BACHELOR

Using Random Graphs to Model the Network of Countries

van Veldhuizen, Aron

Award date:
2023

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science

Using Random Graphs to Model the Network of Countries

Bachelor's Thesis

Aron van Veldhuizen

Supervisors:
Benoît Corsini
Rowel C. Gündlach

Final version

Eindhoven, July 2023

Abstract

In this project, we analyze the network of bordering countries and territories of the world, which we refer to as the *worldgraph*. We attempt to identify a random geometric graph model with properties that match those of the worldgraph in expectation, focusing our attention on the number of edges, the number of triangles, and the degree distribution.

In this thesis, we look into three families of graphs: the ε-neighborhood graph, the k-nearest-neighbors graph, and neighborhood graph models. We describe these models and prove theoretical results concerning their properties. Using stochastic simulation, we generate sample graphs of all models, and we analyze the effect of certain parameters on the properties of the graphs.

We conclude that the worldgraph can be best described by the ε-neighborhood graph model in terms of the number of edges, number of triangles, and average degree distribution. Alternatively, the k-nearest-neighbors graph model with maximal edge length is a good visual model for the worldgraph, as the triangulation pattern observed in the worldgraph is more evident here. We discuss reasons why the neighborhood graph models we studied are less suitable as models for the worldgraph, and we hypothesize a few improvements that can be made.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Consider the set V of all countries in the world and the set E of all links between a country and its neighbors, that is the countries with a shared border. We identify the pair (V, E) with a network of worldwide countries G which we call the *worldgraph*. Figure 3.1 in Chapter 3 shows a representation of this network. A quick glance can reveal a lot of interesting facts about this graph: Russia and China have many neighbors, Africa has a lot of triangles, and Europe has lots of very small links. In this report, we attempt to approximate this particular graph with random graphs, by using geometrical rules that tell us how to link nodes together. We can summarize the problem we are trying to solve as follows:

Can we define a random graph model embedded on the surface of the Earth, that generates random graphs with properties (e.g. degree distribution, number of edges, number of triangles) matching those of the worldgraph in expectation?

Our strategy to answer this question is as follows. First, we analyze some relevant properties of the worldgraph we want to recreate. Second, we define how we model the Earth, generate random nodes on its surface, and compute distances between nodes. Third, we select a few random graph models that we suspect to resemble the worldgraph. Finally, we generate a great number of samples for each model and compare their behavior to that of the worldgraph. We complete our experiments with a theoretical analysis of the models to validate our results and eventually pinpoint the best model to approximate the worldgraph.

The remainder of this report is organized as follows. In Chapter 2 we provide some basic definitions, along with a glossary defining some less common terms and notation used throughout the report. In Chapter 3 we define the *worldgraph*. In particular, we specify the set of countries and territories included as its nodes. We further list some basic properties of this graph. In Chapter 4 we discuss some geometrical models for the Earth and provide reasons why we chose the spherical model. We define the distance function on the sphere and explain our sampling method.

Having covered all this groundwork, Chapter 5 allows us to introduce a variety of random graph models, provided with concise descriptions and sample plots of each model. Specifically the class of *neighborhood graphs* is introduced, along with some relationships between these graph models. Chapter 6 is dedicated to proving some preliminary functions useful for theoretical calculations, then used to show results on the ε-neighborhood graph model and on neighborhood graphs. In Chapter 7 we present and analyze simulated results for all random graph models. Chapter 8 contains a list of potential improvements and extensions to this project, such as other random graph models, node distributions, other real-life networks to recreate, and theoretical results left to prove. Finally, in Chapter 9 we provide a conclusion, discussing which graph models best suit our search for a model of the worldgraph, and why we chose them.

Chapter 2

Notations

A *graph* (also *network*) is a pair $G = (V, E)$, where V is a set of elements called *vertices* (also *nodes* or *points*), and $E \subseteq V \times V$ is an unordered set of paired vertices called *edges* (also *links* or *connections*). A *simple graph* is a graph where all edges are distinct, and no edge connects a vertex to itself. A *planar graph* is a graph that can be drawn on the plane in such a way that no edges cross. The *degree* of a vertex $v_i \in V$ is the number of edges it is part of. A *triangle* is a triple of vertices that are pairwise connected. A *subgraph* of a graph $G = (V, E)$ is a graph $G' = (V', E')$ such that $V' \subseteq V$ and $E' \subseteq E$. A graph is *connected* when there exists a path of edges from any vertex to any other vertex in the graph. A *connected component* of a graph is a connected subgraph that is not part of any larger connected subgraph. The *complete* graph K_n is the graph on n vertices, where all vertices are pairwise connected. The *complete bipartite graph* $K_{m,n}$ is the graph with two sets of vertices, one of size m and one of size n, where each vertex in one set is connected to each vertex in the other set.

Term	Definition
$G = (V, E)$	Graph G composed of a set of vertices V and a set of edges $E \subseteq V \times V$
Insular node	Node with degree 0
Δ	Number of triangles
$r_\oplus$	Arithmetic mean of Earth's radius, approximately 6 371.009 km
$\mathbb{S}^2_\oplus$	Earth's surface, modelled as a sphere of radius $r_\oplus$
$\text{dist}(v_i, v_j)$	Distance in km between points v_i and v_j along Earth's surface modelled as $\mathbb{S}^2_\oplus$
$\mathcal{B}_\oplus(v, \varepsilon)$	Set of points on $\mathbb{S}^2_\oplus$ within ε km from the point v
$\mathcal{N}(v_i, v_j)$	(5) Neighborhood of the edge (v_i, v_j)
εN	(5.1) ε-neighborhood graph
kNN	(5.2) k-nearest neighbors graph
ε-kNN	(5.2) k-nearest neighbors graph, max. edge-length ε km
DT	(5.3) Delaunay triangulation, dual of the Voronoi diagram
RNG	(5.4) Relative neighborhood graph
MST	(5.4) Minimal spanning tree
λ-RNG	(5.4) Generalized relative neighborhood graph
βS	(5.5) Beta skeleton graph
GG	(5.5) Gabriel graph
d_{TV}	Total variation distance of two probability distributions

Table 2.1: Glossary of terms and abbreviations.

Chapter 3

The worldgraph

In this chapter, we take a first look at the worldgraph. We start by precisely defining the set of nodes, composed of countries and territories around the world. Then we look into some simple properties of the graph, including its degree distribution and planarity.

Defining the worldgraph

The *worldgraph* is the graph $\mathrm{WG} = (V_{\mathrm{WG}}, E_{\mathrm{WG}})$, composed of the set of vertices V_{WG} corresponding to a selection of countries and territories in the world (to be specified), and a set of edges E_{WG} consisting of all pairs of countries/territories who share a border. Every vertex in V_{WG} is coupled with a pair of coordinates, which correspond to the *centroid* (that is, the arithmetic mean or the center of gravity) of the associated area of the world[1]. The set of edges E_{WG}[2] contains mostly land borders, though it does also connect a few countries separated by a relatively small stretch of water, e.g. France and the United Kingdom. See Figure 3.1 for a world map showing the embedding of the worldgraph.

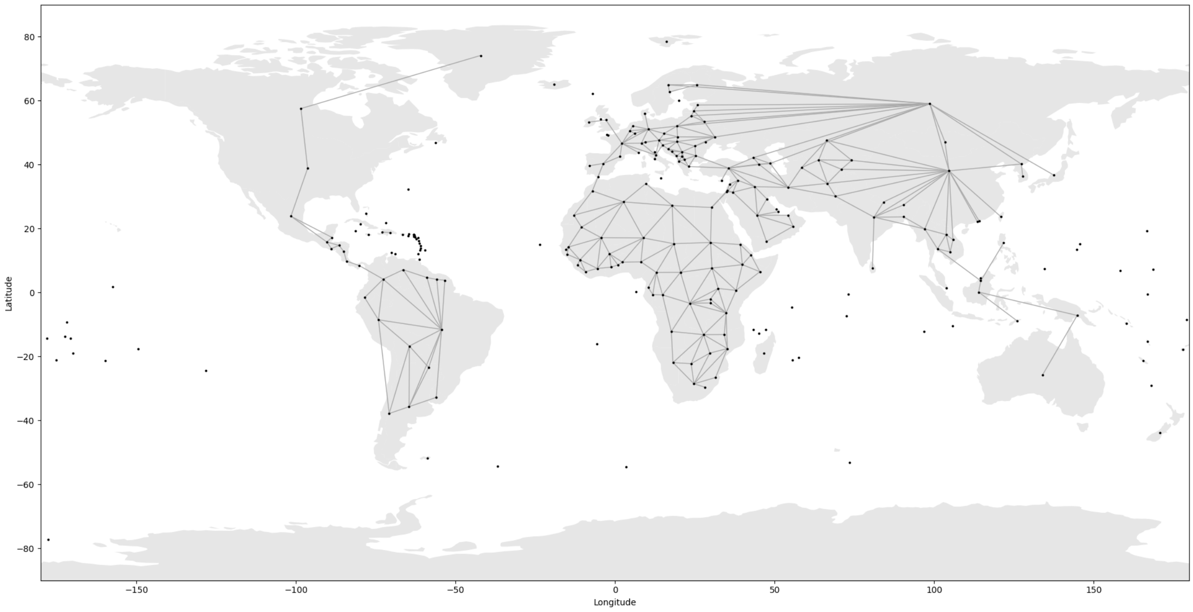


Figure 3.1: Worldgraph embedded on the world map.

[1] github.com/gavinr/world-countries-centroids/blob/master/dist/countries.csv
[2] github.com/geodatasource/country-borders/blob/master/GEODATASOURCE-COUNTRY-BORDERS.CSV

Defining the set of countries and territories

It is important to have a clearly defined set of countries that is consistent throughout the research. This set of worldwide countries is not exactly set in stone: a lot of territories are recognized as sovereign, independent countries by some number of countries or institutions, and not recognized by others (e.g. Kosovo and Northern Cyprus). It also seems wrong not to include certain autonomous or dependent territories: Greenland is part of the Kingdom of Denmark and about 50 times as big as 'Denmark proper'; French Guyana is a department of France and over 7 000 kilometers distant from Metropolitan France.

Various online databases are loaded and converted into Python dictionary objects. These databases are expected to cooperate in order to produce interesting graphs, but they are heavily inconsistent. As an example, what one database calls 'Laos', another one calls 'Lao People's Democratic Republic'. Luckily all databases include a two-letter country code for each country or territory. The ISO 3166-1 alpha-2[3] standard is used for country codes, which assigns two-letter codes to a total of 249 countries, (e.g. NL for the Netherlands). Though this standard does not include it, we chose to also include Kosovo (code: XK) in this list, since it is recognized by a good number of countries, and many databases include it anyway. This then gives us a total of 250 countries and territories. We refer to this set of countries and territories as V_{WG}, the set of vertices of the worldgraph. If any database contains entries about other territories, they will simply be ignored. If any database lacks entries from this list, they will be added manually with the help of other sources. See Table C.1 in the Appendix for a comprehensive list of all countries and territories.

Planarity

It would be natural to expect that the worldgraph is a planar graph, but this is not the case. More precisely, the worldgraph is very nearly a *map graph*, an undirected graph formed as the intersection graph of finitely many simply connected, internally disjoint regions of the plane. Even more precisely, it is very nearly a *3-map graph*, which means that at most 3 regions can meet at any point on the map. There are cases where a quadripoint border almost exists (two separate tripoints exist about 150 meters apart), but they are not present in the worldgraph. A map graph cannot contain (subdivisions of) the $K_{3,3}$ graph due to the regions being simply connected, and a 3-map graph, in turn, cannot contain (subdivisions of) the K_5 graph: hence the 3-map graph is a planar graph due to *Kuratowski's theorem* [3], stated below.

Theorem 3.1 (Kuratowski's theorem) *Let G be a graph. Then G is nonplanar if and only if G contains a subgraph that is a subdivision of either $K_{3,3}$ or K_5.*

A *subdivision* of a graph G is another graph constructed from G, where any edge can be subdivided in a string of consecutive edges, separated by vertices. Now, the worldgraph does not contain any K_5 subgraphs (or subdivisions thereof), but it does contain one subdivision of a $K_{3,3}$ graph. By Kuratowski it is therefore not a planar graph. The reason for containing this subdivision is an exclave of Azerbaijan called the *Nakhchivan Autonomous Republic* that causes there to be a border between Azerbaijan and Turkey, which in turn completes the $K_{3,3}$ subgraph. Hence the worldgraph does not qualify as a map graph, since Azerbaijan is not a simply connected region. Barring this particular exclave the worldgraph is a map graph and thus also a planar graph. We are therefore still interested in generating planar graphs, as they are meant to model 3-map graphs and can ignore exclaves as a first approximation.

[3] www.iso.org/iso-3166-country-codes.html

Degree distribution

The worldgraph has many insular nodes (that is, nodes without connections to any other nodes) and a large proportion of nodes with degrees between 1 and 5. For degrees higher than 5, the number of nodes slowly tapers off, up to a pair of nodes of degrees 15 and 17. See Table 3.1 and Figure 3.2 below for the precise degree distribution of the worldgraph.

Degree	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**17**
Number	79	26	31	28	26	26	13	10	4	4	1	1	1

Table 3.1: Degree distribution of the worldgraph.

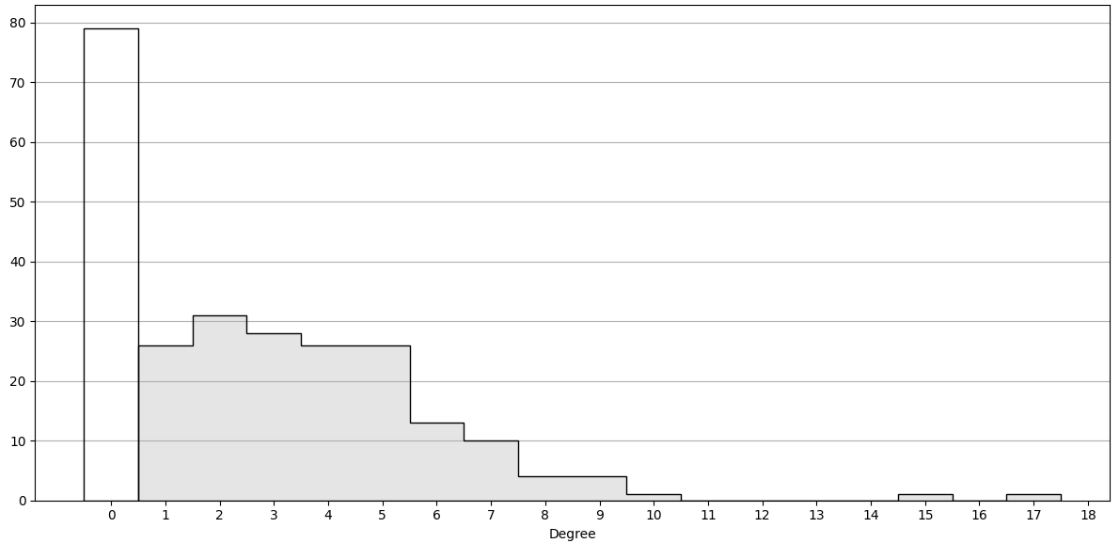


Figure 3.2: Degree distribution of the worldgraph.

From now on, when referring to the worldgraph, we do not include the 79 insular nodes that are part of it anymore. This is because insular nodes are not particularly interesting for the purpose of recreating the network of bordering countries and territories. These insular nodes are generally small islands, except for Antarctica, New Zealand, Madagascar, and a few others. Also, some countries technically surrounded by water are connected to nearby countries, e.g. Australia to Papua New Guinea. Hence we feel that not much information is lost by deleting these nodes from our set V_{WG}, which now reduces to 171 nodes. The new degree distribution is shown as the shaded part of the histogram in Figure 3.2.

The mean of the distribution is approximately 3.883, and the variance is approximately 6.092. Purely from a visual perspective, it doesn't seem too unreasonable to fit a binomial distribution (with matching mean and variance) to the degree distribution. This gives the bell-curve shape observed in Figure 3.3.

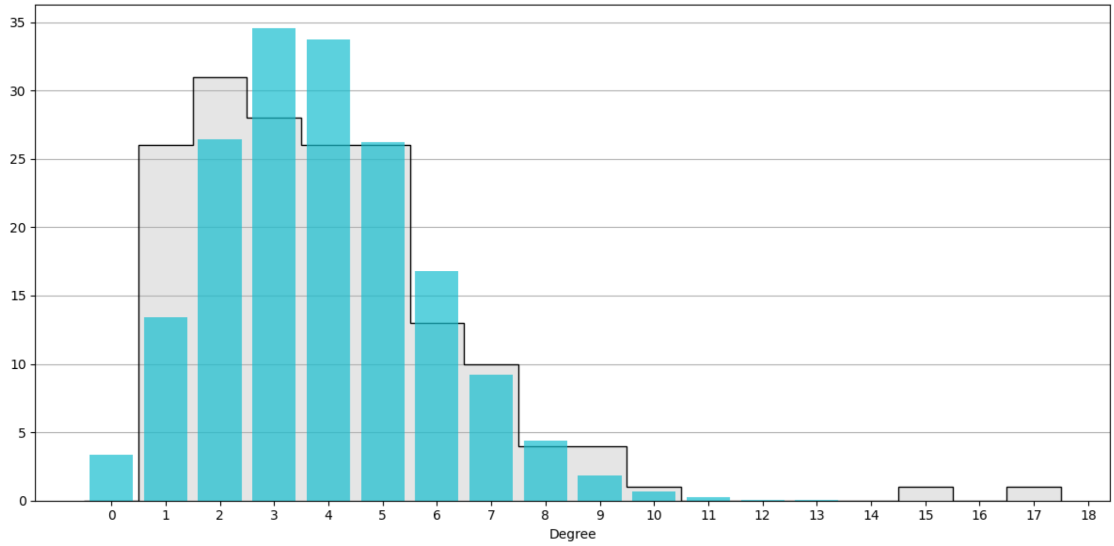


Figure 3.3: Binomial distribution fit to the degree distribution.

Other properties

Table 3.2 below lists some general properties of the worldgraph. Of particular importance are the number of edges (332) and the number of triangles (173), which together with the degree distribution form the properties which we will try to optimize when evaluating our random graph models.

Property	Value	Remark
nodes	250	
edges	332	
triangles	173	
connected components	83	
non-insular connected components	4	Americas, Afro-Eurasia, Saint Martin, Hispaniola
mean degree	2.656	
maximum degree	17	China
insular nodes	79	
mean edge length (km)	1 006.857	
std edge length (km)	885.389	
minimum edge length (km)	4.296	Saint Martin (FR) - Sint Maarten (NL)
maximum edge length (km)	4 743.131	Poland - Russia (via Kaliningrad exclave)

Table 3.2: Some properties of the worldgraph.

The two largest connected components are comprised of North and South America on one hand, with a total of 24 countries and territories, and Africa, Europe, Asia, and part of Oceania on the other hand, with a total of 143 countries and territories. Saint Martin and Hispaniola both contain two nodes, and the rest are all islands.

Chapter 4

Nodes on the sphere

In this chapter, we explain how we sample points on the Earth. In Section 4.1, we choose a geometrical model for the Earth and define a distance function for pairs of nodes, with a brief consideration of the error our simplified model produces. Then, in Sections 4.2 and 4.3, we look at ways of generating uniformly distributed nodes on the Earth, respectively with spherical coordinates and with the more commonly used (latitude, longitude) coordinates.

4.1 Modeling the Earth

There are three options for how to model the shape of the Earth:

- The simplest option is to model the Earth as a sphere. The sphere is centered at the Earth's center of mass, and the radius is stipulated by the *International Union of Geodesy and Geophysics* (R_1 in [6]) as approximately 6 371 008 meters. This value is defined by $\frac{2a+b}{3}$, where a and b are respectively the semi-major and semi-minor axes of the Earth. There are concise equations and efficient computational methods for calculating the great-circle distance between two points on the surface of a sphere.

- The next option is to model the Earth as a spheroid, which is an ellipsoid of revolution. The spheroid is centered at the Earth's center of mass, has a semi-major equatorial axis of constant length $b = 6\,356\,752.314\,245$ meters, and a semi-minor polar axis of length $a = 6\,378\,137$ meters. Several computational methods for the geodesic distance between two points are available, notably one by Karney [1] which improves on the more widely used method by Vincenty [9].

- The most accurate option is to model the Earth as an irregular spheroid constructed from the geoid. The geoid describes the combined effects of irregularities in elevation and density of the Earth's surface and the Earth's rotation on the gravitational field. To obtain the distances between any two points on the surface, there are methods that compute the geodesic distance, then correct it by taking into account the topology of the Earth with the help of observed data. Naturally, this method is very computationally costly and should be avoided for repeated queries.

