MSc Project Report

Path Length Distribution in Random Directed Acyclic Graphs

Jens Petersen

Supervisor: Dr. Timothy S. Evans

Imperial College London – September 15, 2014

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Physics and the Diploma of Imperial College London.

The first part of this work will present numerical results for both the path length distribution and the longest path in random discrete intervals of multi-dimensional Cube Space and Minkowski Space. Analytic formulae exist, but it will be shown that only the occurence of short paths can be predicted accurately. However, the longest path can be predicted with reasonable accuracy using either a greedy algorithm or with the knowledge of the scale length $L_{scale} = N^{1/D}$.
The second part will generalise the setting of the first to one where edges are only present with a probability $p < 1$. Analytic formulae give a very poor estimate of the path length distribution, but it will be shown that results for the longest path can be reproduced by introducing a modified scale length $L'_{scale} = (Np)^{1/D}$. The work will conclude with a proposal of how the existing analytic formulae could be modified empirically to reproduce numerical results.

Preface

I'd like to use a couple of lines to put this work into context.
This project was carried out over a period of three months as the final project of my MSc programme at Imperial College London. Naturally in such a short time, it was not possible to tackle the problem in its entirety and the reader will find themselves on several occasions rightfully asking: "Why was this or that not done?" I will try and highlight some of these shortcomings in the final chapter so that others may pick up the ideas for future work.
As a programming language for this work I chose Python. Using C++ or others would have allowed for faster simulations and hence bigger networks, but Python is remarkably accessible, so that, despite my almost non-existent experience with it, it was certainly the most efficient choice. I decided to focus on having working code in order to do simulations rather than making it as fast as possible, and Python seemed ideal for this.
I'd like to point out that while I had some examples to work by all code used in this work was written by me, meaning that not the entire three months were available to produce actual results.

Contents

List of Figures

List of Tables

1. Introduction

In broad terms, a network (or graph) is just a collection of points that are in some way connected. The points are referred to as *vertices* or *nodes* and the connections as *edges*. This concept can be applied to a multitude of problems and has indeed become an important tool in a variety of research fields. The most prominent example of a network is probably the internet, which is both an information network and a technological network and has in turn become a host for other networks. To give an example of the latter and of what the study of networks can be used for, take a look at figure 1.1, which is a visualisation of the author's Facebook network. The vertices represent all the author's contacts and the edges represent a Facebook friendship between two contacts. Clearly there are some groups that are highly interconnected, called *communities*. Community detection in networks is still being investigated very actively [13, 29, 12]. To understand why, take the two highlighted communities in figure 1.1. One represents the author's high school friendships, the other represents connections made during undergraduate studies. From the fact that there are virtually no shared contacts between the two communitites it can be inferred that the author went to a college or university relatively far from home, or at least one where few of his friends chose to go. Clearly there is information contained in the network that can be retrieved by such an analysis. This is not directly relevant to this particular work, but it highlights one important factor of the motivation behind network studies: A network is more than the sum of its parts.

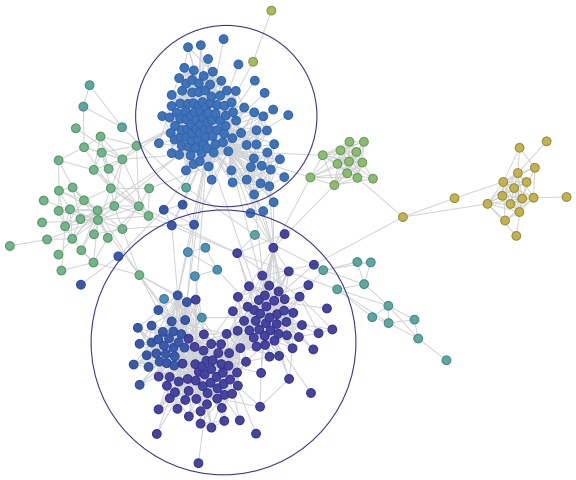


Figure 1.1.: The author's Facebook network with clearly visible community structure, for example connections from high school (top) or undergraduate studies (bottom). These communities show a high level of interconnectedness. Graphic generated using WolframAlpha (www.wolframalpha.com: "facebook report")

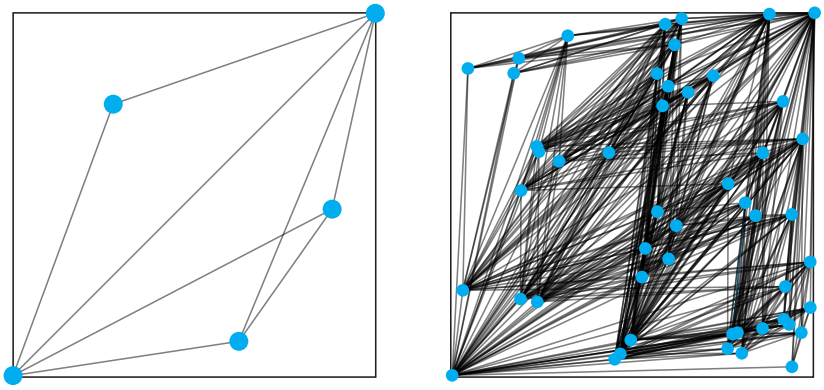


Figure 1.2.: All possible paths from bottom left to top right corner of a square for $N = 5$ and $N = 50$ points. One point connects to the next only if both coordinates increase in the step.

The study of social networks has a comparatively long history [35] and has become part of pop culture in the form of the *six degrees of separation*. The term stems from a 1990 play by John Guare [16, 27] and subsequent film and refers to the idea that even in large networks any two people are connected via at most 5 other people. This notion is also found as *small world network* in the literature following a paper by Milgram [24].
Other important fields that use a network approach include biology, for example food-chains [22] and more recently brain simulations[1] [14], and of course Physics, ranging from Solid State Physics [1] to Particle Physics [26].
What this project sets out to do is, in simple terms, to find all possible paths across the diagonal of a square, when one can only step to certain points along the way. This is illustrated in figure 1.2. A number of points are scattered randomly in the square with bottom left and top right point fixed. Now one starts in the bottom left corner and in each step advances to a point that has higher coordinates in both directions. For $N = 5$ there are only very few possible ways to reach the top right corner, but for $N = 50$ it is already hard to count them manually. The number of possible paths does of course depend on the relative placement of the points.
This concept will be generalised to higher dimensions and special attention will be paid to the longest path, which is of significance in quantum gravity (see section 2.4). It will be tested if the longest path can be approximated by a greedy algorithm.
In a next step it will be investigated how the situation changes when each possible edge is only present with a certain probability $p < 1$. This structure is found in citation networks [28, 6], for example. Imagine an academic paper in a given field, meaning in a given set of papers. If all papers that appeared earlier in time are place in a volume, for example a cone, then the paper at the top of that cone will be connected to some other papers, but certainly not all of them.

[1] The Human Brain Project (www.humanbrainproject.eu)
Blue Brain Project (http://bluebrain.epfl.ch/)

2. Fundamentals

2.1. Network Properties

This section will go into a bit more mathematical detail on networks, called graphs, and introduce some basic properties.
A graph G is an ordered pair $\mathcal{G} = (V, E)$ of a set of vertices V and set of edges E [2]. The elements of E are two-subsets of V, meaning that every edge represents a connection of exactly two vertices. That pair of vertices can either be unordered, giving an *undirected graph*, or ordered, defining a direction and thus giving a *directed graph* (or *digraph*) [27]. Additionally, edges can carry weights, for example the Euclidean distance between the two vertices that is explained in section 2.2. The number of vertices in a graph is called its *order* $|V|$ [9], which is not to be confused with the order also defined in section 2.2. The following will list a number of concepts that are not all directly relevant to this work, but are useful for a better understanding of networks in general. They can be found in any textbook about networks or graph theory, from the more accessible such as Newman [27] to the more mathematical, for example Harary [17] or Diestel [9].

Path/Chain

A path, sometimes also called a chain, is a sequence of consecutively *adjacent* vertices, where adjacent means connected by an edge. The length of a path is the number of edges in the path. In general a path can traverse any edge and vertex multiple times. The path length distribution is just the number of paths of length n as a function of the path length n.

Adjacency matrix

The adjacency matrix is a $|V| \times |V|$ matrix $\mathbf{A}$ whose entries are

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{else} \end{cases} \tag{2.1}$$

where (i, j) can be defined as a connection from i to j or the other way round. Weighted edges can of course have other entries than just 0 and 1. An adjacency matrix has an empty diagonal if there are no *self-edges*, edges to and from the same vertex. Undirected networks have a symmetric adjacency matrix, while directed networks generally don't. The format in which a network is saved on a computer will usually resemble an adjacency matrix.

Degree

The degree of a vertex is the number of edges that are connected to it. For directed graphs a distinction can be made between *in-degree* and *out-degree*. Similar to the path length distribution and much more commonly used is the degree distribution. The latter will play no role in this work, but it might be of interest how the two are related.

2.2. Box Spaces and Intervals

The basis for a box-space is a *measure space*, which is just any space with a defined measure μ [36]. While there are numerous examples, the most intuitive one is the Euclidean Space with the Euclidean Distance as a measure [8]:

$$\mu(x, y) = \sqrt{\sum_{i \in D} (x_i - y_i)^2} \tag{2.2}$$

This is what will be used in this work, because both Cube Space and Minkowski Space – the two will be introduced shortly – can be described as a Euclidean Space.
Now the goal is to only have a discrete set of connected points. Which point may connect to which is defined by a *partial order* $\prec$ [4]. Why not a "full" order? Because not all points need to be related. An example will be given with the definitions of the orders in the two spaces. Applying such a partial order to a measure space M allows one to define an interval $\langle x, y \rangle = \{z \in M \,|\, x \preceq z \preceq y\}$. *All graphs analysed in this work are intervals in either Cube Space or Minkowski Space.* A *box-space* is just such an interval that fulfils $\mu\langle x, y \rangle = \mu(x, y) = 1$. Bollobás & Brightwell [4] additionally demand that the measure space be homogeneous, but the Euclidean Space is, so this requirement shall be of no concern.

2.2.1. Cube Space

The *Cube Space* is simply a D-dimensional cube with edge length 1. This shape follows directly from the order relation, which is [4]:

$$x \prec y \iff x_i < y_i \;\; \forall\, i \in D \qquad \forall\, x, y \in \mathbb{R}^D \tag{2.3}$$

This just means that x is smaller than y in terms of the order $\prec$ if and only if all components of x are smaller than the corresponding component of y. Two points are not related if they have both smaller and larger components than their counterpart. Hence it's a partial order. Now given a point $0 = (0, 0, ..., 0)$ and a point $1 = (1, 1, ..., 1)$ it is easy to understand that all points in the interval $\langle 0, 1 \rangle$ must lie within a cube of edge length 1. This interval is the Cube Space used for this work. Note that it's not a box-space as $\mu\langle 0, 1 \rangle = \sqrt{D}$, but it can of course be normalised if necessary. It was decided to connect points backwards, meaning 1 is the starting point and each point connects to all points of lower order.

2.2.2. Minkowski Space

The Minkowski Space can be thought of as two intersecting cones. Again this is a direct consequence of the order for this space [4]:

$$x \prec y \iff (x_0 - y_0)^2 - \sum_{i=1}^{D} (x_i - y_i)^2 \geq 0 \;\wedge\; x_0 < y_0 \qquad \forall\, x, y \in \mathbb{R} \times \mathbb{R}^{D-1} \tag{2.4}$$

This means two points can only be connected if they lie within each others lightcones – if their spatial distance is smaller than their temporal distance (for convenience the

lightspeed was chosen to be $c = 1$). Such points are said to be *causally connected* [18]. Now take the points $0 = (0, 0, ..., 0)$ and $1 = (1, 0, ..., 0)$, whose interval $\langle 0, 1 \rangle$ will then certainly contain all the points that are in the forward lightcone of the one and in the backward lightcone of the other. This interval is already a box-space and will be used in this work referred to as *Minkowski Space*. Again connections were made backwards, so that 1 is the starting point and every point connects to all points in its backward lightcone.

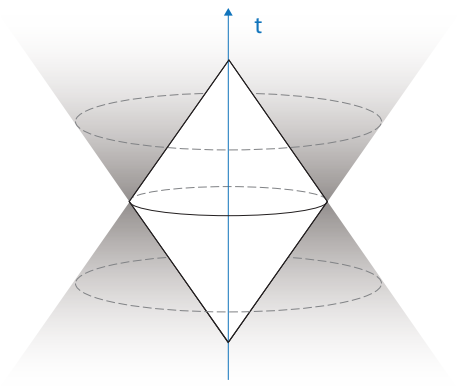


Figure 2.1.: Illustration of $D = 3$ Minkowski Space comprised of two intersecting lightcones

Quite obviously, both connection rules will result in a directed graph without loops (see next section). In fact, the two spaces are identical in two dimensions, which will be shown in chapter 3. In spite of that fact, results will always be presented for both spaces to check that the data is not faulty. One more important quantity has to be introduced: The set of points a given point x *directly* connects to, which will be denoted as $\Omega(x)$. Note that Ω does not include the points x receives connections from, neither does it represent all points it is causally connected to – a causal connection can exist via intermediate points. For complete intervals this distinction is irrelevant, but for incomplete graphs it is of great importance.

2.3. Directed Acyclic Graphs and Completeness

A cycle or loop in directed graphs is a path that contains at least one vertex twice [34], and a directed graph not containing any such cycles is called a *directed acyclic graph*, often abbreviated as *DAG*. All networks analysed in this work are DAGs, as a necessary consequence of the order relation under which they are created.
A *complete graph* is usually defined as one in which there is an edge connecting each pair of vertices [15], but this definition is not very useful for the type of work that will be presented in this report. Instead, whenever this text talks about complete graphs or complete intervals, it will refer to one that contains all edges allowed by the order relations defined in the previous section, which is the same as *transitive completeness* [27]. All others will be called *incomplete graphs*. Note that in an undirected graph of N vertices there will be exactly $\frac{1}{2}N(N-1)$ edges if the graph is complete. For the kind of graphs used in this work a prediction of the exact number of edges is generally not possible, because it depends on the placement of the vertices, which is chosen randomly.

2.4. Longest Path and Greedy Path

The concept of the longest path L is very simple in principle: It is the path in the network that contains the highest number of edges or vertices. The longest path need not be unique, as shown in figure 2.2. For undirected graphs or directed graphs containing cycles the additional limitation has to be imposed that either each edge or each vertex may be passed only once, and the latter is most commonly used definition [27, 9]. Otherwise an infinitely long path could be found. In finite directed acyclic graphs, however, – the ones this work is concerned with – there will automatically be a finite longest path. Finding the longest path in an undirected graph can generally not be done in polynomial time [31], for DAGs on the other hand it is possible [32].

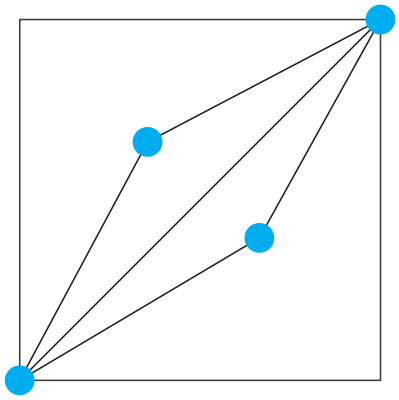


Figure 2.2.: $D = 2$ Cube Space with two longest paths of length 2

The longest path is of significance for quantum gravity, specifically in discrete spacetimes such as the aforementioned Minkowski Space. Myrheim [25] conjectured that the longest path approaches the geodesic of the corresponding continuous spacetime in the limit $\rho \to \infty$, a proof of which exists for flat spacetimes by Brightwell & Gregory [5]. Numerical work for flat spacetimes can be found for example with Rideout & Wallden [30] or Thompson [33], for curved spacetimes with Ilie et al. [19].

Most relevant for this project is the work of Bollobás & Brightwell [4], who established that in the limit $N \to \infty$ the longest path L is proportional to the scale length $L_{scale} = N^{1/D}$ and derived bounds on the proportionality constant m_X, called the *maximal chain constant*. The index X is either Cu_D for the D-dimensional Cube Space or Mi_D for Minkowski Space. As these bounds will be used later in this work, they are given here in detail. The upper bounds, also called the *chain constants* c_X, are:

$$\text{Cube Space:} \quad c_{Cu_D} = e \tag{2.5}$$

$$\text{Minkowski Space:} \quad c_{Mi_D} = e\,\frac{2^{1-1/D}\big(\Gamma(D+1)\big)^{1/D}}{D} \tag{2.6}$$

The lower bounds follow from the upper bounds, so that:

$$c_X \geq m_X \geq \frac{c_X D}{e\big(\Gamma(D+1)\big)^{1/D}\Gamma(1+1/D)} \tag{2.7}$$

It will be tested if these bounds are only valid in the large N limit or if they extend to much smaller N.

The greedy path G is a path that starts at a given point and always proceeds to the point that is closest in terms of a measure μ [3]. For this work the measure will simply be the euclidian distance. The path ends of course when it arrives at a point with no outgoing connections, a *sink*.

The greedy path will only be relevant for the sections concerned with complete graphs, since there would otherwise be multiple possible starting points, or *sources*. For a complete interval the greedy path will always connect the two end points, so that on average it should be irrelevant which one is assigned as the starting point. Note however that for a single network it can well make a difference.
It is possible that the greedy path is a good approximation of the longest path. Figure 2.3 shows two dimensional Minkowski Spaces with increasing N, comparing the top to bottom greedy path with a longest path. Evidently both approximate the diagonal, which is the geodesic in this case, quite well, so the greedy path could be a good predictor of the longest path length. In general greedy algorithms do not give an optimal solution for a given problem [7], but the advantage of such algorithms is that they are extremely simple and fast. Karger et al. [20] show that for undirected dense networks a greedy algorithm can find the longest path. This work will investigate a possible relationship between greedy path and longest path in the given spaces.

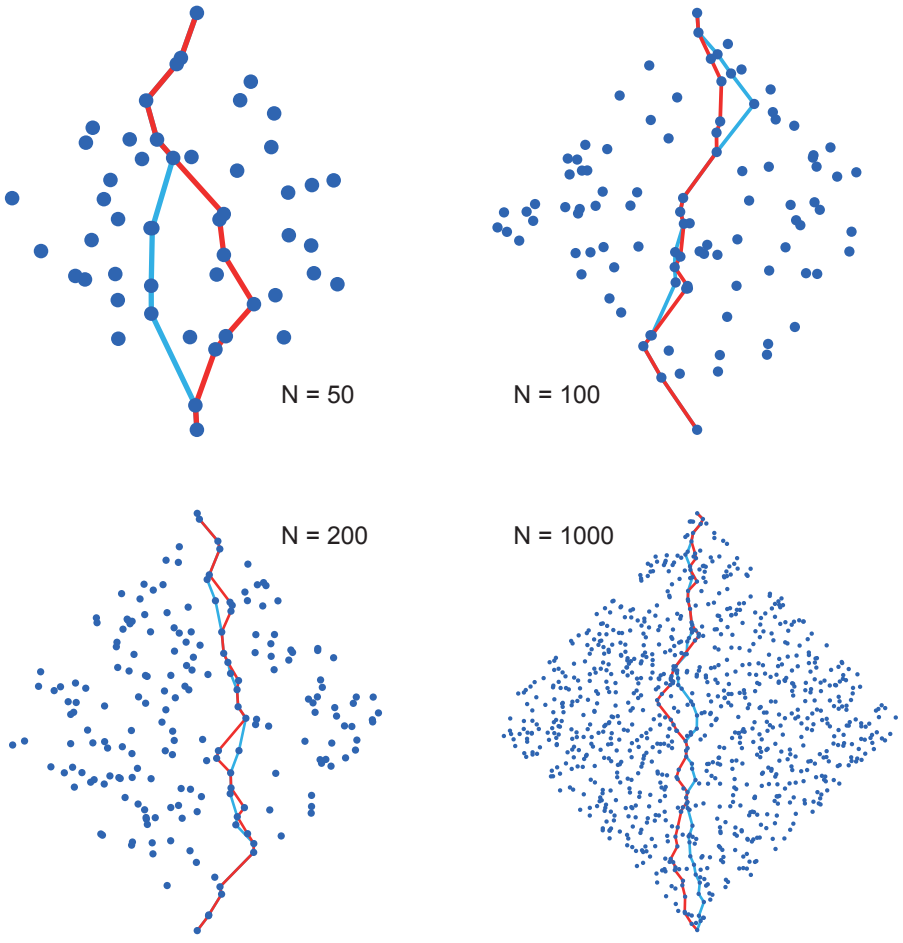


Figure 2.3.: Top-to-bottom greedy path (blue or light grey) and one longest path (red or dark grey) in $D = 2$ Minkowski Space with $N = 50, 100, 200, 1000$. Both approximate the diagonal, also the geodesic in this case, increasingly well. Hence the greedy path could be a good approximation of the longest path.

