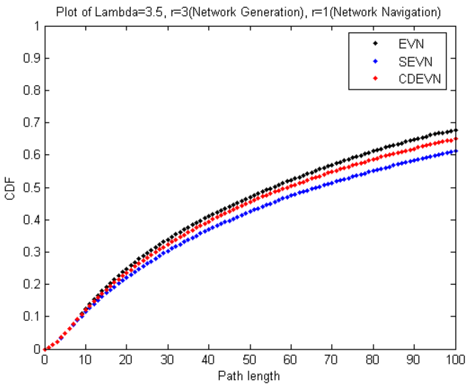
(d) Poisson Network- Lambda=3.5, r=2



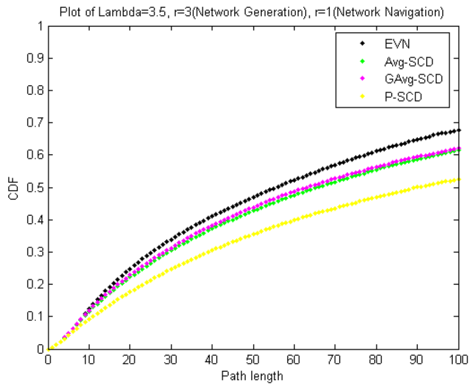
(a) Poisson Network- Lambda=3.5, r=3



(b) Poisson Network- Lambda=3.5, r=3