For this project, the choice of modeling the Earth as a perfect sphere works best. Though the Python library *GeoPy*[1] provides methods for both the great-circle distance and the geodesic distance, the former is notably more time-efficient. Computational results are derived via stochastic simulation with a considerable number of runs to improve confidence intervals. To construct random graphs from n nodes typically all pairwise distances need to be computed, so the method is

[1] Release 2.3.0, https://geopy.readthedocs.io/en/stable/

run in $O\left(n^2\right)$ time. For these reasons, it is unfeasible to use the more accurate geodesic distance method.

In terms of analytical derivations, it is much simpler to work with the spherical model of the Earth: the rotational symmetry ensures the neighborhood of each node is identical regardless of its position on the surface. Indeed, in this case, the area and shape of the neighborhood of an edge are only given by the length of the edge and not by the position of its incident nodes. Furthermore, there are formulas for computing the area of intersection between two neighborhoods on a sphere, but not on a spheroid.

To understand the difference between the simplest model of the sphere and the most accurate one of the irregular spheroid, we show in Figure 4.1 the difference between these two distances compared with the distance between the two points, and in Figure 4.2 the corresponding relative error. As one can see, the relative error is higher for smaller distances, topping at roughly 0.5%, which remains very low and precise enough for this project. These figures were created by generating a set of 1 000 uniform nodes, computing all pairwise distances for both models, and computing the absolute and relative difference.

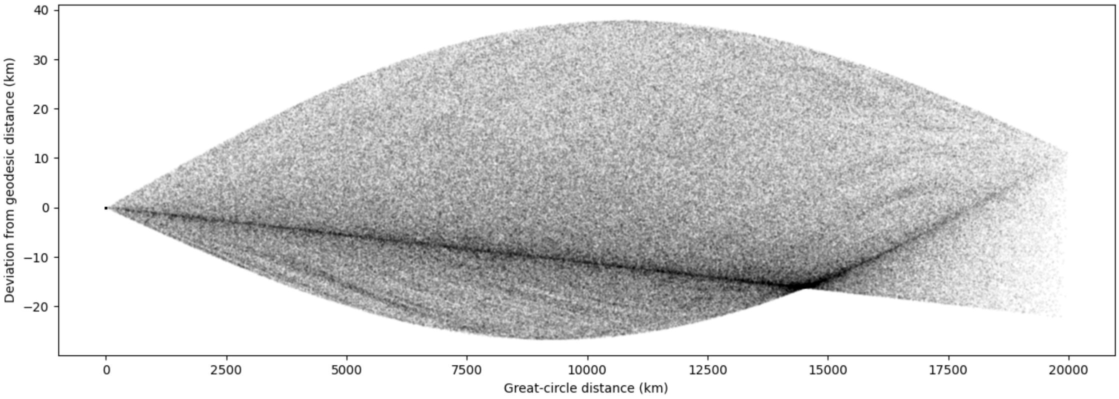


Figure 4.1: Error in distance due to spherical model.

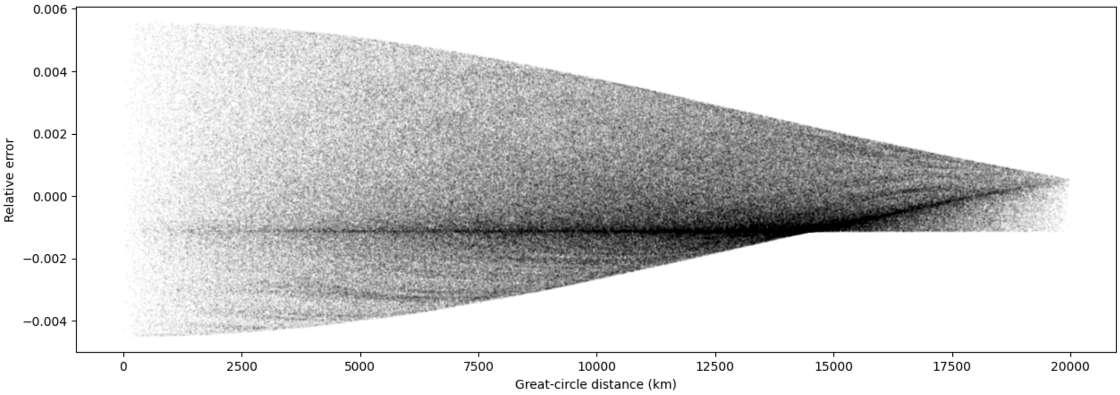


Figure 4.2: Relative error in distance due to spherical model.

4.2 Uniform distribution on a sphere

Let $\mathbb{S}^2_\oplus$ be the 2-sphere, that is the set of points in 3-dimensional Euclidean space which are at a fixed distance $r_\oplus > 0$ from the origin. In order to uniformly sample points on $\mathbb{S}^2_\oplus$, we are looking for a constant probability density function f such that

$$\int_{\mathbb{S}^2_\oplus} f(\Omega)\, d\Omega = 1,$$

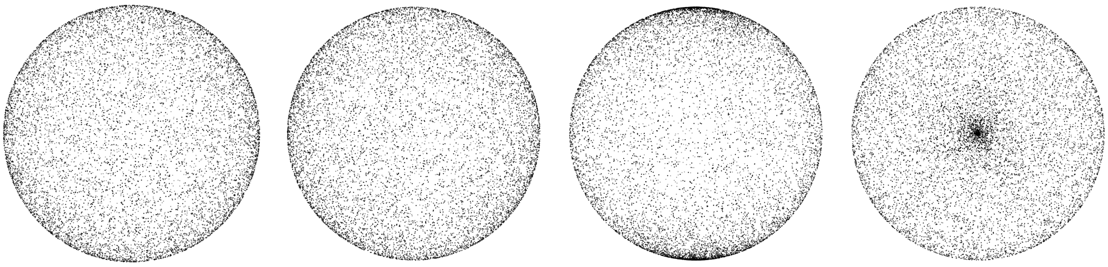
where $d\Omega$ is the volume element defined on all measurable subsets of $\mathbb{S}^2_\oplus$. The area of $\mathbb{S}^2_\oplus$ is equal to $4\pi r^2_\oplus$, so for normality and uniform density we want to define

$$f(\Omega) := \frac{1}{4\pi r^2_\oplus}.$$

We wish to transform from Cartesian coordinates to spherical coordinates. See Appendix B.1 for details on the derivation of the Jacobian determinant below. By the Change of Variables formula, we have

$$d\Omega = \left| \frac{\delta(x,y,z)}{\delta(r,\theta,\phi)} \right| \cdot dr d\theta d\phi = r^2 \sin(\theta) \cdot dr d\theta d\phi.$$

The term $\sin(\theta)$ in the Jacobian determinant slightly skews the distribution so we cannot just sample θ (the *inclination*) uniformly on $[0, \pi]$: many samples picked this way would accumulate near the poles of the globe (see Figure 4.3 below). Note that it is however fine to sample ϕ (the *azimuth*) uniformly on $[0, 2\pi]$.



Correct — sideview Correct — topview Wrong — sideview Wrong — topview

Figure 4.3: Correct uniform distribution on the left compared to the wrong distribution on the right where θ (the inclination) is uniformly sampled on $[0, \pi]$.

To confirm that this definition of $d\Omega$ is correct we substitute in the previous equation and obtain that

$$\begin{aligned}\int_{\mathbb{S}^2_\oplus} f(\Omega)\, d\Omega &= \left(\int_0^{2\pi} \int_0^{\pi} \frac{1}{4\pi r^2_\oplus} \cdot r^2_\oplus \sin(\theta)\, d\theta\, d\phi \right) \\ &= \int_0^{2\pi} \int_0^{\pi} \frac{\sin(\theta)}{4\pi}\, d\theta\, d\phi = 1.\end{aligned}$$

Using the previous computations, we see that $\tilde{f} := \frac{\sin(\theta)}{4\pi}$ is a probability density function over the space $(\theta, \phi) \in [0, \pi] \times [0, 2\pi)$. We can compute the marginal density functions of θ and ϕ by integrating respectively over all values of ϕ and θ, as such

$$\tilde{f}_{\text{inc}}(\theta) = \int_0^{2\pi} \frac{\sin(\theta)}{4\pi}\, d\phi = 2\pi \cdot \frac{\sin(\theta)}{4\pi} = \frac{\sin(\theta)}{2},$$

and

$$\tilde{f}_{\text{az}}(\phi) = \int_0^{\pi} \frac{\sin(\theta)}{4\pi}\, d\theta = \frac{1}{2\pi}.$$

Now we compute the corresponding cumulative distribution functions:

$$\tilde{F}_{\text{inc}}(\theta) = \int_0^\theta \tilde{f}_{\text{inc}}(\theta)\, d\theta = \int_0^\theta \frac{\sin(\theta)}{2}\, d\theta = \left(-\frac{\cos(\theta)}{2}\right)\Bigg|_0^\theta = \frac{1-\cos(\theta)}{2},$$

and

$$\tilde{F}_{\text{az}}(\phi) = \int_0^\phi f_{\tilde{U}}(\phi)\, d\phi = \int_0^\phi \frac{1}{2\pi}\, d\phi = \left(\frac{1}{2\pi}\right)\Bigg|_0^\phi = \frac{\phi}{2\pi}.$$

Now that we have the CDFs, we can use Inverse Transform Sampling to generate samples of the two marginal probability distributions above, and thus generate uniformly distributed samples on the space $\mathbb{S}^2_\oplus$. Let $u = \tilde{F}_{\text{inc}}(\theta)$, and $v = \tilde{F}_{\text{az}}(\phi)$ be two independent uniform random variables on $[0,1]$. Both these functions are invertible, so solving for θ and ϕ gives

$$\theta = \tilde{F}^{-1}_{\text{inc}}(u) = \cos^{-1}(1-2u),$$

and

$$\phi = \tilde{F}^{-1}_{\text{az}}(v) = 2\pi v.$$

In summary, a uniformly random point on $\mathbb{S}^2_\oplus$ has spherical coordinates $(r_\oplus, \cos^{-1}(1-2U), 2\pi V)$, where U and V are two independently distributed uniform random variables on $[0,1]$.

4.3 Uniform distribution on Earth

If we are modeling the Earth as $\mathbb{S}^2_\oplus$, then we can define the following transformation in terms of latitudes and longitudes:

$$\begin{cases} \text{lat}(\theta) = \left(\frac{\pi}{2} - \theta\right) \cdot \frac{180°}{\pi}, \\ \text{lon}(\phi) = (\phi - \pi) \cdot \frac{180°}{\pi}, \\ r_\oplus \approx 6\,371.009 \text{ km}. \end{cases}$$

Latitudes take values in the interval $[-90°, 90°]$, and longitudes in the interval $[-180°, 180°)$. The south pole has latitude $-90°$, the equator has latitude $0°$, and the north pole has latitude $90°$.

It may be desirable to generate points within particular subsets of latitudes and/or longitudes. For two measurable subsets $\mathcal{A} \subseteq [-90°, 90°]$, $\mathcal{B} \subseteq [-180°, 180°)$, we can easily generate points uniformly at random on the space $\{x \in \mathbb{S}^2 : \text{lat}(x) \in \mathcal{A}, \text{lon}(x) \in \mathcal{B}\}$ as follows.

Let $\tilde{\mathcal{A}} := \tilde{F}_{\text{inc}}\left(\text{lat}^{-1}(\mathcal{A})\right) \subseteq [0,1]$ and $\tilde{\mathcal{B}} := \tilde{F}_{\text{az}}\left(\text{lon}^{-1}(\mathcal{B})\right) \subseteq [0,1)$ be the corresponding sets in $[0,1] \times [0,1)$. Now a uniformly random point on the space $\{x \in \mathbb{S}^2 : \text{lat}(x) \in \mathcal{A}, \text{lon}(x) \in \mathcal{B}\}$ has spherical coordinates distributed as $(r_\oplus, \cos^{-1}(1-2U), 2\pi V)$, where U is uniform on $\tilde{A}$, V is uniform on $\tilde{B}$, and they are independent of each other.

The interest of the previous sub-sampling can be multiple. One could for example use $\mathcal{A} = [50°, 54°]$ and $\mathcal{B} = [3°, 8°]$ to sample points around the region of the Netherlands. Similarly, one could use $\mathcal{A} = [-66°, 66°]$ and $\mathcal{B} = [-180°, 180°)$ to reproduce the fact that very few people live within both Arctic circles.

Chapter 5

Graph models

In this chapter, several random graph models are defined, described and some samples are shown. The chapter ends with a short discussion on the relationships between these graph models, in Section 5.6.

All graphs are built starting from a set of points $V = \{v_1, \ldots, v_n\} \subset \mathbb{S}^2_\oplus$. It is worth noting that this can be applied to any set of points on $\mathbb{S}^2_\oplus$, however all samples will be obtained by uniformly placing $n = 171$ nodes on $\mathbb{S}^2_\oplus$. Each random graph model is characterized by a specific procedure that defines which nodes are to be connected via edges. Contrary to what the name 'random graph model' might suggest, this procedure is actually rather deterministic: given a fixed set of points V and a fixed model, the resulting graph $G = (V, E)$ is uniquely determined. The randomness is only given by the distribution of the nodes on the globe.

We present a couple of simple random graph models, the *ε-neighborhood graph* in Section 5.1 and the *k-nearest neighbors graph* in Section 5.2. The former connects nodes if they are within a certain distance of each other, and the latter connects nodes if one node is part of the k nearest neighbors of the other node.

So-called *neighborhood graphs* form a large class of random graph models. These graph models define a way to associate a 'neighborhood' $\mathcal{N}(v_i, v_j) \subseteq \mathbb{S}^2_\oplus$ for each distinct pair of nodes $v_i, v_j \in V$. Such a neighborhood $\mathcal{N}(v_i, v_j)$ is typically a connected area situated around the nodes v_i and v_j, scaling in size with the distance between the nodes. If there are no nodes within this neighborhood other than v_i and v_j, then the neighborhood graph model dictates that (v_i, v_j) is an edge of the graph. For many known neighborhood graphs, there are alternative definitions in terms of distances between nodes which are usually much easier to work with for simulations. Sections 5.3, 5.4, and 5.5 provide some examples of neighborhood graphs.

All random graph models in this chapter have one or more parameters that can be varied to influence the procedure of creating edges and by extension influence the behavior of the sampled graphs. The inclusion of such parameters allows for greater control of the properties of the resulting graphs, making it easier to compare them with the worldgraph.

5.1 Epsilon neighborhood graph

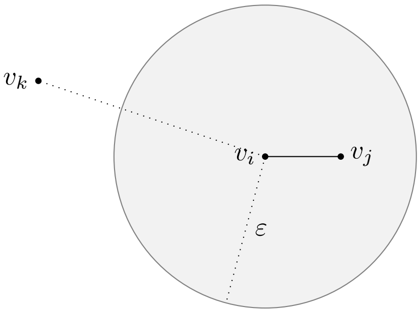


Figure 5.1: The ε-neighborhood of v_i.

The *ε-neighborhood graph* (εN) on a set of points $V = \{v_1, \ldots, v_n\} \subset \mathbb{S}^2_\oplus$ is obtained by connecting all distinct unordered pairs of nodes (v_i, v_j) if and only if their distance is less than ε kilometers, for some $\varepsilon > 0$ of choice, as represented in Figure 5.1. See Figures 5.2 and 5.3 below for some examples of εN graphs. The set of edges E is thus defined as

$$E := \{(v_i, v_j) \in V \times V : \operatorname{dist}(v_i, v_j) < \varepsilon\}.$$

Recall that we only consider simple graphs and thus never connect a node to itself, even though $\operatorname{dist}(v_i, v_i) = 0 < \varepsilon$. For small enough ε the graph is empty, and for large enough ε the graph is complete. Moreover, given $\varepsilon_1 < \varepsilon_2$, the ε_1NN graph is a subgraph of the ε_2NN graph since any edge within ε_1 kilometers length is also within ε_2 kilometers length.

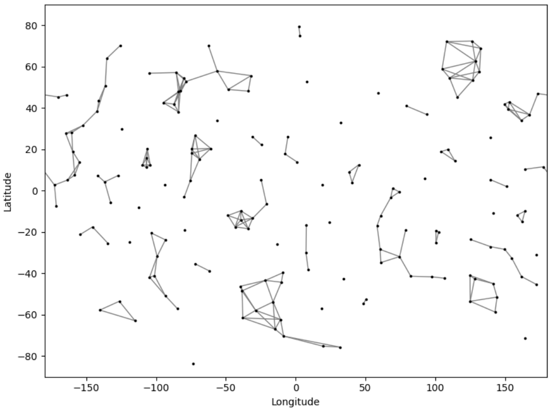


Figure 5.2: Sample εN graph, $\varepsilon = 1\,500$ km.

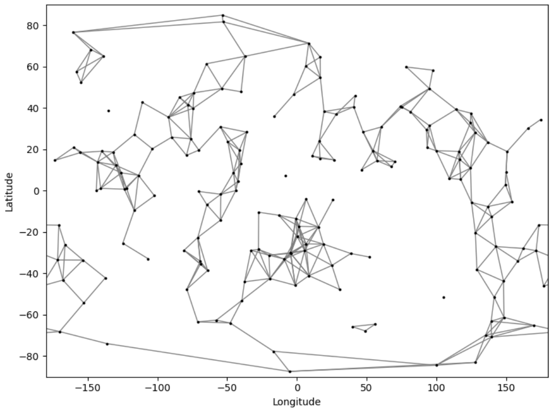


Figure 5.3: Sample εN graph, $\varepsilon = 2\,000$ km.

The εN graph is the simplest model for the worldgraph and is based on the assumption that countries are connected if and only if their centroids are close enough.

5.2 K-nearest neighbors graph

The *k-nearest neighbors graph* (kNN) on a set of points $V = \{v_1, \dots, v_n\} \subset \mathbb{S}^2_\oplus$ is obtained by including (v_i, v_j) as an edge whenever v_i is one of the k nearest neighbors of v_j or v_j is one of the k nearest neighbors of v_i, for some $k \in \mathbb{N}$ of choice. In the case of a tie we can use the lexicographic ordering to identify the k neighbors, though this happens with zero probability for uniformly distributed points V. See Figures 5.4, 5.5, 5.6, and 5.7 below for some examples of kNN graphs with different values of k. The kNN graph is a simple graph as a node is not considered to be its own neighbor. Moreover, given $k_1 < k_2$, the k_1NN graph is a subgraph of the k_2NN graph since one of the k_1 nearest neighbors of a node is also one of the k_2 nearest neighbors.

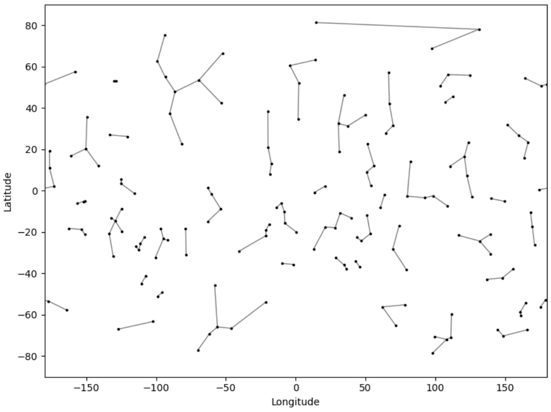


Figure 5.4: Sample kNN graph, $k = 1$.

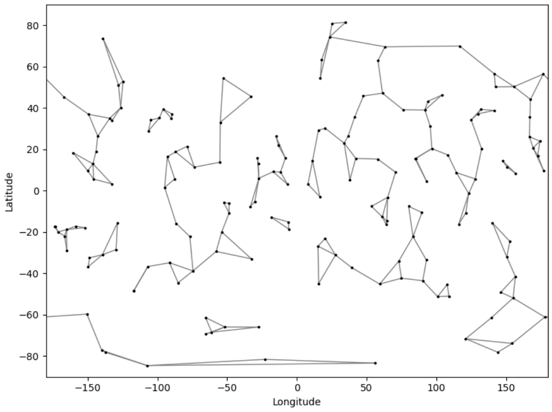


Figure 5.5: Sample kNN graph, $k = 2$.

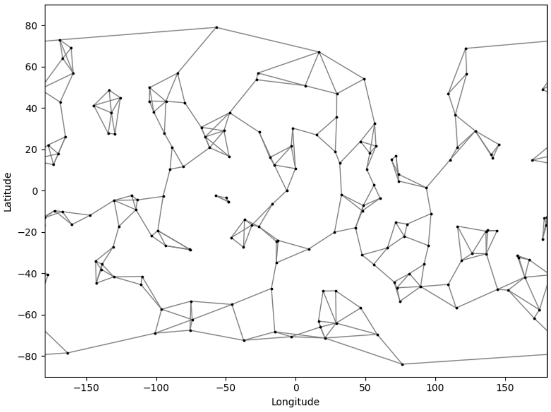


Figure 5.6: Sample kNN graph, $k = 3$.

