Random Walks: Diffusion Limited Aggregates

Matthew Rollings

December 13, 2019

Abstract

Random walks are heavily studied mathematical objects with a nature corresponding to that of a stochastic process. In this investigation a program was created to analyse the process of simple symmetric random walks and the interesting analytical features associated with them. This was done in up to 3 dimensions, with 1 dimensional distributions of the distance a walk travels found to follow that of a Gaussian distribution, while 2 and 3 dimensional distributions of the radial distance a walk travels were found to follow that of a Rayleigh distribution.

1 Introduction

A random walk is a general type of random process involving taking a series of steps; the direction of which is determined probabilistically. The succession of these steps on a mathematical space creates a path, or mathematical object, referred to as a stochastic process (a collection of random events). The most basic example of a random walk can be found using the integers as a number line centred on zero and flipping a coin to determine which direction to move. In this case, each step has equal probability of $\frac{1}{2}$, of being a move of positive one or negative one from the starting point.

The possible applications to the real world that random walks have is innumerable. Random walks are used in all aspects of life, ranging from modelling biological systems, notably in ecology (animal movements) and pathophysiology (cell movements) [1], to being used as web-crawlers in search engines [2]. Even the simplest example of a random walk described previous is an example of a Markov chain, with each integer representing a different state. In physics, the most significant relation comes in modelling the random motion of particles suspended in a fluid (or gas) as a result of their collisions with the fast-moving molecules in that fluid (or gas), otherwise known as Brownian Motion. This application of a random walk to describe Brownian Motion is referred to as diffusion limited aggregation.

Self-avoiding random walks is an extension of the idea used in random walks with the added complexity of not being allowed to return to a position on the lattice the path has already passed through. This characteristic provides a use for random walks in more biological processes such as topological and knot-theoretic behaviour in proteins [3]. Although, many of the properties self-avoiding random walks express are not yet extensively known, meaning this is still a topic of high interest in the world of mathematical programming and science.

This leads on to the investigation nicely, in which the programming of multiple types of random walks is attempted in different numbers of dimensions. Followed by the analysis of the qualitative and quantitative properties that these walks possess, with comparisons to current, accepted literature.

2 Background Theory

The most common random walk to model is that on a regular lattice, where at each step the position jumps to another lattice site according to a probability distribution. In a simple random walk, like the one investigated in this project, the position can only jump to a neighbouring lattice site thus forming a lattice path. When the probabilities of the position passing to each of its immediate neighbours are the same, the random walk is considered symmetric. If the path is limited to a finite number of possible positions, the random walk is referred to as a simple bordered symmetric random walk. This means the transition probabilities are now dependent on the location of state, with each of the lattice movements on the borders no longer having the same probability.[4]

The most highly-investigated example of the random walk is on the d-dimensional integer lattice (otherwise known as the hypercubic lattice) $\mathbb{Z}^d$. The most elementary example of this integer lattice is the 1-dimensional lattice (mentioned briefly in the introduction), creating a random walk on the integer number line, $\mathbb{Z}$. The premise of this random walk is that the path starts at 0 and moves +1 or -1 at each step with equal probability; a decision easily replicated by the flipping of a fair coin. Thus meaning a landing on heads represents a move of 1 to the right on the integer number line and a landing on tails represents a move of 1 to the left.

In this investigation, a value for the average distance travelled by a random walk after N steps (where N is any number) is calculated. To start the theory behind it, the random walk must be defined formally. To do this, the independent random variables are taken to be $Z_1, Z_2...$, with each variable representing either +1 or -1, with 50% probability for each value. The series representing the simple random walk on $\mathbb{Z}$ can therefore be defined as

$$S_N = \sum_{j=1}^{N} Z_j, \tag{1}$$

where $S_0 = 0$ and the value of S_N will be the distance the walk is from S_0 after n steps. From this, the expectation value for S_N can be found as it will be the sum of the expectation values for each independent step Z_j. As $E(Z_j) = 0$; found from the probability of each move +1 or -1 being 50% and therefore having an average of 0; then

$$E(S_N) = \sum_{j=1}^{N} E(Z_j) = 0. \tag{2}$$

This is not particularly surprising, as $E(S_N)$ is the average location after N steps and since each step is equally likely to move forwards or backwards, it is expected the distance moved should be 0, on average.