(c) Poisson Network- Lambda=3.5, r=3



(d) Poisson Network- Lambda=3.5, r=3

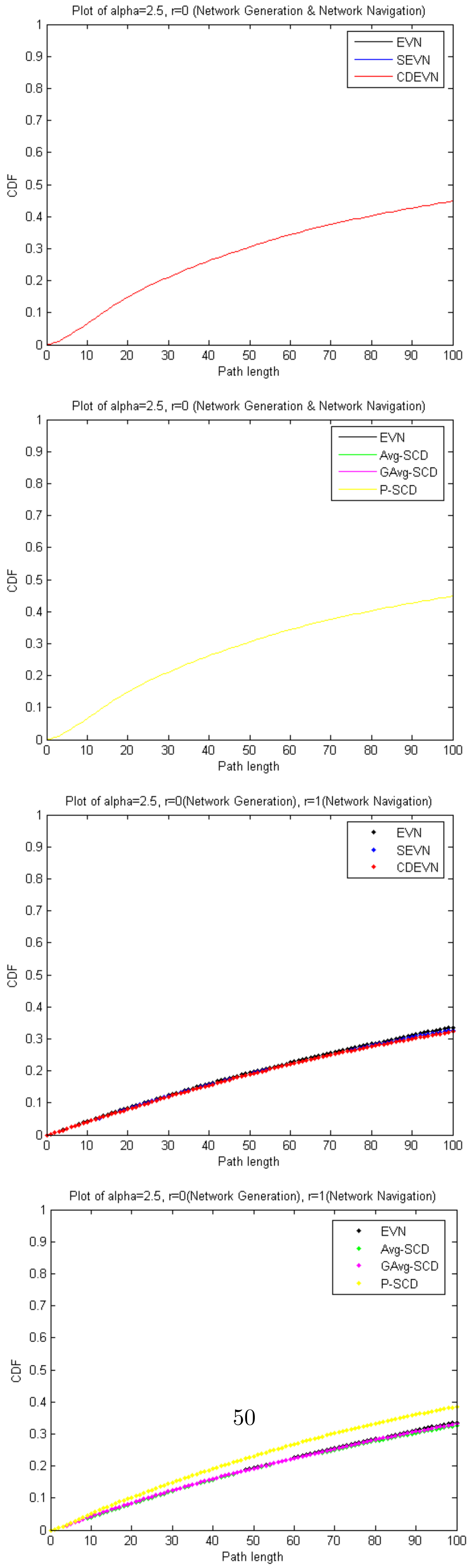


Figure A.1: Power Law Network- alpha=2.5, r=0

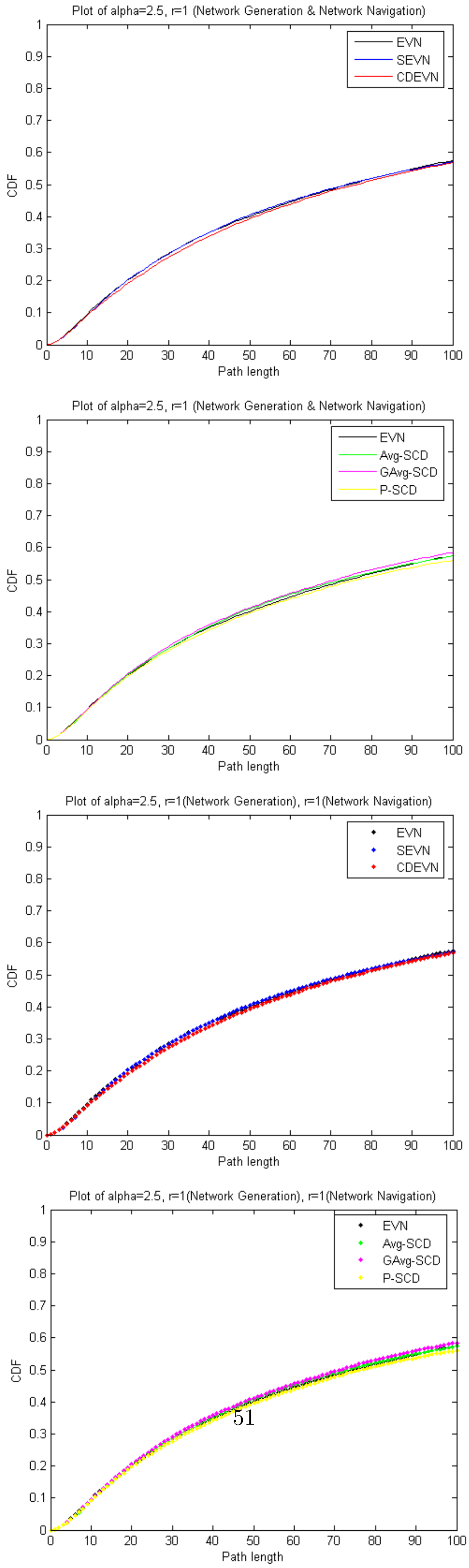