3. Analytic Groundwork

This chapter will introduce analytic formulae that give the path length distribution – and consequently the longest path – for random intervals in the aforementioned spaces. The following two sections are adapted from an unpublished work of T.S. Evans[1].

Take an interval from 0 to x with all possible edges present, meaning $\Omega(x) = \{y \,|\, 0 \prec y \prec x\}$. The number of paths of length n in that interval will then equal the number of paths of length $n-1$ in the interval from 0 to x', summed over all points x' in the interval $\Omega(x)$:

$$C_n(x) = \sum_{x' \in \Omega(x)} C_{n-1}(x') \tag{3.1}$$

To find an analytic formula for $C_n(x)$ from this recursion relation, one approach is to employ a *mean field approximation*, i.e. to assume a constant point density, and to transfer equations to continuous space. A solution can be found for both Cube and Minkowski Space that solves the resulting integral and also ensures $C_1(x) = 1$, which is the direct connection between 0 and x. Two things should be noted: If equation (3.1) is expressed as a continuous integral, integrating over $\Omega(x)$ and $V(x)$ has the same meaning. And, more importantly, the approximation should be strictly true in the limit $N \to \infty$. To test the validity of the formulae established in the following two sections for *finite* N is part of this work.

3.1. Cube Space

Assuming the number of points $N(x)$ in the interval is large and the points are uniformly distributed, it is possible to integrate over the entire volume between points 0 and x and incorporate the constant[2] point density $\rho = N(x)/V(x)$ into the expression. Equation (3.1) can then be written as:

$$C_n(x) = \int_{\Omega(x)} \mathrm{d}x' \; \rho \cdot C_{n-1}(x') \tag{3.2}$$

where $0, x, x' \in \mathbb{R}^D$. Then take the ansatz

$$C_n(x) = \frac{(N(x))^{n-1}}{((n-1)!)^D} = \frac{(\rho V(x))^{n-1}}{((n-1)!)^D} \tag{3.3}$$

which evidently fulfils $C_1(x) = 1$.

[1] My supervisor's work will not be cited, as it is very much work in progress.

[2] Generally it need not be constant, but in the simulation point coordinates will be taken from a uniform distribution, so the assumption is justified

The recursion relation (3.2) then becomes

$$C_{n+1}(x) = \int_{\Omega(x)} \mathrm{d}x' \, \rho \cdot \frac{(\rho V(x'))^{n-1}}{((n-1)!)^D} \tag{3.4}$$

$$= \rho^n \prod_{i=1}^{D} \left(\int_0^{x_i} \mathrm{d}x_i' \, \frac{(x_i')^{n-1}}{(n-1)!} \right)$$

$$= \rho^n \prod_{i=1}^{D} \left(\frac{x_i^n}{n(n-1)!} \right)$$

$$= \rho^n \prod_{i=1}^{D} \frac{x_i^n}{n!}$$

$$= \frac{(\rho V(x))^n}{(n!)^D} = \frac{(N(x))^n}{(n!)^D} \tag{3.5}$$

as required. The first step makes use of the fact that in Cube Space $V(x) = \prod_{i=1}^{D} x_i$. Equation (3.3) is one of the equations the simulation data will be tested against. The next step is to find the longest path L from the equation. For finite N there is no reason to assume that there is exactly one longest path, instead take that number to be N_{max}. Then

$$N_{max} = C_L(x) = \frac{(N(x))^{L-1}}{((L-1)!)^D} \tag{3.6}$$

To get an estimate of the longest path imagine one were to take the average of infinitely many random DAGs with N points. n would then become a continuous variable and it would be reasonable to take the average longest path where $C_L(x) = 1$, so that:

$$1 = \frac{\left[(N(x))^{1/D}\right]^{L-1}}{\Gamma(L)} \tag{3.7}$$

$$\Gamma(L) = (L_{scale})^{L-1} \tag{3.8}$$

In this approximation L only depends on a single variable, the scale length $L_{scale} = N^{1/D}$. One can make use of the Stirling formula [10] to solve for L:

$$L\ln(L) - L - \frac{1}{2}\ln\left(\frac{L}{2\pi}\right) + O(L^{-1}) \approx (L-1)\ln(L_{scale}) \tag{3.9}$$

Now keep only the terms of orders $O(L\ln(L))$, $O(L)$ and take L to be sufficiently large so that $L - 1 \approx L$. Then equation (3.9) reduces to:

$$L \approx e\, L_{scale} \tag{3.10}$$

or more accurately:

$$L \approx e\, L_{scale} - \frac{1}{2}\ln\left(\frac{2\pi L_{scale}}{e}\right) \tag{3.11}$$

For a derivation of the correction in equation (3.11) see appendix A.1[3]. Interestingly, the derived slope is in first approximation, which is only strictly valid for $N \to \infty$, the same as the upper bound introduced in section 2.4.

[3]Not part of Evans' work

3.2. Minkowski Space

The approach for Minkowski Space is very similar to the one for Cube Space, indeed the recursion relation is the same:

$$C_n(x) = \int\limits_{\Omega(x)} \mathrm{d}x' \; \rho \cdot C_{n-1}(x') \tag{3.12}$$

only that here $x, x' \in \mathbb{R} \times \mathbb{R}^{D-1}$, meaning that the first component is a time $x_0 = \tau$ and the $d = D - 1$ spatial components form a d-dimensional spherical subspace. Hence one must integrate over time and for each time integrate over the subspace. It is useful to remember that for spherical symmetry in any dimension d [11]:

$$V_d(r) = \int \mathrm{d}r \; S_d \cdot r^{d-1} \tag{3.13}$$

where S_d is the unit surface in a d-dimensional space. Note that in the literature one often finds the index being the dimension of the surface itself [11] rather than the dimension of the space it resides in. For example, a circle would then have dimension 1, while in this work it will be 2.
Rewriting equation (3.12) with separate time and space integrals:

$$C_n(x) = \int\limits_{\Omega_\tau(x),\, \underline{\Omega}(t')} \mathrm{d}t'\mathrm{d}r' \; S_d \cdot (r')^{d-1} \cdot \rho \cdot C_{n-1}(x') \tag{3.14}$$

where $\Omega_\tau(x)$ is the time interval and $\underline{\Omega}(t')$ the subspace for each time t'. In most cases it is convenient to just choose $\Omega_\tau(x) = [0, \tau]$.
To find an ansatz for $C_n(x)$ first note that for fixed speed of light $c = 1$ the volume of the space is proportional to τ^D. Similar to the Cube Space assume that:

$$C_n(x) \propto (\rho V)^{n-1} \propto \left(\rho \tau^D\right)^{n-1} \tag{3.15}$$

$$\implies C_n(x) = A_n(D) \left(\rho \tau^D\right)^{n-1} \tag{3.16}$$

$A_n(D)$ is a proportionality constant that should only depend on n and D – as it does for the Cube Space. The idea is now to plug this ansatz into equation (3.14) and get a recursion relation for $A_n(D)$. Together with the condition that $A_1(D) = 1$, which is dictated by the need for $C_1(x) = 1$, one arrives at an expression for $A_n(D)$ and hence for $C_n(x)$ in Minkowski Space.
The derivation is rather lengthy, so only the result is given here, which is:

$$C_n(x) = \left(\frac{S_d \Gamma(d) \rho \tau^D}{2^D}\right)^{n-1} \frac{n\Gamma\left((D/2)+1\right)}{\Gamma\left(D(n-1)/2+1\right)\Gamma(Dn/2+1)} \tag{3.17}$$

For simulation purposes it is more useful to have N as a variable:

$$\begin{aligned} N(x) &= \int\limits_{\Omega(x)} \mathrm{d}x' \; \rho = \int\limits_{\Omega_\tau(x),\, \underline{\Omega}(t')} \mathrm{d}t'\mathrm{d}r' \; \rho \cdot S_d \cdot (r')^{d-1} \\ &= 2\rho S_d \int\limits_0^{\tau/2} \mathrm{d}t' \int\limits_0^{t'} \mathrm{d}r' \; (r')^{d-1} \\ N(x) &= \rho \tau^D \frac{S_d}{2^d D(D-1)} \end{aligned} \tag{3.18}$$

Equation (3.17) then becomes:

$$C_n(x) = \left(\frac{N(x)\Gamma(D+1)}{2}\right)^{n-1} \frac{\Gamma(D/2)}{\Gamma\left(D(n-1)/2+1\right)\Gamma(Dn/2)} \tag{3.19}$$

This is the equation data from Minkowski Space simulations will be compared to. As mentioned in section 2.2, the distributions for Cube Space and Minkowski Space should be the same in two dimensions, and indeed equation (3.3) and equation (3.19) are identical for $D = 2$.
Equation (3.19) can also be found with Myrheim and Meyer [25, 23] and can be used to find an expression for the longest path L. Take the logarithm and make use of the Stirling approximation:

$$\begin{aligned}\ln\left(C_L(x)\right) &= (L-1)\ln\left(\frac{N(x)\Gamma(D+1)}{2}\right) \\ &\quad + \ln\left(\Gamma\left(\frac{D}{2}\right)\right) - \ln\left(\Gamma\left(\frac{D(L-1)}{2}+1\right)\right) - \ln\left(\Gamma\left(\frac{DL}{2}\right)\right) \qquad (3.20)\\ &\approx L\ln\left(\frac{N(x)\Gamma(D+1)}{2}\right) - DL\ln\left(\frac{DL}{2}\right) + DL \qquad (3.21)\end{aligned}$$

where only terms of orders $O(L), O(L\ln(L))$ and $O(L\ln(N))$ were kept. Then:

$$\begin{aligned}\ln\left(C_L(x)\right) &\approx DL\ln\left(\left(N(x)\right)^{1/D}\left(\frac{\Gamma(D+1)}{2}\right)^{1/D}\frac{2e}{DL}\right) \qquad (3.22)\\ &\approx DL\ln\left(\frac{c_{Co(D)}L_{scale}}{L}\right) \qquad (3.23)\end{aligned}$$

where $L_{scale} = N^{1/D}$ and $c_{Co(D)}$ is the chain constant for the Minkowski Space as introduced by Bollobás & Brightwell [4]. Now again make the assumption that $C_L(x) = 1$ and it follows immediately that.

$$L \approx c_{Co(D)}\ L_{scale} \tag{3.24}$$

So, as seen before for Cube Space, in the limit of large N the longest path should scale linearly with L_{scale} by a proportionality constant that Bollobás & Brightwell [4] established is the upper bound on their maximal chain constant (see section 2.4).

3.3. Incomplete Graphs

One question this work will seek to answer is what happens if edges are only present with a certain probability. The concept is the same as outlined before, but now point x is not automatically connected to all points in $\langle 0, x\rangle$. Instead, for each $x' \in \langle 0, x\rangle$ a connection will only be present with probability $p < 1$. J. Clough[4] conjectures that in such a situation it is sufficient to take the established formulae and just replace N with Np, which is reasonable, because the number of points a given point connects to should be proportional to N. For example, the origin will always connect to (or receive a connection from) all other points, while in Cube Space a point in the centre of the interval

[4]Special thanks to my colleague James Clough for this suggestion

will be able to connect to 2^{-D} of all points on average. Including an edge probability this proportionality should simply be in relation to Np instead of N. Accordingly, there would be a new scale length $L_{scale} = (Np)^{1/D}$ and the path length distributions would be given by:

$$\text{Cube Space:} \qquad C_n(x) = \frac{\big(p\,N(x)\big)^{n-1}}{\big((n-1)!\big)^{D}} \tag{3.25}$$

$$\text{Minkowski Space:} \qquad C_n(x) = \left(\frac{p\,N(x)\,\Gamma(D+1)}{2}\right)^{n-1} \frac{\Gamma(D/2)}{\Gamma\left(D(n-1)/2+1\right)\Gamma(Dn/2)} \tag{3.26}$$

Quite obviously, these equations fulfil the most basic requirement of being consistent with the $p = 1$ case. Note however that these equations still predict exactly one path of length 1, which seems very unrealistic, although it will of course depend on how one defines a path in this new setting. One could only count those that actually connect the bounds of the interval – which would by the way mean an average of p paths of length 1 – but for this work it was decided to count all existing paths in the interval.

4. Methods & Algorithms

4.1. Path Length Distribution Algorithm

This section will delineate the algorithm used to determine the path length distribution for a given DAG. Information about the latter was stored in a way that for each node or vertex a list of nodes it connects to is provided. The steps were as follows:

1. Determine the order of each node. The order of a sink, i.e. a node with empty connection list, was defined to be zero: $o(x) = 0 \Leftrightarrow \Omega(x) = \{\}$, while the order of any other node is one more than the highest order of the nodes it connects to: $o(x) = \max(o(x') \,|\, x' \in \Omega(x)) + 1$. This process is similar to topological sorting [7], with the difference that multiple nodes can have the same order in this case.

2. Looping through orders starting at 1, calculate the path length distribution for all nodes of each order. For first order nodes, the path length distribution is simply $C_1(x) = |\Omega(x)|$ the magnitude of its $\Omega(x)$. For higher order nodes the distribution is obtained via equation (3.1). The increasing order ensures that when the distribution is to be computed for one node, the distributions for all nodes it connects to are already available.

3. The final path length distribution is the sum of the distributions of the highest order nodes. For complete intervals there is of course exactly one node with highest order.

For too large N the path count becomes too high to be saved as integer value. To prevent that from happening, everything was saved and computed in logarithms of base 10. It is useful to remember that

$$\log(a+b) = \log(a) + \log\left(1 + 10^{\log(b)-\log(a)}\right) \tag{4.1}$$

The algorithm is generally faster the more sparse the networks become for a given N. To give an estimate of its efficiency, the computation time for complete networks in two dimensions is roughly proportional to a power of N: $t \propto N^{2.55}$ (timing analysis was performed up to $N = 2000$)

4.2. Measurement Statistics

Because this work deals with random networks, every simulation has to be repeated a number of times to then average the results, and it was found that 500 is a reasonable number of repititions. To quantify the quality of any statistical measurement, it is necessary to provide an error, and the most common error measure is the *standard deviation of the measurement*, which is defined as follows [37]:

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{N}(x_i - \bar{x})^2}{N}} \tag{4.2}$$

where N is the number of data points and $\bar{x}$ their arithmetic mean.
An important distinction has to be made here. Normally one measures a physical quantity and wants to be able to say that the true value lies in an interval of $\pm 1\sigma(, 2\sigma, 3\sigma, ...)$ around the calculated mean with a probability of $68.27\%(, 95.45\%, 99.73\%, ...)$ [21]. When counting paths in a random network, the count is always exact, and there is no true value one should find for a given path length n. Instead, when averaging a large number of simulations, one seeks to find the distribution all values for a path length n follow. This will in good approximation be a normal distribution, and then the standard deviation is a measure of the width of that distribution and will usually not become smaller for an increasing number of data points.
All statements in this work will implicitly take the centre of said distribution as its representative, so rather than the width of the distribution, one would like to have a measure of the exactitude with which the position of the centre of the distribution is known, which is characterised by the *standard error of the measurement* or *standard error of the mean* [37]:

$$SE = \frac{\sqrt{\Sigma_{i=1}^{N}(x_i - \bar{x})^2}}{N} = \frac{\sigma}{\sqrt{N}} \tag{4.3}$$

This is the error most figures in this work will provide. Note that there are multiple names for these errors and the standard deviation can occasionally be found identified simply as the standard error.

5. Results: Complete Graphs

This chapter will present the results of a number of simulations that used the algorithm outlined in section 4.1 to find the path length distribution and the longest path in a random DAG. Only complete networks were analysed, meaning all allowed edges were present with probability $p = 1$. The results for Cube Space and Minkowski Space turned out to be very similar, so in most cases figures will be shown for one space and be made available in the appendix for the other. Because this work consists of a relatively high number of separate sections, it was decided to discuss results directly, so each part can be self-contained, rather than letting the reader skim through large parts of the work to find figures. A summary of the important findings will be given at the end.
A quick outline of this chapter:

Section 5.1 will evaluate how well actual data is predicted by the formulae derived in chapter 3. The influence on the results of three parameters in particular will be tested: The number of averaged runs, the number of points N and the dimension of the space D.

Section 5.2 will test the validity of the formulae that predict the length of the longest path L, especially for small L_{scale}. It will be investigated if the bounds from section 2.4 apply in this limit. In the limit $N \to \infty$ L is a linear function of L_{scale} [4], so the same might be true for finite N.

Section 5.3 will repeat the measurements of section 5.2 for the greedy path. As this should serve as a lower bound on L, a comparison will be made to values of the lower bound introduced in section 2.4. The behaviour of the ratio of greedy path length G to longest path length L for large N will be evaluated. It is known that for some network types the greedy path can predict the longest path [20].

5.1. The Path Length Distribution

As established in chapter 3, equations (3.3) and (3.19) give an estimate for the path length distribution, but were derived using a mean field approximation, so the first question that needs to be answered is how well simulation data is represented by these formulae.

Averaging an increasing Number of Runs

As with any experiment that deals with random inputs, it is imperative to take the average of a large enough number of measurements for any given quantity before findings can be said to be reliable. An example measurement for the Cube Space with $N = 100$ in two dimensions averaged over 1000 runs is shown in figure 5.1. The errors already indicate that there's a significant discrepancy between the prediction and the actual measurement, which has fewer long paths and fewer total paths than predicted. This notion is reinforced by figure 5.2; even with a hundred times as many repititions, the measured path length distribution does not change significantly. Note that the logarithm of the distribution is displayed, so naturally small variations towards the long end appear amplified. This was tested for a number of different starting parameters and each time the number of averaged runs had little effect on the measured path length distribution, both in Cube Space and in Minkowski Space. In two dimensions Cube Space and Minkowski Space should give the same results (see chapter 3) and that is indeed the case. The corresponding figures for Minkowski Space are figures A.1 and A.2.

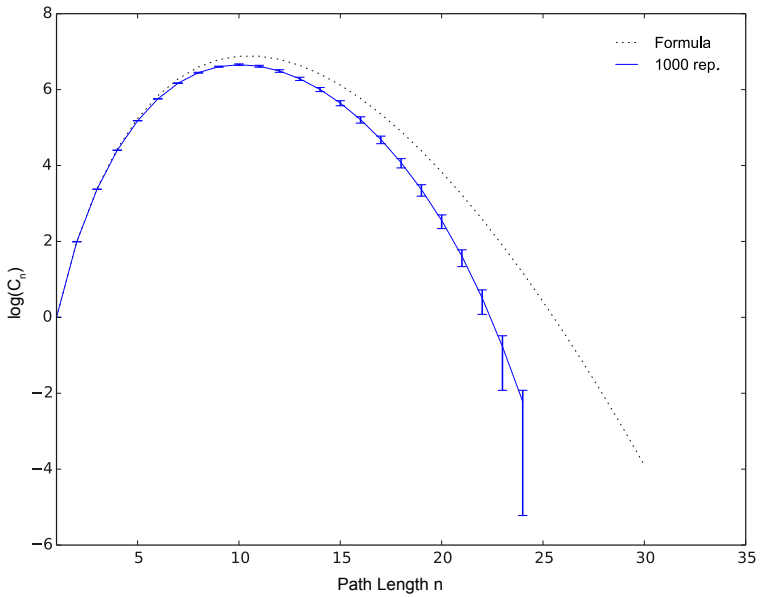


Figure 5.1.: Path length distribution predicted by equation (3.3) (dotted) and actual data in Cube Space with $N = 100$, $D = 2$, 1000 runs averaged on a log scale. Error bars show standard error of measurement. The occurence of long paths and the total number of paths is overestimated by the formula, but short paths are predicted accurately.