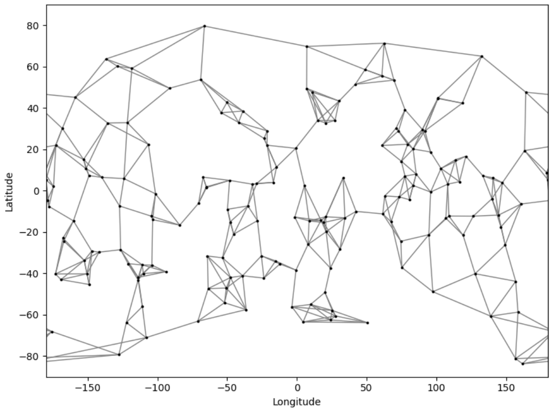


Figure 5.7: Sample kNN graph, $k = 4$.

The kNN graph is a reasonable choice as a model for the worldgraph, based on the assumption that a country is more likely to border another country if there are few other countries closer to it, and much less likely if that other country has many other countries closer to it. However, as shown in Figures 5.4, 5.5, 5.6, and 5.7, it tends to create very long edges, something we want to avoid. For this reason we define the next model, mixing the kNN graph with the ε-neighborhood graph.

Epsilon-k-nearest neighbors graph

The kNN graph can be filtered to only include edges up to a certain length. The ε-kNN graph on a set of points $V = \{v_1, \dots, v_n\} \subset \mathbb{S}^2_{\oplus}$ is obtained by including (v_i, v_j) as an edge if v_i and v_j are in the kNN graph and their distance is less than ε, for some $\varepsilon > 0$ of choice. See Figures 5.8 and 5.9 for some examples of this extension of the kNN graph model. For large enough ε the graph is equivalent to the kNN graph, while for large enough k the graph is equivalent to the εN graph. Given $\varepsilon_1 < \varepsilon_2$, the ε_1-kNN graph is a subgraph of the ε_2-kNN graph, and similarly, given $k_1 < k_2$, the ε-k_1NN graph is a subgraph of the ε-k_2NN graph.

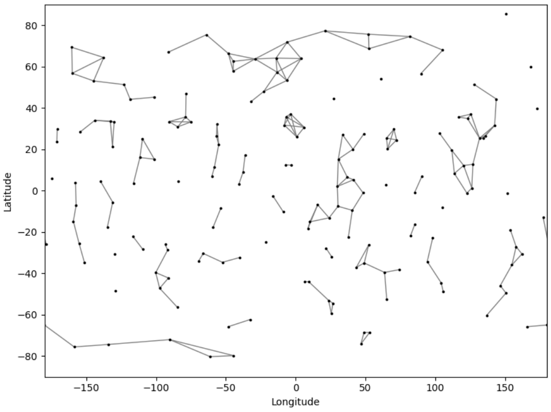


Figure 5.8: Sample ε-4NN graph, $\varepsilon = 1\,500$ km.

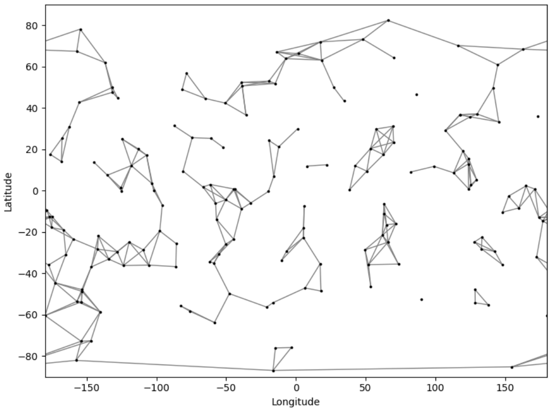


Figure 5.9: Sample ε-4NN graph, $\varepsilon = 2\,000$ km.

We will consider ε-kNN as a family of graphs indexed by $k \in \{1, 2, \dots\}$. The adaptation of the kNN graphs to include a varying parameter ε allows for greater control of the properties of the resulting graphs.

5.3 Delaunay triangulation

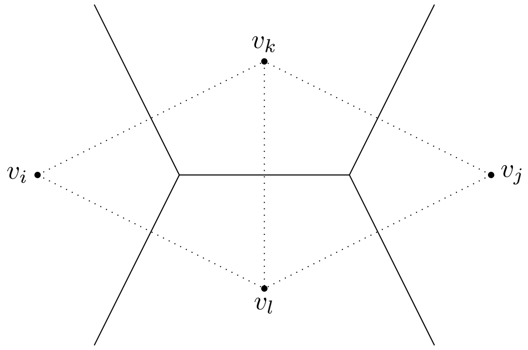


Figure 5.10: Voronoi diagram.

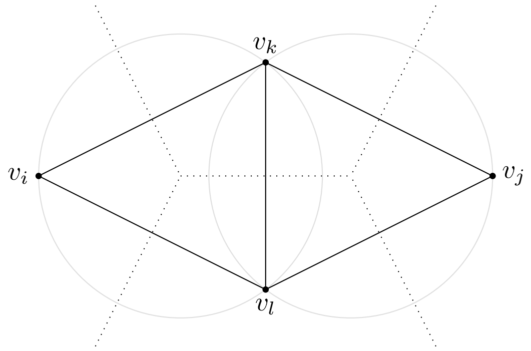


Figure 5.11: Delaunay triangulation.

The *Voronoi diagram* (shown in Figure 5.10) is a partition of the surface $\mathbb{S}^2_\oplus$ into regions (called Voronoi *cells*), such that each region contains exactly one node v_i, together with every point that is closer to v_i than to any other node in V. The *Delaunay triangulation* (shortened to DT and shown in Figure 5.11) is the dual of the Voronoi diagram: nodes are connected via an edge if their respective Voronoi cells share an edge. One can choose to model the nodes in V as country centroids and the Voronoi cells as the countries themselves; then the DT is a natural way to model the graph of bordering countries. There are a few other ways to define this graph. For instance, a triple of nodes (v_i, v_j, v_k) forms a triangle of the DT if there are no other nodes within the circle passing through this triple, as can be seen in Figure 5.11). Alternatively, a pair of nodes (v_i, v_j) forms an edge of the DT if there exists a circle of any size passing through both nodes that contains no other nodes.

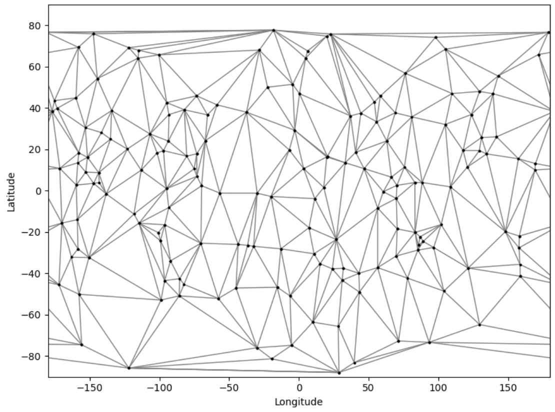


Figure 5.12: Sample Delaunay triangulation.

The Delaunay triangulation is not a good model for the worldgraph as it is composed solely of triangles and all nodes are of degree at least 3, as is evident in Figure 5.12; in reality, the world landmass is not as connected and contains many countries bordering strictly less than three others. The DT works better in the absence of oceans — accurate results can be achieved by restricting nodes to the world landmass with edges that cross relatively little ocean surface.

5.4 Relative neighborhood graph

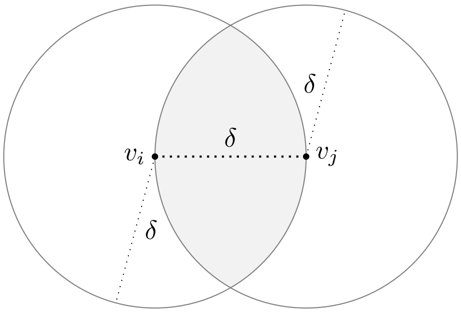


Figure 5.13: $\mathcal{N}(v_i, v_j)$ in RNG.

The *relative neighborhood graph* (RNG) on a set of points $V = \{v_1, \ldots, v_n\} \subset \mathbb{S}^2_\oplus$ is a type of neighborhood graph. Given a pair of nodes (v_i, v_j) at distance δ, define the neighborhood $\mathcal{N}(v_i, v_j)$ as the intersection of two balls centered at v_i and v_j with radius δ, like in Figure 5.13. If $\mathcal{N}(v_i, v_j)$ contains no other nodes $v_k \in V \setminus \{v_i, v_j\}$, (v_i, v_j) is included as an edge of the RNG. Equivalently, the RNG is obtained by including (v_i, v_j) as an edge when

$$\text{dist}(v_i, v_j) \leq \max\left(\text{dist}(v_i, v_k), \text{dist}(v_j, v_k)\right) \quad \text{for all } v_k \in V \setminus \{v_i, v_j\}.$$

The RNG was defined in 1980 by Toussaint [8], who proved it is a supergraph of the *minimum spanning tree* (Theorem 1) and a subgraph of the *Delaunay triangulation* (Theorem 2). Figures 5.14, 5.15, and 5.16 show samples of the MST, RNG, and DT from the same distribution of nodes, and highlight this relationship. Given a set of points $V \subset \mathbb{S}^2_\oplus$, the minimum spanning tree (MST) is the graph that minimizes the total length of its edges while maintaining connectedness (using lexicographic ordering as a tiebreaker). The MST contains no triangles since any edge can be removed while remaining connected, and hence it is not appropriate for modeling the worldgraph.

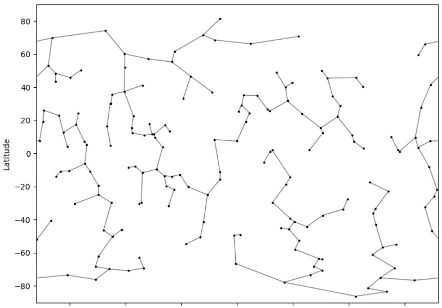


Figure 5.14: Sample MST.

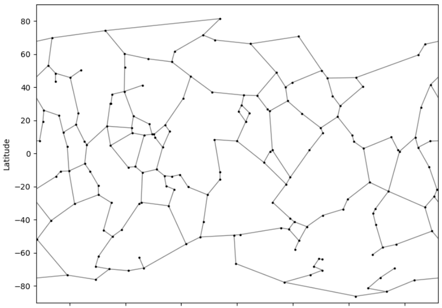


Figure 5.15: Sample RNG.

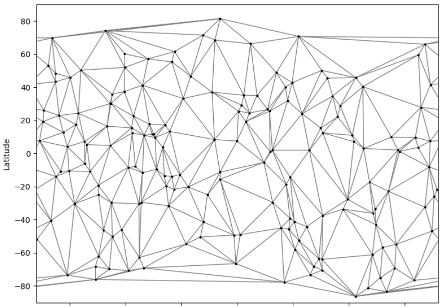


Figure 5.16: Sample DT.

The RNG is a suitable intermediate between the MST and the DT in terms of modeling the worldgraph. The same paper by Toussaint [8] discusses its ability to extract a 'perceptually meaningful' structure from the set of points V. It is suggested that the RNG is a powerful model of low-level visual processes involved in the perception of dot patterns.

Generalized relative neighborhood graph

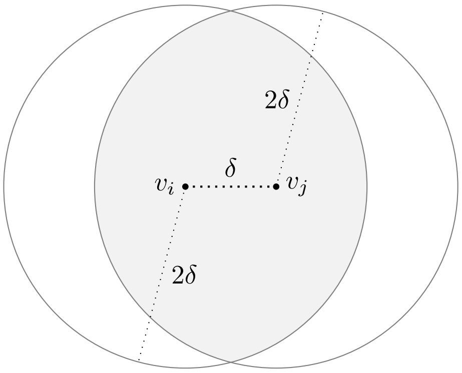


Figure 5.17: $\mathcal{N}(v_i, v_j)$ in λ-RNG, $\lambda = 0.5$.

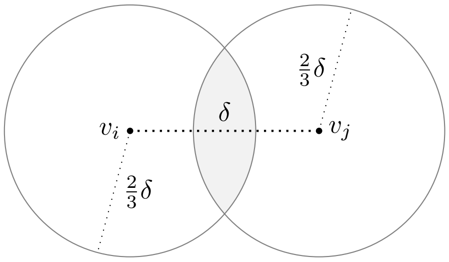
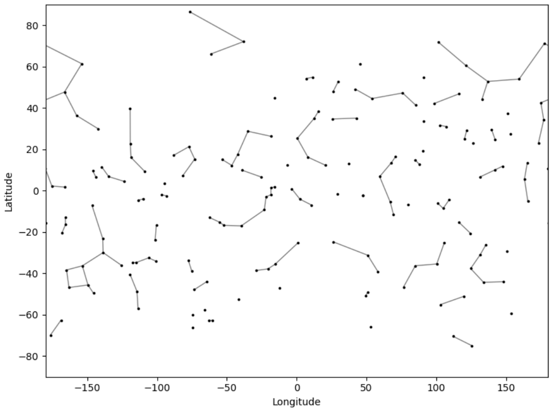


Figure 5.18: $\mathcal{N}(v_i, v_j)$ in λ-RNG, $\lambda = 1.5$.

The RNG is uniquely defined given a set of points V. It can be adapted to include a parameter λ to allow for more control of its properties. We define the λ-RNG as a generalized relative neighborhood graph, where the neighborhood $\mathcal{N}(v_i, v_j)$ is now the intersection of two balls centered at v_i and v_j with radius δ/λ, as shown in Figures 5.17 and 5.18. If $\mathcal{N}(v_i, v_j)$ contains no other nodes $v_k \in V \setminus \{v_i, v_j\}$, the edge (v_i, v_j) is part of the λ-RNG. Equivalently, the λ-RNG contains all edges (v_i, v_j) such that

$$\operatorname{dist}(v_i, v_j) \leq \lambda \cdot \max[\operatorname{dist}(v_i, v_k), \operatorname{dist}(v_j, v_k)] \quad \text{for all } v_k \in V \setminus \{v_i, v_j\}.$$

Figures 5.19 and 5.20 below show some samples of the λ-RNG. For $\lambda = 1$ this is the regular RNG. For $\lambda \geq 2$, the neighborhood $\mathcal{N}$ of any pair of nodes is an empty set, hence trivially containing no other nodes, which results in it being a complete graph. Given $\lambda_1 > \lambda_2$, the λ_1-RNG is a subgraph of the λ_2-RNG.



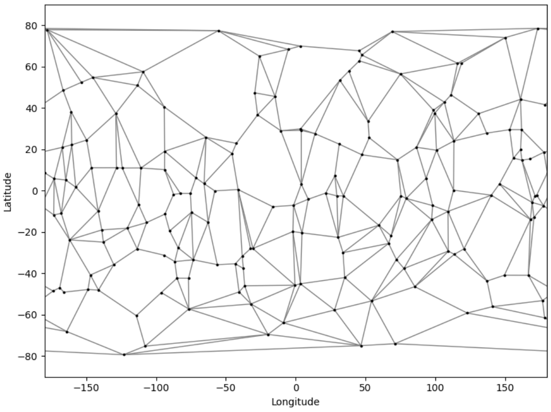
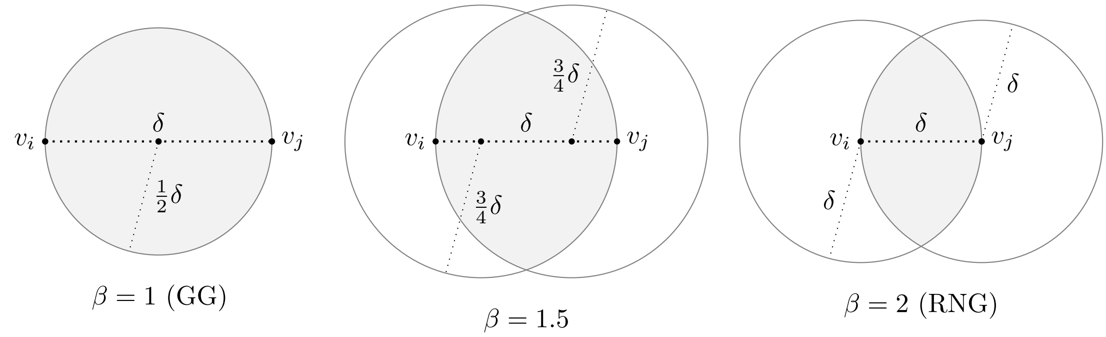


Figure 5.19: Sample λ-RNG, $\lambda = 0.8$.

Figure 5.20: Sample λ-RNG, $\lambda = 1.2$.

The adaptation of the RNG to include a varying parameter λ allows for greater control of the properties of the resulting graphs.

5.5 Beta skeleton graph



$\beta = 1$ (GG) $\beta = 1.5$ $\beta = 2$ (RNG)

Figure 5.21: $\mathcal{N}(v_i, v_j)$ in βS.

The *Beta skeleton graph* (βS) on a set of points $V = \{v_1, \dots, v_n\} \subset \mathbb{S}^2_{\oplus}$ is a class of neighborhood graphs. Though different variants exist, we solely focus on the 'lune-based' version. Given a pair of nodes (v_i, v_j) at distance δ, we define the neighborhood $\mathcal{N}(v_i, v_j)$ as the intersection of two balls, as follows.

$$\mathcal{N}(v_i, v_j) := \mathcal{B}_{\oplus}\left(\left(1 - \frac{\beta}{2}\right) \cdot v_i + \frac{\beta}{2} \cdot v_j, \frac{\beta}{2}\delta\right) \cap \mathcal{B}_{\oplus}\left(\left(1 - \frac{\beta}{2}\right) \cdot v_j + \frac{\beta}{2} \cdot v_i, \frac{\beta}{2}\delta\right).$$

The parameter β can take values in the interval $[1, 2]$. For $\beta = 2$ the Beta skeleton graph is equivalent to the RNG, as shown in Figure 5.21. The graph for $\beta = 1$ is commonly referred to as the *Gabriel graph* (GG), where the two balls coincide which makes $\mathcal{N}$ a ball centered at the midpoint of v_i and v_j that passes through both points. A sample of the Gabriel graph is shown in Figure 5.23. As β increases from 1 to 2, the centers of the balls linearly move towards the endpoints of the edge (see Figure 5.21).

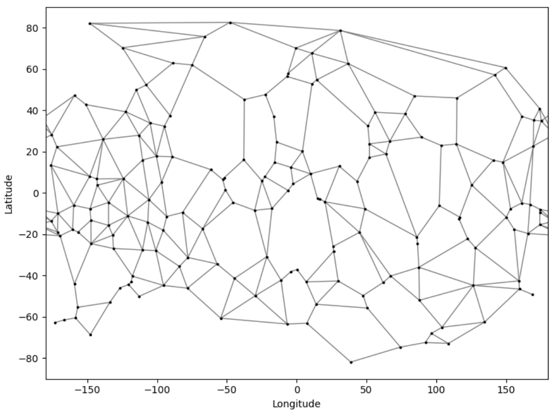


Figure 5.22: Sample βS graph, $\beta = 1$ (GG)

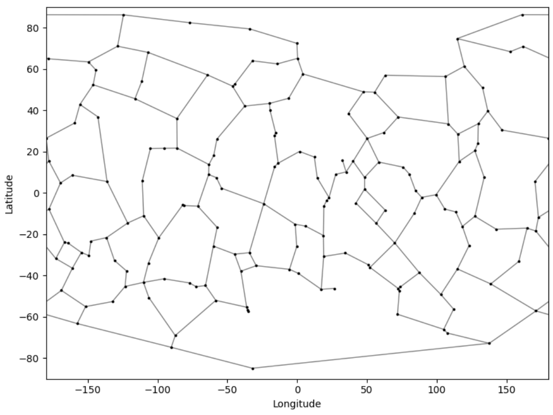


Figure 5.23: Sample βS graph, $\beta = 2$ (RNG)

The Beta skeleton graph was defined in 1985 by Kirkpatrick and Radke [2]. It is created with the assumption that all nodes of some empirical network are equally significant, and hence *neighborliness* is the dominant factor determining connections. More information on this notion of neighborliness is given in [2, Section 3.2]. By comparing the connections of the βS graph and those of the real network, it is possible to focus attention on parts of the network where forces other than neighborliness are at work. The paper shows success in this methodology in artificially constructed networks and a few empirical networks, both planar road networks and non-planar airline networks. When applied to model the worldgraph and compared to the real network of countries, it might give interesting insights into how influential cultural/administrative/topological factors are in determining borders, compared to simply neighborliness.