Figure A.2: Power Law Network- alpha=2.5, r=1

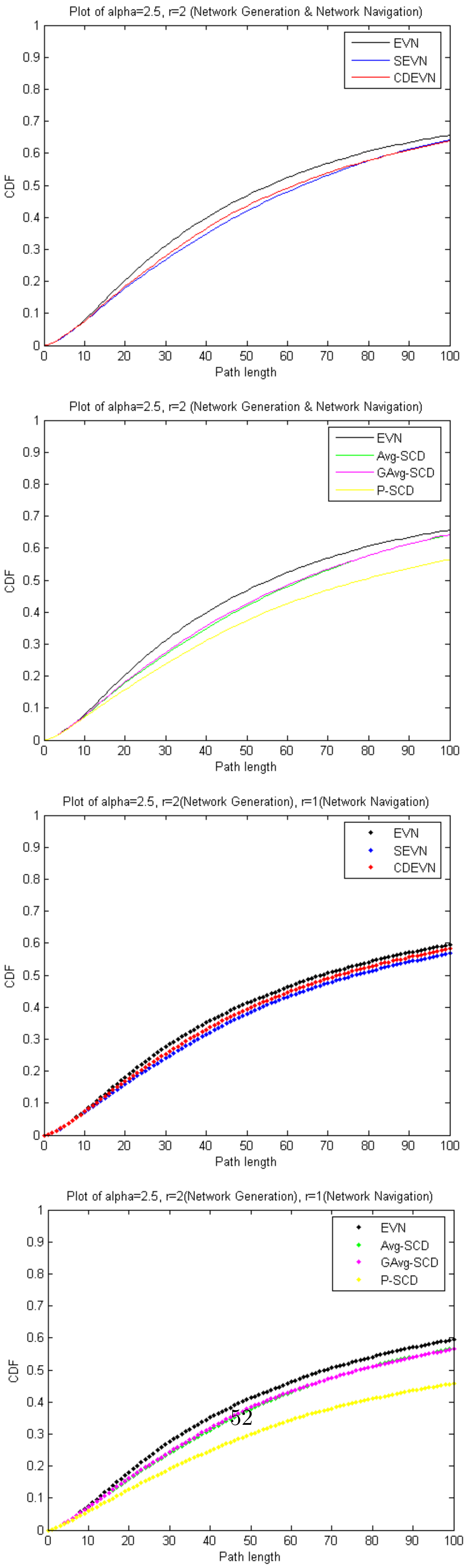


Figure A.3: Power Law Network- alpha=2.5, r=2

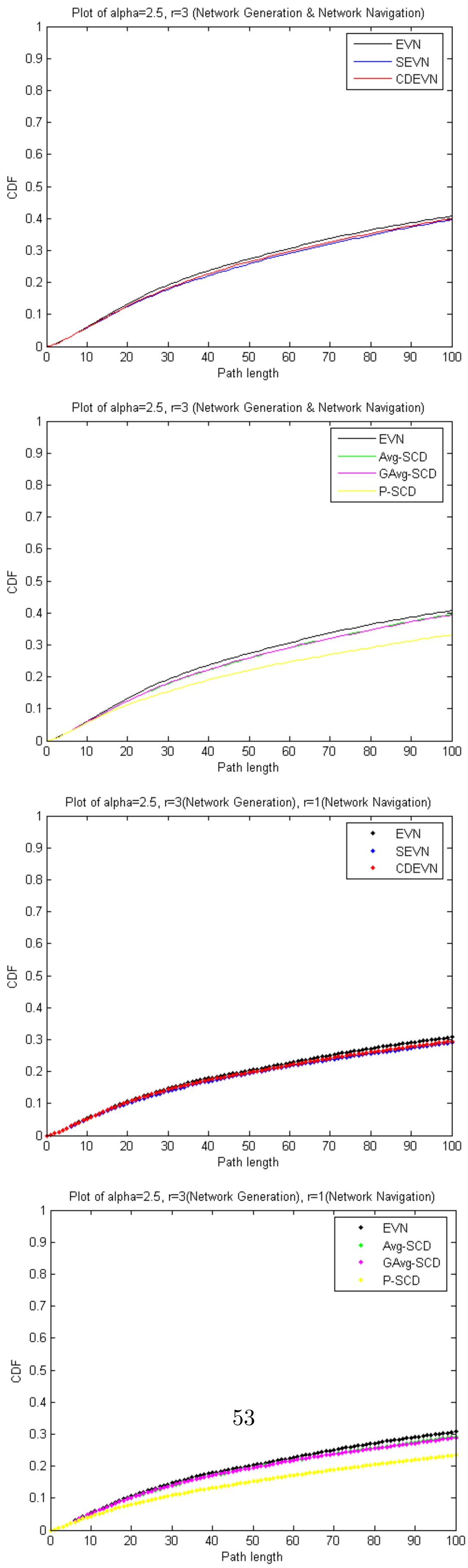


Figure A.4: Power Law Network- alpha=2.5, r=3