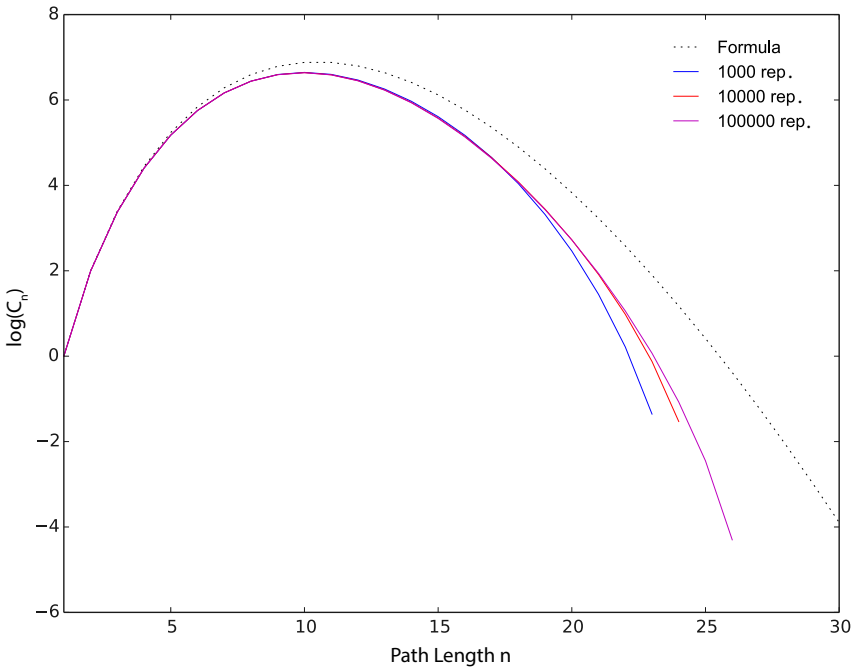


Figure 5.2.: Path length distribution predicted by equation (3.3) (dotted) and actual data in Cube Space with $N = 100$, $D = 2$ on a log scale. Error bars are omitted for better readability. Increasing the number of repititions from 1000 to 100000 (left to right) does not significantly change the measured path length distribution.

Increasing the Number of Points N

Because equations (3.3) and (3.19) were derived from a mean field approximation, one should expect that with increasing N the measurements and the predicted distributions tend to be in better agreement. This could not be observed, as seen in figure 5.3, which shows a comparison of data and predicted distribution in two-dimensional Cube Space for $N = 500, 1000, 2000$. Instead, the measured distribution seems to follow roughly the same curve compared to the formula regardless of N, which means that the number of short paths (smaller than about half the maximum path length) is always very accurately predicted, while long paths are evidently overestimated by the formulae. For completeness the result for Minkowski Space is given in figure A.3, which is of course identical to the one for Cube Space.
The discrepancy between data and prediction can not be assessed by eye, so to determine whether increasing N results at least in a small improvement, one can look at the ratio of data to predicted distribution. This was done in figure 5.4 for two-dimensional Cube Space. Ideally all data should follow the displayed straight line, meaning a perfect fit, and it appears that larger N do that a little bit better. This is an indicator that for $N \to \infty$ theory and data will coincide, but it would be useful to see the results for even bigger N. Note also that the errors for short paths are extremely small, which means there is virtually no variance among the single measurements. For completeness the result for Minkowski Space is given in figure A.4.

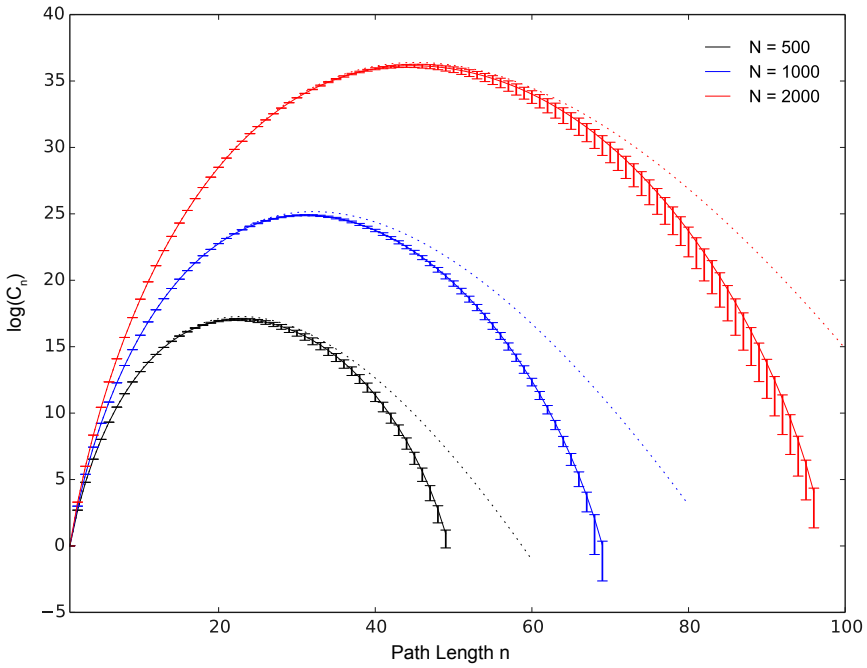


Figure 5.3.: Effect of increasing N ($N = 500, 1000, 2000$ left to right) comparing prediction from equation (3.3) (dotted) and simulation data in Cube Space with $D = 2$ from 500 repititions. Error bars show standard error of measurement. Increase of N does not seem to bring prediction and data closer together.

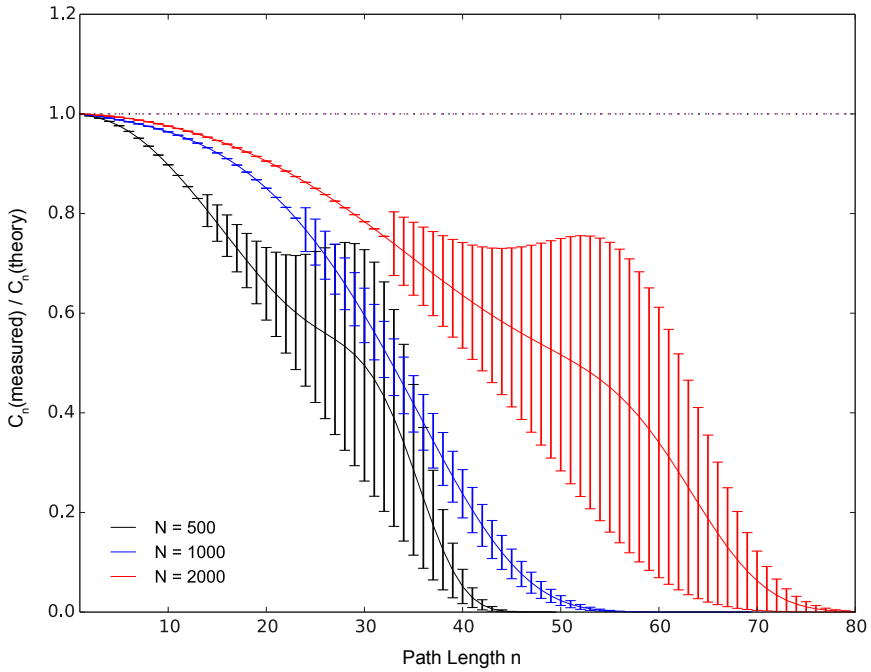
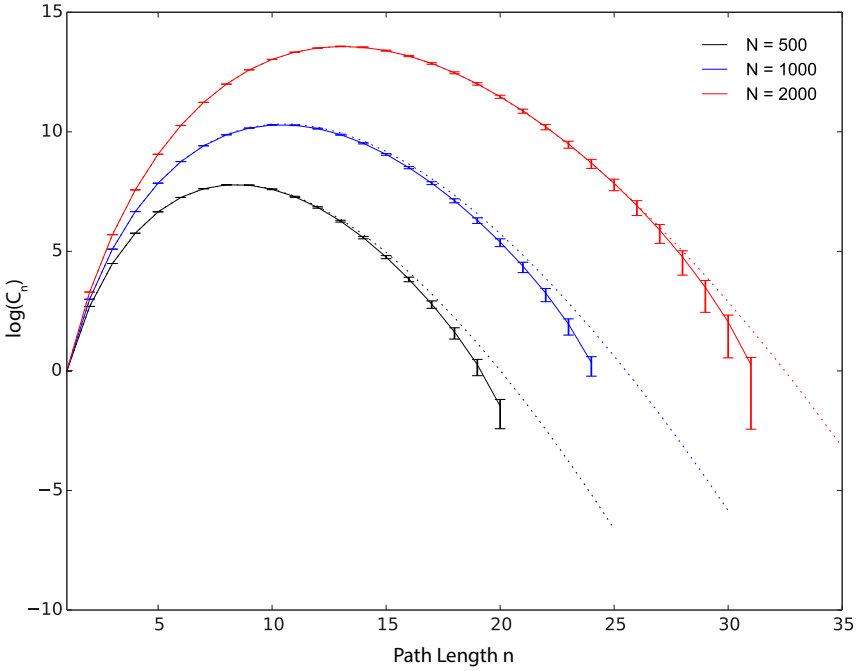


Figure 5.4.: Ratio of measured distributions to predicted distributions for $N = 500, 1000, 2000$ (left to right) in $D = 2$ Cube Space. Data is the same as in figure 5.3 and should ideally follow straight dotted line. Higher N seem to be match unity ratio a little better.

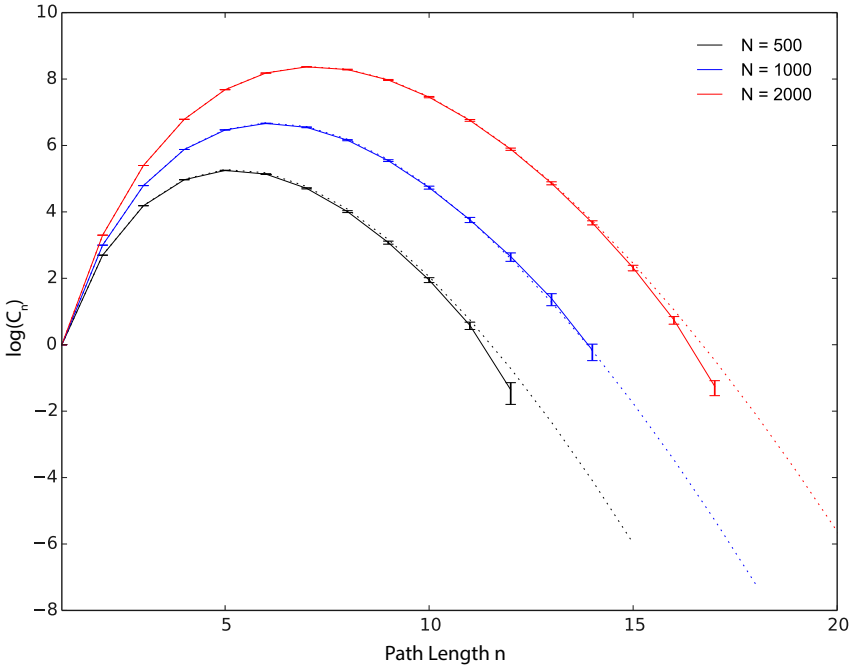
Increasing the Dimension D

The previous measurement, increasing the number of points N, was repeated for different dimensions. It was found that increasing the dimension D resulted in a much better agreement of the data with the formula. Figures 5.5a and 5.5b show that in higher dimensions $D = 3, 4$ of Cube Space the formula appears to predict the measured curves quite well, certainly much better than in two dimensions (figure 5.3). This was not tested for higher dimensions, because the path length distribution or rather the longest path become much shorter, and one can't infer too much information from only a handful of data points to compare. So to produce figures of similar quality as the ones displayed, larger N would have been necessary but could not be realised in reasonable time. Results for Minkowski Space were almost identical to the ones in Cube Space (see figures A.3, A.5a and A.5b).

The key point to take away from this section is that in both Cube Space and Minkowski Space the formulae used to predict the path length distribution appear to overestimate the number of long paths and the total number of paths. On the other hand the prediction for short paths is very accurate. The degree of accuracy or inaccuracy depends on the parameters used:
The first observation was that the number of simulations one takes the average of appears to be of little importance, as long as this number is sufficiently high ($\sim$500) so that variations inherent to random processes are suppressed. This was expected, but is nevertheless reassuring in that it provides confidence for other observations.
Increasing the number of points N in the network had little influence on the quality of the prediction. It could be shown that higher N fit the prediction slightly better, but the improvement is small compared to the discrepancy between data and predicted distributions. It is important to note that all tested N can still be considered small. After all, the formulae describe the result of a mean field approximation. It is likely that only for extremely large N the approximation delivers much better results. So the only conclusion that can be drawn from the data is that predictions of the number of long paths are poor in two dimensions for N up to 2000 and probably a little higher, while short paths are predicted accurately.
For higher dimensions – specifically $D = 3$ and $D = 4$ were tested – predictions apparently become increasingly accurate. It is not immediately clear why that should be the case. The reason could be that for a given N the number of paths of length n decreases with increasing D, both in Cube Space and Minkowski Space, and that the prediction is just more accurate for shorter paths in absolute terms. It should be interesting to see the behaviour for higher N and D and it is not unreasonable to think that then the number of long paths would again be overestimated significantly.



(a) $D = 3$



(b) $D = 4$

Figure 5.5.: Reproduction of figure 5.3 with $D = 3$ and $D = 4$ instead of $D = 2$. Evidently in higher dimensions predicted values (dotted) match data more closely.

5.2. The Longest Path

This section will present measured longest path lengths L as a function of $L_{scale} = N^{1/D}$. Because of limited time and resources, the maximum scale length used was $L_{scale} = 50$, and that only in two dimensions, which is obviously a very small number. The Stirling formula used to derive a proportionality between L and L_{scale} is only a valid approximation for large numbers (relative error $\sim 5\%$ at $L = 20$), so for most data points recorded here its use is unjustified. Consequently, no well-grounded assumption can be made about the behaviour of $L(L_{scale})$ in this regime. Only an indication is given by the work of Bollobás & Brightwell, who prove that for $N \to \infty$ the behaviour is linear.

Figures 5.6a and 5.6b show L as a function of L_{scale} in Cube Space and Minkowski Space, where two observations are most striking.

The first is that there appears to be a clear difference between dimensions, although one would expect that for example $L_{scale} = 1000^{1/3}$ and $L_{scale} = 100^{1/2}$ give the same result. On the other hand, it was just established that especially for $D = 2$ the formulae poorly predict the path length distribution and hence the longest path length.

The other is that for each dimension a linear dependence between L and L_{scale} emerges. A closeup view of the lower left hand sections is given in figures A.6a and A.6b.

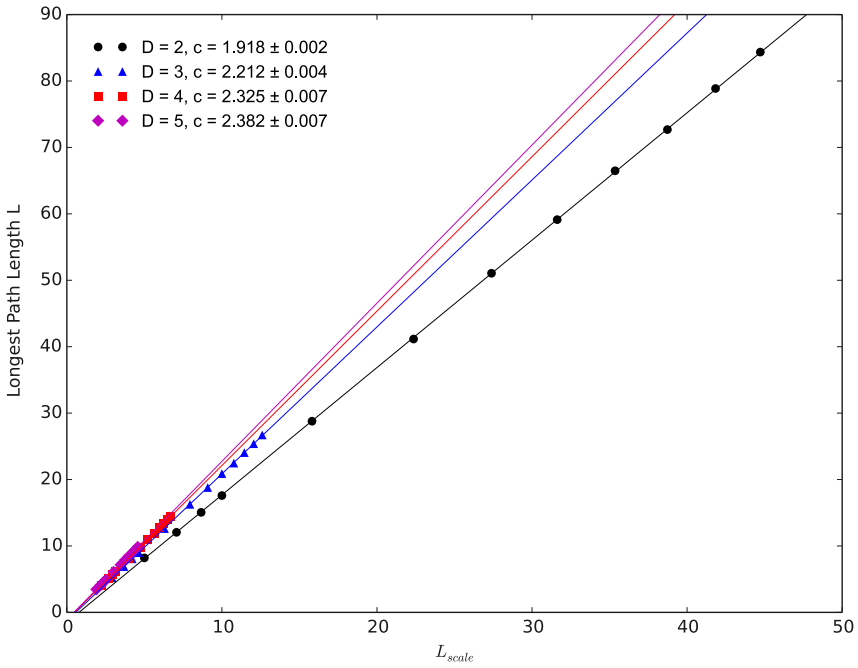
D	Lower bound	Slope	Upper bound
2	1.596	1.918 ± 0.002	2.718
3	1.849	2.212 ± 0.004	2.718
4	1.994	2.325 ± 0.007	2.718
5	2.090	2.382 ± 0.007	2.718

(a) Cube Space

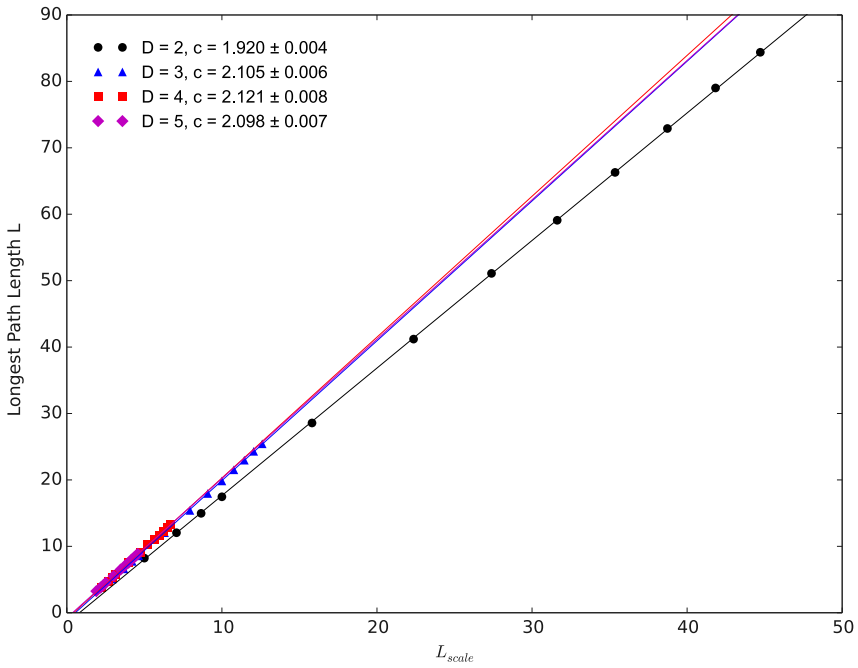
D	Lower bound	Slope	Upper bound
2	1.596	1.920 ± 0.004	2.718
3	1.778	2.105 ± 0.006	2.614
4	1.855	2.121 ± 0.008	2.530
5	1.896	2.098 ± 0.007	2.466

(b) Minkowski Space

Table 5.1.: Measured slopes of $L(L_{scale})$ in Cube Space and Minkowski Space along with bounds calculated from equation (2.7). Errors are 1σ standard deviation. Evidently even in the low L_{scale} regime bounds remain valid.



(a) Cube Space



(b) Minkowski Space

Figure 5.6.: Longest path length L as a function of L_{scale} in Cube Space and Minkowski Space, data points from 500 run average. Error bars smaller than symbols. L shows linear behaviour with slope c given in legend. Points should ideally be on one curve, but there's clearly a dimension dependence, at least for small D.