5.6 Relations between graph models

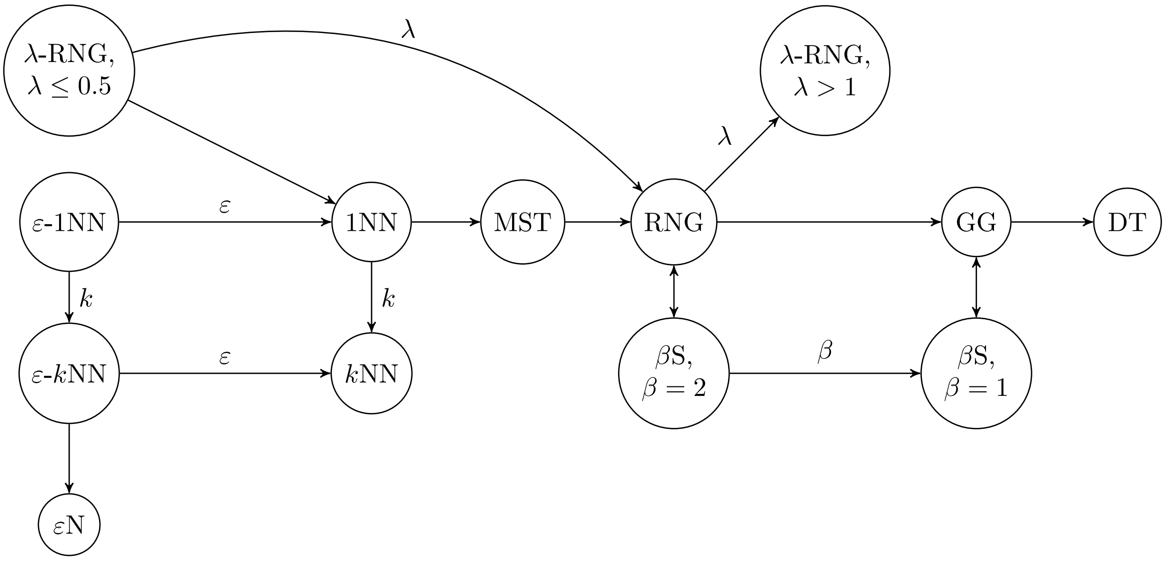


Figure 5.24: Diagram showing relations between random graph models.

See Figure 5.24 above for a quick overview of how the random graph models relate to each other (given a common set of points $V \subset \mathbb{S}^2_\oplus$). An arrow $A \to B$ indicates that A is a subgraph of B. A double arrow $A \leftrightarrow B$ indicates that A and B are equivalent graphs. An arrow $A \xrightarrow{\pi} B$ indicates that by continuously varying the parameter π, common to both graphs A and B, the sample graphs will also continuously vary from graph A to graph B.

Given a set of nodes V uniformly distributed on the surface $\mathbb{S}^2_\oplus$, all these models uniquely construct a graph — all distances of pairs of nodes are a.s. distinct so there is no ambiguity in the construction. Many of these relations are easily verified by considering the neighborhoods of both models: say graph A defines a neighborhood $\mathcal{N}_A$ for each edge, and graph B defines a neighborhood $\mathcal{N}_B$ for each edge. If $\mathcal{N}_A \supseteq \mathcal{N}_B$ for each edge, then A is a subgraph of B. The remaining relations are shown here.

Proof that 1NN $\subseteq$ MST: *Let $V = \{v_1, \ldots, v_n\}$ be a set of nodes uniformly distributed on $\mathbb{S}^2_\oplus$. Suppose there is an edge $e = (v_i, v_j)$ in 1NN that is not in MST. Either v_i is the nearest neighbor of v_j, or vice versa. Suppose w.l.o.g. that v_i is the nearest neighbor of v_j. Since MST is a connected graph, v_j must connect to another node $v_k \neq v_i$. But we know that $dist(v_j, v_i) < dist(v_j, v_k)$. Hence exchanging the edge (v_j, v_k) for the edge (v_j, v_i) in the MST gives a strictly smaller spanning tree, contradicting the minimality of MST.*

Proof that MST $\subseteq$ RNG: *Let $V = \{v_1, \ldots, v_n\}$ be a set of nodes uniformly distributed on $\mathbb{S}^2_\oplus$. Suppose there is an edge $e = (v_i, v_j)$ in MST that is not in RNG. Deletion of e from MST yields two subtrees, T_i containing v_i and T_j containing v_j. Each node in V is either in T_i or T_j.*

Since e is not in RNG, there must be another node $v_k \in \mathcal{B}_\oplus(v_i, \delta) \cap \mathcal{B}_\oplus(v_j, \delta)$, where $\delta = dist(v_i, v_j)$. If v_k is part of T_i, then $T_i \cup T_j \cup \{(v_k, v_j)\}$ is a strictly smaller spanning tree than MST, since $dist(v_k, v_j) < dist(v_i, v_j) = \delta$. The same argument holds if v_k is part of T_j. This contradicts the minimality of MST.

Proof that GG $\subseteq$ DT: *Let $V = \{v_1, \ldots, v_n\}$ be a set of nodes uniformly distributed on $\mathbb{S}^2_{\oplus}$. Consider an edge $e = (v_i, v_j)$ of GG. That means that there is no other node v_k inside the ball $\mathcal{B}_{\oplus}(m, \delta/2)$, where m is the midpoint of e, and $\delta = dist(v_i, v_j)$. The boundary of this ball passes through both v_i and v_j. By definition of how DT is constructed, this edge e must therefore also be part of DT.*

A clear picture of these relationships is useful when optimizing computational methods for creating sample graphs. For example, when constructing the RNG from a set of points $V = \{v_1, \ldots, v_n\}$, one can naively choose to define the neighborhood $\mathcal{N}$ for each pair of nodes ($O\left(n^2\right)$ operations) and verify they are empty. Alternatively, given that RNG $\subseteq$ GG $\subseteq$ DT, one can more efficiently create the DT first ($O\left(n \log n\right)$ operations) and then filter edges to arrive at the RNG (DT has $O\left(n\right)$ edges). Another advantage is that it can allow one to easily extrapolate the theoretical results of one graph model to a related one.

Given that the DT is a triangulation, all its random graph samples are planar embeddings on the surface $\mathbb{S}^2_{\oplus}$. As is evident from Figure 5.24, given a common set of points $V \subset \mathbb{S}^2_{\oplus}$, most random graph models generate subgraphs of the DT. Hence they are also planar graphs. The only exceptions are the εN graph, the kNN or ε-kNN graphs for $k \geq 2$, and the λ-RNG graph for $\lambda > 1$.

Chapter 6

Theoretical results

For the sake of legibility and conciseness, it is useful to define a few functions that are often used in deriving other results. The results are presented in this introductory section, and the derivations can be found in Appendices B.2, B.3, B.4, and B.5.

In Section 6.1 we discuss how the considered random graph models behave when varying the number of nodes $|V| = n$, in particular for large n. We wish for consistent local behavior in the graph samples regardless of the size of n: the neighborhood graph models scale naturally with n, while the other graph models must have their parameters manually adjusted to retain consistency. The remainder of this chapter is dedicated to a variety of theoretical results of properties of the random graph models, particularly the εN graph in Section 6.2.

Area of a neighborhood

Given a point $x \in \mathbb{S}^2_\oplus$ and a radius $\varepsilon > 0$, we define the *ε-neighborhood* of x as the set of points on $\mathbb{S}^2_\oplus$ which are within a great-circle distance of ε kilometers from x. We denote this neighborhood by $\mathcal{B}_\oplus(x, \varepsilon)$. The area of the neighborhood is

$$|\mathcal{B}_\oplus(x, \varepsilon)| = 2\pi r_\oplus^2 \left(1 - \cos\left(\frac{\varepsilon}{r_\oplus}\right)\right).$$

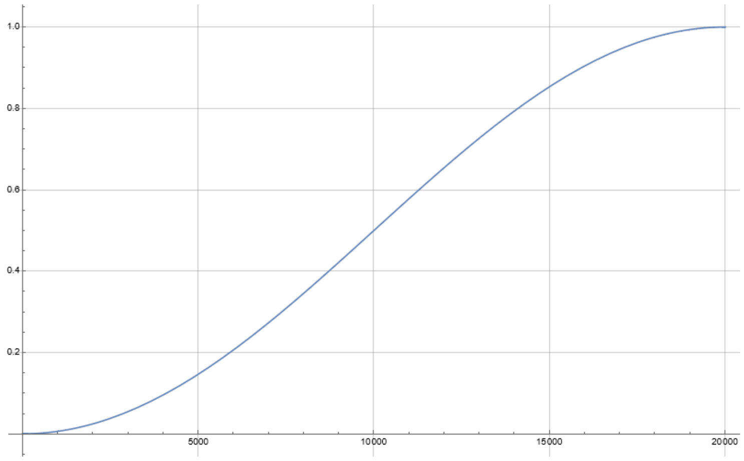


Figure 6.1: Plot of $p(\varepsilon)$, ε from 0 to $\pi r_\oplus \approx 20\,000$ kilometers.

See Appendix B.2 for the full derivation. Henceforth we define the function $p : \mathbb{R} \to \mathbb{R}$ as the proportion of the globe covered by an ε-neighborhood, as shown in Figure 6.1 above.

$$p(\varepsilon) := \frac{1}{2}\left(1 - \cos\left(\frac{\varepsilon}{r_\oplus}\right)\right).$$

Area of intersection of neighborhoods

Let $\varepsilon > 0$ and u, v be two distinct points on $\mathbb{S}^2_\oplus$. Consider their two neighborhoods $\mathcal{B}_\oplus(u,\varepsilon)$ and $\mathcal{B}_\oplus(v,\varepsilon)$ and denote by δ the great-circle distance between u and v. We use the following definitions for computing the area of the intersection.

$$\theta := \frac{\varepsilon}{r_\oplus}, \qquad \theta_v := \frac{\delta}{r_\oplus},$$

$$\theta_{\min} := \tan^{-1}\left(\frac{1}{\sin(\theta_v)} - \frac{1}{\tan(\theta_v)}\right),$$

Given these definitions, the area of the intersection is

$$|\mathcal{B}_\oplus(u,\varepsilon) \cap \mathcal{B}_\oplus(v,\varepsilon)| = 2\pi r_\oplus^2 \cdot \int_{\theta_{\min}}^{\theta} \sin(\phi) \cdot I\left(1 - \left(\frac{\tan(\theta_{\min})}{\tan(\phi)}\right)^2, \frac{1}{2}, \frac{1}{2}\right) d\phi,$$

where $I(z,a,b) = I_z(a,b)$ is the regularized incomplete beta function. See Appendix B.3 for the full derivation. The formula is accurate on the condition that $\theta \in [\theta_v/2, \pi/2)$, which entails that the intersection of neighborhoods is not empty and that both neighborhoods are strictly smaller than half the globe (roughly speaking, ε must be smaller than 10 000 kilometers).

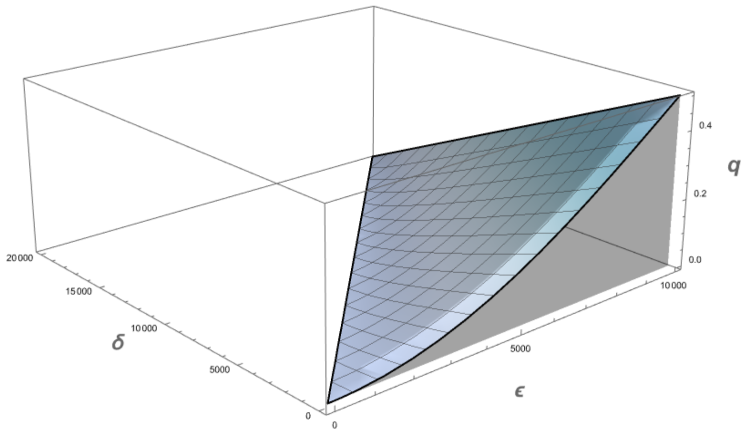


Figure 6.2: Plot of $q(\varepsilon,\delta)$, ε from 0 to $\pi r_\oplus/2$ kilometers, δ from 0 to $\pi r_\oplus$ kilometers.

Henceforth we define the function $q : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as the proportion of the globe covered by the intersection of neighborhoods, as shown in Figure 6.2 above. Keeping the same definitions for θ, θ_v, and $\theta_{\min}$ as above,

$$q(\varepsilon,\delta) := \frac{1}{2} \cdot \int_{\theta_{\min}}^{\theta} \sin(\phi) \cdot I\left(1 - \left(\frac{\tan(\theta_{\min})}{\tan(\phi)}\right)^2, \frac{1}{2}, \frac{1}{2}\right) d\phi.$$

Probability of closing a triangle

Given a neighborhood $\mathcal{B}_\oplus(u,\varepsilon)$ for some $u \in \mathbb{S}^2_\oplus$ with positive radius $\varepsilon > 0$, and two other points v and w that are uniformly distributed in the neighborhood $\mathcal{B}_\oplus(u,\varepsilon)$, the probability that v and

w are also in each others' respective ε-neighborhoods is

$$\frac{1}{(1-\cos(\varepsilon))^2} \cdot \int_0^{\varepsilon} \sin(\theta) \int_{\theta_{\min}(\theta)}^{\varepsilon} \sin(\psi) \cdot I\left(1 - \left(\frac{\tan(\theta_{\min}(\theta))}{\tan(\psi)}\right)^2, \frac{1}{2}, \frac{1}{2}\right) d\psi\, d\theta,$$

where $I(z, a, b) = I_z(a, b)$ is the regularized incomplete beta function, and $\theta, \theta_v, \theta_{\min}$ are defined as in the previous section. See Appendix B.4 for the full derivation. We denote this probability as the function $\tilde{p}(\varepsilon) : \mathbb{R} \to \mathbb{R}$ shown in Figure 6.3.

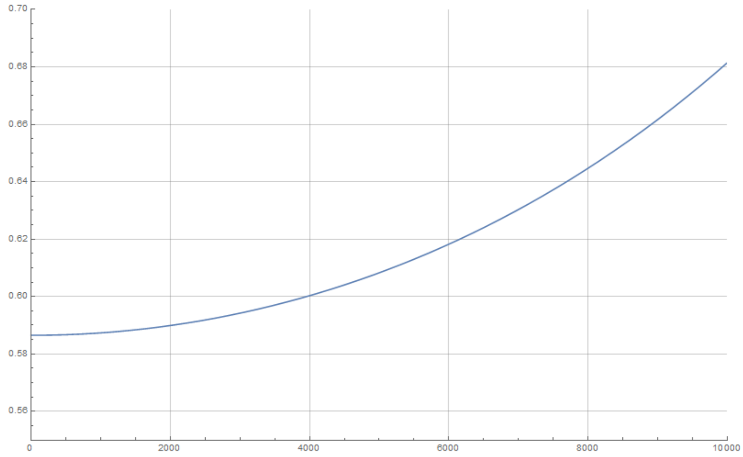


Figure 6.3: Plot of $\tilde{p}(\varepsilon)$, ε from 0 to $\pi r_{\oplus}/2 \approx 10\,000$ kilometers.

Minimum distance of nodes

Given a set of points $V = \{v_1, \ldots, v_n\} \subset \mathbb{S}^2_{\oplus}$ for some $n \geq 2$, the expectation of the great-circle distance in kilometers from an arbitrary node v_i to its nearest neighbor is

$$\rho = \frac{\pi r_{\oplus}}{4^{n-1}} \binom{2n-2}{n-1}.$$

See Appendix B.5 for the full derivation, and Figure 6.4 below for its plot. As $n \to \infty$, ρ converges to zero at a rate of order $n^{-1/2}$.

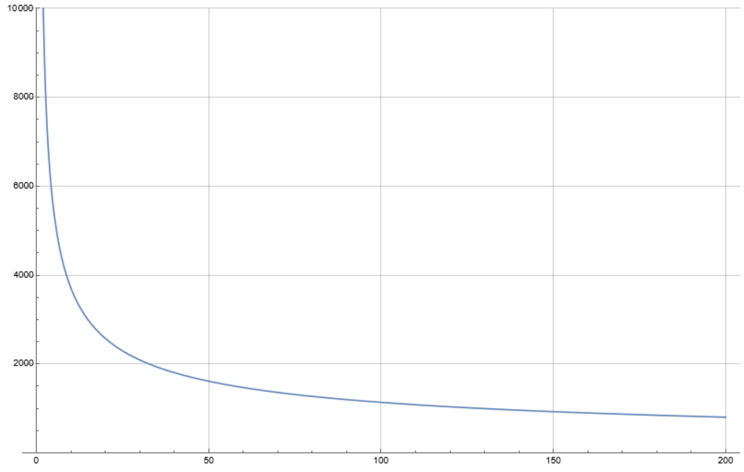


Figure 6.4: Plot of $\rho(n)$, n from 2 to 200 nodes.

6.1 Scaling of neighborhoods

6.1.1 Neighborhood graphs

The neighborhood graph models define a precise neighborhood $\mathcal{N}(v_i, v_j)$ for each pair of nodes v_i and v_j in V. Though the position of the neighborhood is determined by the position of these nodes, the shape and size of the neighborhood are solely determined by the great-circle distance $\operatorname{dist}(v_i, v_j)$ due to the rotational symmetry of $\mathbb{S}^2_\oplus$. We denote by $\mathcal{N}_\delta$ the *relative size* of a neighborhood of a pair of nodes at a distance of δ kilometers. By relative size we mean the proportion of the globe covered by the neighborhood, which can be stated in terms of the functions p and q defined at the start of this Chapter 6. Table 6.1 below lists the relative sizes of the neighborhoods of all neighborhood graph models.

Neighborhood graph model	**Relative size $\mathcal{N}_\delta$**	**Integral bounds $\mathcal{I}$**
RNG	$q(\delta, \delta)$	$[0, \pi r_\oplus/2]$
λ-RNG	$q\left(\frac{\delta}{\lambda}, \delta\right)$	$[0, \lambda\pi r_\oplus/2]$
GG	$p\left(\frac{\delta}{2}\right)$	$[0, \pi r_\oplus]$
βS	$q\left(\frac{\beta\delta}{2}, (\beta-1)\delta\right)$	$[0, \pi r_\oplus/\beta]$
εN	0	$[0, \varepsilon]$

Table 6.1: Relative neighborhood sizes of an edge with length δ km for various models.

The column *Integral bounds* refers to the bounds of the integral in $(*)$ in Section 6.3.1 below. This is the interval of distances δ that we condition on to determine the expected number of edges for a given neighborhood graph model. The εN graph with fixed ε, though not a real neighborhood graph model, can also be expressed in these terms by integrating over distances up to ε kilometers. The relative size $\mathcal{N}_\delta$ is zero in this case because the interaction with other nodes has no influence on determining edges.

6.1.2 Other graph models

As the number of nodes n increases, the nodes become more densely packed on $\mathbb{S}^2_\oplus$. This can be seen by considering $\rho(n)$, the expected minimum distance of a node's nearest neighbor, defined before. This minimum distance decreases at a rate of order $n^{-1/2}$. It is important to adjust the random graph models according to n to make sure the local behavior remains somewhat consistent. Take for example an εN graph with fixed ε, and let n grow to infinity: any node will asymptotically connect to linearly many other nodes.

All the neighborhood graphs and the kNN graph take care of this naturally. Indeed, the edges tend to have small localized neighborhoods, since large edges are bound to have other nodes situated in their neighborhoods. Similarly, the kNN graph only creates localized edges, since the distance of a node to any k nearest neighbors scales along with $\rho(n)$ at a rate of order $n^{-1/2}$.

The εN and the ε-kNN graphs must be manually adjusted according to n to retain consistent local behavior. Hence we must find an appropriate function $\varepsilon(n)$ to scale the maximal edge length down as n grows to infinity. If $\varepsilon(n)$ converges to 0 too quickly, the resulting εN and ε-kNN graphs are empty graphs for large enough n. Conversely, if $\varepsilon(n)$ converges to 0 too slowly, the resulting εN graph has node degrees growing to infinity, and the ε-kNN converges to the regular kNN graph. A good middle-ground is to scale $\varepsilon(n)$ linearly in $\rho(n)$:

$$\varepsilon(n) := c \cdot \rho(n) = c \cdot \frac{\pi r_\oplus}{4^{n-1}} \binom{2n-2}{n-1},$$

for some constant $c > 0$ of choice. The presence of a binomial coefficient can be computationally problematic for large n (Mathematica gives false results for n larger than about 500). However, there is a handy approximation derived from *Stirling's formula* which shows that

$$\varepsilon(n) \approx c \cdot \frac{\sqrt{\pi} r_\oplus}{\sqrt{n}}.$$

6.2 Epsilon neighborhood graph

In many aspects, the εN graph model is the simplest one. This holds especially true in terms of theoretical analysis of its properties. Whereas all other random graph models (kNN and all neighborhood graphs) must take the positioning of other nodes into account when generating edges between two nodes, edges in the εN graphs are selected solely based on their length.

In this section, we consider a set of points $V = \{v_1, \ldots, v_n\}$ of n nodes uniformly distributed on the surface $\mathbb{S}^2_\oplus$. Given a parameter $\varepsilon > 0$, we construct the εN graph on V and refer to it as $G = (V, E)$, where E is the set of edges. We then prove some general results of the properties of this graph, such as the number of edges, the number of triangles, and the degree distribution. See Table 6.2 below for a quick summary of the obtained results.

Property	**Value**
Distribution of node's degree	$\mathrm{Bin}(n-1, p(\varepsilon))$
Expected number of edges	$\binom{n}{2} \cdot p(\varepsilon)$
Expected number of triangles	$\binom{n}{3} \cdot \tilde{p}(\varepsilon)p^2(\varepsilon)$

Table 6.2: Summary of εN graph theoretical results.