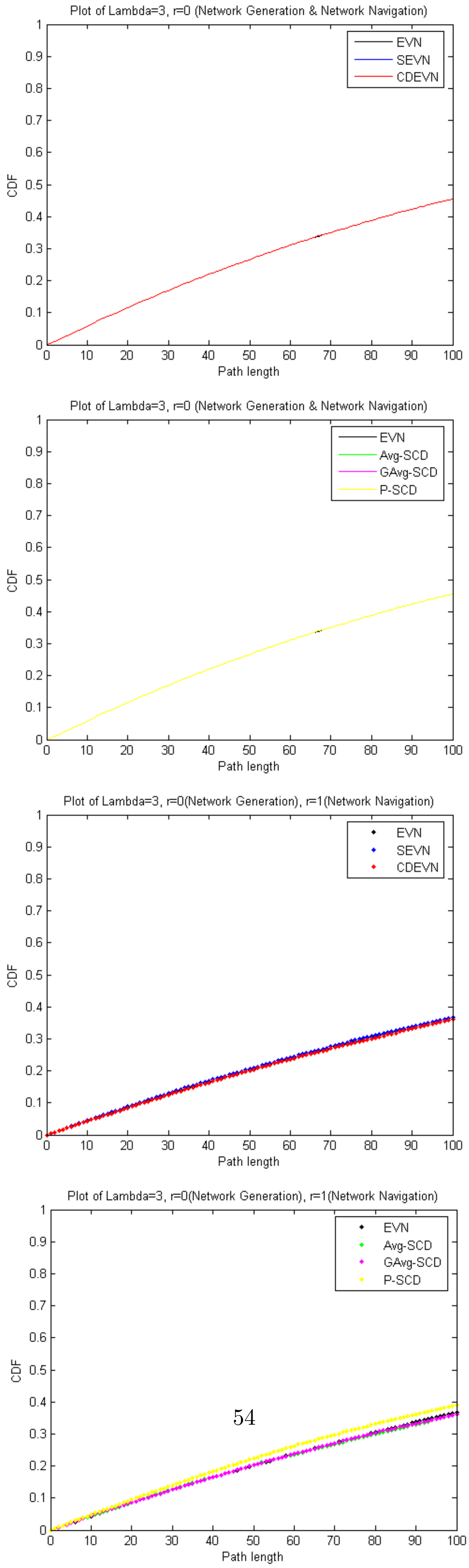


Figure A.5: Poisson Network- Lambda=3, r=0

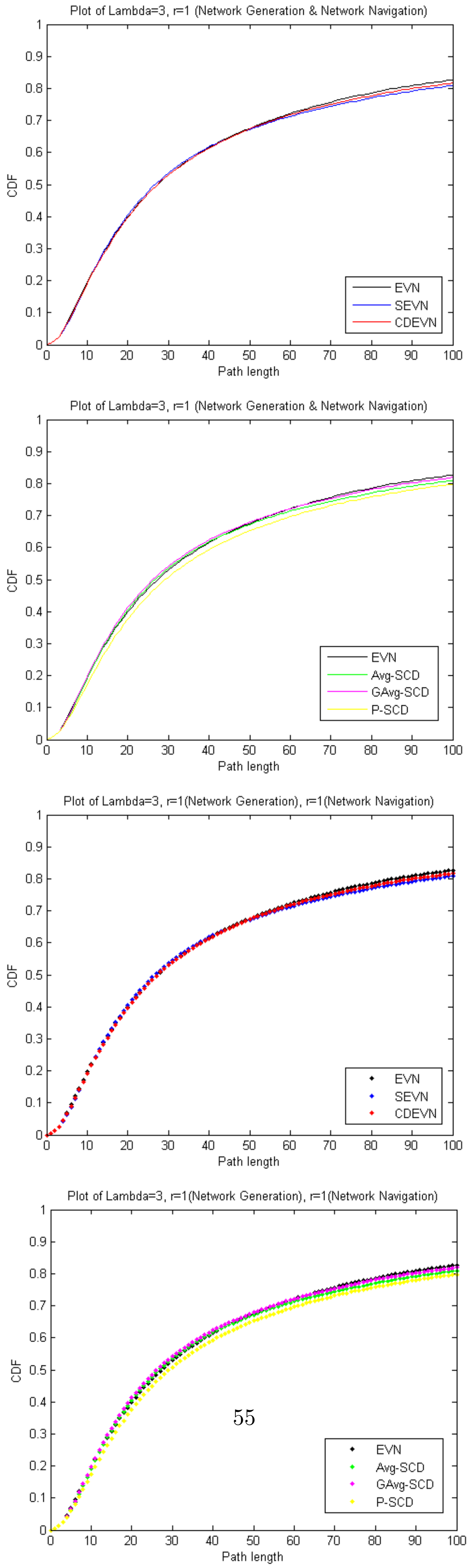


Figure A.6: Poisson Network- Lambda=3, r=1

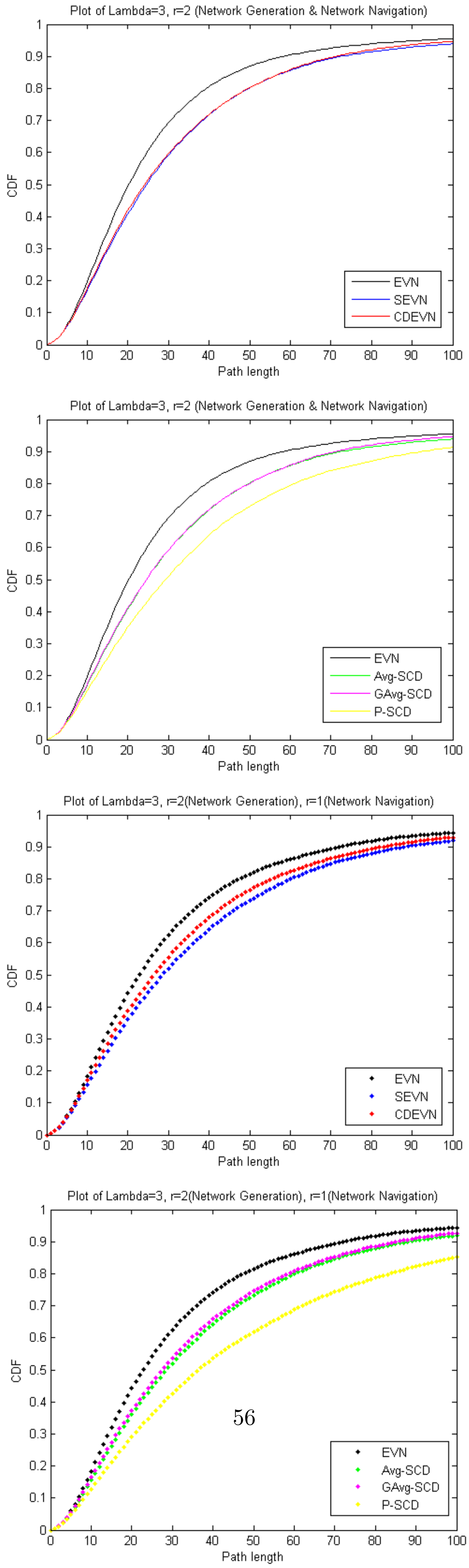