Table 5.1 lists the slopes for each dimension in the two spaces along with the upper and lower bounds calculated from equation (2.7). Interestingly, the bounds that Bollobás & Brightwell [4] established for the limit $N \to \infty$ still hold in the low L_{scale} regime. It was not possible to find an empiric formula that would allow one to estimate the slope for an arbitrary dimension. However, note that for both spaces the slopes appear to approach a limit (The decrease from $D = 4$ to $D = 5$ in Minkowski Space is not significant[1]). Looking back to section 5.1 where it was found that with increasing dimension the path length distribution also approaches a limit – the formula prediction in that case – it is not unreasonable to expect slopes for higher dimensions to be very close to the one for $D = 5$. Unfortunately this could not be tested, as for higher dimensions N must also be increased to achieve at least moderately high L_{scale}.
To summarise, L is an approximately linear function of L_{scale} for $L_{scale} \leq 50$ and probably higher, which had only been shown for infinitely large N [4]. The slope depends on the dimension D, but high dimensions can be expected to have about the same slope as $D = 5$. Bounds derived for the limit $N \to \infty$ also hold in this regime. It is also reassuring to find that Cube Space and Minkowski Space give almost identical results in two dimensions, precisely what they should do as established in chapter 3. In terms of applicability, knowing that L is a linear function of L_{scale} one can easily estimate L for a given set of N and D.

[1] There are multiple definitions for a *significant deviation*, but a common one is to take the difference of two values and the error of the difference. If the former is more than three times the latter, the two points differ significantly, because the statistical probability for that to happen is $< 0.3\%$. Note here $\Delta = 0.023$ and $\delta(1\sigma) = \sqrt{0.008^2 + 0.007^2} = 0.011$ so that $\Delta \approx \delta(2\sigma)$

5.3. The Greedy Path

This section will showcase the results of measurements of the greedy path. For a given interval this was always taken from point 1 to point 0 rather than the other way round. While for a single graph the direction can make a difference, on average it will not, because the two spaces are symmetric. As a first step the simulations from section 5.2 for the longest path were repeated for the greedy path, meaning that the greedy path length G was measured as a function of L_{scale}.
Figures 5.7a and 5.7b show that G presents linear behaviour like L, but with a smaller slope, in both Cube Space and Minkowski Space. The fact that the slope is smaller than for L is trivial, but there was no reason to expect linear behaviour in the first place. Closeup views of the lower L_{scale} part of the figures are again available in the appendix (figures A.7a and A.7b). The slopes determined from these measurements are listed in table 5.2 together with the calculated lower bounds on L. The fact that these are exclusively bigger than the values obtained in simulation poses the question if maybe the lower bound on L at the same time serves as an upper bound on G. Recall that Bollobás & Brightwell [4] used the concept of the greedy path to derive lower bounds on L, so one might expect the slopes to match those bounds, but again the derivations treat the $N \to \infty$ limit and are therefore a poor benchmark for the measurements at hand. It would have been surprising to find the values in close agreement. Interestingly, the ratio of the slopes of G and L for each dimension is remarkably constant between 0.77 and 0.81. Indeed, with L and G both showing linear behaviour, examining their ratio is the next logical step.

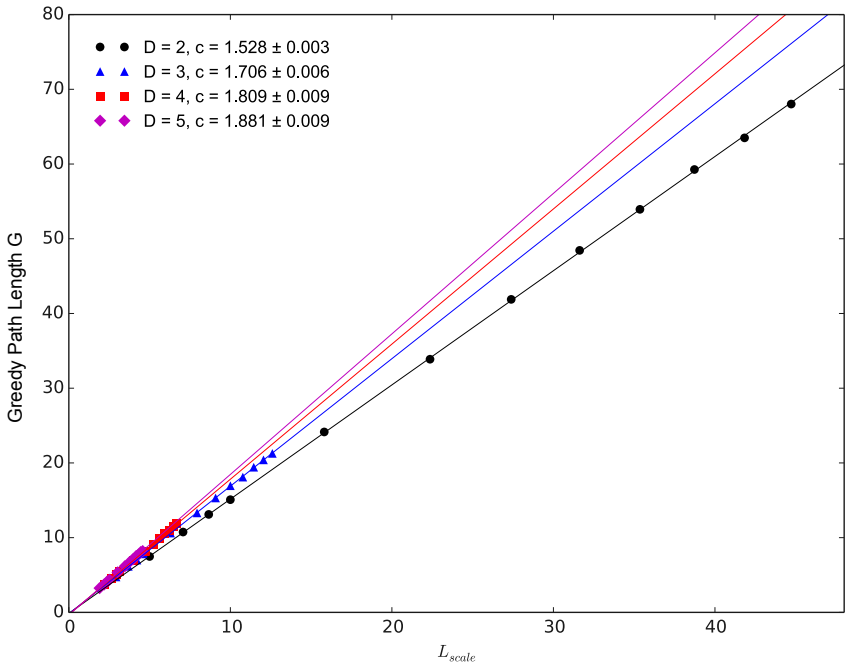
D	Slope	Bound
2	1.528 ± 0.003	1.596
3	1.706 ± 0.006	1.849
4	1.809 ± 0.009	1.994
5	1.881 ± 0.009	2.090

(a) Cube Space

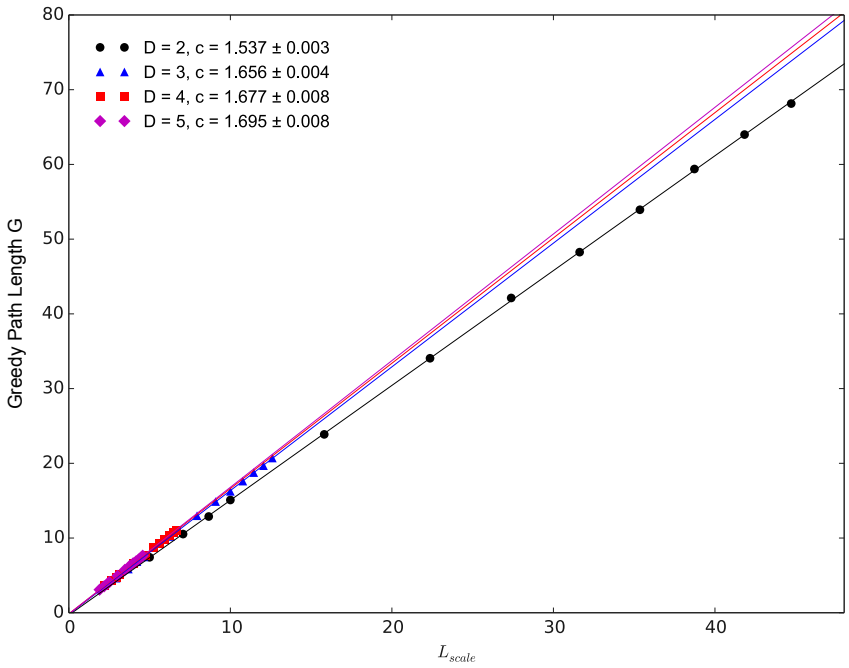
D	Slope	Bound
2	1.537 ± 0.003	1.596
3	1.656 ± 0.004	1.778
4	1.677 ± 0.008	1.855
5	1.695 ± 0.008	1.896

(b) Minkowski Space

Table 5.2.: Measured slopes of $G(L_{scale})$ in Cube Space and Minkowski Space along with bounds calculated from equation (2.7). Errors are 1σ standard deviation. Evidently all slopes are smaller than the lower bound for L.



(a) Cube Space



(b) Minkowski Space

Figure 5.7.: Greedy path length G as a function of L_{scale} in Cube Space and Minkowski Space, data points from 500 run average. Error bars smaller than symbols. G shows linear behaviour with slope c given in legend. Behaviour almost identical to L, but with smaller slope.

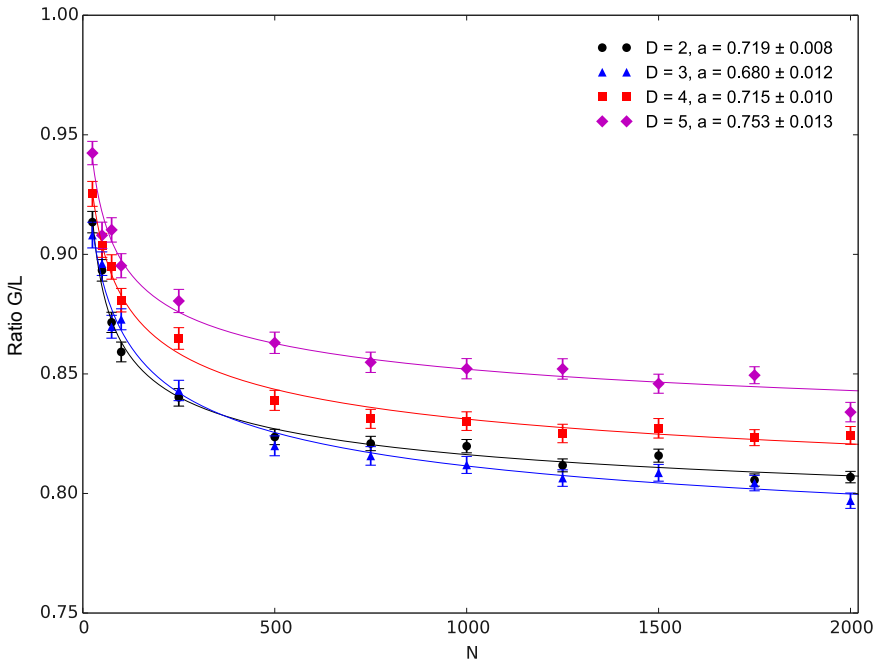
D	a (Cube Space)	a (Minkowski Space)
2	0.719 ± 0.008	0.730 ± 0.007
3	0.680 ± 0.012	0.701 ± 0.013
4	0.715 ± 0.010	0.731 ± 0.010
5	0.753 ± 0.013	0.766 ± 0.015

Table 5.3.: Estimates of the roughly constant ratio a of greedy path length G and longest path length L for large N.

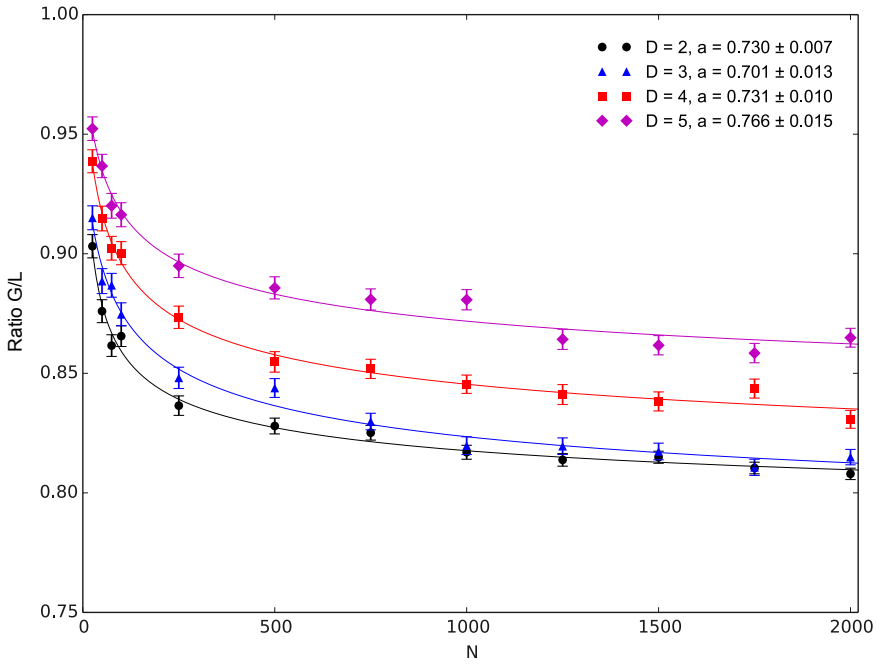
Unfortunately, just plotting G/L as a function of L_{scale} does not reveal much, because most values of L_{scale} are too small. For completeness these figures can be found in the appendix (figures A.8a and A.8b). It is more informative to display the ratio as a function of N, which is done in figures 5.8a and 5.8b. For both Cube Space and Minkowski Space the ratio of greedy path length to longest path length appears to approach a constant value for large N that depends on the dimension D. To find that constant a curve fit was performed with

$$f(N) = a + \frac{b}{\ln(N+c)} \quad , \quad a, b, c \text{ fit parameters} \tag{5.1}$$

which was chosen by eye. A polynomial with negative powers was also tested, but gave poorer results. The function is not meant to accurately describe the behaviour, only to give an estimate of the seemingly constant ratio for larger N. It is undefined for $N = -c$, but that is of little importance here, because the focus lies on the larger N regime. The arbitrariness of the fit function will introduce an error that can outweigh the statistical error of the fit. Nevertheless, only the latter is given in table 5.3, which lists the constants a obtained from the fits.
Much like in the other experiments, the data for $D = 2$ behaves a bit "out of the ordinary", while the curves for higher dimensions appear nicely ordered with a increasing almost linearly with D. This can of course not be the case when going to even higher dimensions, as the ratio cannot exceed unity. Instead, it is reasonable to assume that the $D = 5$ case is a good approximation of the behaviour in higher dimensions, as seen in earlier sections. Going back to figures A.8a and A.8b, which show the ratio G/L as a function of L_{scale} instead of N, one could easily guess that for higher dimensions the ratio roughly approaches a value of 0.75 for large L_{scale}, which is not too different from the ratios of the slopes found earlier, ranging from 0.77 to 0.81
It should be noted that talking about "large" N (or L_{scale}) is a bit misleading here, because these values of N can still be considered small. It is just an assumption that the data will keep following the trend observed in the figures. It might in fact seem trivial to say it will tend towards a constant, after all for any finite N the ratio must be greater than 0. But that does not necessarily mean it will be in the limit $N \to \infty$.
The key finding of this section is that the ratio of G/L most likely approaches a constant value for large N, which is about 0.75 for $D \geq 5$. This could help to quickly estimate L for a given network, as finding the greedy path takes little computational effort. This had been shown for other types of networks before [20], but not for two particular cases discussed here.



(a) Cube Space



(b) Minkowski Space

Figure 5.8.: Ratio of greedy path length G to longest path length L as a function of N. Data from 500 averaged runs. Error bars show standard error of measurement. Ratio appears to approach constant value a for large N.

6. Results: Incomplete Graphs

The second part of the results addresses the subject of incomplete graphs. Random networks were created in the same manner as before, but allowed edges were only inserted with a probability $p < 1$. The algorithm to determine the path length distribution is outlined in section 4.1. To the best of the author's knowledge no similar work has been undertaken.
Unlike before, for any interval $\langle 0, x \rangle$ point x will generally not be the only source and 0 not be the only sink in that interval. As a consequence, the idea of a greedy path is not well defined in this setting and it was decided to not investigate it here.
The results for Cube Space and Minkowski Space were again very similar, so only one of the two will be shown were appropriate. This chapter will have the following structure:

Section 6.1 will investigate whether it is N and p as independent variables or only their product Np that determines the path length distribution. To this end, a normalisation will be introduced to the formulae for the path length distributions, so that ultimately they're only a function of Np. Equations (3.25) and (3.26) predict that only Np is an independent variable.

Section 6.2 will try to determine if $(Np)^{1/D}$ serves as a modified scale length. The longest path length L will be measured to see if it is a function of only $(Np)^{1/D}$. From the findings in chapter 5 it should be expected that there is some dependence on D but that L is a linear function of $L'_{scale} = (Np)^{1/D}$.

Section 6.3 will investigate the change of the path length distribution. From equations (3.25) and (3.26) it should be expected that for a given Np the combination of p and N is irrelevant to the resulting distribution. However, chapter 5 already showed that the formulae do not give an accurate description of the data for finite N, so it would be suprising if it were different for incomplete graphs.

6.1. Np as a single Variable

A necessary requirement for there to be a modified scale length is that rather than the combination of p and N, only the product Np determines the path length distribution. If the latter is the case, it must be possible to introduce a normalisation that will result in a "data collapse", meaning that Np is the only independent variable and any choice of p and N for a given Np gives the same result. To achieve this, one modifies equations (3.25) and (3.26) and plots the following as a function of the path length n:

$$\text{for Cube Space:} \qquad \log\big(C_n(x)\big) + D \cdot \log\big((n-1)!\big) \tag{6.1}$$

$$\text{for Minkowski Space:} \qquad \log\big(C_n(x)\big) - \log\left(\frac{\Gamma(D+1)^{n-1}\Gamma(D/2)}{2^{n-1}\Gamma\big(D(n-1)/2\big)\Gamma(Dn/2)}\right) \tag{6.2}$$

which should in both cases be a straight line with slope $\log(Np)$. This was done in figure 6.1, which shows data collapse for three different values of Np. For each Np a total of 9 distributions were computed with $D \in \{2, 3, 4\}$, $p \in \{1.0, 0.5, 0.1\}$ and $N = Np/p$. The observed data collapse indicates that as hypothesised the only independent variable is Np. There is some variation towards the upper end of the distribution, which is again due to the fact that in the unnormalised distribution there are only very few paths (< 1 on average) of that length. While figure 6.1 is only a qualitative plot, one can readily check that the slopes do match $\log(Np)$ quite accurately.
The plot for Minkowski Space, shown in figure A.9, looks almost identical and because of the normalisation there should indeed be no difference at all between the two. It should also be noted that for smaller path lengths the lines are not completely congruent, unlike seen for $p = 1$. The reason for that is that for a complete graph there will always be exactly one path of length 1, while for $p < 1$ there will likely be more than one, which also mean there's more room for variance among the single measurements. Note also that equations (3.25) and (3.26) predict exactly one path of length 1 regardless of the choice of N and p (as long as they are positive). This will be discussed in more detail in section 6.3.
To summarise, it is possible to introduce a normalisation to equations (3.25) and (3.26), so that for a given product Np data for all combinations of N, p and D approximately follow a straight line with slope $\log(Np)$, which indicates that only Np is an independent variable.

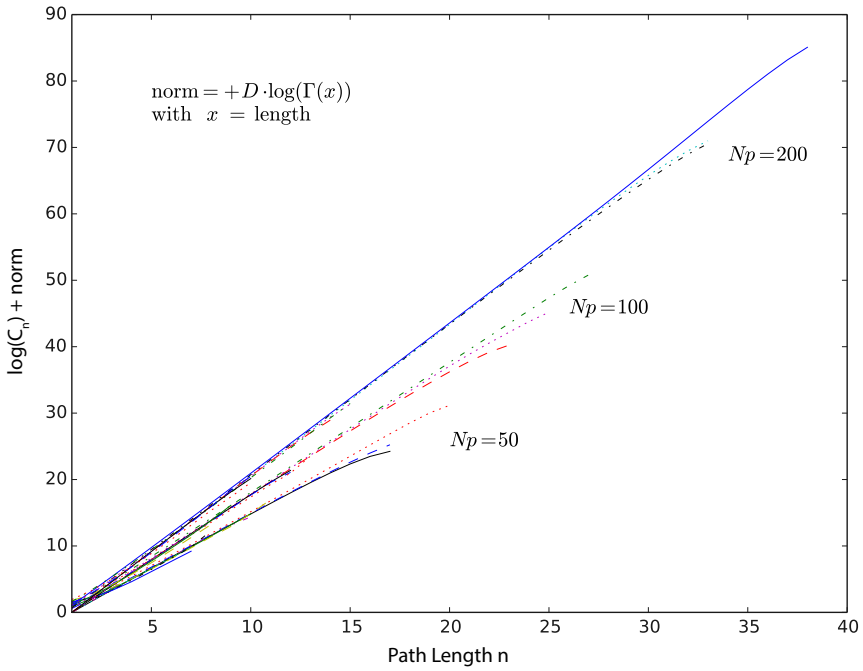


Figure 6.1.: Plot of $\log(C_n(x)) + \text{norm} = (n-1)\log(Np)$ in Cube Space, data from 1000 run average. 9 different distributions for each Np with $D \in \{2, 3, 4\}$, $p \in \{1.0, 0.5, 0.1\}$ and $N = Np/p$. Error bars omitted for better readability. Data collapse indicates that only the product Np is an independent variable.

6.2. A Modified Scale Length

This section aims to ascertain the validity of J. Clough's suggestion, who speculates that for incomplete graphs with edge probability p there is a scale length $L'_{scale} = (Np)^{1/D}$, which would be consistent with the scale length for $p = 1$. Now that Np appears to be a single variable, it is necessary to establish with more confidence that there is indeed such a modified scale length.