6.2.1 Degree distribution

Consider an arbitrary node $v_i \in V$. The ε-neighborhood of node v_i is $\mathcal{B}_\oplus(v_i, \varepsilon)$, and the probability that a node uniformly distributed on $\mathbb{S}^2_\oplus$ is within this neighborhood is equal to $p(\varepsilon)$. The $n-1$ nodes in $V \setminus \{v_i\}$ are i.i.d. uniformly distributed, and hence the probability distribution of the degree of v_i is the sum of $n-1$ independent Bernoulli variables with probability $p(\varepsilon)$:

$$\deg(v_i) \sim \mathrm{Bin}(n-1, p(\varepsilon)).$$

The same holds for the degree of every node in V. Though the degrees of the n nodes are identically distributed, they are not independent. One can easily imagine a situation where one node has an extremely high degree which affects the probability distribution of the degree of the other nodes. The degrees of nodes are also not pairwise uncorrelated. Consider for example a simple case where $n = 3$, and let X, Y, Z be the degrees of the three nodes:

$$\mathbb{E}[X] \cdot \mathbb{E}[Y] = 4p^2(\varepsilon),$$

whereas, since it is possible that $X = Y = 1$, we have

$$\mathbb{E}[XY] = \sum_{k=0}^{4} k \cdot \mathbb{P}(XY = k) > \sum_{k=2}^{4} k \cdot \mathbb{P}(XY = k)$$

and moreover,

$$\begin{aligned}\sum_{k=2}^{4} k \cdot \mathbb{P}(XY = k) &= 2 \cdot \mathbb{P}(X = 1, Y = 2)\\ &\quad + 2 \cdot \mathbb{P}(X = 2, Y = 1)\\ &\quad + 4 \cdot \mathbb{P}(X = 2, Y = 2)\\ &= 4 \cdot \mathbb{P}(X = 2, Y = 1)\\ &\quad + 4 \cdot \mathbb{P}(X = 2, Y = 2)\\ &= 4 \cdot \mathbb{P}(X = 2),\end{aligned}$$

where the last equality follows from the fact that $X = 2$ implies $Y \geq 1$. This eventually leads to

$$\mathbb{E}[XY] > 4 \cdot \mathbb{P}(X = 2) = 4 \cdot p^2(\varepsilon) = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

6.2.2 Number of edges

We wish to know the expectation of the number of edges $|E|$. Express $|E|$ as the sum of indicator functions over all distinct pairs of nodes, and take expectations on both sides to obtain

$$|E| = \sum_{v_i \neq v_j \in V} \mathbb{1}\{(v_i, v_j) \in E\},$$

and

$$\mathbb{E}[|E|] = \sum_{v_i \neq v_j \in V} \mathbb{P}((v_i, v_j) \in E).$$

Due to the uniform distribution of nodes, this probability is equal for any pair of nodes, hence we can simplify the sum to

$$\mathbb{E}[|E|] = \binom{n}{2} \mathbb{P}((u, v) \in E),$$

where u and v are now arbitrary nodes uniformly distributed on $\mathbb{S}^2_\oplus$. To compute the probability we must condition on the distance $\text{dist}(u, v) = \delta$ as follows.

$$\mathbb{E}[|E|] = \binom{n}{2} \int_0^{\pi r_\oplus} \mathbb{P}((u, v) \in E \,|\, \text{dist}(u, v) = \delta) \cdot f_{\text{dist}}(\delta)\, d\delta \qquad (*)$$

where $f_{\text{dist}}(\delta) = \frac{\sin(\delta / r_\oplus)}{2 r_\oplus}$ is the p.d.f. of the distance between two uniformly distributed nodes on $\mathbb{S}^2_\oplus$. The conditional probability can now be simplified to the indicator function.

$$\begin{aligned}\mathbb{E}[|E|] &= \binom{n}{2} \int_0^{\pi r_\oplus} \mathbb{1}\{\delta < \varepsilon\} \cdot f_{\text{dist}}(\delta)\, d\delta\\ &= \binom{n}{2} \int_0^{\varepsilon} f_{\text{dist}}(\delta)\, d\delta\\ &= \binom{n}{2} \cdot F_{\text{dist}}(\varepsilon)\\ &= \binom{n}{2} \cdot p(\varepsilon).\end{aligned}$$

The *Handshaking lemma* allows us to verify that, for arbitrary $v_i \in V$,

$$\mathbb{E}[\deg(v_i)] = \frac{2}{n} \cdot \mathbb{E}[|E|] = \frac{2}{n} \binom{n}{2} p(\varepsilon) = (n - 1) p(\varepsilon);$$

which confirms results from Section 6.2.1.

6.2.3 Number of triangles

We denote by Δ the total number of triangles in G, and we denote by $\Delta(v_i)$ the number of triangles that node $v_i \in V$ is a part of. Node v_i is part of a triangle when two of its neighbors are themselves connected, which means that they are also within ε kilometers of each other. The probability that two vertices, uniformly distributed in the ε-neighborhood of v_i, are within each others' ε-neighborhoods is denoted by $\tilde{p}(\varepsilon)$. Hence the expected number of triangles that node v_i is a part of is

$$\begin{aligned}\mathbb{E}[\Delta(v_i)] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{u,v\in\mathcal{B}_\oplus(v_i,\varepsilon)} \mathbb{1}\{(u,v)\in E\} \,\middle|\, \deg(v_i)\right]\right]\\ &= \mathbb{E}\left[\binom{\deg(v_i)}{2}\cdot\tilde{p}(\varepsilon)\right]\\ &= \frac{\mathbb{E}[\deg(v_i)\cdot(\deg(v_i)-1)]\cdot\tilde{p}(\varepsilon)}{2}.\end{aligned}$$

Since node v_i was taken arbitrarily, we can compute the expectation of the total number of triangles by summing over all nodes, with an added factor of $1/3$ because each triangle is counted three times.

$$\begin{aligned}\mathbb{E}[\Delta] &= \frac{1}{3}\cdot\sum_{v\in V}\mathbb{E}\left[\Delta(v)\right]\\ &= \frac{n}{3}\cdot\mathbb{E}[\Delta(v)]\\ &= \frac{n\tilde{p}_\varepsilon}{6}\cdot\big(\mathbb{E}[\deg(v)^2]-\mathbb{E}[\deg(v)]\big)\\ &= \frac{n\tilde{p}(\varepsilon)}{6}\cdot\big((n-1)p(\varepsilon)(1-p(\varepsilon))+(n-1)^2p^2(\varepsilon)-(n-1)p(\varepsilon)\big)\\ &= \binom{n}{3}\cdot\tilde{p}(\varepsilon)p^2(\varepsilon).\end{aligned}$$

6.3 Neighborhood graphs

In this section, we provide some theoretical results that hold for general neighborhood graphs. We consider a set of points $V = \{v_1,\dots,v_n\}$ of n nodes uniformly distributed on the surface $\mathbb{S}^2_\oplus$. Given a neighborhood model of choice that defines a neighborhood $\mathcal{N}(v_i,v_j)$ for every distinct pair of nodes $v_i, v_j \in V$, we construct the neighborhood graph on V and refer to it as $G = (V,E)$. Then we prove some general results regarding the properties of this graph.

6.3.1 Number of edges

We wish to know the expectation of the number of edges $|E|$. We can start similarly to what we did for the εN graph in Section 6.2.2. Specifically, we start from the expression labeled by $(*)$.

$$\mathbb{E}[|E|] = \binom{n}{2}\int_0^{\pi r_\oplus}\mathbb{P}((u,v)\in E\,|\,\mathrm{dist}(u,v)=\delta)\cdot f_{\mathrm{dist}}(\delta)\,d\delta.$$

The size of the neighborhood $\mathcal{N}(u,v)$ is uniquely determined by the distance $\mathrm{dist}(u,v)=\delta$. We can refer back to Table 6.1 for the *relative sizes* of the neighborhoods $\mathcal{N}_\delta$, that is the proportion of the sphere $\mathbb{S}^2_\oplus$ covered by the neighborhood.

Given that the distance between nodes u and v is δ, they form a neighborhood of size $\mathcal{N}_\delta$. Since there are $n-2$ other nodes which are uniformly distributed on $\mathbb{S}^2_\oplus$, we have that

$$\mathbb{P}((u,v)\in E\,|\,\mathrm{dist}(u,v)=\delta) = (1-\mathcal{N}_\delta)^{n-2}.$$

When substituting this into the integral, it is important to realize that some neighborhoods are measured with the function q, which is only accurate for neighborhoods smaller than half the globe. Hence the integral bounds must be restricted to those stated in Table 6.1. This means we are only counting edges with lengths below a certain threshold. Note however that is threshold is at least $\pi r_\oplus/2$ and that the average nearest neighbor is approximately at distance $\sqrt{\pi} r_\oplus/\sqrt{n}$, so even $n \geq 2$ already provides us with a node at a distance within this threshold. We denote the integral bounds given for the particular neighborhood graph model by $\mathcal{I}$.

$$\mathbb{E}[|E|] \approx \binom{n}{2} \int_{\mathcal{I}} (1 - \mathcal{N}_\delta)^{n-2} \cdot \frac{\sin(\delta/r_\oplus)}{2r_\oplus} \, d\delta.$$

Due to the *Handshaking lemma*, it immediately follows that the average degree of a node $v_i \in V$ is

$$\mathbb{E}[\deg(v_i)] \approx (n-1) \int_{\mathcal{I}} (1 - \mathcal{N}_\delta)^{n-2} \cdot \frac{\sin(\delta/r_\oplus)}{2r_\oplus} \, d\delta.$$

Chapter 7

Simulation results

Stochastic simulation is a powerful tool that has the potential to generate a great variety of accurate results. Though it requires a substantial number of trials ($N = 10\,000$ in our case), through optimization and the use of clever libraries the computational times can be significantly reduced. For this report, we use Python and the *networkx*[1] library. See the public repository *geographic-graphs*[2] on GitHub for more info about the Python code used for the simulations.

In this chapter, we display and discuss the simulation results. In Section 7.1 we discuss some techniques applied during simulation and define some useful metrics for evaluating the results. In Section 7.2 the optimal results are displayed and discussed for every random graph model.

7.1 Methodology

The sampling of uniformly distributed nodes on the sphere is done as described in Section 4.3. For each trial run, a set of n points is generated uniformly on the set $[0, 1]^2$. Then the transformations denoted by $\tilde{F}_{\text{inc}}^{-1}$ and $\tilde{F}_{\text{az}}^{-1}$ are applied to the coordinates of these points. The result is a set V of (latitude, longitude) coordinates which are uniformly distributed on $\mathbb{S}^2_{\oplus}$.

With the help of the *networkx* library, a variety of methods taking the set of nodes V, applying some random graph model procedure, and generating samples of the graph models are defined. The diagram of related graph models in Figure 5.24 is useful for optimizing these sampling methods: if a naive approach of comparing all pairwise candidates for edges is not feasible, one may use a comparable graph model to start from and add/remove edges to generate the desired sample graph. In particular, the Delaunay triangulation (DT) turns out to be very useful since it contains most other graph models considered in this report. With the help of a method (called *scipy.spatial.Delaunay*) the DT graph can be constructed in $O(n \log n)$ time. Then the $O(n)$ edges of the DT can be filtered to fit the desired graph model, giving an overall complexity of $O(n \log n)$. A naive approach that looks at all pairs of nodes has complexity $O\left(n^2\right)$, and is hence much more time-consuming.

To compare a sample degree distribution to the worldgraph's degree distribution, we can measure the total variation distance between them as defined below, provided that we normalize both distributions to probability distributions.

Definition 7.1 (Total variation distance) *Given a measurable space $(\Omega, \mathcal{F})$ and probability measures P and Q defined on $(\Omega, \mathcal{F})$, we define the total variation distance between P and Q as*

$$d_{TV} := \sup_{A \in \mathcal{F}} |P(A) - Q(A)| .$$

Informally, this is the largest possible difference between the probabilities that the two probability distributions P and Q can assign to the same event $A \in \mathcal{F}$. Given that the set $\Omega \subset \mathbb{N}$

[1] Release 3.1, https://networkx.org/documentation/stable/

[2] https://github.com/aronvv1996/geographic-graphs

is discrete, we simply take the power set of Ω as the σ-algebra: $\mathcal{F} = 2^{\Omega}$. Hence, in practice, we could iterate over all possible subsets A of the set Ω of possible degrees attained by either degree distribution and take $\mathrm{d_{TV}}$ to be the highest difference in probability between the two distributions attaining the event A. However, a much more efficient way is to use the following identity which holds when Ω is countable [5, Proposition 4.2].

$$\mathrm{d_{TV}} = \frac{1}{2}\left\|P - Q\right\|_1 = \frac{1}{2}\sum_{\omega\in\Omega}\left|P(\{\omega\}) - Q(\{\omega\})\right|.$$

7.2 Results

In this section, we systematically examine the simulation results for each random graph model, with emphasis on the best-performing parameter values for each model. For the comprehensive lists of results, we refer the reader to Appendix A.

The results for the random graph models are accompanied by figures showing plots of 'optimal' sample graphs. To generate these samples, first a set of 100 distinct configurations of nodes ($n = 171$) on $\mathbb{S}^2_{\oplus}$ is generated. Then each random graph model is applied to these 100 sets of nodes. Finally, for each random graph model, the best-performing sample out of the 100 in terms of total variation distance is then chosen to be shown as a figure in the following sections.

7.2.1 Epsilon neighborhood graph

We have iterated $N = 10\,000$ simulations of εN graphs for values of ε ranging from 100 to 3 000 km in steps of 100 km. See Table 7.1 for optimal results, and Table A.1 for all results.

ε	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1800	287.8939	281.6880	**188.9290**	1322.0830	0.1402
1900	**320.9791**	309.0801	235.4559	1771.1231	**0.1277**
WG	332	-	173	-	-

Table 7.1: Optimal simulation results of εN graph model, $N = 10\,000$.

For $\varepsilon = 1\,800$ km the sample εN graphs optimize the number of triangles. The number of triangles of the worldgraph is within 1 standard deviation (SD) of the mean.

For $\varepsilon = 1\,900$ km the sample εN graphs optimize the number of edges and the total variation distance. The number of edges of the worldgraph is within 1 SD of the mean, and the total variation distance is 0.1277, which is the optimal value over all random graph models in this report.

We can compare these empirical results with the theoretical optimal values that follow from the formulas in Section 6.2. Optimizing the number of edges gives

$$332 = \binom{171}{2}\cdot p(\varepsilon),$$

which resolves to $\varepsilon \approx 1\,933.16$ km. Optimizing the number of triangles gives

$$173 = \binom{171}{3}\cdot \tilde{p}(\varepsilon)p^2(\varepsilon),$$

which resolves to $\varepsilon \approx 1\,759.06$ km. This does not contradict the experimental results.

Figures 7.1 and 7.2 show the average degree distributions of εN graphs, with the sample distribution in blue and the WG distribution in grey. The sample distributions are notably similar to binomial distributions, particularly when comparing them to the binomial fit from Figure 3.3. This gives credit to the notion that the degrees of separate nodes are largely independent of one

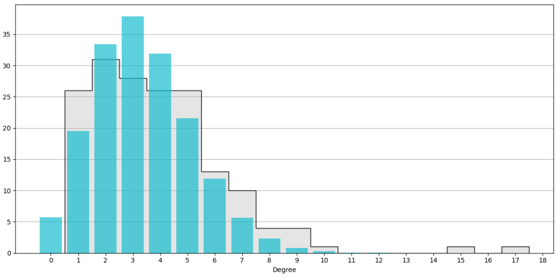


Figure 7.1: Average degree distribution of εN graph, $\varepsilon = 1\,800$ km.

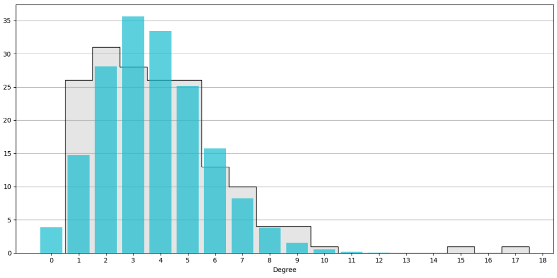


Figure 7.2: Average degree distribution of εN graph, $\varepsilon = 1\,900$ km.

another. The probability distribution of the degree of a single node is $\text{Bin}(n-1, p(\varepsilon))$, and if all degrees in the graph were independent the overall degree distribution would be simply equivalent (when normalized to a probability distribution that is). An argument as to why this dependence of degrees seems to have little effect is that, for large values of n, we have the following relationship.

$$\text{Bin}(n-1, p(\varepsilon)) \xrightarrow{d} \text{Poi}(n \cdot p(\varepsilon)).$$

Hence we can approximate the uniform binomial point process we're using to generate nodes on $\mathbb{S}^2_{\oplus}$ by a much simpler *Poisson point process*. It is important to note that $p(\varepsilon)$ is roughly proportional to ε^2, which in turn scales at a rate of $O\left(n^{-1/2}\right)$ (using $\varepsilon(n)$ as defined in Section 6.1.2). Hence the product $n \cdot p(\varepsilon)$ has a constant limit λ for $n \to \infty$, which is equal to $c^2\pi/4$. The Poisson point process has two important properties: any Borel measurable subset of $\mathbb{S}^2_{\oplus}$ (such as ε-neighborhoods, see Appendix B.2) has a number of nodes given by $\text{Poi}(\lambda)$, which is also the distribution of the degree of a single node. Additionally, the numbers of nodes in two *disjoint* Borel sets are independent, which means that nodes at a distance of at least 2ε km have independent degrees. By allowing some simplifications, we can thus show that the degrees of nodes are only *locally* dependent.

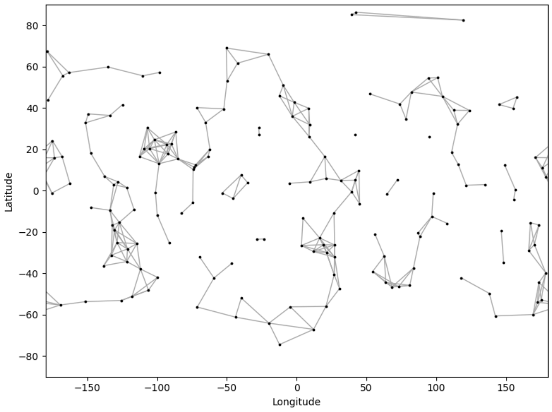


Figure 7.3: Optimal graph sample of εN graph, $\varepsilon = 1\,800$ km.

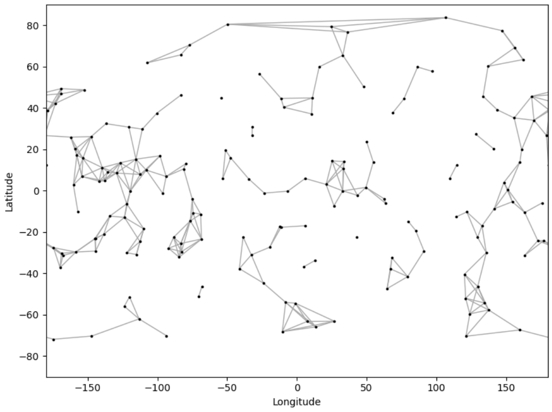


Figure 7.4: Optimal graph sample of εN graph, $\varepsilon = 1\,900$ km.

Figures 7.3 and 7.4 show some samples of graphs that are optimized in terms of total variation distance from the WG degree distribution. These samples also perform fairly well in terms of edges and triangles. However, a large deal of these triangles are situated in large cliques. One can easily spot a set of five or six nodes that are clustered together enough to form a K_5 or K_6 subgraph; seeing as a K_n clique contains $\binom{n}{3}$ triangles, it is clear how easily this number of triangles is achieved. The actual pattern of triangles we aim for is more akin to a Delaunay triangulation pattern: the main difference here is the conservation of planarity.

7.2.2 K-nearest neighbors graph

We have iterated $N = 10\,000$ simulations of ε-kNN graphs for $k \in \{2, 3, 4, 5, 6\}$ and for values of ε ranging from 1 000 to 4 000 km in steps of 100 km. See Table 7.2 for optimal results of each value of $k \in \{2, 3, 4, 5, 6\}$, and Tables A.2-A.6 for all results.

k	ε	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
2	1700	183.6532	32.6435	40.1012	21.5336	**0.4380**
3	1700	226.2579	77.1386	87.6401	98.4846	**0.2912**
4	1900	293.6258	109.7048	**164.4103**	280.0372	**0.2064**
4	2100	**335.1260**	77.3499	203.7909	267.8722	0.2749
5	1800	282.2988	192.2227	**170.7785**	588.4872	**0.1394**
5	2000	**338.8322**	163.9550	236.1235	632.9852	0.1944
6	1800	286.3045	245.1154	**182.7368**	912.1153	0.1420
6	1900	**317.7360**	247.8865	223.2462	1076.2178	**0.1398**
-	**WG**	332	-	173	-	-

Table 7.2: Optimal simulation results of ε-kNN graph model, $N = 10\,000$.