Figure A.7: Poisson Network- Lambda=3, r=2

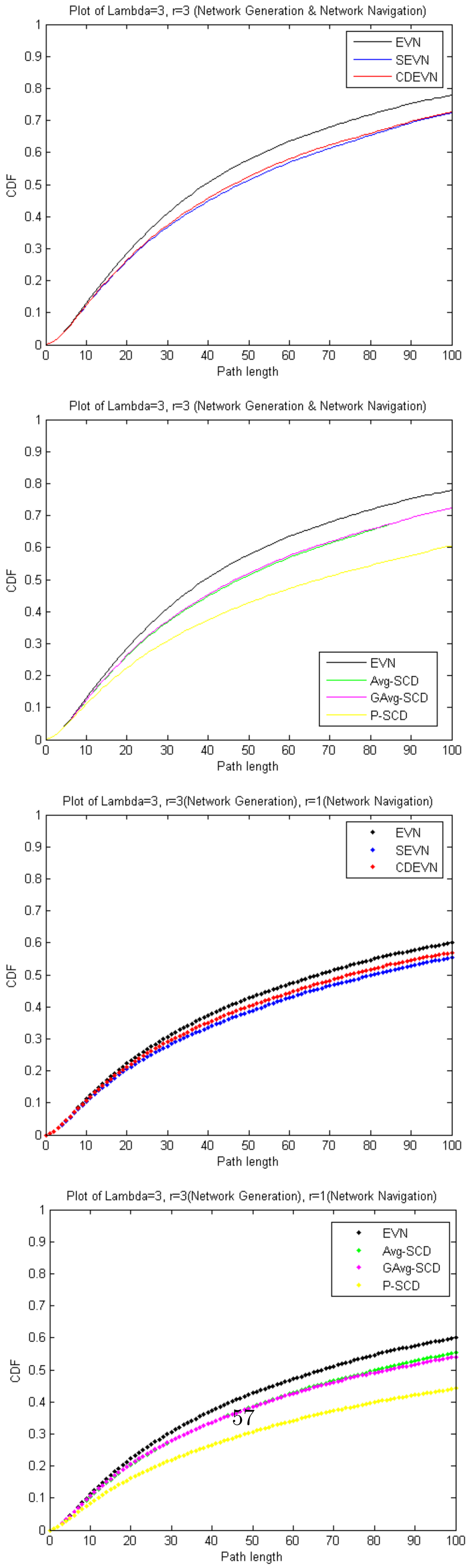


Figure A.8: Poisson Network- Lambda=3, r=3

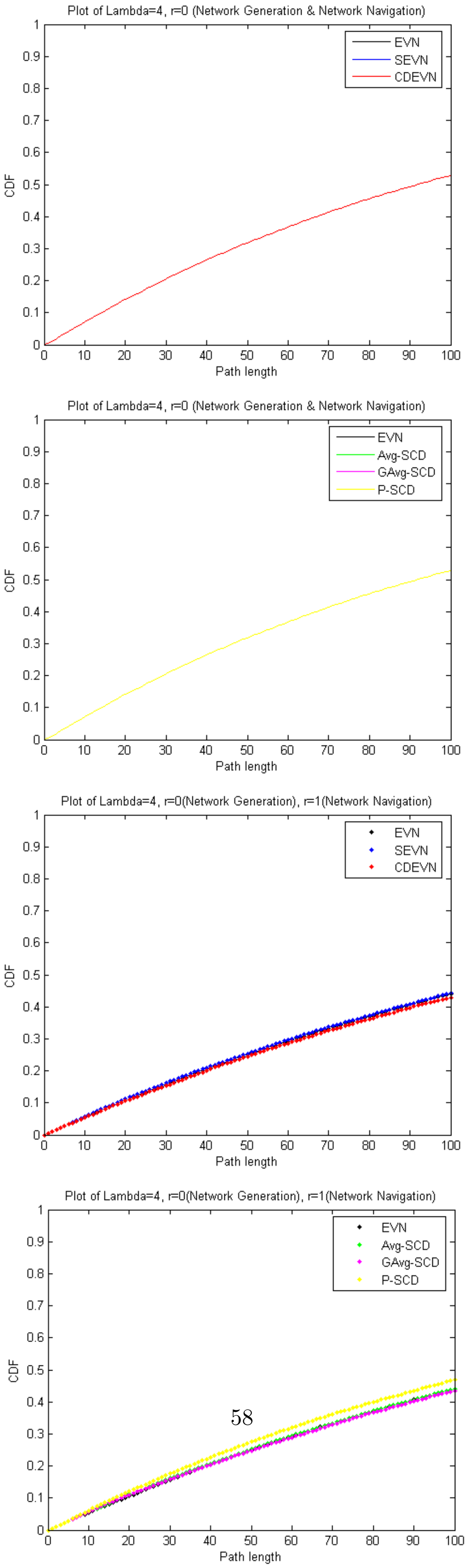


Figure A.9: Poisson Network- Lambda=4, r=0