As a first step, the longest path length L was measured as a function of D for a number of combinations N, p with $Np = 100$. This way it is possible to see again how much difference the choice of N and p makes and if L is proportional to $(Np)^{1/D}$. The result is shown in figure 6.2 for Cube Space and in figure A.10 for Minkowski Space.

The data points for a given D seem to coincide quite well, although there is a rather large variance for $D = 2$. This further indicates that combinations of N and p with the same product Np give roughly the same result. A fit was performed on the averaged data points with the function $f(D) = c \cdot (Np)^{1/D}$ where c and Np are the fit parameters. While there is clearly a $1/D$ dependence, the fit does not return the ideal value of $Np = 100$. Looking back to section 5.2 this is not surprising – while L was found to be a linear function of L_{scale}, there was still a dimension dependence. It is only reasonable to expect the same for incomplete graphs.

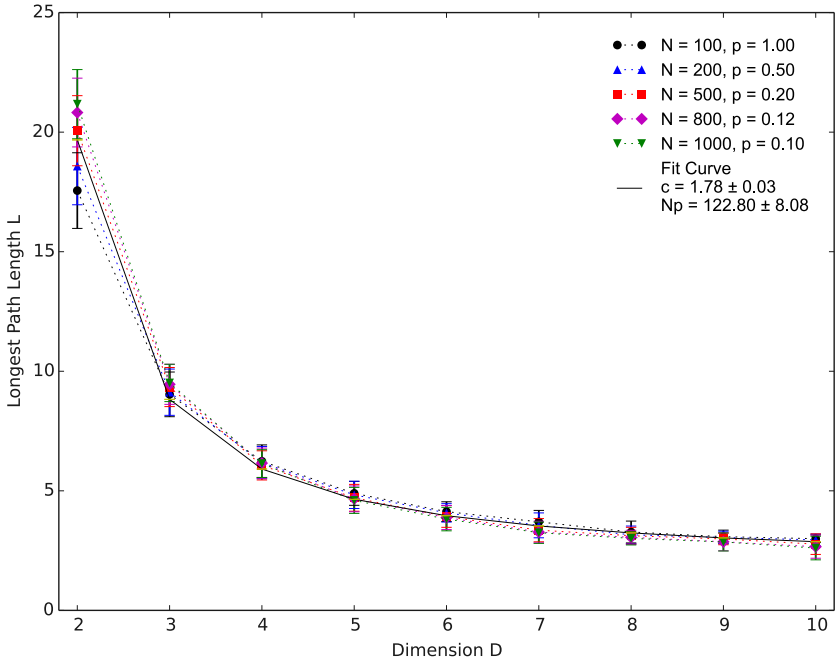


Figure 6.2.: Longest Path L as a function of D for different pairs of N,p with $Np = 100$ in Cube Space, data averaged over 500 runs. Error bars show standard error of measurement. Fit function $c \cdot (Np)^{1/D}$ with c, Np variables. Fit does not return $Np = 100$.

Up to this point it was found that any combination N, p for a given product Np will give very similar results, a necessary requirement for $(Np)^{1/D}$ to be a scale length. Furthermore, L was found to exhibit a $x^{1/D}$ behaviour in D, a second necessary requirement. However, x was significantly different from the given Np. Had that not been the case, the hypothesis could have been reasonably accepted as true, but now one must make a comparison to complete graphs. If L is a linear function in $L'_{scale} = (Np)^{1/D}$ as before, it can be called a scale length.
Figure 6.3 shows L as a function of L'_{scale}. For a fixed range of N six groups of values with $D \in \{2, 3, 4\}$ and $p \in \{0.5, 0.1\}$ were calculated. The data for $D = 2$ and $p = 0.5$ is notably different compared to the other sets, which is similar to what was observed for complete graphs. The rest of the data appears to follow a roughly constant slope. There is some variance among the sets, which is not statistically significant. A closeup view of the lower section of figure 6.3 is given in figure A.12.
Evidently the behaviour of $L(L_{scale})$ is very similar for complete and incomplete graphs. The averaged slopes of the four sets with $D = 3, 4$ are $c_{Cu} = 2.333 \pm 0.035$ in Cube Space and $c_{Mi} = 2.177 \pm 0.034$ in Minkowski Space (shown in figures A.11a and A.11b), which is in good agreement with the slopes found in higher dimensions for complete graphs. Consequently, $L'_{scale} = (Np)^{1/D}$ can be said to serve as a scale length for intervals in Cube Space and Minkowski Space with dimension $D \geq 3$ regardless of the choice of N and p.

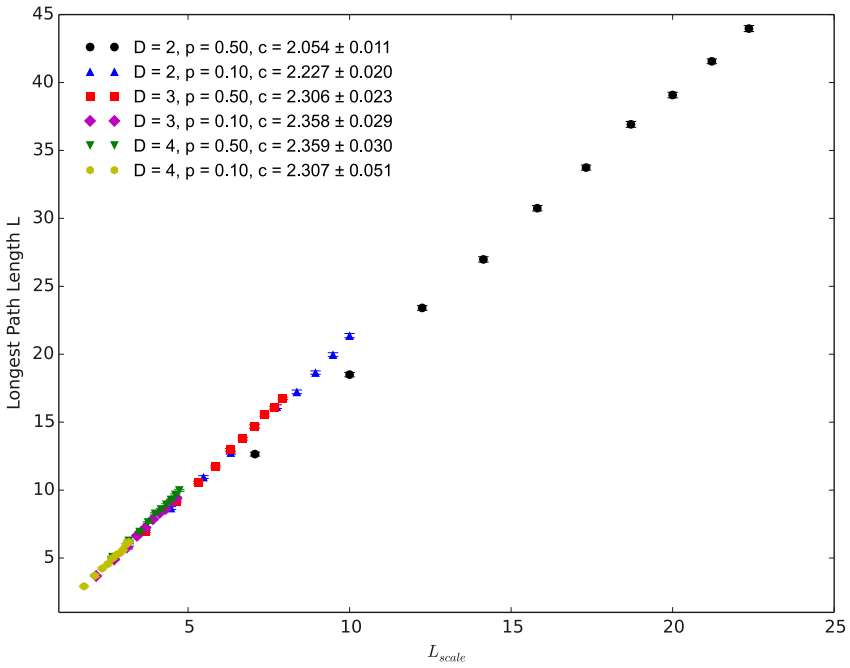


Figure 6.3.: L as a function of the modified scale length $L'_{scale} = (Np)^{1/D}$, data averaged over 100 runs. Error bars show standard error of measurement. Linear fit results given for each set, but line is not displayed. L exhibits linear behaviour with small dependence on p, but roughly constant slope except for $D = 2$.

6.3. Changes to the Path Length Distribution

In section 5.1 it was found that equations (3.3) and (3.19) overestimate the number of long paths and the total number of paths in both Cube Space and Minkowski Space. The prediction was poor in two dimensions but became increasingly accurate in higher dimensions. There is no reason to expect a greater number of longer paths now for $p < 1$, so it is likely that equations (3.25) and (3.26) will overestimate the number of long paths as well.

As briefly mentioned before, in a complete interval there will always be exactly one path of length 1, which leads to a good agreement of data and prediction for shorter path lengths. However, in an incomplete one there are usually numerous sinks and sources, so that it is possible and likely to find more than one path of length 1. Consequently, if the distribution has the same shape as usual, one should expect shorter paths to be underestimated by the formulae.

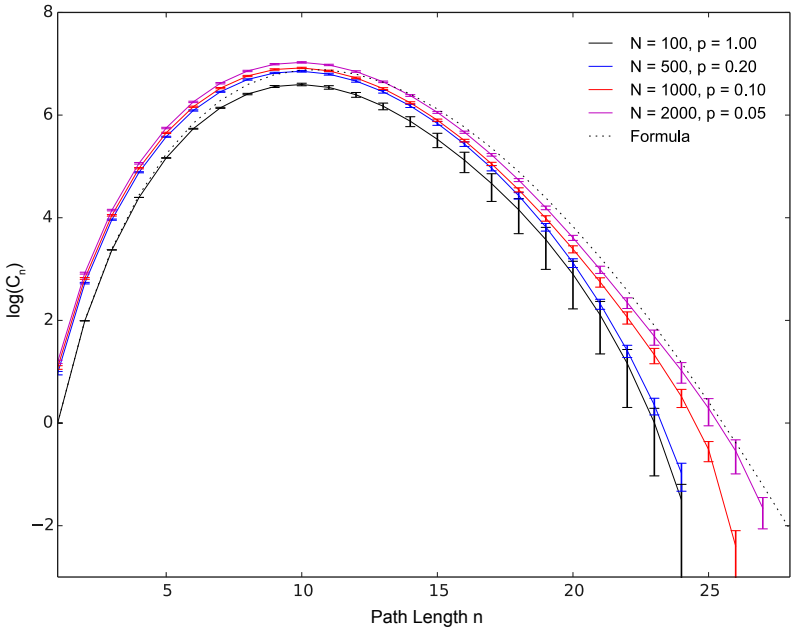
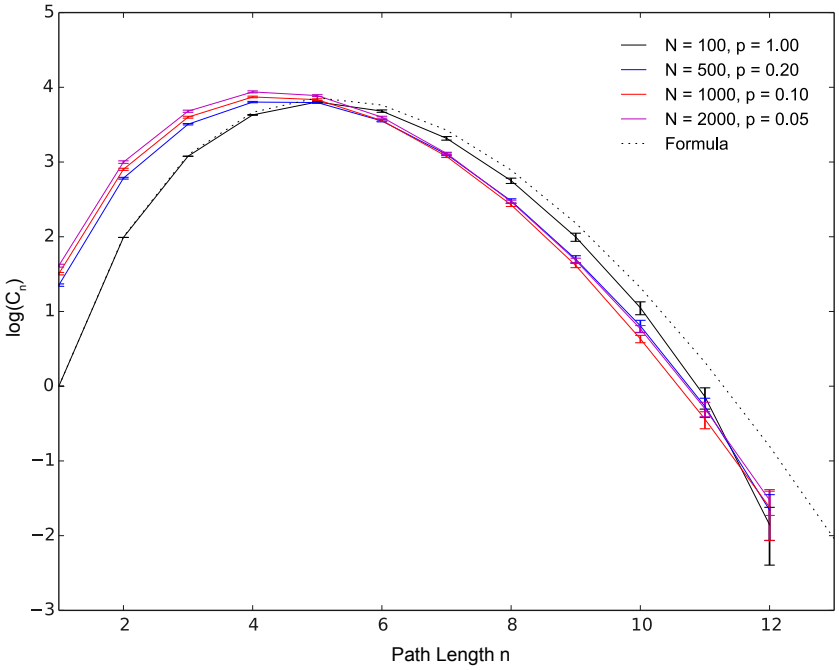
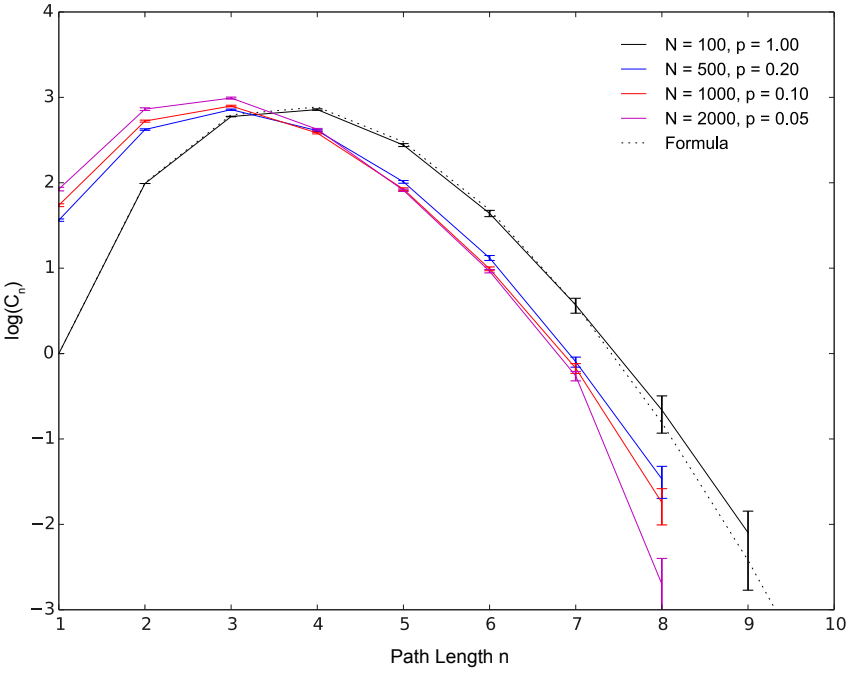


Figure 6.4.: Path Length Distributions in two-dimensional Cube Space with $Np = 100$, data averaged over 500 runs. Error bars show standard error of measurement. For $p < 1$ equation (3.25) underestimates number of short paths, but overestimates number of long paths, as seen for $p = 1$. The total number of paths appears to increase slightly with N.

Figure 6.4 shows a selection of distributions with $Np = 100$ in two-dimensional Cube Space. For comparison $N = 100, p = 1.0$ was included. Two things are immediately obvious. As expected the number of short paths is underestimated by the formula, while the number of long paths is overestimated. Secondly, the total number of paths increases with N. While small, the increase is still significant with respect to the errors.
Knowing that increasing the dimension of the space led to a much better agreement of the data with predicted values for complete graphs, it will be interesting to see if the same is true for incomplete graphs.
The measurements of figure 6.4 were repeated with dimensions $D = 3, 4$, displayed in figures 6.5a and 6.5b. As seen before, the formula predicts the distribution well for $p = 1$. It appears as if there are now many more short paths than predicted compared to $D = 2$, but note that there are far fewer paths overall when the dimension is increased. In fact, the number of short paths is roughly one order of magnitude larger than predicted in all three cases. An increase of the total number of paths can not be observed for the higher dimensions. While it is still significant in the short path regime, the number of long paths appears to decrease in comparison. The upper end of the distribution does not approach the predicted one with increased D, as opposed to the $p = 1$ case. The results are the same for Minkowski Space and are shown in figures A.13, A.14a and A.14b.



(a) $D = 3$



(b) $D = 4$

Figure 6.5.: Reproduction of figure 6.4 with $D = 3, 4$, data averaged over 500 runs. Error bars show standard error of measurement. For $p < 1$ equation (3.25) underestimates number of short paths, but overestimates number of long paths, as seen for $p = 1$. The total number of paths appears to increase slightly with N.

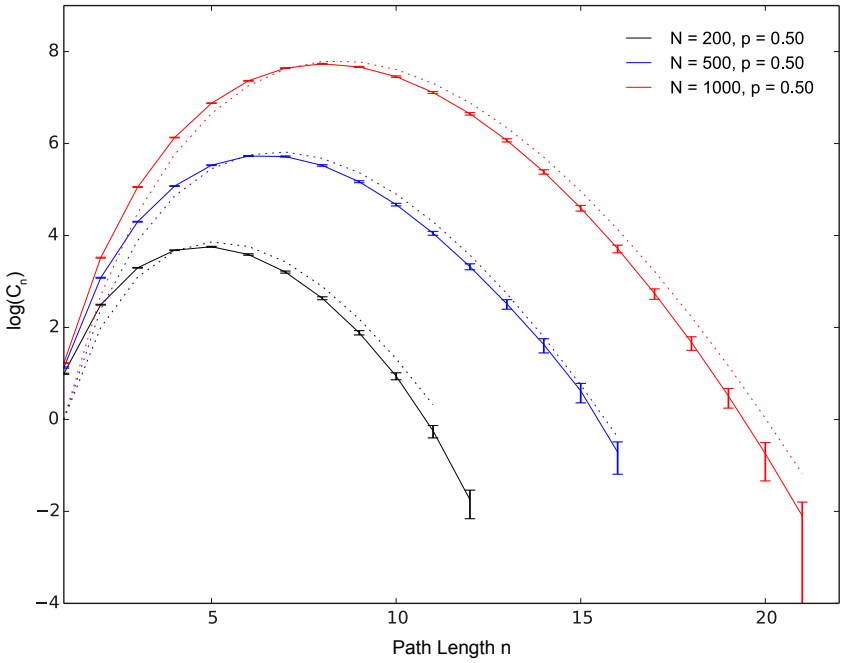


Figure 6.6.: Comparison of measured and predicted (dotted) distributions in three-dimensional Cube Space with $p = 0.5$ and $N = 200, 500, 1000$ (left to right), data averaged from 500 runs. Error bars show standard error of measurement. Discrepancy between data and formula does not change when N is increased.

Because increasing N had little effect on the quality of the prediction for complete graphs, it can be expected that above results remain the same for larger combinations Np. As an example figure 6.6 (and figure A.15 for Minkowski Space) shows three distributions with increasing N and $p = 0.5$ in three dimensions. Evidently the discrepancy between prediction and real data is the same regardless of the choice of N when p and D are held constant.

The key finding of this section is that in incomplete graphs the number of long paths is still overestimated by equations (3.25) and (3.26). In addition the number of short paths is underestimated. As for complete graphs increasing Np makes little difference, but contrary to before increasing the dimension D does not bring data and formulae to better agreement.

7. Empiric Formulae

With the combined results from chapters 5 and 6 it should be possible to modify equations (3.25) and (3.26) to give better estimates for the actual path length distributions. The following is in no way intended to result in an accurate description of the distributions, only to show how one could improve equations (3.25) and (3.26). To find a quantitatively accurate formula obviously goes beyond the scope of this work.
To do the latter, it might not be the best approach to modify the existing formulae, but rather to use a polynomial with a high enough order. Then one could fit the polynomial to a large number of data sets until the behaviour of the coefficients in N, p and D emerges. The following is only one of many possible methods.

7.1. Cube Space

The two most striking observations made in chapters 5 and 6 were that for complete graphs longer paths were being overestimated and that for incomplete graphs shorter paths were also being underestimated. But because the distributions always have more or less the same shape, it should be possible to get a better estimate by just rescaling the axes and by moving the prediction curve appropriately. Remember the proposed formula for the distribution in Cube Space is equation (3.25), which was:

$$C_n(x) = \frac{\big(p\,N(x)\big)^{n-1}}{\big(\Gamma(n)\big)^D} \tag{7.1}$$

Now to rescale the horizontal axis it is necessary to replace $(n-1) \to \alpha\,(n-1)$, while vertical scaling can be done by multiplying $C_n(x)$ with a factor β. To displace the curve to the left just add γ to $(n-1)$. Hence the ansatz for this problem was:

$$C_n(x) = \frac{\big(p\,N(x)\big)^{\alpha(n-1)+\gamma}}{\big[\Gamma\big(\alpha(n-1)+\gamma+1\big)\big]^D} \cdot \beta \tag{7.2}$$

α, β, γ are of course not just constants but functions of possibly all three parameters N, p and D. With some trial and error it was possible to roughly incorporate these dependencies and the following expressions were found to give reasonable results:

$$\alpha = 1 + \frac{0.2}{D^2} \tag{7.3}$$

$$\beta = 1 - \frac{0.4p}{D^2} \tag{7.4}$$

$$\gamma = \frac{2}{3}(1-p) \tag{7.5}$$

The quality of the match in each case was only assessed by eye, so it should be easy to get more accurate expressions with a little more time. Some examples comparing data with the modified prediction are given in figure 7.1a. Obviously the predictions can still not be considered good, but they are better than the original ones, which was the only objective.