As valuable as proving the above is, it has not really given much useful information as to how far away the random walk gets from its origin after N steps. So in order to get some useful information, the expectation value for the square of the distance is found and, since the square of +1 and -1 are both positive and 1, it can no longer average out to 0.

$$E(S_N^2) = (\sum_{i=1}^{N} E(Z_i))(\sum_{j=1}^{N} E(Z_j)) = \sum_{i=1}^{N} E(Z_i^2) + 2 \sum_{1 \leq i < j \leq N}^{N} E(Z_i Z_j) \tag{3}$$

Since Z_i can only be +1 or -1, the expectation value of the square of the independent random valuables is $E(Z_i^2) = 1$. Now, the $Z_i Z_j$ term is considered and there are only 4 possible combinations of Z_i and Z_j that can be made, each with equal probability:

Z_i	Z_j	$Z_i Z_j$
1	1	1
1	-1	-1
-1	1	-1
-1	-1	1

Therefore $Z_i Z_j$ has an equal chance of being +1 or -1 so will have an expectation value of 0, $E(Z_i Z_j) = 0$. Since this is true for any combination of i and j, a value for the expectation value for the square of the distance has been found.

$$E(S_N^2) = N \tag{4}$$

This means the average of the square of the distance is equal to the number of steps, N. If the square root of this equation is taken

$$\sqrt{E(S_N^2)} = E(|S_N|) = \sqrt{N} \tag{5}$$

The $\sqrt{E(S_N^2)}$ term is actually the root-mean-squared distance, representing the average positive distance away from 0 after N steps. Thus it is expected that after N steps, the walk should be, on average, $\sqrt{N}$ steps from where it started. In fact as N tends to infinity,

$$\lim_{N \to \infty} \frac{E(|S_N|)}{\sqrt{N}} = \sqrt{\frac{2}{\pi}} \tag{6}$$

The number of different walks of length N steps, where each step is +1 or -1, is 2^N with each path being equally likely. For S_N to be equal to a number k, the number of +1 variables must exceed that of the -1 variables by k. In order for this to happen +1 must appear $\frac{N+k}{2}$ times throughout the walk. Hence the number of walks which satisfy $S_N = k$ is equal to the number of ways of choosing $\frac{N+k}{2}$ elements from a set of N elements. This is represented as $\binom{N}{\frac{N+k}{2}}$. The conditions required for this to be true is that $N + k$ must be an even integer, meaning N and k both have to be even or both have to be odd. This makes the probability that $S_N = k$ is

$$2^{-N}\binom{N}{\frac{N+k}{2}}. \tag{7}$$

Good estimates for this probability, for large values of N, can be found using Stirling's approximation.

The central limit theorem (CLT) states that, even if a probability distribution is not normal, the distribution of the mean values of samples from the distribution will approximate a normal distribution for a large sample size[5]. In reference to our investigation, this means that the probabilities for $S_N = k$ will tend to the normal distribution, as N increases. Thus meaning the distribution of distances from the origin should fit a Gaussian curve with a mean, μ, of 0 and a variance, σ^2, corresponding to the equation:

$$P(x) \propto A \exp \frac{-x^2}{\sigma^2}. \tag{8}$$

For the case of the 1 dimensional random walk, with each step the walk makes a time, t, passes, so it follows that the variance of the walk should increase with time also. The proof for this comes from the idea that each step of the walk is stochastic and therefore its variance is statistically independent of all other steps. This is derived below:[6]

$$\begin{aligned} Var(R_t) &= Var(R_1 + R_2 + R_3) + ... \\ &= Var(R_1) + Var(R_2) + Var(R_3) + ... \\ &= \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + ... \end{aligned} \tag{9}$$

where R represents a 1 dimensional random walk and $R_1, ..., R_t$ is said random walk at time, t.

This is almost true for higher dimensions too. For example the case of the two dimensional random walk can be described through the superposition of the two 1 dimension lattice paths. For the square lattice used in this investigation, that is represented by the x and y movements through the lattice space. As each directional movement is simply representative of a one dimensional symmetric walk, with each step direction being of equal probability, each x and y distance follow a normal distribution curve for large n, similar to how the 1 dimensional walk did previously. To

superimpose the two lattice parameters, a change in basis to polar coordinates is required, where $r^2 = x^2 + y^2$. This transformation is performed on the probability distribution through the substitution below.