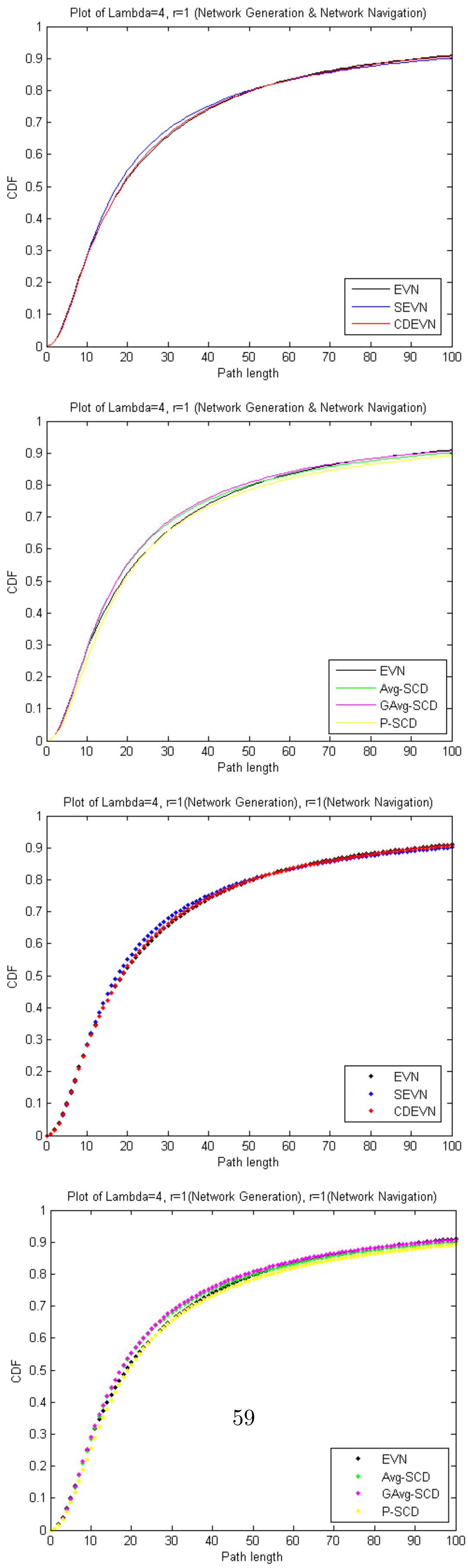


Figure A.10: Poisson Network- Lambda=4, r=1

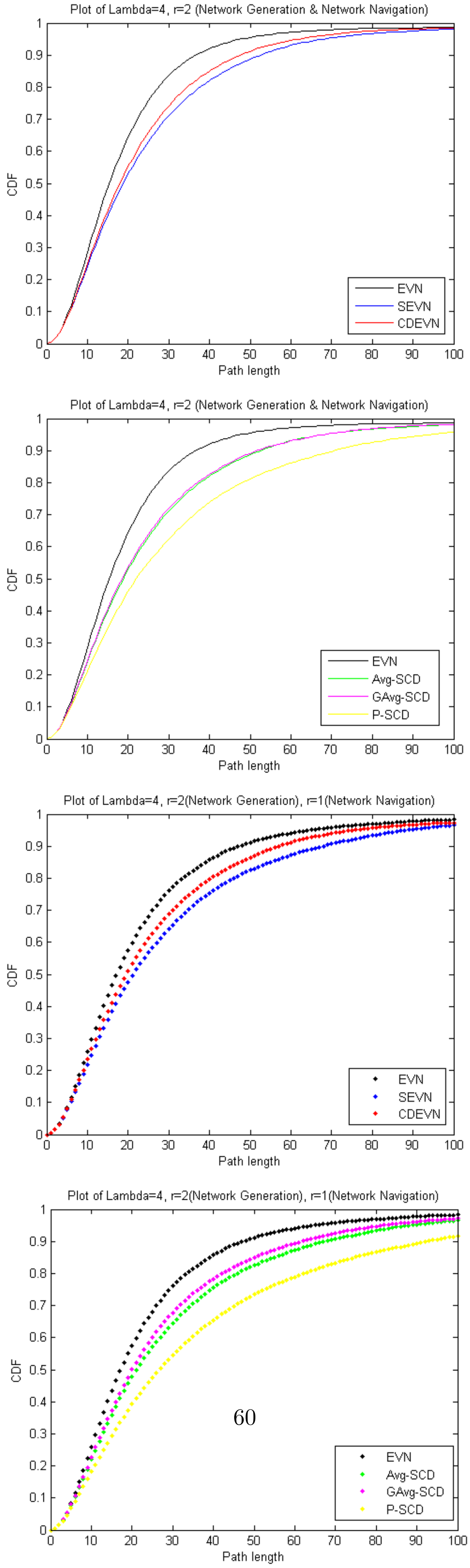


Figure A.11: Poisson Network- Lambda=4, r=2

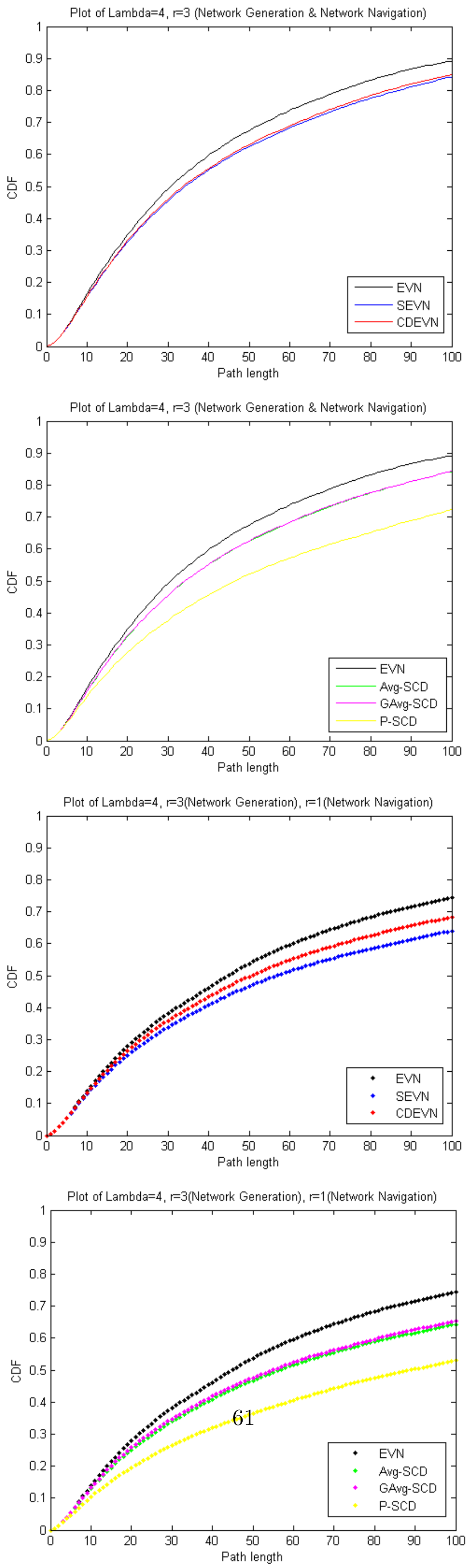