7.2. Minkowski Space

The approach was the same as for the Cube Space, so the ansatz formula for Minkowski Space was:

$$C_n(x) = \left(\frac{p\,N(x)\,\Gamma(D+1)}{2}\right)^{\alpha(n-1)+\gamma} \Gamma(D/2) \cdot \left[\Gamma\left(\frac{D}{2}\big(\alpha(n-1)+\gamma\big)+1\right)\Gamma\left(\frac{D}{2}\big(\alpha(n-1)+\gamma+1\big)\right)\right]^{-1} \tag{7.6}$$

Again with some trial and error the following could be found:

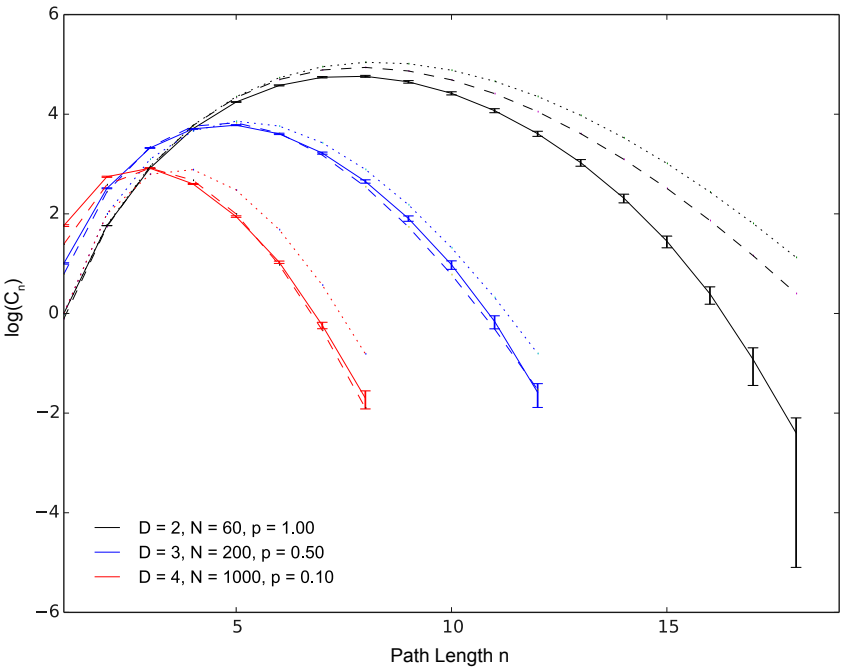
$$\alpha = 1 + \frac{0.8p}{D^3} \tag{7.7}$$

$$\beta = 1 - \frac{0.1p}{D^2} \tag{7.8}$$

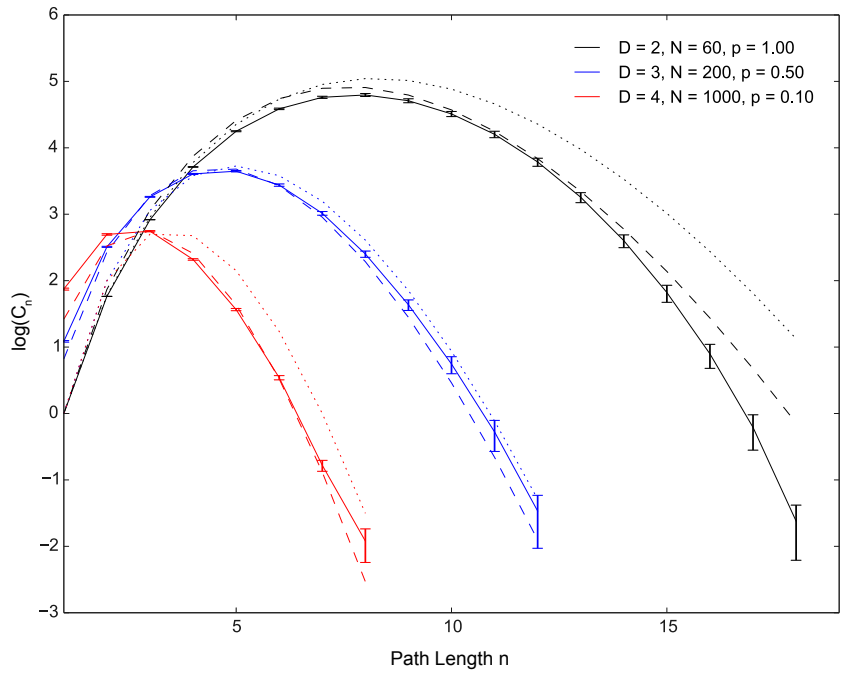
$$\gamma = \frac{2}{3}(1-p) \tag{7.9}$$

Some examplary distributions are shown in figure 7.1b.

It should be noted that a good fit in the centre of the distribution is of highest importance, because the numbers of paths are orders higher than on the wings. Quite obviously there is little physical meaning in the way the parameters are presented and the influence of p and D was only determined by eye, as mentioned before. On a final note, there is no reason to assume that the new formulae give better results for every configuration, but especially for small p they seem to do a much better job. What they do show is that the selected approach can be used to emprirircally improve the existing formulae.



(a) Cube Space



(b) Minkowski Space

Figure 7.1.: Comparison of a selection of path length distributions with original formula (dotted) and new formula (dashed), data averaged from 500 runs. Error bars show standard error of measurement. Modified formulae still don't give a good prediction, but a better one than the original.

8. Summary & Outlook

The objective of this work was to answer how many paths of any length n are present in a random discrete interval of N points in D-dimensional ordered Cube Space and Minkowski Space, where $D \geq 2$. Particular attention was paid to the longest path. The order of Cube Space was that a point can only connect to another, if the latter has a smaller component in all dimensions. The order of Minkowski Space was that a point can only connect to points earlier in time that also lie within its lightcone. Both resulted in a Directed Acyclic Graph (DAG).
A distinction was made between two cases. One where all possible connections are present – a complete interval – and the other where every possible edge is only present with a probability $p < 1$ – an incomplete interval.

For complete graphs analytic formulae exist, which were derived in a mean field approximation, meaning that they assume a constant point density $\rho = N/V$ rather than discrete points. It was investigated how well simulation data is predicted by these formulae. In the next step the dependence of the longest path L on N and D was probed and finally the relation between greedy path and longest path was examined. The following could be found:

1. It is sufficient to take the average of circa 500 runs to see a signifcant deviation of the data from the formulae. Averaging more runs made errors smaller, but did not change the measured path length distribution. This was incorporated in the rest of the experiments.

2. The prediction of the formulae is very accurate for short paths, but becomes increasingly inaccurate for longer paths. The formulae significantly overestimate the occurence of long paths, where long paths means roughly the longer half of the distribution.

3. The previous finding was qualitatively the same for all tested N. While it was found that with higher N prediction and data matched slightly better, this behaviour was expected to be expressed much strongly. It can only be seen as an indication that for $N \to \infty$ data and prediction will be in agreement. More rigorous testing with higher N is necessary to fully confirm this.

4. The prediction becomes more accurate in higher dimensions. Because for a given N a higher dimensions results in fewer and shorter paths, it could just be that the formulae always predict the number of short paths well in absolute terms rather than relative to the whole distribution. However, it is more likely that the formulae perform poorly in low dimensions ($D = 2, 3$) and give better results in higher dimensions, which is supported by later findings.

5. The longest path length L is a linear function of the scale length $L_{scale} = N^{1/D}$, with the limitation that the slope also depends on the dimension. It increases with D, but appears to become constant from around $D = 5$, which supports the notion of the previous point. The measured slopes for $D = 5$ were $c_{Cu} = 2.382 \pm 0.007$ in Cube Space and $c_{Mi} = 2.098 \pm 0.007$ in Minkowski Space

6. Upper and lower bounds on the slope derived for the limit $N \to \infty$ [4] were found to be valid in the tested regime $L_{scale} \leq 50$.

7. The ratio of greedy path length to longest path length G/L appears to approach a constant value for large N that depends on the dimension. Again this D dependence became insignificant for $D \geq 5$, where the constant is roughly 0.75 for both Cube Space and Minkowski Space.

There is no analytic derivation of formulae for incomplete intervals. However, it was conjectured that it should be sufficient to replace N with Np in the existing formulae. It was investigated if the choice of N and p is irrelevant for a given product Np. It was further tested if $L'_{scale} = (Np)^{1/D}$ also serves as a scale length. Again the quality of the prediction of the modified formulae was examined:

1. It is possible to normalise the formulae in a way that for a given Np all choices of N, p and D result in approximately the same straight line with slope $\log(Np)$, when the normalised distribution is displayed as logarithm. This indicates that indeed only Np is an independent variable, not N and p separately.

2. L is a linear function of $L'_{scale} = (Np)^{1/D}$. Except for $D = 2$ the slope is roughly constant for all dimensions with $c_{Cu} = 2.333 \pm 0.035$ in Cube Space and $c_{Mi} = 2.177 \pm 0.034$ in Minkowski Space, values very similar to the ones obtained for complete intervals. Consequently it can be said that L'_{scale} serves as a scale length for incomplete intervals.

3. The modified formulae continue to overestimate the occurence of long paths, but additionally underestimate the number of short paths.

4. Np seems to define the path length distribution, with only a small dependence on N and p separately.

5. Increasing Np is again with little effect, the measured distribution deviates from the prediction in the same manner for all tested Np.

6. The prediction does not become better in higher dimensions. It appears to be about the same for all tested dimensions in terms of order of magnitude.

The findings relating measured path length distributions and prediction were used to modify the existing formulae. It should be noted again that this was only done to highlight one possible approach to empirically improving predictions and is not meant as a serious attempt. It was found that even though data and formula were often in poor agreement, the distributions had a similar shape. So to find a more accurate expression the approach was to rescale the axes and to displace the curve. This meant replacing $(n-1) \to \alpha(n-1) + \gamma$ and $C_n \to \beta C_n$ in the formulae. It was found that the

following coefficients resulted in a better – but still not good – prediction of path length distributions:

$$\begin{aligned} \text{Cube Space:} \quad \alpha &= 1 + \frac{0.2}{D^2} \\ \beta &= 1 - \frac{0.4p}{D^2} \\ \gamma &= \frac{2}{3}(1-p) \\ \text{Minkowski Space:} \quad \alpha &= 1 + \frac{0.8p}{D^3} \\ \beta &= 1 - \frac{0.1p}{D^2} \\ \gamma &= \frac{2}{3}(1-p) \end{aligned}$$

Some of the findings, specifically the linearity of L as a function of L_{scale} and the constant ratio of G and L for large N, could prove useful to predict or estimate the longest path or even the path length distribution for very large directed acyclic networks without the need to perform lengthy computational analyses. On the other hand, there are still numerous open ends to this work.

All the simulations were performed for relatively small N because of limited computing power. Verifying the claims for larger N is no doubt necessary for every section of this work. Especially the behaviour of L for larger L_{scale} in higher dimensions should be interesting. With a more powerful computer or simply more time it should be an overall easy task.

Somewhat more demanding would be to find an empirical formula that gives accurate predictions for both complete and incomplete graphs. This work only highlighted one possible approach that could prove adequate while there are no doubt numerous others. All of them have in common that they will require a very large data set. Approaches using a different ansatz than the one from this work will face an additional challenge in that they will have to guess an appropriate shape – for example the order of the polynomial – as a first error source and will then have to find the right expression to describe the behaviour of all coefficients in N, p and D – another error source.

It should also be interesting to see what other characteristics of a network can be extracted from the path length distribution, such as the average shortest path, clustering, average degree or the degree distribution. If that's possible an accurate prediction becomes even more valuable and the path length distribution could become a relevant tool of network analysis.

Acknowledgements

First and foremost I'd like to thank my supervisor Dr. Tim Evans for giving me the opportunity to carry out this project and for invaluable support and advice. Thanks also to Prof. Kim Christensen for agreeing to be my assessor.
For ideas, help and also code examples I want to thank James Clough. Finally I'd like to thank the whole Complexity Science group for the relaxed yet stimulating atmosphere.

Bibliography

[1] Balaban, A.T. *Chemical Applications of Graph Theory*. Academic Press, 1976.

[2] Biggs, N. *Algebraic Graph Theory*. Cambridge University Press, 1993.

[3] Black, P.E. *Greedy Algorithms*. Dictionary of Algorithms and Data Structures (online). National Institute of Standards and Technology, 2005. URL: http://xlinux.nist.gov/dads//HTML/greedyalgo.html (visited on 09/12/2014).

[4] Bollobás, B. and Brightwell, G. "Box-spaces and random partial orders". In: *Transactions of the American Mathematical Society* 324.1 (1991), pp. 59–72.

[5] Brightwell, G. and Gregory, R. "Structure of random discrete spacetime". In: *Physical Review Letters* 66.3 (Jan. 1991), pp. 260–263. DOI: 10.1103/PhysRevLett.66.260.

[6] Clough, J.R. and Evans, T.S. *What is the dimension of citation space?* Aug. 2014. arXiv: 1408.1274. (Visited on 09/10/2014).

[7] Cormen, T.H. et al. *Introduction to Algorithms*. 3rd ed. Cambridge: The MIT Press, 2009.

[8] Deza, E. and Deza, M.M. *Encyclopedia of distances*. Berlin, Heidelberg: Springer-Verlag, 2009. DOI: 10.1007/978-3-642-00234-2.

[9] Diestel, R. *Graph Theory*. Springer, 2005. DOI: 10.4171/OWR/2005/03.

[10] Feller, W. *An Introduction to Probability Theory and Its Applications,* 3rd ed. Vol. 1. New York: Wiley, 1968.

[11] Flanders, H. *Differential Forms with Applications to the Physical Sciences*. Dover, 2003.

[12] Fortunato, S. "Community detection in graphs". In: *Physics Reports* 486.3-5 (Feb. 2010), pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002.

[13] Girvan, M. and Newman, M.E.J. "Community structure in social and biological networks." In: *Proceedings of the National Academy of Sciences of the United States of America* 99.12 (June 2002), pp. 7821–7826. DOI: 10.1073/pnas.122653799.

[14] Graham-Rowe, D. *Mission to build a simulated brain begins*. 2005. URL: http://www.newscientist.com/article/dn7470-mission-to-build-a-simulated-brain-begins.html (visited on 09/07/2014).

[15] Gries, D. and Schneider, F.B. *A Logical Approach to Discrete Math*. Springer Science & Business Media, 1993.

[16] Guare, J. *Six Degrees of Separation: A play*. New York: Random House, 1990.

[17] Harary, F. *Graph Theory*. Addison-Wesley, 1969.

[18] Hawking, S.W. *The Large Scale Structure of Space-Time*. Cambridge University Press, 1973.

[19] Ilie, R., Thompson, G.B., and Reid, D.D. "A numerical study of the correspondence between paths in a causal set and geodesics in the continuum". In: *Classical and Quantum Gravity* 23 (Dec. 2005), pp. 3275–3286. DOI: 10.1088/0264-9381/23/10/002.

[20] Karger, D., Motwani, R., and Ramkumar, G.D.S. "On approximating the longest path in a graph". In: *Algorithmica* 18.1 (May 1997), pp. 82–98. DOI: 10.1007/BF02523689.

[21] Kenney, J.F. and Keeping, E.S. *Mathematics of Statistics*. van Nostrand, 1963.

[22] Martinez, N. "Constant Connectance in Community Food Webs". In: *The American Naturalist* 139.6 (1992), pp. 1208–1218.

[23] Meyer, D.A. "The dimension of causal sets". PhD Thesis. MIT, 1988.

[24] Milgram, S. "The small world problem". In: *Psychology Today* 2 (1967), pp. 60–67.

[25] Myrheim, J. *Statistical Geometry*. Tech. rep. Geneva: CERN preprint TH-2538, 1978.

[26] Nakanishi, N. *Graph Theory and Feynman Integrals*. Vol. 11. Mathematics and its applications. Gordon and Breach, 1971.

[27] Newman, M.E.J. *Networks: An Introduction*. New York: Oxford University Press, 2010.

[28] Page, L. et al. *The PageRank Citation Ranking: Bringing Order to the Web*. Tech. rep. Stanford InfoLab, 1999.

[29] Porter, M.A., Onnela, J.P., and Mucha, P.J. "Communities in networks". In: *Notices of the American Mathematical Society* 56.9 (2009).

[30] Rideout, D. and Wallden, P. "Spacelike distance from discrete causal order". In: *Classical and Quantum Gravity* 26.15 (Aug. 2009), p. 155013. DOI: 10.1088/0264-9381/26/15/155013.

[31] Schrijver, A. *Combinatorial Optimization: Polyhedra and Efficiency*. Vol. 1. Algorithms & Combinatorics. Springer Science & Business Media, 2003, p. 114.

[32] Sedgewick, R. and Wayne, K. *Algorithms*. 4th ed. Addison-Wesley Professional, 2011, pp. 661–666.

[33] Thompson, G.B. "Paths in discrete spactime : A numerical investigation of the Myrheim length conjecture". MSc Thesis. Eastern Michigan University, 2003.

[34] Thulasiraman, K. *Graphs: Theory and Algorithms*. John Wiley & Sons, 1992, p. 460.

[35] Tönnies, F. *Gemeinschaft und Gesellschaft*. Leipzig: Fues's Verlag, 1887.

[36] Weisstein, E.W. *Measure Space*. URL: http://mathworld.wolfram.com/MeasureSpace.html (visited on 09/05/2014).

[37] Zwillinger, D. *Standard Mathematical Tables and Formulae*. 32nd ed. CRC Press, 2011.

A. Appendix

A.1. Derivation of the Corrective Term for $L(L_{scale})$ in Cube Space

The aim is to find a solution L for equation (3.9), which was

$$L\ln(L) - L - \frac{1}{2}\ln\left(\frac{L}{2\pi}\right) + O(L^{-1}) \approx (L-1)\ln(L_{scale}) \tag{A.1}$$

and the ansatz is $L = eL_{scale} + f$, where f is the corrective term that needs to be found. Neglect terms of order $O(L^{-1})$ and lower, so that with left hand side and right hand side switched:

$$(eL_{scale} + f - 1)\ln(L_{scale}) \approx \left(eL_{scale} + f - \frac{1}{2}\right)\ln(eL_{scale} + f) - eL_{scale} - f + \frac{1}{2}\ln(2\pi) \tag{A.2}$$

$$\approx \left(eL_{scale} + f - \frac{1}{2}\right)\left(1 + \ln(L_{scale}) + \ln\left(1 + \frac{f}{eL_{scale}}\right)\right) - eL_{scale} - f + \frac{1}{2}\ln(2\pi) \tag{A.3}$$

$$-\frac{1}{2}\ln(L_{scale}) \approx \left(eL_{scale} + f - \frac{1}{2}\right)\left(1 + \ln\left(1 + \frac{f}{eL_{scale}}\right)\right) - eL_{scale} - f + \frac{1}{2}\ln(2\pi) \tag{A.4}$$

$$\approx \left(eL_{scale} + f - \frac{1}{2}\right)\ln\left(1 + \frac{f}{eL_{scale}}\right) + \frac{1}{2}\ln(2\pi) - \frac{1}{2} \tag{A.5}$$

$$-\frac{1}{2}\ln\left(\frac{2\pi L_{scale}}{e}\right) \approx \left(eL_{scale} + f - \frac{1}{2}\right)\ln\left(1 + \frac{f}{eL_{scale}}\right) \tag{A.6}$$

Assume now that f is sufficiently small compared to eL_{scale} so one can write

$$\ln\left(1 + \frac{f}{eL_{scale}}\right) \approx \frac{f}{eL_{scale}} \tag{A.7}$$

which, neglecting terms of order $O(f^2)$, yields

$$-\frac{1}{2}\ln\left(\frac{2\pi L_{scale}}{e}\right) \approx f - \frac{f}{2eL_{scale}} \tag{A.8}$$

Hence:

$$f \approx -\frac{eL_{scale}\ln\left(\frac{2\pi L_{scale}}{e}\right)}{2eL_{scale} - 1} \approx -\frac{1}{2}\ln\left(\frac{2\pi L_{scale}}{e}\right) \tag{A.9}$$

the last step being valid if eL_{scale} is much larger than 1.

A.2. Additional Figures

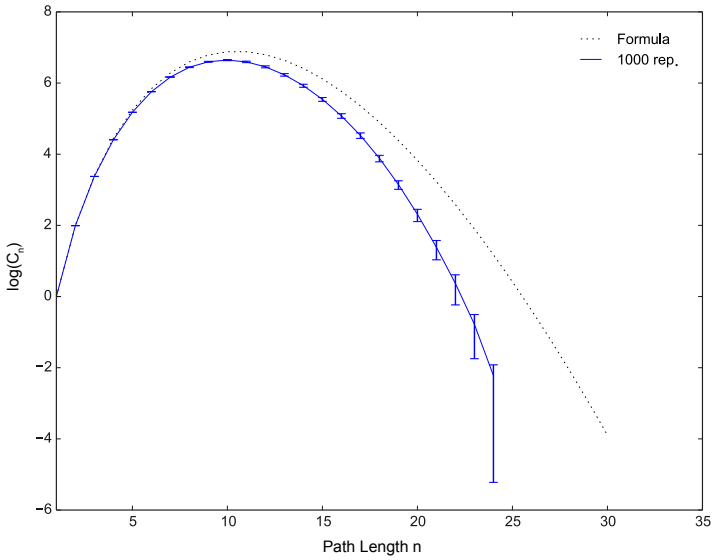


Figure A.1.: Path length distribution predicted by equation (3.19) (dotted) and actual data in Minkowski Space with $N = 100$, $D = 2$, 1000 runs averaged on a log scale. Error bars show standard error of measurement. The occurence of long paths and the total number of paths is overestimated by the formula.