The above Table 7.2 lists all optimal values of ε for each value of k. For $k = 2$ and $k = 3$, the numbers of edges and triangles are simply too low to be reasonably considered. The ε-kNN graphs converge in the limit $\varepsilon \to \pi r_\oplus$ to the simple kNN graphs. The 2NN graph has in expectation only 220 edges and 51 triangles, and the 3NN graph has only 318 edges and 144 triangles. So, regardless of some nice properties such as planarity, the 2NN and 3NN graphs are simply not feasible models of the worldgraph.

One can always optimize ε (by, say, binary search) to perfectly match either the expected number of edges or triangles of the sample graphs with the worldgraph, just like was done for the εN graph. Say we find two optimizing values of ε, ε_E and ε_Δ respectively for the number of edges and triangles. As we increase k, simulations suggest that the gap between ε_E and ε_Δ monotonically decreases. Hence for higher values of k, we can identify a single optimizing value ε that performs reasonably well in both regards. In the limit $k \to n - 1$, we are simply generating εN graphs, which turn out to have the smallest gap between ε_E and ε_Δ (about 174 km).

Figures 7.5-7.12 on the next page show the average degree distributions of ε-kNN graphs, for the values presented in Table 7.2, with the sample distribution in blue and the WG distribution in grey. To see how the degree distribution evolves, we can fix k and gradually increase ε. The first thing to notice is that, for small ε, the graphs behave similarly to the regular εN graphs. To be more specific, for large enough n, the expected distance from a node to its k^{th} nearest neighbor is roughly

$$\begin{aligned}\mathbb{E}[\text{distance to } k^{\text{th}} \text{ nn}] &\approx \frac{2k-1}{2k-2} \cdot \mathbb{E}[\text{distance to } (k-1)^{\text{th}} \text{ nn}] \\ &= \prod_{i=2}^{k} \frac{2i-1}{2i-2} \cdot \rho(n),\end{aligned}$$

where $\rho(n)$ is the expected distance to the nearest neighbor. Hence if ε is smaller than this value for a given k and n, the restriction of connecting to only k nearest neighbors is 'overshadowed' by the restriction of maximum edge length. As ε surpasses this value, the degree distribution resembles more that of a geometric distribution with minimal degree k. This is what causes there to be high peaks in the degree distributions for the smaller values of k, like in Figures 7.5 and 7.6.

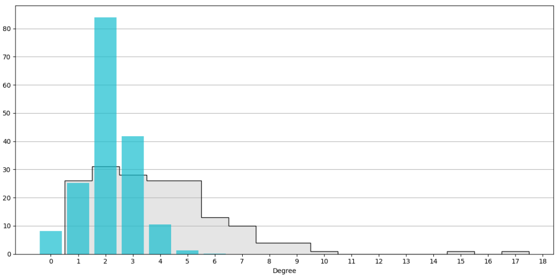


Figure 7.5: Average degree distribution of ε-2NN graph, $\varepsilon = 1\,700$ km.

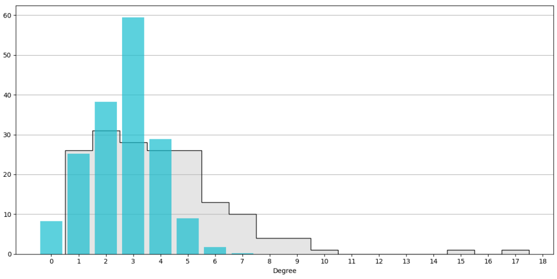


Figure 7.6: Average degree distribution of ε-3NN graph, $\varepsilon = 1\,700$ km.

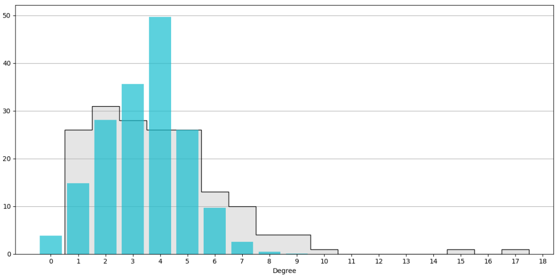


Figure 7.7: Average degree distribution of ε-4NN graph, $\varepsilon = 1\,900$ km.

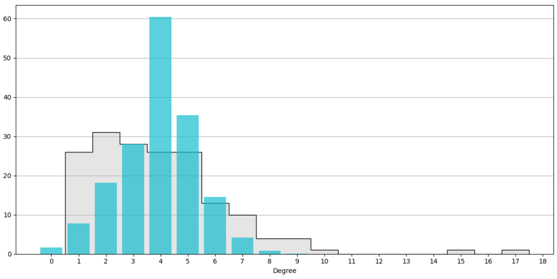


Figure 7.8: Average degree distribution of ε-4NN graph, $\varepsilon = 2\,100$ km.

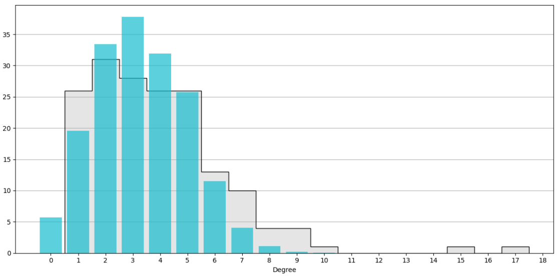


Figure 7.9: Average degree distribution of ε-5NN graph, $\varepsilon = 1\,800$ km.

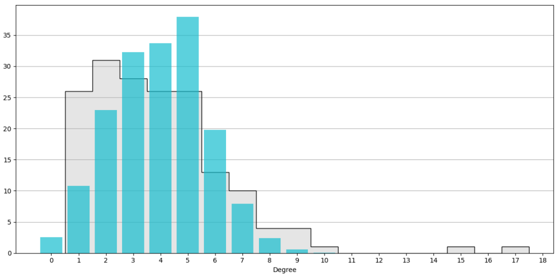


Figure 7.10: Average degree distribution of ε-5NN graph, $\varepsilon = 2\,000$ km.

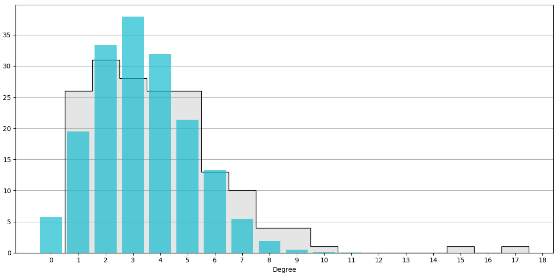


Figure 7.11: Average degree distribution of ε-6NN graph, $\varepsilon = 1\,800$ km.

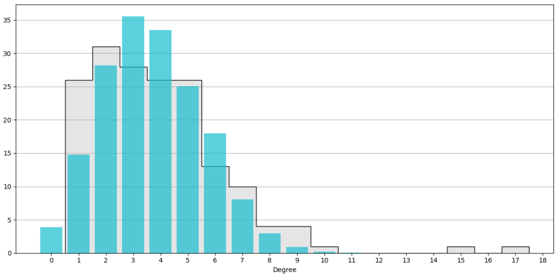


Figure 7.12: Average degree distribution of ε-6NN graph, $\varepsilon = 1\,900$ km.

Figures 7.13-7.20 show some samples of graphs that are optimized in terms of total variation distance from the WG degree distribution. It is evident in these plots how the ε-kNN graphs approach the εN graph as k increases. For given k, one can expect cliques up to K_{k+1} present in the sample graphs. This is partly the reason why we have chosen to investigate the ε-kNN graphs: to try and mitigate the presence of highly connected subgraphs present in the εN graphs. However, it is now clear that in order to preserve planarity and avoid K_5 subgraphs ($K_{3,3}$ subgraphs seem to be hardly present in general), one must sacrifice accuracy in terms of degree distribution. In these terms, the best we can achieve is setting $k = 4$ and ε between 1 900 and 2 100 km: this generates decent-looking graph samples (which can be observed in Figures 7.15 and 7.16), at the cost of a significant increase in total variation distance.

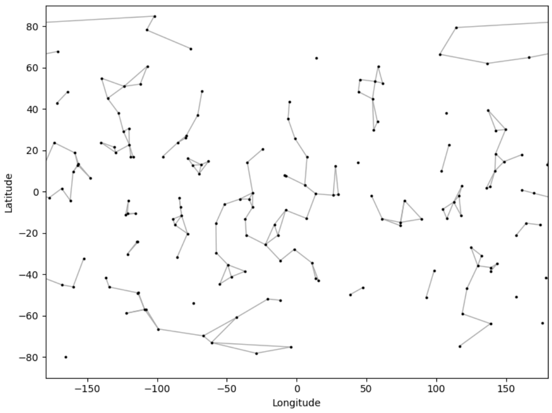


Figure 7.13: Optimal graph sample of ε-2NN graph, $\varepsilon = 1\,700$ km.

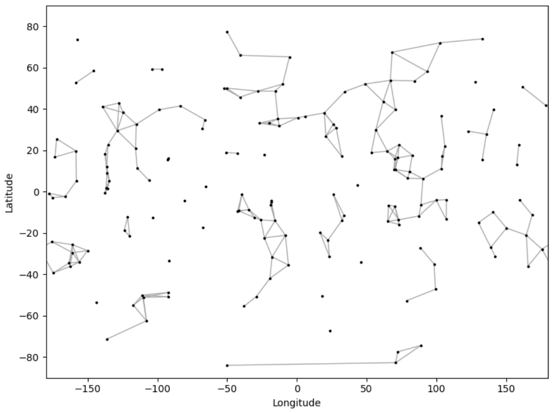


Figure 7.14: Optimal graph sample of ε-3NN graph, $\varepsilon = 1\,700$ km.

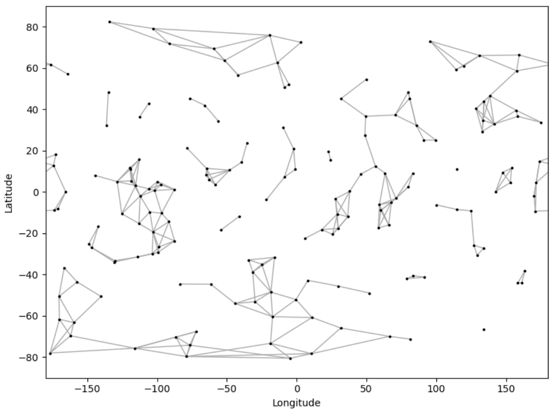


Figure 7.15: Optimal graph sample of ε-4NN graph, $\varepsilon = 1\,900$ km.

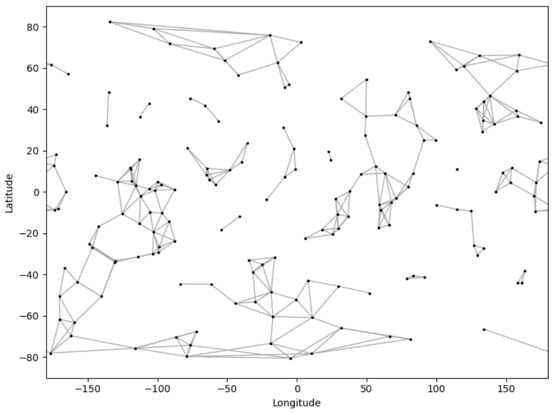


Figure 7.16: Optimal graph sample of ε-4NN graph, $\varepsilon = 2\,100$ km.

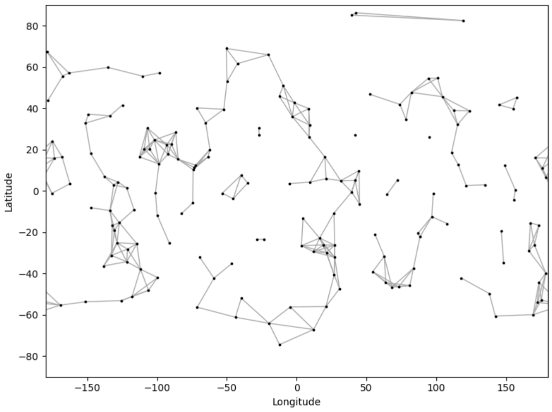


Figure 7.17: Optimal graph sample of ε-5NN graph, $\varepsilon = 1\,800$ km.

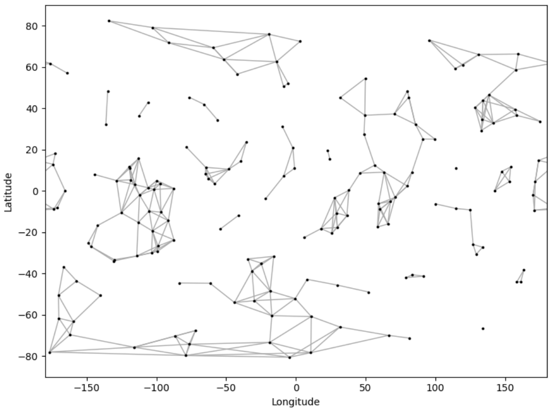


Figure 7.18: Optimal graph sample of ε-5NN graph, $\varepsilon = 2\,000$ km.

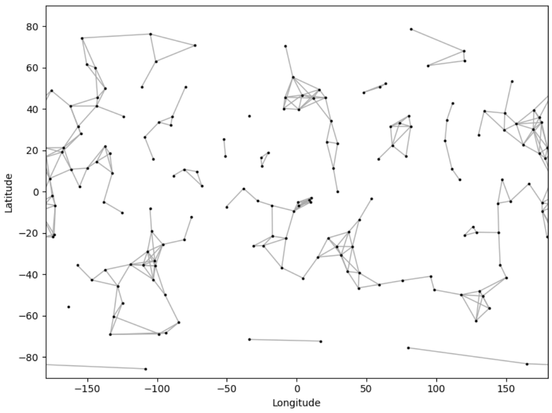


Figure 7.19: Optimal graph sample of ε-6NN graph, $\varepsilon = 1\,800$ km.

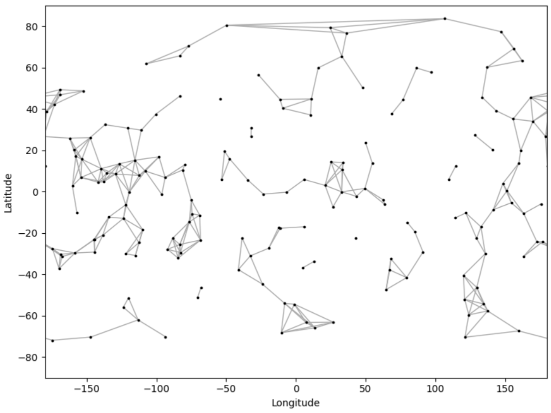


Figure 7.20: Optimal graph sample of ε-6NN graph, $\varepsilon = 1\,900$ km.

7.2.3 Generalized relative neighborhood graph

We have iterated $N = 10\,000$ simulations of λ-RNG graphs for values of λ ranging from 0.5 to 1.5 in steps of 0.05. See Table 7.3 for optimal results, and Table A.7 for all results.

λ	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1.15	**356.8429**	200.7916	**132.1616**	402.7249	**0.3739**
WG	332	-	173	-	-

Table 7.3: Optimal simulation results of λ-RNG graph model, $N = 10\,000$.

We find that $\lambda = 1.15$ is the best-performing parameter in terms of numbers of edges and triangles, *and* degree distribution. Then again, it's evident how distant the simulated mean of the number of triangles (132.16) is from the desired number (173). The next value for $\lambda = 1.2$ is already at 223.19 triangles in expectation. This underlines how quickly this number grows once the threshold of $\lambda = 1$ (with no triangles) is surpassed. It wouldn't be too hard however to find a more appropriate value for λ somewhere in the interval $(1.15, 1.2)$. The main issue is the total variation distance, also minimized for $\lambda = 1.15$, but equal to 0.37 which is considerably higher than the previous random graph models.

We can compare the stochastic results for the number of edges with the theoretical optimal value that results from the formula in Section 6.3.1. We have a relative neighborhood size of $\mathcal{N}_\delta = q\left(\frac{\delta}{\lambda}, \delta\right)$. Hence we must solve for λ in

$$332 = \binom{n}{2} \int_0^{\lambda \pi r_\oplus/2} \left(1 - q\left(\frac{\delta}{\lambda}, \delta\right)\right)^{n-2} \cdot \frac{\sin(\delta/r_\oplus)}{2r_\oplus}\, d\delta,$$

where we substitute $n = 171$, and $r_\oplus \approx 6\,371.009$ km. Solving this numerically with Mathematica gives an optimal value of $\lambda \approx 1.1283$.

Figure 7.22 shows a sample graph that is optimized in terms of total variation distance from the WG degree distribution. Though it has fewer triangles than the optimal εN graph, they are much more visually evident since they behave more like an actual triangulation and less like highly connected subgraphs. So, while from a numerical standpoint the εN graph performs better in terms of triangles, visually the λ-RNG graph is much more adherent to the worldgraph in terms of triangles.

Figure 7.21 shows the average degree distribution of the λ-RNG graph, with the sample distribution in blue and the WG distribution in grey. When one observes the evolution of the degree distribution as λ increases, there is always a clear peak with low variance. This peak slowly moves to the right, and the variance slowly increases; however, the variance only reaches a reasonable size when $\lambda > 1.5$, at which point the graph is not far from being a complete graph (which occurs at $\lambda = 2$). Simply put, the λ-RNG samples are too 'uniform' on the whole surface. This causes most nodes to have a very similar degree, whereas we wish for a similar number of nodes with degrees 1 up to 5.

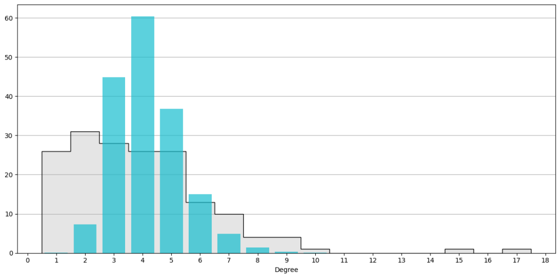


Figure 7.21: Average degree distribution of λ-RNG graph, $\lambda = 1.15$.

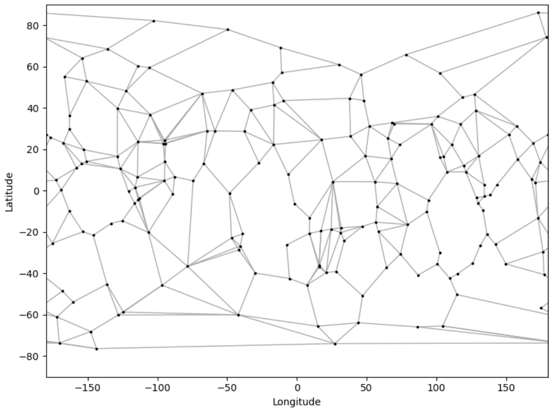


Figure 7.22: Optimal graph sample of λ-RNG graph, $\lambda = 1.15$.

We wish to reiterate that the shape of a neighborhood of a given neighborhood graph model has no influence on the total number of edges of its samples. The size of the neighborhood $\mathcal{N}_\delta$ is the only determining factor. However, the number of triangles is determined by both the shape and the size of the neighborhoods. Hence we are looking for a particular shape of neighborhood when trying to optimize both the number of edges and triangles simultaneously. The neighborhood shape dictated by the λ-RNG is not great. Other neighborhood graph models can provide a larger ratio of triangles to edges, with the Delaunay triangulation as an extreme case.

7.2.4 Beta skeleton graph

We have iterated $N = 10\,000$ simulations of βS graphs for values of β ranging from 1 to 2 in steps of 0.05. See Table 7.4 for optimal results, and Table A.8 for all results.

β	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1.00	**321.9711**	92.0405	**79.2850**	97.2028	**0.4003**
WG	332	-	173	-	-

Table 7.4: Optimal simulation results of βS graph model, $N = 10\,000$.

Even the optimal results are not great in terms of modeling the worldgraph. A look at the full results in Table A.8 suggests that all properties (edges, triangles, and degree distribution) improve gradually as β decreases from 2 (the relative neighborhood graph) down to 1 (the Gabriel graph). It may thus be tempting to increase the range of the parameter β to include values in the interval $(1/2, 1)$ — the definition of the beta skeleton graph can easily be extended to these values. However, this will probably not turn out to be fruitful seeing as for $\beta = 1$ the number of edges is fairly close to the target and the number of triangles is far from the desired number. This makes it difficult to identify a value for β that optimizes both these values.

Figure 7.23 shows the average degree distribution of the Gabriel graph, with the sample distribution in blue and the WG distribution in grey. The same issue arises as with the generalized RNG model: the variance in degrees is too low, which causes a surplus in nodes of degrees 3 and 4 and virtually no nodes of degree 1.

Figure 7.24 shows a sample graph that is optimized in terms of total variation distance from the WG degree distribution. It again seems at first glance to have a larger quantity of triangles when compared to Figure 7.3 of the εN graph. This is due to the absence of highly connected subgraphs, which in turn is due to the planarity of βS graphs (since $\beta\text{S} \subseteq \text{DT}$ when constructed on the same set of points V).