Figure A.12: Poisson Network- Lambda=4, r=3

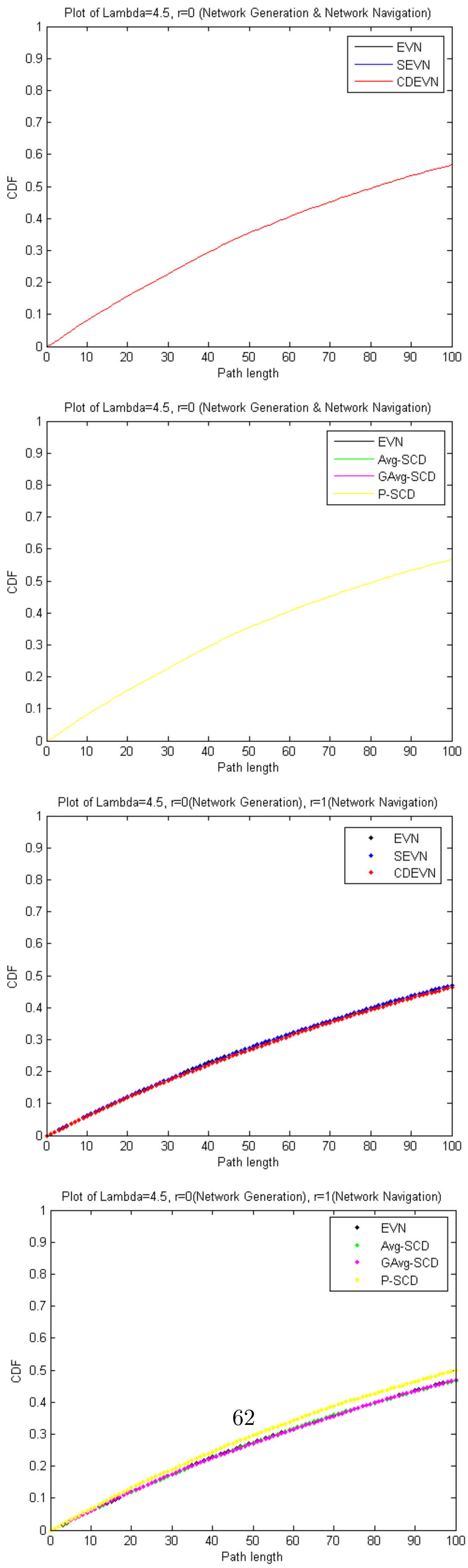
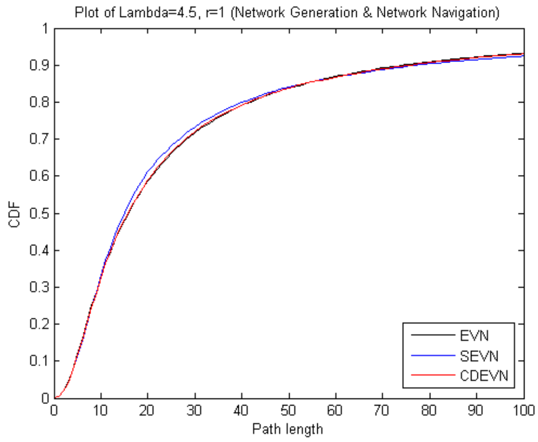
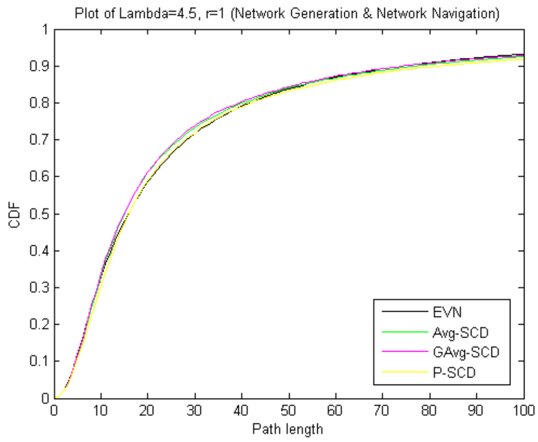
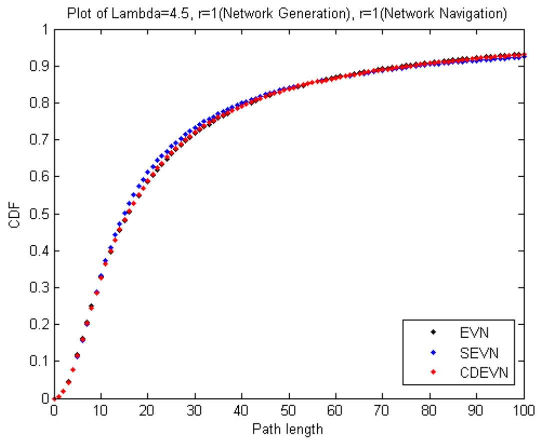