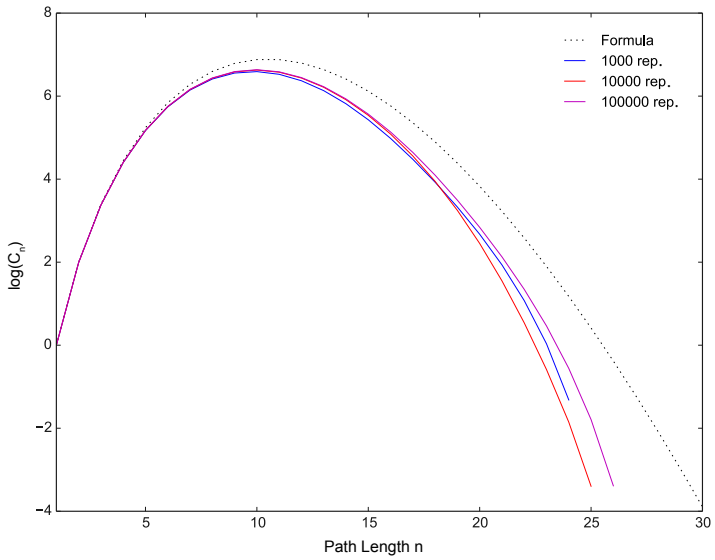


Figure A.2.: Path length distribution predicted by equation (3.19) (dotted) and actual data in Minkowski Space with $N = 100$, $D = 2$ on a log scale. Error bars are omitted for better readability. Increasing the number of repititions from 1000 to 100000 does not significantly change the measured path length distribution.

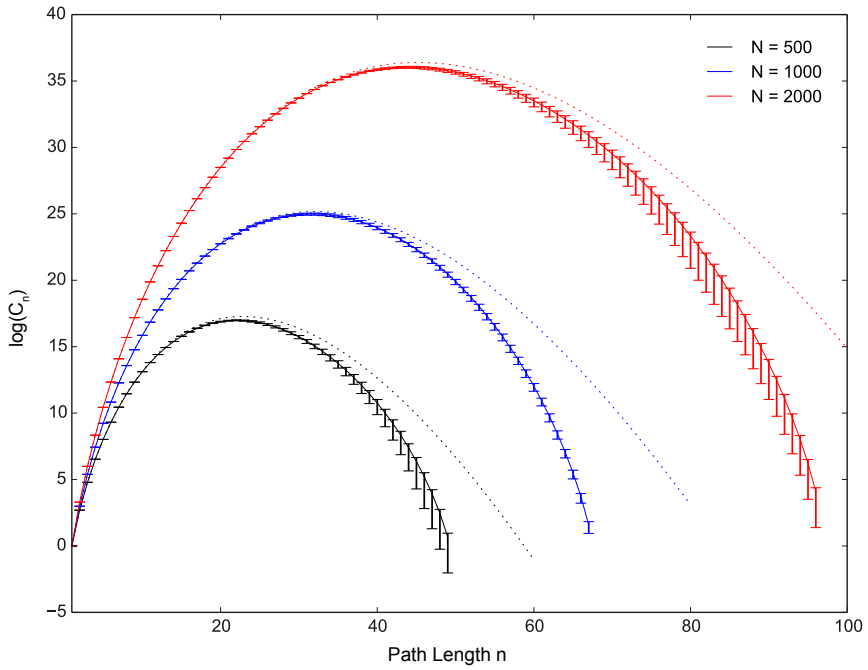


Figure A.3.: Effect of increasing N comparing prediction from equation (3.19) (dotted) and simulation data in Minkowski Space with $D = 2$ from 500 repititions. Error bars show standard error of measurement. Increase of N (left to right) does not seem to bring prediction and data closer together.

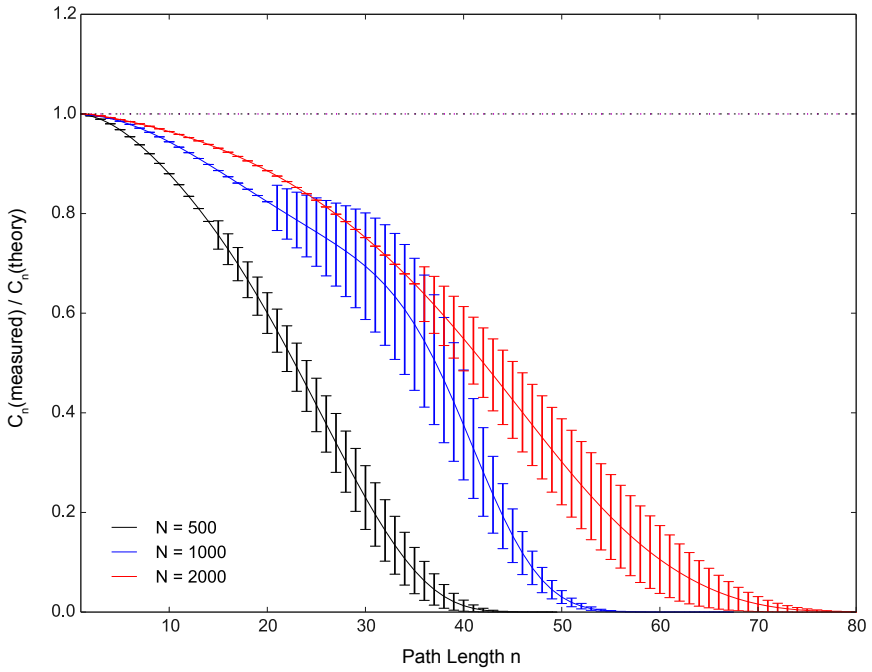
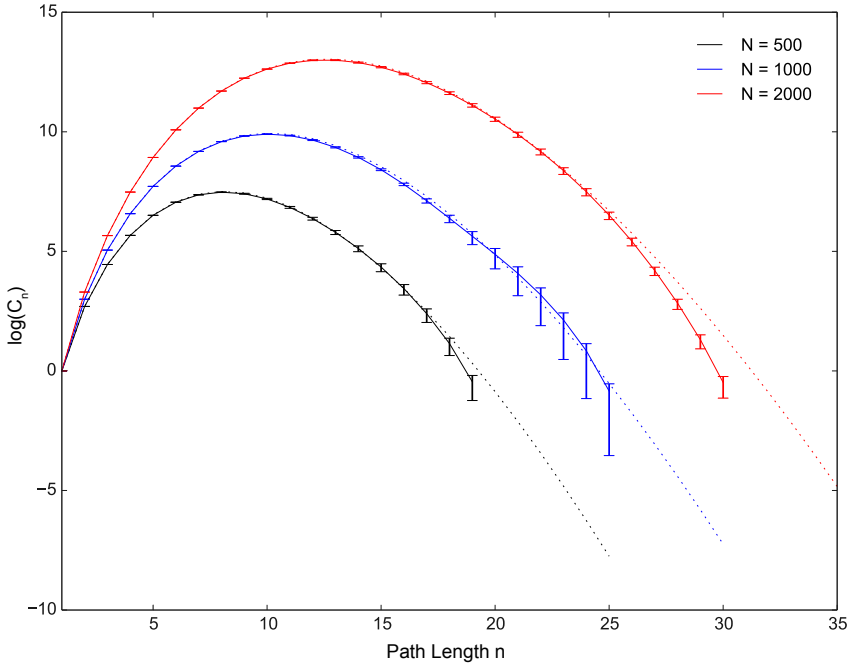
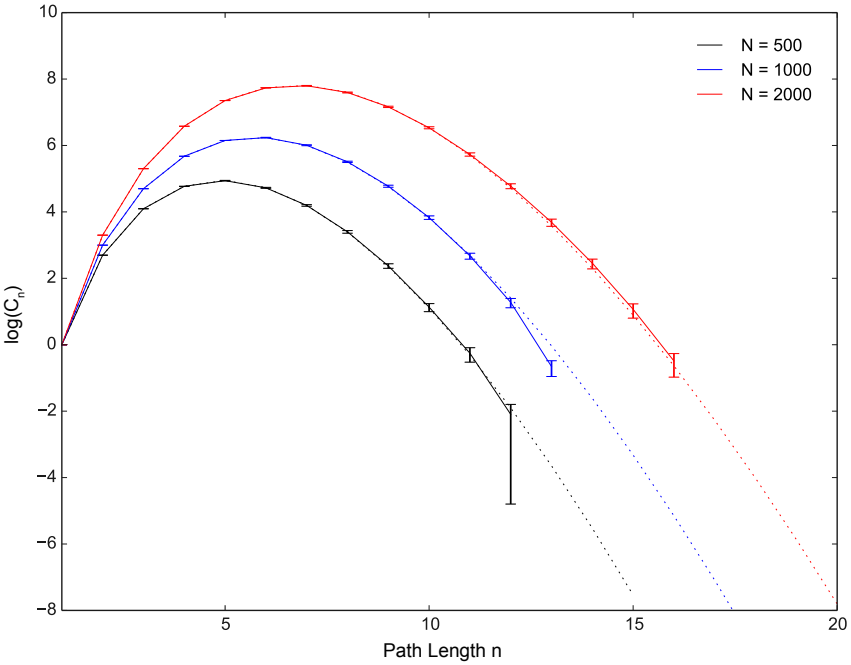


Figure A.4.: Ratio of measured distributions to predicted distributions for $N = 500, 1000, 2000$ (left to right) in $D = 2$ Minkowski Space. Data is the same as in figure 5.3 and should ideally follow straight dotted line. Higher N seem to be match unity ratio a little better.

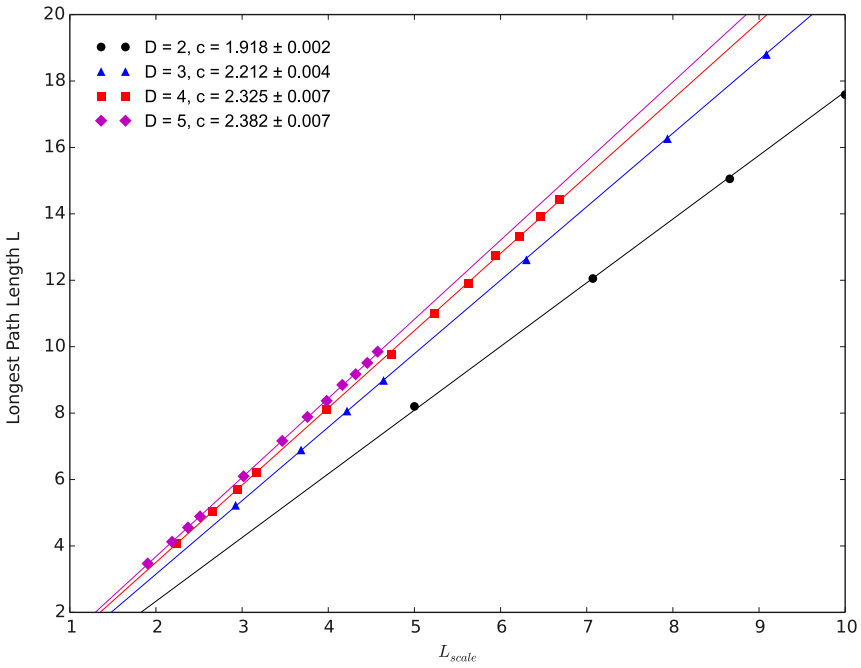


(a) $D = 3$

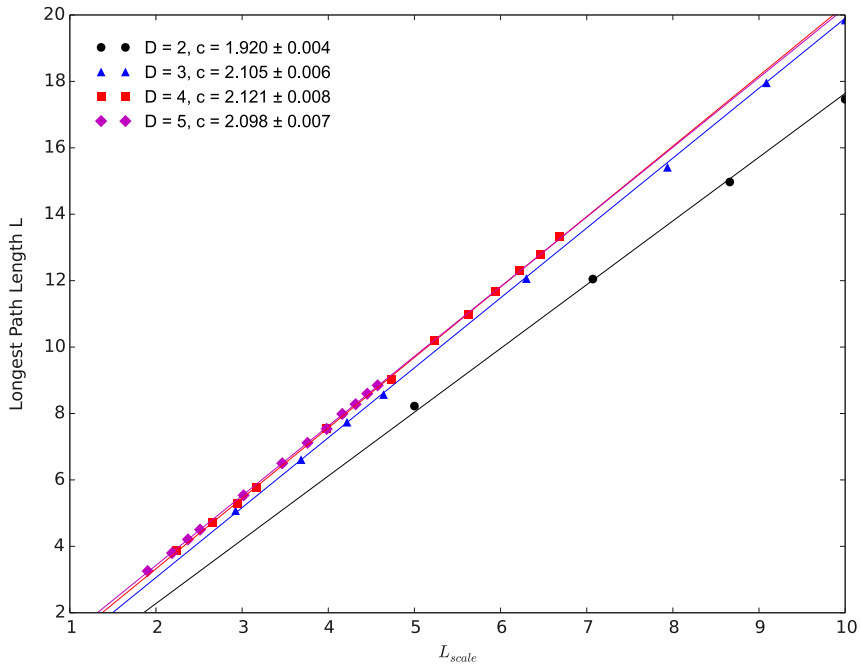


(b) $D = 4$

Figure A.5.: Reproduction of figure A.3 with $D = 3$ and $D = 4$ instead of $D = 2$. Evidently in higher dimensions predicted values (dotted) match data more closely.

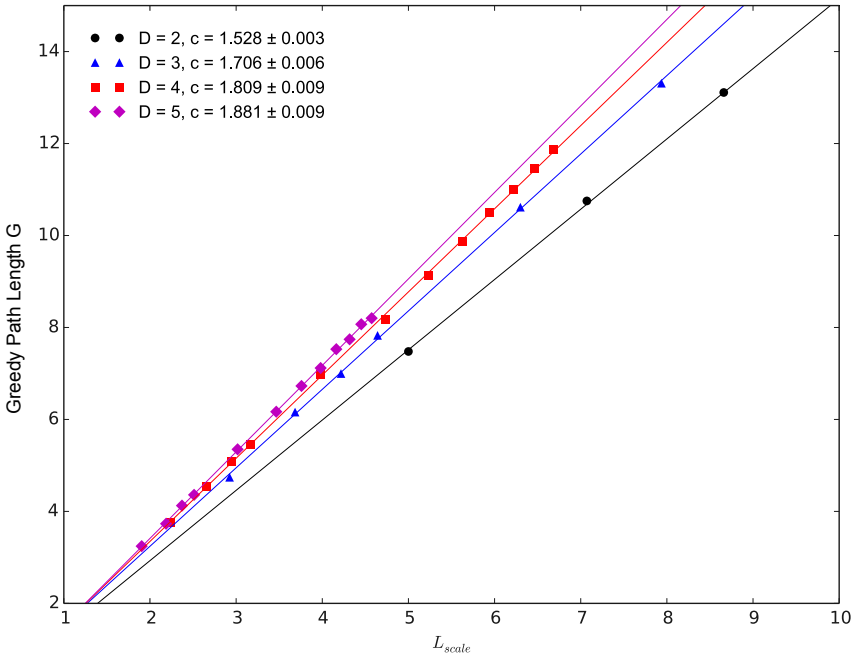


(a) Cube Space

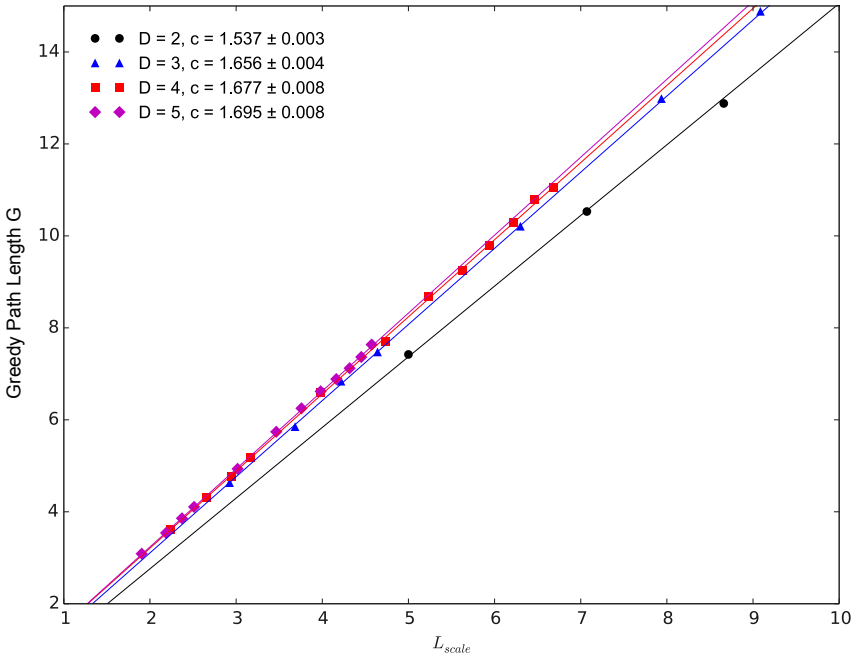


(b) Minkowski Space

Figure A.6.: Longest path length L as a function of L_{scale} in Cube Space and Minkowski Space, data points from 500 run average. Error bars smaller than symbols. L shows linear behaviour with slope c given in legend. Points should ideally be on one curve, but there's clearly a dimension dependence. Closeup view of low L_{scale} area of figures 5.6a and 5.6b.

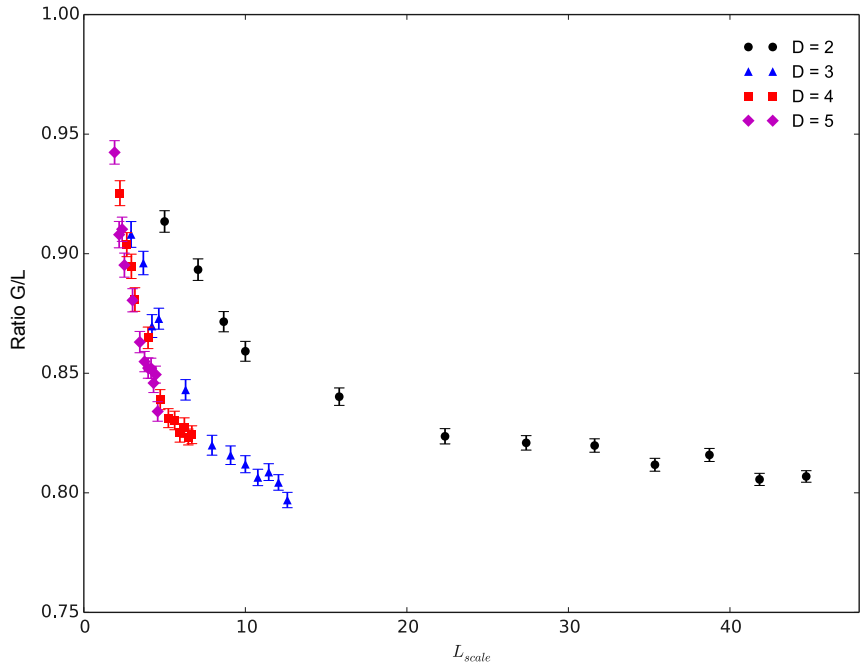


(a) Cube Space

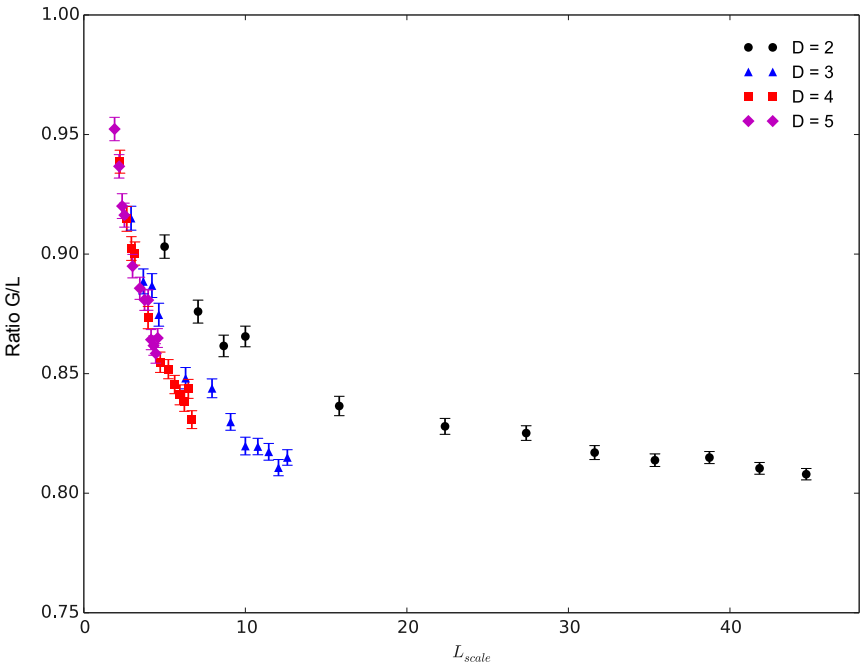


(b) Minkowski Space

Figure A.7.: Greedy path length G as a function of L_{scale} in Cube Space and Minkowski Space, data points from 500 run average. Error bars smaller than symbols. G shows linear behaviour with slope c given in legend. Behaviour almost identical to L, but with smaller slope. Closeup view of low L_{scale} area of figures 5.7a and 5.7b.