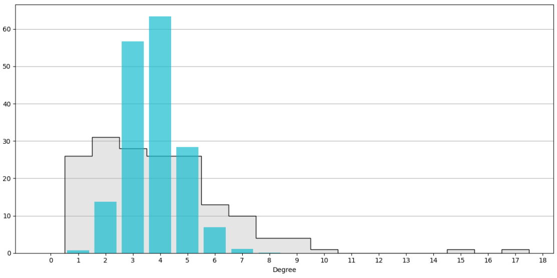


Figure 7.23: Average degree distribution of βS graph, $\beta = 1$.

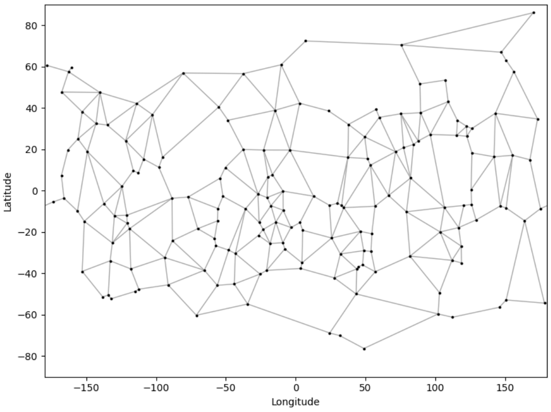


Figure 7.24: Optimal graph sample of βS graph, $\beta = 1$.

Chapter 8

Discussion

We discuss a few potential extensions to the contents of this report.

- **Other random graph models:** the graph models presented in this report are just a handful. There are many other available models, including other neighborhood graph models, such as the *sphere-of-influence graph* (connecting two vertices if their nearest-neighbors-circles intersect) or the *circle-based beta-skeleton.* There are virtually limitless many options for adjusting the existing graph models into new models.

- **Conceiving new neighborhood graph models:** as is discussed in Sections 7.2.3 and 7.2.4, which feature the neighborhood graph models considered in this report, a recurring issue is that the distribution of degrees is too concentrated around the mean — in other words, the graphs are too uniform. A solution to this might be to conceive a neighborhood definition that generates graphs that are not connected: except for λ-RNG graphs with $\lambda < 1$, all neighborhood graphs in this report are supergraphs of the MST and hence connected. The difficulty lies in defining a neighborhood graph model which is not connected but still generates an appropriate amount of edges and triangles.

  A different type of neighborhood shape that could be interesting is one that matches a perfect number of edges *and* a perfect number of triangles for modeling the worldgraph. Though the Delaunay triangulation model can be stated in terms of neighborhoods, it is not strictly a neighborhood graph model like the others as it does not define a unique neighborhood for every pair of nodes. It may be interesting to see what shape of neighborhood would maximize the number of triangles while keeping a constant number of edges. Though any shape of neighborhood can be defined, it is recommended to use neighborhoods defined in terms of distances for ease of use.

- **Other distributions of nodes:** the nodes in this report are always uniformly distributed on the whole surface of the Earth. One glance at the worldgraph in Figure 3.1 shows that this is not the case in reality, as is evident when comparing Europe to North America. The reason is the variance in country sizes on one hand, and the presence of oceans on the other.

  A *point process* is a collection of mathematical points randomly located on a mathematical space. The point process used in this report is a *uniform binomial point process*, meaning a fixed number of points n are i.i.d. uniformly distributed on the surface of the sphere. Other point processes exist, such as the *Poisson point process*, which is similar but has a randomized number of points. Some interesting behavior is shown by *Heterogeneous Poisson* and *Thomas cluster* point processes, which allow for creating less uniform point distributions with visible clusters. For general point processes there is a so-called *avoidance function* defined on a subset B of the underlying space ($\mathbb{S}^2_\oplus$ in our case) as the probability of B containing no points. This is exactly the probability of creating an edge in a neighborhood graph, where B is the neighborhood. A result for Poisson point processes called *Rényi's theorem* states that this probability is given by $e^{-\Lambda(B)} = e^{-n\cdot|B|/\left|\mathbb{S}^2_\oplus\right|}$.

- **Modelling other topological networks:** there are numerous other real-life networks of a geographical nature that can be analyzed and modeled by random graph models. One could model the map graphs of smaller regions like states or provinces, within the confines of countries or continents, and analyze how appropriate random graph models might differ from one part of the world to another. Since these graphs are also map graphs, it seems reasonable to assume they would more or less resemble the worldgraph analyzed in this report. However, confining the scope of the graph to, say, a country rather than the entire world might affect how homogeneous the graph looks: many countries are simply connected, compactly shaped, without many large protrusions, and subdivided fairly uniformly into equally sized regions. The graphs resulting from this kind of countries would probably be more similar to the Delaunay triangulation — it is possible to see this similarity even in the worldgraph (Figure 3.1), as Africa is a fairly compact shape with evenly sized countries, thus displaying a noticeable triangulation pattern.

  A different type of network might be the global network of roads, shipping routes, or air traffic. There are extensive resources freely available for a worldwide network of roads by NASA[1], which are more or less detailed in certain countries. One might choose to smoothen out the resulting graph to generate the simplest subdivision, as the actual shape of the stretches of road is not of interest, but the connections between road crossings are. In addition, this greatly simplifies the graph, as can be seen in Figures 8.1 and 8.2 depicting the road network of the Netherlands.

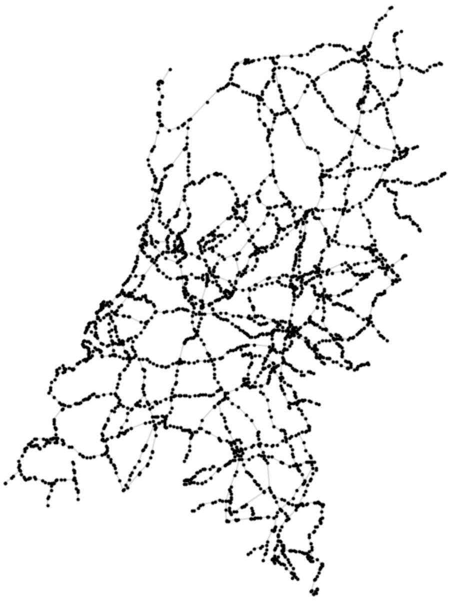


Figure 8.1: Road network of the Netherlands.

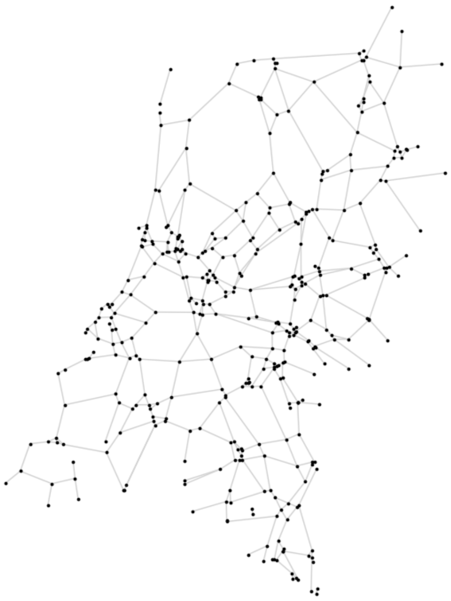


Figure 8.2: Smoothened road network.

The smoothened road network graph seems to feature a large number of cycles of various degrees. It bears a resemblance to the samples generated by the relative neighborhood graph, the Gabriel graph, or other beta-skeleton graphs. Though different from the worldgraph, the road network is also inherently a physical planar network and is therefore naturally modeled by localized graphs such as neighborhood graph models. Shipping- and air traffic routes, on the other hand, feature connections that are less bound to physical restrictions. The links can be much larger and the graphs do not preserve planarity. Instead of the localized

[1] https://sedac.ciesin.columbia.edu/data/set/groads-global-roads-open-access-v1

behavior of the graphs presented in this report, they might more accurately be modeled by scale-free networks, with a power-law degree distribution and several nodes of extremely high degree (*hubs*) identified with large cities.

- **Beta-skeleton analysis:** as was described in the original paper [2] that defined the beta-skeleton, the purpose of this model is to analyze where factors other than simple neighborliness are at play in defining the links between nodes. By comparing the actual network with the beta-skeleton generated on its nodes, one can identify these links as being somehow logistically/culturally/administratively significant. It might lead to some interesting insights when applied to various geographical networks around the globe.

- **Theoretical results:** there are a number of theoretical results left unproven, in particular concerning the first and second moments of various properties (number of edges, number of triangles, degree distribution) of the random graph models. Though many results of this kind exist for the same graph models defined on random point processes in Euclidean space, some work must be done to extend these results to our case of the 2-sphere. A paper by Penrose and Yukich [7] provides a very useful CLT-like result for functionals on Poisson and binomial point processes on a very general class of regions. In the same paper, CLT-like results are then shown to hold for the number of edges, components, and total edge length of various graph models. A few technical conditions on the type of graph model and the considered region (in our case, the 2-sphere $\mathbb{S}^2_{\oplus}$) make it difficult to extend those results to our case. The stochastic results seem to approach some normal distributions, which seems to suggest there might be similar results in our case, but this would require a more detailed understanding of the proofs in [7].

Chapter 9

Conclusion

We have considered a number of different random graph models and evaluated their potential to model the worldgraph. The three graph properties that we have studied are the number of edges, the number of triangles, and the degree distribution.

The four random graph models that we have analyzed are the epsilon-neighborhood graph, the epsilon-k-nearest neighbors graph with maximal edge length ε, the generalized relative neighborhood graph, and the beta skeleton graph. The feature shared by all these models is a parameter that takes a continuous range of values (ε, λ, and β). Since the expected number of edges and triangles are continuous functions dependent on the value of these parameters, by the Intermediate value theorem we can pinpoint a value for each parameter such that the expected number of edges or triangles matches perfectly with that of the worldgraph. That is, as long as this particular value is within the range of considered values for the parameter. This optimizing value can be found by solving the formula for number of edges or triangles from Sections 6.2 and 6.3, if such a formula is available. Otherwise, it can be found fairly easily with a binary search procedure. Hence we can find random graph models that perfectly match either the number of edges or triangles of the worldgraph. The exceptions are the ε-kNN graph for $k = 2, 3$ and the βS graph, though the latter can easily be adjusted with values of $\beta < 1$.

The difficulty lies in finding a random graph model with particular parameters that can optimize the number of edges and triangles *simultaneously*. If we denote by π_E and π_Δ respectively the optimizing values of the parameter π for number of edges and triangles, we wish for $|\pi_E - \pi_\Delta|$ to be minimized. The minimal difference we can achieve is with the εN graph, which gives $|\varepsilon_E - \varepsilon_\Delta| \approx 174.10$ km. Setting $\varepsilon = 1836.20$ km provides a good balance between optimizing both the number of edges and triangles.

Though the εN graph outperforms all other random graph models from a numerical perspective, from a visual perspective it has a clear issue. The presence of highly connected cliques makes the graph non-planar and artificially inflate the number of triangles. The ε-kNN graph with $k = 4$ and ε in the range $[1900, 2100]$ is a good alternative, as it performs fairly well in numerical terms and also avoids these highly connected cliques. The triangles in these graphs are visually more evident, as together they look more like actual triangulations — the sort that we would expect in a 3-map graph.

Regarding neighborhood graph models, we know that the number of edges is determined solely by the size of the neighborhoods, whereas the number of triangles also depends on their shape. Since there are neighborhood graph models with relatively few triangles and relatively many triangles, we hypothesize that we could find a shape of neighborhood that matches the number of edges *and* triangles of the worldgraph *exactly*; by continuously varying from one shape to another, the Intermediate value theorem again says there is some optimal intermediate shape. The challenge is defining how to continuously vary the shape while keeping the area constant and the shape simple (that is, able to be defined in terms of distances).

The last property to consider is the degree distribution. We have measured the accuracy of degree distribution by the total variation distance from the worldgraph's degree distribution. The

ε-neighborhood graph with $\varepsilon = 1900$ km is the best-performing model in this regard, with a total variation distance of about 0.1277. The ε-kNN graphs generally perform worse, though for higher values of k they approach the εN graph. The neighborhood graphs considered in this report are not able to approximate the worldgraph's degree distribution, given their total variation distances of around 0.40. In all cases, the degree distribution is far too concentrated around the mean. It is not presently clear to us how one could define a neighborhood graph model that solves this issue, or whether it is possible at all.

Bibliography

[1] Charles F.F. Karney. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55, 2013. 7

[2] David G. Kirkpatrick and John D. Radke. A Framework for Computational Morphology. In *Machine Intelligence and Pattern Recognition*, volume 2, pages 217–248. 1 1985. 18, 40

[3] Casimir Kuratowski. Sur le problème des courbes gauches en Topologie. *Fundamenta Mathematicae*, 15:271–283, 1930. 4

[4] Yongjae Lee and Woo Chang Kim. Concise Formulas for the Surface Area of the Intersection of Two Hyperspherical Caps. *KAIST Technical Report*, pages 1–22, 2014. 53

[5] David A Levin, Yuval Peres, and Elizabeth L Wilmer. *Markov Chains and Mixing Times, second edition.* American Mathematical Society, second edition, 2017. 30

[6] H Moritz. Geodetic reference system 1980. *Bulletin Géodésique*, 54(3):395–405, 1980. 7

[7] Mathew D. Penrose and J. E. Yukich. Central limit theorems for some graphs in computational geometry. *Annals of Applied Probability*, 11(4):1005–1041, 11 2001. 40

[8] Godfried T. Toussaint. The relative neighbourhood graph of a finite planar set. *Pattern Recognition*, 12(4):261–268, 1980. 16

[9] T. Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 23(176):88–93, 1975. 7

Appendix A

Results

A.1 Epsilon neighborhood graph

ε	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
100	0.9195	0.9090	0.0021	0.0021	0.9893
200	3.5761	3.5590	0.0279	0.0291	0.9590
300	8.0699	8.2764	0.1528	0.1985	0.9100
400	14.2814	13.9836	0.4690	0.5862	0.8461
500	22.3408	22.5805	1.1412	1.6947	0.8194
600	32.2402	31.9625	2.3787	4.0269	0.7926
700	43.8465	43.2957	4.3679	8.1525	0.7542
800	57.2911	57.5750	7.4551	16.4416	0.7026
900	72.4350	71.6928	12.0024	31.0720	0.6397
1000	89.1912	89.6206	18.0253	52.8933	0.5792
1100	107.9395	105.9166	26.4292	86.8868	0.5332
1200	128.5479	125.1823	37.6160	139.9831	0.4752
1300	150.6618	149.4644	51.5222	215.8281	0.4076
1400	174.5109	175.3987	69.4054	331.7655	0.3537
1500	200.5592	195.5111	91.7535	474.3973	0.2935
1600	228.4119	229.3670	119.3066	728.3410	0.2239
1700	257.3943	252.1802	151.5313	998.5898	0.1703
1800	287.8939	281.6880	**188.9290**	1322.0830	0.1402
1900	**320.9791**	309.0801	235.4559	1771.1231	**0.1277**
2000	354.9458	345.6787	287.7680	2455.7004	0.1554
2100	391.1673	376.9563	349.8490	3139.9980	0.1932
2200	429.0839	430.6805	421.5947	4420.5868	0.2566
2300	467.9676	453.9808	501.1981	5373.2283	0.3125
2400	509.5436	481.7977	595.1218	6848.1428	0.3768
2500	552.1842	532.0329	699.1325	8763.2215	0.4416
2600	597.2765	568.4616	819.6954	11054.5184	0.5125
2700	643.1303	603.6695	950.8765	13502.1300	0.5735
2800	690.7085	666.4313	1097.5889	16973.1911	0.6279
2900	739.4750	685.2276	1257.0520	19724.5531	0.6717
3000	791.0739	754.8052	1441.4340	25188.5084	0.7058
WG	332	-	173	-	-

Table A.1: Simulation results of εN graph model, $N = 10\,000$.

A.2 K-nearest neighbors graph

A.2.1 2NN

ε	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1000	85.4920	61.9501	12.2927	12.2188	0.6051
1100	101.0602	60.5412	16.1604	14.0265	0.5774
1200	116.9272	61.4835	20.4693	16.3501	0.5435
1300	132.4591	55.7115	24.8579	18.0473	0.5054
1400	147.1744	49.5266	29.1291	19.5688	0.4660
1500	160.8988	43.1954	33.2739	20.5759	0.4540
1600	173.0287	36.8813	36.8884	21.5947	0.4436
1700	183.6532	32.6435	40.1012	21.5336	**0.4380**
1800	192.4780	28.0909	42.7538	22.3726	0.4615
1900	199.8098	24.5766	45.0488	22.5934	0.4811
2000	205.5520	22.5805	46.6447	23.1707	0.4970
2100	209.8771	19.9812	47.8894	22.6412	0.5102
2200	213.2572	19.3354	48.9022	24.0432	0.5192
2300	215.6268	18.3641	49.5333	24.3293	0.5264
2400	217.2318	17.5605	49.9206	24.4865	0.5318
2500	218.3729	17.2616	50.3522	25.3868	0.5352
2600	219.1588	17.1736	50.5392	25.0655	0.5378
2700	219.6376	17.2463	50.6524	24.6620	0.5391
2800	220.0247	16.9097	50.6948	24.9517	0.5399
2900	220.2098	16.7038	50.8541	25.1194	0.5403
3000	220.3801	16.9330	50.8797	24.7784	0.5405
3100	220.3815	16.4212	50.7987	24.9214	0.5413
3200	220.3152	16.9474	50.8088	25.6550	0.5418
3300	220.4377	16.9149	50.8105	25.5150	0.5412
3400	220.4023	16.9435	50.8920	24.7477	0.5418
3500	220.3760	16.6232	50.8171	25.3234	0.5417
3600	220.3775	17.0532	50.8258	24.8439	0.5418
3700	220.4132	16.4399	50.7995	24.8301	0.5413
3800	220.5229	17.2867	50.8997	24.7396	0.5411
3900	220.4550	16.9450	50.8332	25.0576	0.5414
4000	220.4466	16.7147	50.8727	25.7447	0.5415
2NN	220.4401	17.0096	50.8328	24.6812	0.5414
WG	332	-	173	-	-

Table A.2: Simulation results of ε-2NN graph model, $N = 10\,000$.

A.2.2 3NN

ε	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1000	88.8976	80.2535	17.0333	33.2744	0.5781
1100	107.0543	93.0882	24.1073	47.5396	0.5319
1200	126.0432	100.4001	32.2894	60.2776	0.4759
1300	146.0772	104.9900	42.1602	74.6163	0.4270
1400	166.9557	101.3501	53.1722	83.9677	0.3940
1500	187.4749	95.5836	64.5199	90.0522	0.3561
1600	207.4332	87.6649	76.1818	95.3755	0.3144
1700	226.2579	77.1386	87.6401	98.4846	**0.2912**
1800	243.6841	63.1879	98.3822	95.4937	0.3081
1900	259.0955	54.1654	107.9548	96.0806	0.3370
2000	272.6394	45.1744	116.4969	96.2994	0.3752
2100	283.8091	38.9701	123.4319	93.6586	0.4076
2200	293.0336	34.8301	129.1139	91.9453	0.4344
2300	300.1082	32.4789	133.4986	92.1130	0.4555
2400	305.6089	29.7589	136.8466	94.4743	0.4721
2500	309.7990	28.3602	139.5450	93.0890	0.4836
2600	312.7009	27.5442	141.0713	95.8518	0.4931
2700	314.6989	26.2492	142.3080	94.5565	0.4992
2800	316.1305	25.4711	143.0437	93.1034	0.5037
2900	316.9731	26.1676	143.5322	96.9204	0.5067
3000	317.4849	25.6642	143.8576	98.4093	0.5086
3100	317.9759	25.9589	144.1058	98.9558	0.5090
3200	318.2176	25.7799	144.2020	96.1304	0.5103
3300	318.2753	25.9791	144.3413	96.0052	0.5106
3400	318.5053	25.3728	144.4627	96.0710	0.5103
3500	318.4698	25.4551	144.3653	93.6993	0.5108
3600	318.5217	25.9695	144.4054	95.2335	0.5104
3700	318.5133	26.0856	144.4609	96.4359	0.5107
3800	318.5141	25.5450	144.4457	93.6917	0.5109
3900	318.5370	25.6906	144.4026	96.4041	0.5108
4000	318.5454	25.6985	144.4634	96.3233	0.5107
3NN	318.3983	26.8457	144.3700	97.6133	0.5112
WG	332	-	173	-	-

Table A.3: Simulation results of ε-3NN graph model, $N = 10\,000$.