Figure A.13: Poisson Network- Lambda=4.5, r=0







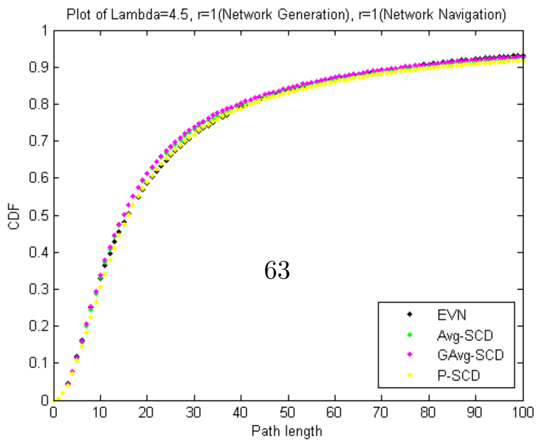


Figure A.14: Poisson Network- Lambda=4.5, r=1

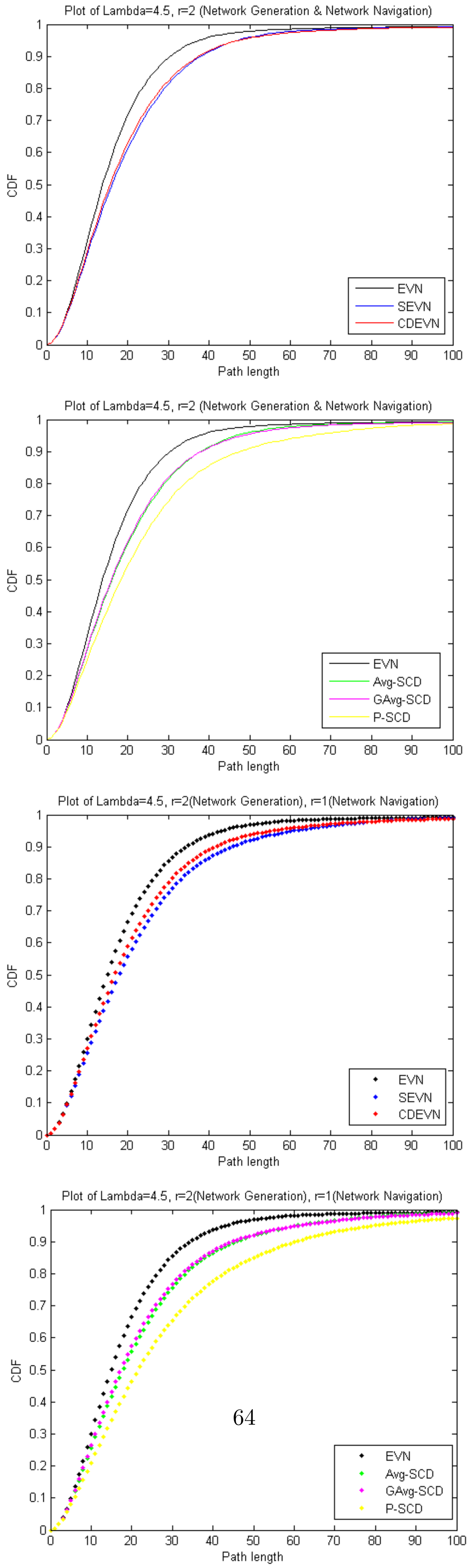


Figure A.15: Poisson Network- Lambda=4.5, r=2

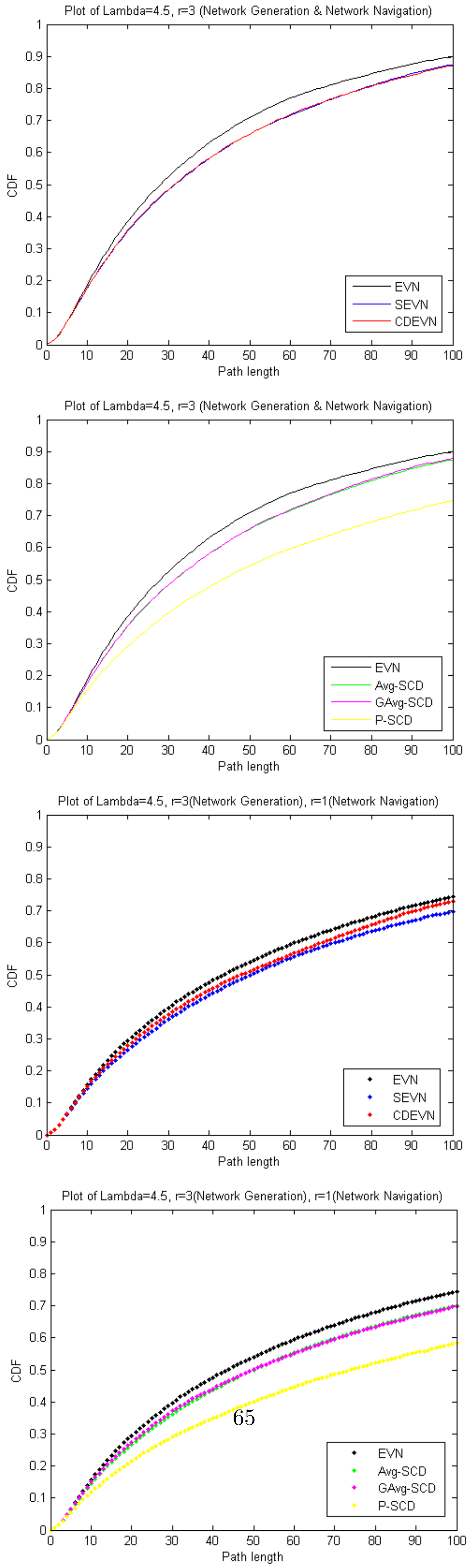


Figure A.16: Poisson Network- Lambda=4.5, r=3

Bibliography

[1] Scale-free network. Online.

[2] W. Aiello, F. Chung, and L. Lu . . In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing* , pages 171–180, New York, NY, USA, 2000.

[3] Abdallah, S. A New Methodology for Generalizing Unweighted Network Measures. *ArXiv e-prints*, may 2009.

[4] Adamic, Lada A. and Lukose, Rajan M. and Puniyani, Amit R. and Huberman, Bernardo A. Search in power-law networks. *Physical Review E*, 64, Sep 2001.

[5] Albert, Réka and Barabási, Albert L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, Jan 2002.

[6] Dodds, Peter S. and Muhamad, Roby and Watts, Duncan J. An Experimental Study of Search in Global Social Networks. *Science*, 301(5634):827–829, August 2003.

[7] Hay Kranen. The Long Tail, as in use by the book of Chris Anderson., 2006.

[8] Kleinberg, J. Small-world phenomena and the dynamics of information, 2001.

[9] Kleinberg, J. M. Navigation in a small world. *Nature*, 406(6798), August 2000.

[10] Laurens van Lieshout. Six degrees of separation, 2007.

[11] Lazega, E. The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership. volume 21. European Sociological Review 2005, 2001.

[12] Leskovec, Jure and Horvitz, Eric. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 915–924, New York, NY, USA, 2008. ACM.

[13] Macskassy, Sofus A. and Provost, Foster. Classification in Networked Data: A Toolkit and a Univariate Case Study. *Machine Learning Research*, 8:935–983, 2007.

[14] Perspective, An A. and Kleinberg, Jon. The Small-World Phenomenon:. In *in Proceedings of the 32nd ACM Symposium on Theory of Computing*, pages 163–170, 2000.

[15] Şimşek, Özgür and Jensen. Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences*, 2008.

[16] Skbkekas. Plot of the probability mass function for the Poisson distribution., 2010.

[17] Travers, Jeffrey and Milgram, Stanley. An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443, 1969.

[18] Watts, D. J. and Dodds, P. S. and Newman, M. E. J. Identity and Search in Social Networks. May 2002.