$$dxdy = rdr \tag{10}$$

$$P(r) \propto A \exp \frac{-x^2}{\sigma^2} \exp \frac{-y^2}{\sigma^2} dxdy \propto Ar \exp \frac{-r^2}{\sigma^2} dr \tag{11}$$

Equation 11 actually takes the form of a Rayleigh function, representing the probability distribution of the radial distance, r, from the origin at $x = y = 0$. The nature of the radial calculation, in squaring the component coordinates, means that this will be a positive distribution, with the mean and variance of a Rayleigh random variable is calculated using the following relationships:

$$\mu = 1.253\sigma \tag{12}$$

$$Var = 0.429\sigma^2 \tag{13}$$

3 Algorithm

The first step in this investigation was programming the random walk numerically for the steps and dimensions specified in the parameter setting section, from 1 to 3 dimensions. This code can be found in 'RW.py'. The walk starts at the origin, x=y=z=0, and operates on a square lattice where, for each step n, the path could move to any of the six neighbouring lattice sites with equal probability (up, down, forward, backward, left or right). To implement this randomisation, the in-built Python pseudo-random number generator 'random' was used, and more specifically the 'randint' function. Python uses the Mersenne Twister as the core generator for the random numbers which is one of the most extensively tested random number generators in existence [7]. The random numbers generated corresponded to a translation of +1 or -1 along one of the three axes at each step, n. These were simultaneously indexed into an ndarray, called 'data', recording the position of the walk on the square lattice after each step. Looping over this n times produces the length n, 1, 2 or 3 dimensional random walk across the square lattice.

To evaluate the distance between all pairs of points at a fixed number of steps apart, another for loop was implemented to minus the distance from the origin of one point on the walk from another point of steps 'fixed_step' further on in the walk. This was enough to find the distance moved for the 1 dimensional walk but for 2 and 3 dimensions a further calculation was required to find the radial distance moved. This was done using simple Pythagoras. This means that each ndarray of distances will be 'steps-fixed_step' long so a mean value was found for the distance at each fixed step size and this value was plotted in the Rayleigh distributions in 'Rayleigh.py'. The expected Rayleigh distributions were plotted using the 'scipy.stats.rayleigh.pdf' function which takes the idealised parameters for the data and plots the expected results. The 1 dimensional distributions in the 2 and 3 dimensional walks (meaning the distance moved along the coordinate axes) are also recorded and can be plotted using the 'Gauss.py' program. The expected Gaussian distributions were plotted in a similar way to the Rayleigh distribution except the 'scipy.stats.norm.pdf' function was used.

To assess the assertion that the variance scales as N for N steps, the radial data was used. By increasing the fixed step size between which the distances were calculated, the time the walk was active for was analogously increased. So the variance was found for the distance travelled at each fixed step size and plotted against its respective fixed step size to find the relationship between the two.

4 Results and Analysis

To start with the 2 dimensional symmetric random walks, shown in Figures 1 and 2, have been analysed as they show the square lattice structure the most clearly. For both low and high step random walks, self-symmetry is clearly visible. The time taken for the program to generate the random walks was relative to the number of steps it was required to calculate. This comes as a result of the while loop used to iterate through the steps is only dependent on the total number of steps.