(a) Cube Space



(b) Minkowski Space

Figure A.8.: Ratio of greedy path length G to longest path length L as a function of L_{scale}. Data from 500 averaged runs. Error bars show standard error of measurement. Most data points for too small L_{scale} to see any significant behaviour with the exception of $D = 2$.

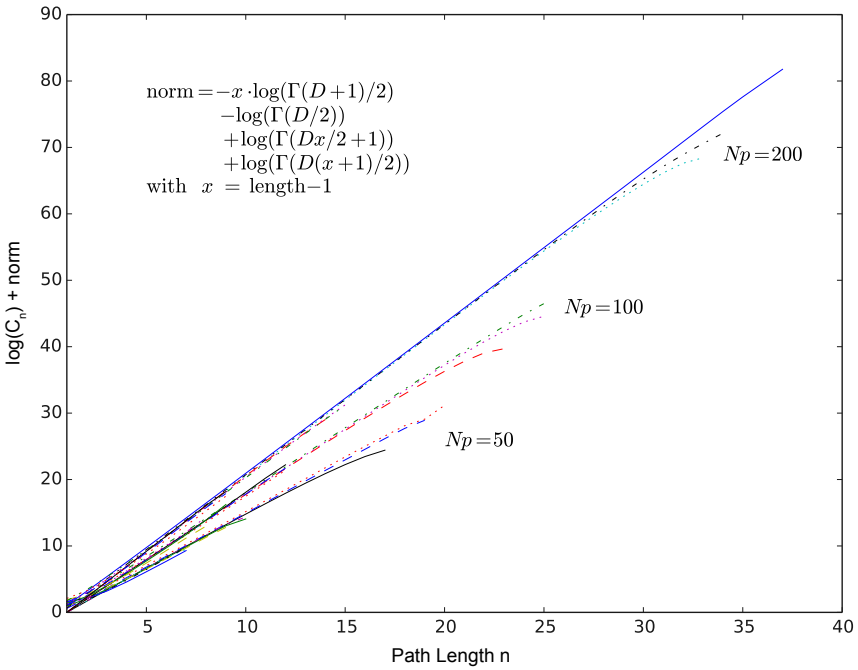


Figure A.9.: Plot of $\log(C_n(x)) + \text{norm} = (n-1)\log(Np)$ in Minkowski Space, data from 1000 run average. 9 different distributions for each Np with $D \in \{2,3,4\}$, $p \in \{1.0, 0.5, 0.1\}$ and $N = Np/p$. Error bars omitted for better readability. Data collapse indicates that only the product Np is an independent variable.

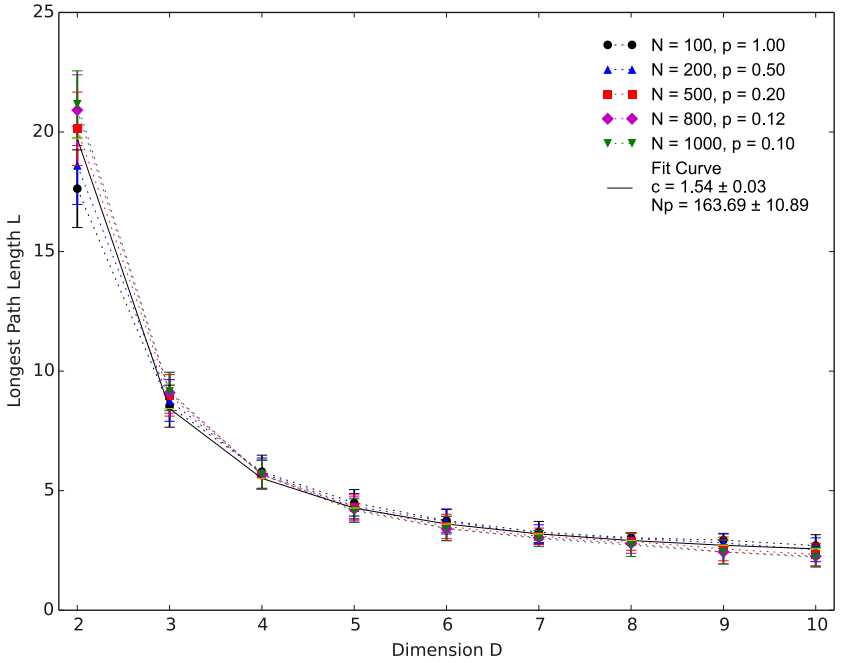
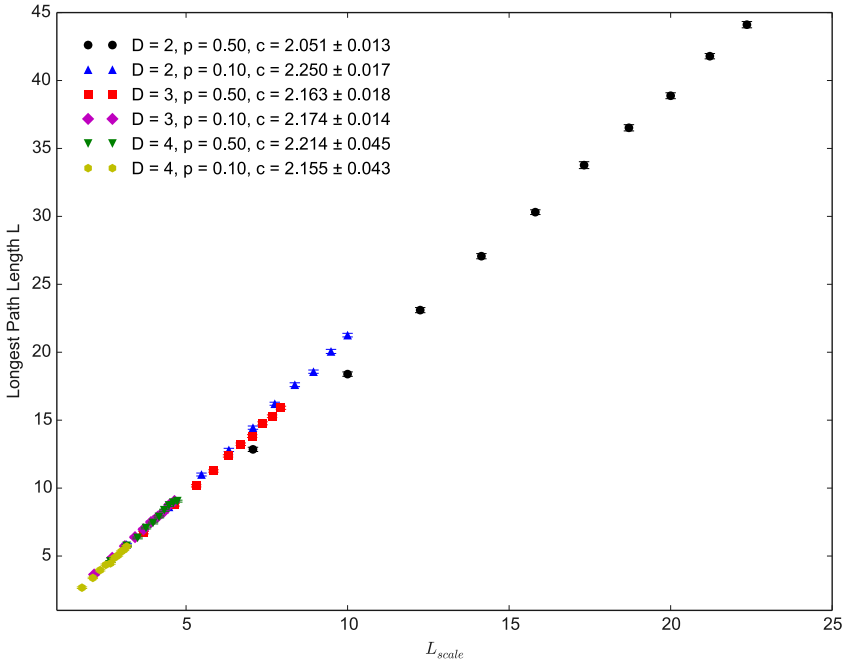
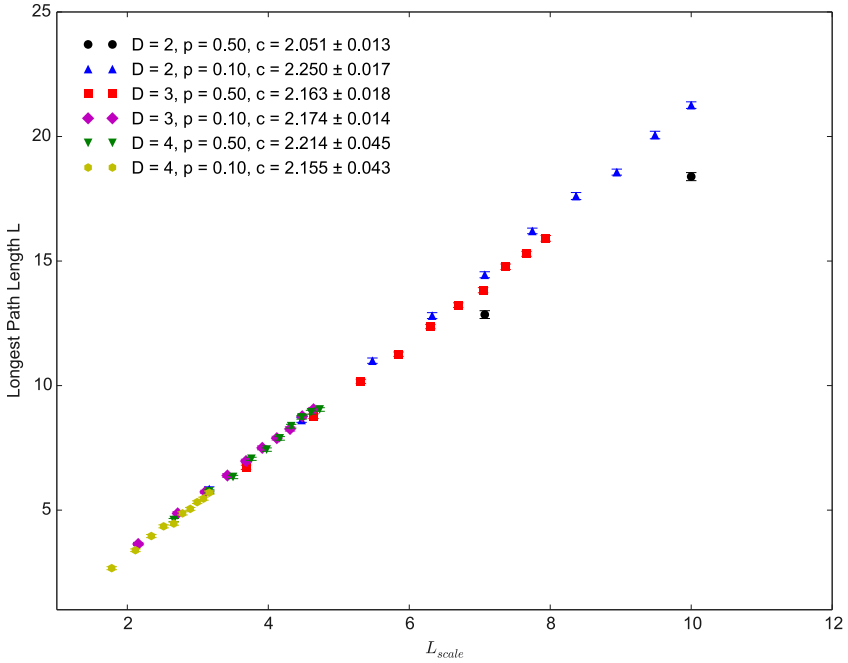


Figure A.10.: Longest Path L as a function of D for different pairs of N,p with $Np = 100$ in Minkowski Space, data averaged over 500 runs. Error bars show standard error of measurement. Fit function $c \cdot (Np)^{1/D}$ with c, Np variables. Fit does not return $Np = 100$.



(a) Full view



(b) Close-up for small L'_{scale}.

Figure A.11.: L as a function of the modified scale length $L'_{scale} = (Np)^{1/D}$ in Minkowski Space, data averaged over 100 runs. Error bars show standard error of measurement. Linear fit results given for each set, but line is not displayed. L exhibits linear behaviour with small dependence on p, but roughly constant slope except for $D = 2$.

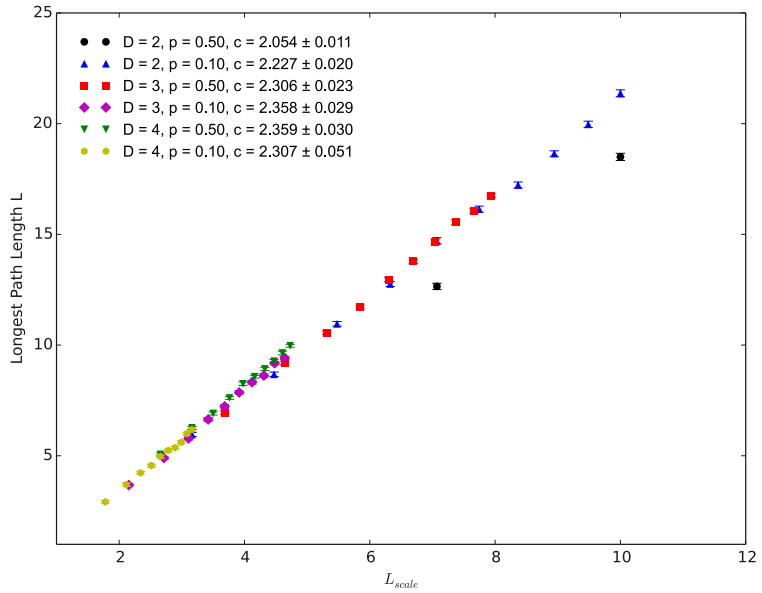


Figure A.12.: Closeup view of figure 6.3: L as a function of the modified scale length $L'_{scale} = (Np)^{1/D}$, data averaged over 100 runs. Error bars show standard error of measurement. Linear fit results given for each set, but line is not displayed. L exhibits linear behaviour with small dependence on p, but roughly constant slope except for $D = 2$.

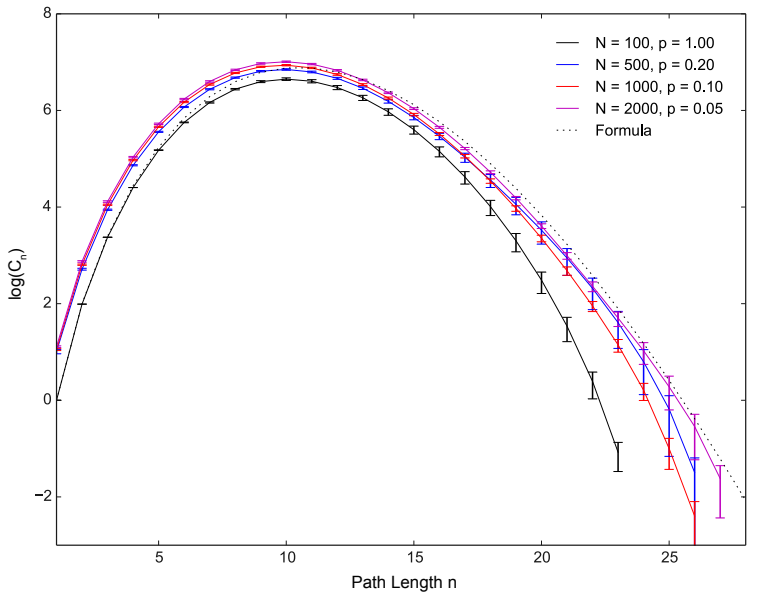
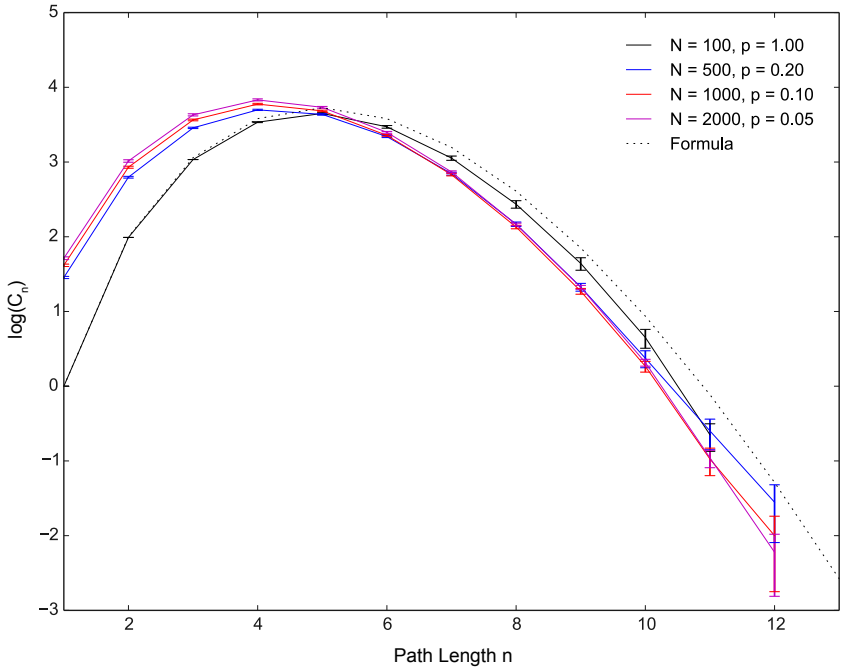
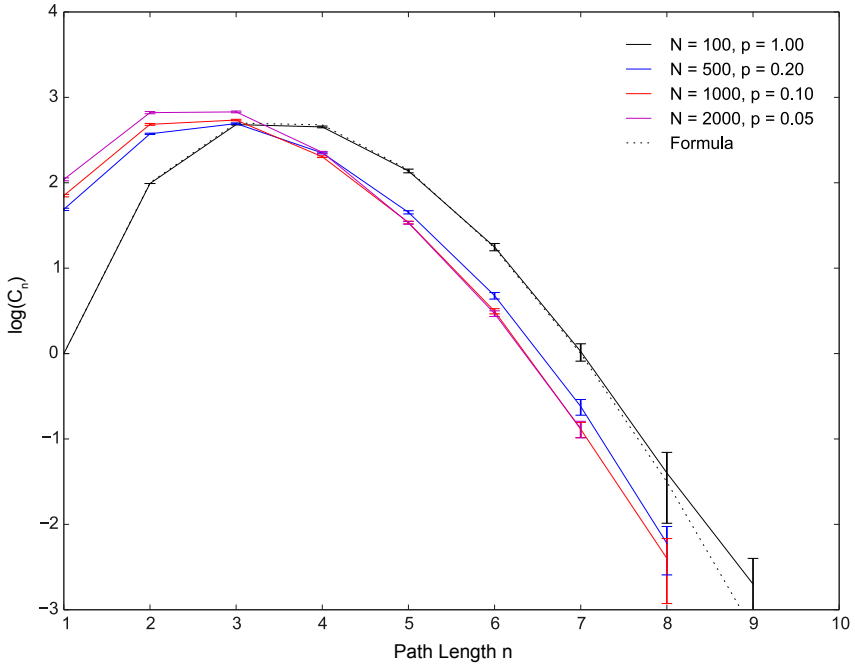


Figure A.13.: Path Length Distributions in two-dimensional Minkowski Space with $Np = 100$, data averaged over 500 runs. Error bars show standard error of measurement. For $p < 1$ equation (3.26) underestimates number of short paths, but overestimates number of long paths, as seen for $p = 1$. The total number of paths appears to increase slightly with N.



(a) $D = 3$



(b) $D = 4$

Figure A.14.: Reproduction of figure A.13 with $D = 3, 4$, data averaged over 500 runs. Error bars show standard error of measurement. For $p < 1$ equation (3.26) underestimates number of short paths, but overestimates number of long paths, as seen for $p = 1$. The total number of paths appears to increase slightly with N.

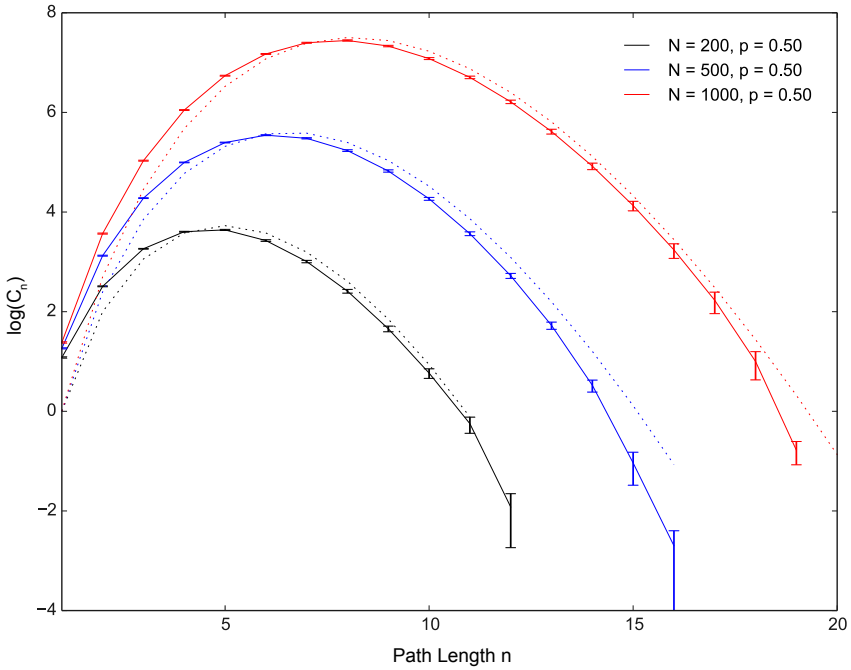


Figure A.15.: Comparison of measured and predicted (dotted) distributions in three-dimensional Minkowski Space with $p = 0.5$ and $N = 200, 500, 100$, data averaged from 500 runs. Error bars show standard error of measurement. Discrepancy between data and formula does not change when N is increased.