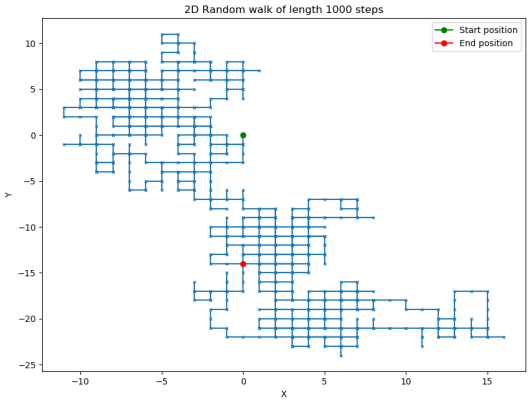

Figure 1: 2D walk of length 1000 steps

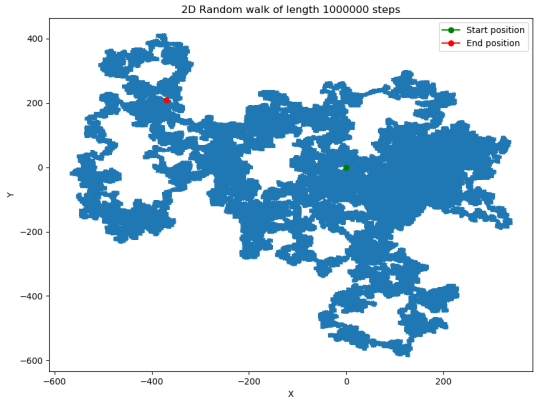

Figure 2: 2D walk of length 1,000,000 steps

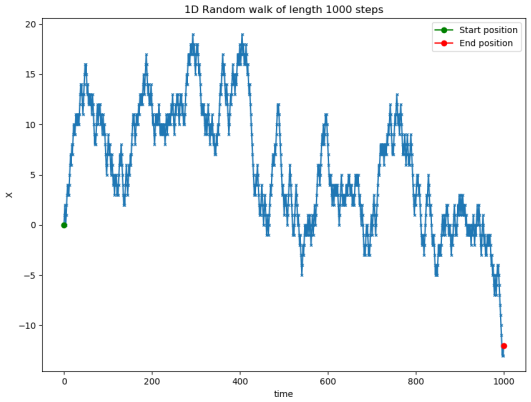

Figure 3: 1D walk of length 1000 steps

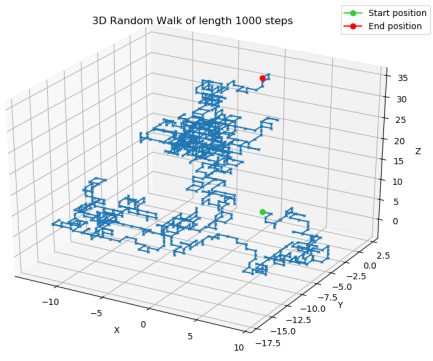

Figure 4: 3D walk of length 1000 steps

The 1 and 3 dimensional symmetric random walks, shown in Figures 3 and 4, both took a similar time to execute as the 2 dimensional walk shown in Figure 1. This is because the random number generation occurs in only one dimension at a time, to achieve the square or cubic lattice, so no more numbers are calculated in any number of dimensions.

From 1 through 3 dimensional random walks, the single directional components (coordinate axes components) were analysed. The data found proved that as the number of steps was increased, the component distance from the origin followed a normal distribution more and more accurately, as shown in Figures 5 and 6. The inaccuracy of the distribution as it approaches zero comes in having to round the distances to the nearest lattice site meaning the normalised count for each integer includes any distances within ± 0.5 from it. This rounding had to be implemented due to the distances being a mean of multiple distances measured for each fixed step and therefore the values could be very specific (to lots of decimal places) and would result in almost all distances being found only once and therefore not following the distribution as desired.

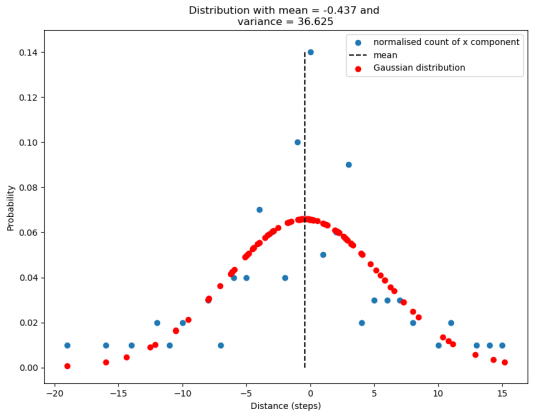

Figure 5: 1D x-component for 100 steps

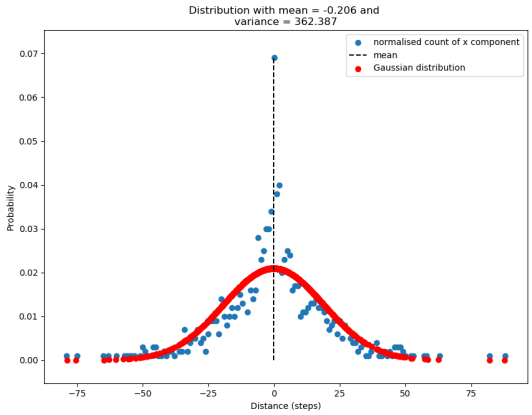

Figure 6: 1D x-component for 1000 steps

A similar pattern was found for the component axes of the 2 and 3 dimensional walks, as shown in Figures 7 and 8. In these cases, the distances varied a little more in comparison to the appropriate Gaussian distribution. This could be a result of the cubic lattice being used for the path of the walk. The cubic lattice means that the walk can only move in one of the x, y or z directions at each step so there would actually be a probability of the component to not move at each step which has not been taken into consideration in the formation of the expected Gaussian distribution lines. The expected value for the mean is ≈ 0 as shown by how the distributions are centred around it, which agrees with the predictions made in the theory section.

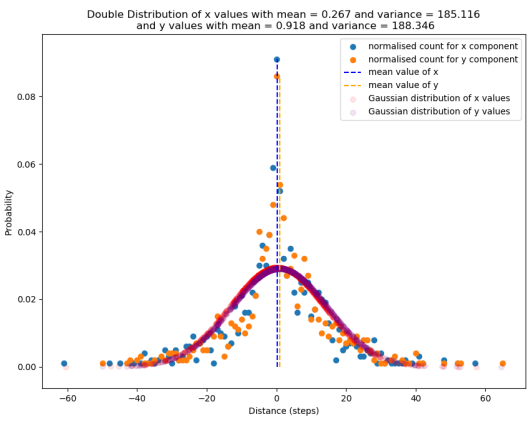

Figure 7: 2D axes-components for 1000 steps

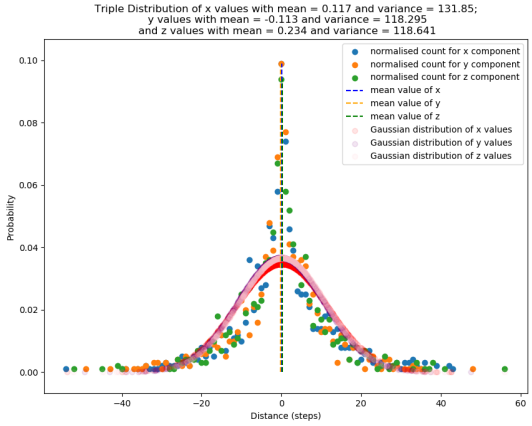

Figure 8: 3D axes-components for 1000 steps

In contrast to the component axes distances, the radial distances found for the 2 and 3 dimensional random walks follow a Rayleigh distribution, as can be seen in Figures 9 and 10. As the number of steps increased, the distribution of distances became closer to the theoretical distribution; shifting towards 0. This means that the walks were, on average, of radial distance closer to 0 for higher numbers of steps, therefore confirming the central limit theorem. This also suggests

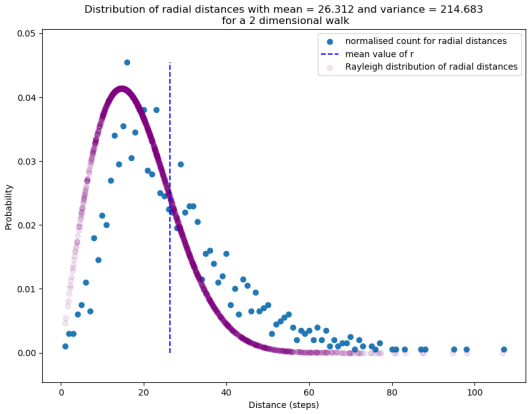

Figure 9: 2D Rayleigh distribution, 2000 steps

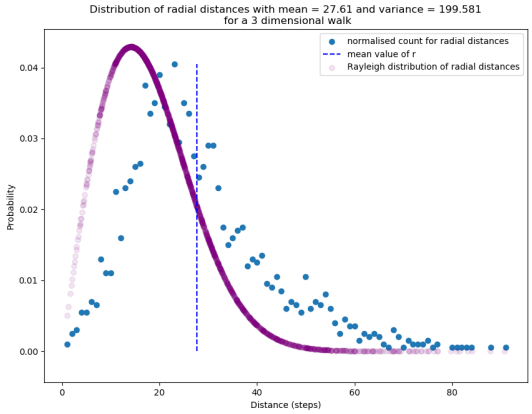

Figure 10: 3D Rayleigh distribution, 2000 steps

that the radial distances should tend to the Rayleigh distribution at high values of N but this was not able to be analysed due to the process taking hours to execute. This comes from inefficiency when calculating the radial distances, having to calculate the distances between each fixed step for

each coordinate axes then perform Pythagoras for each step. A computer with a higher number of cores or a better processor (or maybe a better way of coding the calculation although I could not think of one within the time from of this investigation) might be able to execute this faster but it was decided that the desired effects could be seen for a low enough number of steps that this was not necessary.

For the 1 dimensional Gaussian distributions the variance was coded to behave in accordance with equation 9 as the step size was increased. It was read and theorised that this and the 2 dimensional Gaussian distribution was supposed to give a linear relationship where the variance scaled as N for N steps. However, as is clear from Figures 11 and 12, this was not the case for either of the 1 dimensional or 2 dimensional random walks. In fact it appears that the variance is about a factor of 100 too large which suggests this is very much an issue with the implementation of the variance calculation. And, due to time constraints, this was unfortunately not able to be corrected.

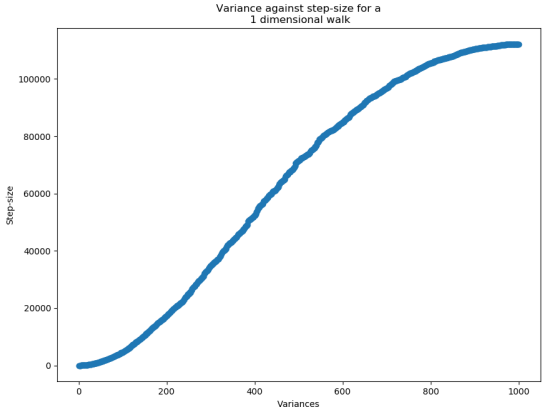

Figure 11: Variance relationship for 1D

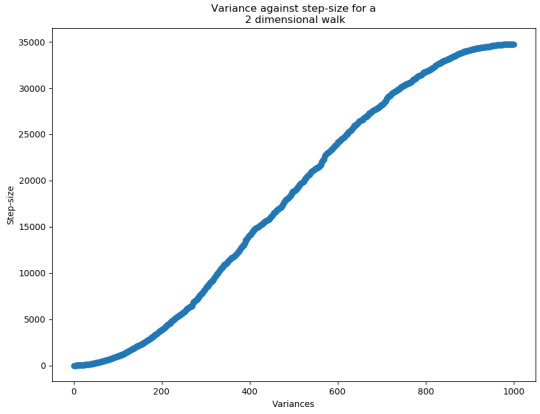

Figure 12: Variance relationship for 2D

5 Conclusion

After a thorough investigation into the behaviour of simple symmetric random walks, it was found that the distance travelled from the origin tends to well defined distributions. For 1 dimensional distributions of the coordinate axes this was found to be the Gaussian distribution and for 2 and 3 dimensional distributions of the radial distances this was found to be the Rayleigh distribution. Furthermore this conformity became more rigid as the number of steps the walks were executed for increased. This suggests that the stochastic nature of the random walks is the cause as the accuracy of the distances is increasing as N increases.

Although the expected relationship between the variance and step size was not established, the increasing variance as step size increases is still proof of the stochastic nature of the random walk as the variance on each step is clearly increasing with time, as theorised. I believe the issue with the variance calculations may have come in computing the distance calculations for every possible pair of points on the random walk, rather than choosing the end of the previous step to be the start of the next. This means that each distance calculated for the step size is not totally independent of the other.

6 Acknowledgements

Thank you to Alan Watson, Juraj Bracinik, Roman Lietava, Juergen Thomas, Angus Hollands, Christopher Oliver and Walter Van Rossem.

References

[1] Michael J Plank Edward A Codling and Simon Benhamou. “Random walk models in biology”. In: *Journal of The Royal Society Interface* (2008).

[2] Sergey Brin and Lawrence Page. “The Anatomy of a Large-Scale Hypertextual Web Search Engine”. In: *Computer Networks and ISDN Systems* (1998).

[3] Rahul K. Das and Rohit V. Pappu. “Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues”. In: *Proceedings of the National Academy of Sciences of the United States of America* (2013).

[4] *Random Walk*. URL: `https://en.wikipedia.org/wiki/Random_walk`.

[5] Andrew P. King and Robert J. Eckersley. “Inferential Statistics I: Basic Concepts”. In: *Statistics for Biomedical Engineers and Scientists* (2019).

[6] Oliver C. (Oliver Chukwudi) Ibe. *Elements of random walk and diffusion processes / Oliver C. Ibe.* eng. Wiley series in operations research and management science. 2013.

[7] M. Matsumoto and T. Nishimura. “Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator”. In: *ACM Transactions on Modeling and Computer Simulation Vol. 8, No. 1* (January 1998), pp.3–30.