A.2.3 4NN

ε	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1000	89.1496	86.3178	18.0306	47.3461	0.5784
1100	107.8054	103.4229	26.0523	72.9250	0.5330
1200	128.2281	118.1765	36.6051	106.3462	0.4750
1300	149.7802	133.6463	49.2142	143.3549	0.4078
1400	172.8394	143.3856	64.4847	186.2944	0.3532
1500	197.1747	151.1702	82.4141	227.5468	0.2936
1600	221.6933	146.5660	101.7074	256.8988	0.2502
1700	246.4018	139.7468	122.4726	273.6964	0.2171
1800	270.5258	126.8025	143.2310	283.4590	0.2080
1900	293.6258	109.7048	**164.4103**	280.0372	**0.2064**
2000	315.4864	91.4946	185.0792	270.9131	0.2391
2100	**335.1260**	77.3499	203.7909	267.8722	0.2749
2200	352.3806	65.6801	220.6179	255.0733	0.3319
2300	366.9253	55.7525	234.3797	250.3543	0.3830
2400	379.0776	48.1354	246.0342	243.3692	0.4260
2500	388.7064	44.6632	255.1035	240.6430	0.4604
2600	396.2416	40.4220	262.5677	234.7768	0.4878
2700	401.8930	38.1224	267.9498	241.4603	0.5094
2800	406.0732	36.0058	271.6305	237.5652	0.5245
2900	409.0023	34.8847	274.3726	236.6082	0.5355
3000	410.9553	35.4437	276.1672	241.6560	0.5430
3100	412.2934	35.0549	277.4117	243.4288	0.5480
3200	413.1850	34.5956	278.1218	240.6554	0.5516
3300	413.6772	35.3778	278.5245	247.0344	0.5539
3400	414.1317	36.0548	279.0630	251.8934	0.5546
3500	414.1452	36.3463	278.8282	249.1109	0.5559
3600	414.1850	36.5468	278.8218	251.3360	0.5566
3700	414.2556	36.7473	278.7236	248.4854	0.5566
3800	414.3644	36.4432	279.0724	253.3818	0.5568
3900	414.3793	36.6658	278.8848	242.6611	0.5570
4000	414.4062	37.1106	279.1527	252.8428	0.5568
4NN	414.3902	35.9845	278.8011	248.1925	0.5567
WG	332	-	173	-	-

Table A.4: Simulation results of ε-4NN graph model, $N = 10\,000$.

A.2.4 5NN

ε	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1000	89.2584	89.6802	18.0960	52.1848	0.5786
1100	108.0901	104.1628	26.5634	83.4334	0.5324
1200	128.2785	122.7065	37.2414	129.0507	0.4759
1300	150.7241	146.8646	51.3872	195.4859	0.4068
1400	174.3818	164.1470	68.4165	268.2760	0.3534
1500	199.4900	178.8501	88.7065	350.4050	0.2949
1600	226.7601	196.4657	113.7948	464.0161	0.2237
1700	253.9958	202.3416	140.6269	539.0213	0.1705
1800	282.2988	192.2227	**170.7785**	588.4872	**0.1394**
1900	310.8313	180.1676	202.9436	616.1526	0.1598
2000	**338.8322**	163.9550	236.1235	632.9852	0.1944
2100	365.3918	139.7923	268.7726	621.0637	0.2212
2200	390.6247	123.1782	300.6323	612.7967	0.2784
2300	413.6419	100.3629	330.4017	566.3583	0.3330
2400	433.7459	82.9189	356.5369	542.5488	0.3949
2500	451.1818	70.2177	379.2315	507.1483	0.4552
2600	465.8443	63.5993	398.7054	511.6702	0.5061
2700	477.4844	55.7666	413.8319	487.5828	0.5485
2800	486.6965	52.0220	425.9940	490.6890	0.5825
2900	493.5793	49.4479	434.7234	503.4343	0.6081
3000	498.6603	48.2733	441.2890	494.6841	0.6272
3100	502.1322	46.8291	445.5828	492.9697	0.6409
3200	504.6729	46.3171	449.1198	516.1808	0.6510
3300	506.2970	46.1176	450.9078	506.4329	0.6571
3400	507.3831	47.5285	452.1156	511.4128	0.6616
3500	507.9652	46.1308	452.6353	498.9719	0.6643
3600	508.4731	47.9871	453.6036	513.0431	0.6660
3700	508.6987	47.6925	453.5277	514.7080	0.6670
3800	509.0016	48.4146	454.4241	520.6942	0.6674
3900	509.0276	48.0038	454.1369	509.0352	0.6678
4000	509.0385	48.9802	454.0572	533.9853	0.6681
5NN	509.0393	48.8056	454.2125	520.1785	0.6681
WG	332	-	173	-	-

Table A.5: Simulation results of ε-5NN graph model, $N = 10\,000$.

A.2.5 6NN

ε	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1000	89.4180	87.3067	18.2370	52.6812	0.5782
1100	108.0286	105.3556	26.5059	84.9254	0.5327
1200	128.4398	128.3174	37.4556	136.2024	0.4756
1300	150.8207	149.3378	51.7951	216.1129	0.4073
1400	174.6230	169.6063	69.3329	308.7937	0.3528
1500	200.3017	191.9239	90.9164	430.3584	0.2940
1600	227.4872	216.6002	116.9684	584.3448	0.2246
1700	256.2652	230.8367	147.6723	749.6393	0.1715
1800	286.3045	245.1154	**182.7368**	912.1153	0.1420
1900	**317.7360**	247.8865	223.2462	1076.2178	**0.1398**
2000	349.0150	237.5122	265.3880	1169.4667	0.1659
2100	381.0151	219.8873	311.7405	1206.7780	0.2069
2200	412.3150	201.3346	359.1912	1215.6512	0.2639
2300	442.1815	175.8128	405.7001	1208.8656	0.3132
2400	470.0823	142.1631	450.3330	1087.7381	0.3775
2500	495.6599	115.4966	492.0296	1018.4051	0.4435
2600	518.4481	100.2993	529.9884	990.9355	0.5159
2700	537.9246	86.0193	562.3086	952.1520	0.5803
2800	554.1896	74.6555	589.0157	908.7457	0.6352
2900	567.3999	69.2256	611.2687	877.2465	0.6813
3000	577.6020	64.3846	628.1383	887.0264	0.7175
3100	585.4231	61.5193	641.3832	903.7600	0.7462
3200	591.1569	60.3425	650.2755	923.7002	0.7675
3300	595.0636	56.9576	656.6551	883.0789	0.7829
3400	597.8221	57.7659	661.1194	917.0277	0.7935
3500	599.7436	59.2417	664.0271	923.6008	0.8006
3600	600.8608	59.9756	665.6703	952.5488	0.8054
3700	601.4882	60.0023	666.5531	934.6434	0.8083
3800	602.1446	61.0929	667.7760	939.4100	0.8103
3900	602.3129	62.9756	668.0601	952.6101	0.8114
4000	602.4522	62.3487	668.3469	948.8860	0.8121
6NN	602.6693	63.3613	668.4256	940.2173	0.8129
WG	332	-	173	-	-

Table A.6: Simulation results of ε-6NN graph model, $N = 10\,000$.

A.3 Generalized relative neighborhood graph

λ	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
0.50	31.2208	15.6250	-	-	0.8480
0.55	39.5483	17.5445	-	-	0.8409
0.60	49.2790	19.2852	-	-	0.8201
0.65	60.8248	22.4873	-	-	0.7802
0.70	74.2705	27.8755	-	-	0.7149
0.75	89.8198	32.8375	-	-	0.6588
0.80	108.0308	37.0061	-	-	0.6418
0.85	129.1893	41.2625	-	-	0.6020
0.90	153.7553	42.4274	-	-	0.5227
0.95	182.4430	35.9970	-	-	0.4952
1.00	215.5010	18.4384	-	-	0.5876
1.05	255.7518	54.8198	27.4962	39.0328	0.4640
1.10	302.0098	109.8019	69.3286	138.8364	0.4247
1.15	**356.8429**	200.7916	**132.1616**	402.7249	**0.3739**
1.20	421.8129	323.7791	223.1866	1002.1660	0.3791
1.25	499.2141	512.3987	354.3028	2294.6913	0.4560
1.30	592.3572	791.3772	542.4894	5098.0821	0.5697
1.35	705.7187	1203.7286	813.7904	11024.6489	0.7070
1.40	842.3748	1777.4387	1199.6006	22896.0589	0.7896
1.45	1015.1351	2732.4862	1777.9757	49885.8163	0.8780
1.50	1228.1240	3969.9650	2618.8958	104364.6577	0.9341
WG	332	-	173	-	-

Table A.7: Simulation results of λ-RNG graph model, $N = 10\,000$.

A.4 Beta skeleton graph

β	**Edges**	**Var edges**	**Triangles**	**Var triangles**	$\mathbf{d_{TV}}$
1.00	**321.9711**	92.0405	**79.2850**	97.2028	**0.4003**
1.05	318.7092	94.4490	74.4443	94.7473	0.4088
1.10	315.6950	87.3720	71.1877	87.9193	0.4142
1.15	310.7042	81.7655	66.2472	80.4371	0.4231
1.20	305.1693	80.5268	60.9773	75.5376	0.4284
1.25	299.8445	74.3479	56.1006	67.6795	0.4318
1.30	294.3574	70.4875	51.4940	63.3100	0.4313
1.35	289.3792	69.5688	47.4248	59.7967	0.4272
1.40	284.1860	65.9464	43.4370	54.5108	0.4202
1.45	279.4799	61.8022	39.9512	49.7112	0.4175
1.50	274.5234	58.3595	36.5477	46.1705	0.4243
1.55	269.7637	55.7195	33.2696	42.4615	0.4301
1.60	264.8063	52.7060	30.1148	37.9056	0.4355
1.65	259.8147	48.6858	27.0320	33.5068	0.4399
1.70	255.1322	45.8681	24.1819	31.1000	0.4436
1.75	250.0819	42.7502	21.3222	26.4740	0.4594
1.80	245.1202	38.3772	18.4393	22.8605	0.4771
1.85	240.0620	34.4660	15.5849	18.8912	0.4943
1.90	234.4390	30.6373	12.3437	14.6306	0.5139
1.95	226.9626	27.4538	7.5779	8.9639	0.5430
2.00	215.4726	18.6076	-	-	0.5873
WG	332	-	173	-	-

Table A.8: Simulation results of βS graph model, $N = 10\,000$.

Appendix B

Derivations

B.1 Jacobian determinant of spherical to Cartesian mapping

We wish to transform a particular integral from Cartesian coordinates to spherical coordinates (as defined according to ISO standard 80 000-2:2019[1]). In short, spherical coordinates are described by the tuple (r, θ, ϕ), resp. the *radial distance*, *polar angle*, and *azimuthal angle*. These take values in

$$r \geq 0, \qquad \theta \in [0, \pi], \qquad \phi \in [0, 2\pi).$$

Cartesian coordinates may be retrieved from spherical coordinates by

$$\begin{cases} x = r\sin(\theta)\cos(\phi), \\ y = r\sin(\theta)\sin(\phi), \\ z = r\cos(\theta). \end{cases}$$

When applying an integral change of variables from Cartesian to spherical, we need to multiply by the Jacobian determinant

$$\begin{aligned} \left|\frac{\delta(x,y,z)}{\delta(r,\theta,\phi)}\right| &= \begin{vmatrix} \sin(\theta)\cos(\phi) & r\cos(\theta)\cos(\phi) & -r\sin(\theta)\sin(\phi) \\ \sin(\theta)\sin(\phi) & r\cos(\theta)\sin(\phi) & r\sin(\theta)\cos(\phi) \\ \cos(\theta) & -r\sin(\theta) & 0 \end{vmatrix} \\ &= r^2\left(\cos^2(\theta)\sin(\theta)\cos^2(\phi) + \cos^2(\theta)\sin(\theta)\sin^2(\phi) + \sin^3(\theta)\cos^2(\phi) + \sin^3(\theta)\sin^2(\phi)\right) \\ &= r^2\sin(\theta). \end{aligned}$$

B.2 Area of a neighborhood

Define $\mathcal{B}_\oplus(x, \varepsilon) \subseteq \mathbb{S}^2_\oplus$ as the set of points on the globe which are within a great-circle distance of ε km from the point $x \in \mathbb{S}^2_\oplus$. This is the ε-neighborhood of x, and it is a measurable set: by rotational symmetry of $\mathbb{S}^2_\oplus$, we can rotate x onto the 'north pole' of the globe (where the inclination θ is zero); then the set $\mathcal{B}_\oplus(x, \varepsilon)$ is the preimage of $[0, \alpha)$ under the function $\tilde{F}^{-1}_{\text{inc}}$ defined in Section 4.2. The angle $\alpha = \varepsilon / r_\oplus$ is the *angular distance*, that is the angle at the center of the globe that subtends the radius of the ε-neighborhood.

The ε-neighborhood is the intersection of a cone of angle α, with its apex at the center of $\mathbb{S}^2_\oplus$, and the surface $\mathbb{S}^2_\oplus$. Given this angle α, we can integrate the volume element $d\Omega = r^2\sin(\theta) \cdot drd\theta d\phi$

[1] https://www.iso.org/standard/64973.html, or freely accessible on https://en.wikipedia.org/wiki/Spherical_coordinate_system.

over the appropriate subset of the globe:

$$\begin{aligned} |\mathcal{B}_{\oplus}(x,\varepsilon)| &= \left(\int_{\mathcal{B}_{\oplus}(x,\varepsilon)} d\Omega\right) \\ &= \int_0^{2\pi}\int_0^{\alpha} r_{\oplus}^2 \sin(\theta)\, d\theta\, d\phi \\ &= 2\pi r_{\oplus}^2 \left(1-\cos\left(\frac{\varepsilon}{r_{\oplus}}\right)\right). \end{aligned}$$

We define the function $p:\mathbb{R}\to\mathbb{R}$ as the area of the neighborhood over the total area of the globe. It is also the probability that a vertex taken uniformly from the surface of the globe is within a given ε-neighborhood. By rotational symmetry of $\mathbb{S}^2_{\oplus}$ there is no need to specify the center point of the neighborhood.

$$p(\varepsilon) := \frac{|\mathcal{B}_{\oplus}(v,\varepsilon)|}{4\pi r_{\oplus}^2} = \frac{1}{2}\left(1-\cos\left(\frac{\varepsilon}{r_{\oplus}}\right)\right).$$

B.3 Area of intersection of neighborhoods

Let $\mathcal{B}_{\oplus}(u,\varepsilon_u)$ and $\mathcal{B}_{\oplus}(v,\varepsilon_v)$ be two neighborhoods on the same surface of the sphere. If u and v are close enough, these neighborhoods will intersect in a lune-shaped area. The shape and area of this intersection depend on both the radii ε_u and ε_v and the distance between the centers u and v. By rotational symmetry, the actual positions of u and v on the sphere are not important. Only the great-circle distance $\delta = \operatorname{dist}(u,v)$ is of influence.

A paper by Lee and Kim gives a comprehensive list of formulas [4, Table 1] for the surface area of the intersection of two hyperspherical caps, which we typically call neighborhoods. Depending on the values of ε_u, ε_v, and δ, it provides many formulas based on case distinction. Case 8 from their paper is the appropriate choice if we slightly restrict the values of our parameters, and corresponds to:

- $\delta < \varepsilon_u + \varepsilon_v$ (the two neighborhoods intersect, and the intersecting area is nonzero);
- $\varepsilon_u + \varepsilon_v \leq 2\pi r_{\oplus} - \delta$ (the two neighborhoods do not cover the entire sphere together);
- $\varepsilon_u, \varepsilon_v \in [0, \pi r_{\oplus}/2)$ (both neighborhoods are smaller than a hemisphere);
- $\varepsilon_u = \varepsilon_v$ (the two neighborhoods have the same shape and area).

Typically we are dealing with relatively small intersecting identical neighborhoods, so these assumptions are naturally respected. For Case 8, the paper provides the following definitions.

$$\theta_1 = \theta_2 := \frac{\varepsilon_u}{r_{\oplus}} = \frac{\varepsilon_v}{r_{\oplus}}, \qquad \theta_v := \frac{\delta}{r_{\oplus}}, \tag{1}$$

$$\theta_{\min}(\theta_v) := \tan^{-1}\left(\frac{\cos(\theta_1)}{\cos(\theta_2)\sin(\theta_v)} - \frac{1}{\tan(\theta_v)}\right) = \tan^{-1}\left(\frac{1}{\sin(\theta_v)} - \frac{1}{\tan(\theta_v)}\right). \tag{2}$$

Given these definitons, the formula for the area of the intersection is

$$|\mathcal{B}_{\oplus}(u,\varepsilon_u)\cap\mathcal{B}_{\oplus}(v,\varepsilon_v)| = 2\pi r_{\oplus}^2 \cdot \int_{\theta_{\min}}^{\theta_2} \sin(\phi)\cdot I\left(1-\left(\frac{\tan(\theta_{\min})}{\tan(\phi)}\right)^2, \frac{1}{2}, \frac{1}{2}\right) d\phi,$$

where $I(z,a,b) = I_z(a,b)$ is the regularized incomplete beta function. The formula works only if the area is nonzero, and $\theta_1 = \theta_2 \in [\theta_v/2, \pi/2)$, which means that ε must be smaller than roughly 10 000 km. On this domain, the function is smooth, as can be seen in Figure 6.2.

We define the function $q : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as the area of the intersection of two neighborhoods over the total area of the globe. It is also the probability that a vertex taken uniformly from the surface of the globe is within both neighborhoods.

$$q(\varepsilon, \delta) := \frac{|\mathcal{B}_\oplus(u,\varepsilon) \cap \mathcal{B}_\oplus(v,\varepsilon)|}{4\pi r_\oplus^2} = \frac{1}{2} \cdot \int_{\theta_{\min}}^{\theta_2} \sin(\phi) \cdot I\left(1 - \left(\frac{\tan(\theta_{\min})}{\tan(\phi)}\right)^2, \frac{1}{2}, \frac{1}{2}\right) d\phi.$$

The definitions labeled by (1) and (2) above hold in the above formula.

B.4 Probability of closing a triangle

Let $\mathcal{B}_\oplus(u, \varepsilon)$ be a neighborhood on the surface $\mathbb{S}^2_\oplus$. Suppose there are two other vertices $v, w \in V \setminus \{u\}$ uniformly distributed within this neighborhood. What is the probability that v and w are within each others' ε-neighborhoods?

Define the function $\tilde{p} : \mathbb{R} \to \mathbb{R}$, which takes a distance of ε kilometers and returns this particular probability.

$$\tilde{p}(\varepsilon) := \mathbb{P}\left(\operatorname{dist}(v,w) < \varepsilon \,|\, \operatorname{dist}(u,v) < \varepsilon\,,\ \operatorname{dist}(u,w) < \varepsilon\right).$$

Given $\varepsilon \in (0, \pi/2)$, we can compute $\tilde{p}(\varepsilon)$ by conditioning on the position of, say, vertex v within $\mathcal{B}_\oplus(u, \varepsilon)$. We calculate what fraction of that neighborhood is within the ε-neighborhood of v (by using the formula for the area of intersection in Appendix B.3), which gives the probability that vertex w is within the ε-neighborhood of v:

$$\begin{aligned}
\tilde{p}(\varepsilon) &= \frac{1}{|\mathcal{B}_\oplus(u,\varepsilon)|} \cdot \int_{\mathcal{B}_\oplus(u,\varepsilon)} \mathbb{P}\left(w \in \mathcal{B}_\oplus(v,\varepsilon) \,|\, w \in \mathcal{B}_\oplus(u,\varepsilon)\right) dv \\
&= \frac{1}{|\mathcal{B}_\oplus(u,\varepsilon)|} \cdot \int_{\mathcal{B}_\oplus(u,\varepsilon)} \frac{|\mathcal{B}_\oplus(u,\varepsilon) \cap \mathcal{B}_\oplus(v,\varepsilon)|}{|\mathcal{B}_\oplus(u,\varepsilon)|} dv \\
&= \frac{1}{4\pi^2 r_\oplus^2 (1 - \cos(\varepsilon/r_\oplus))^2} \cdot \int_0^{2\pi} \int_0^{\varepsilon/r_\oplus} \sin(\theta) \cdot |\mathcal{B}_\oplus(u,\varepsilon) \cap \mathcal{B}_\oplus(v_{\theta\phi},\varepsilon)| \, d\theta \, d\phi \\
&= \frac{1}{(1 - \cos(\varepsilon/r_\oplus))^2} \cdot \int_0^{\varepsilon/r_\oplus} \sin(\theta) \int_{\theta_{\min}(\theta)}^{\varepsilon/r_\oplus} \sin(\psi) \cdot I\left(1 - \left(\frac{\tan(\theta_{\min}(\theta))}{\tan(\psi)}\right)^2, \frac{1}{2}, \frac{1}{2}\right) d\psi \, d\theta.
\end{aligned}$$

This is a well-behaved function ranging between roughly 0.58 and 0.69, steadily increasing as ε increases, as one can see in Figure 6.3. The value of $\tilde{p}$ for $\varepsilon = 0$ is undefined due to the presence of $(1 - \cos(\varepsilon/r_\oplus))^2$ in the denominator. However, the limit $\lim_{\varepsilon \to 0} \tilde{p}(\varepsilon)$ is defined and equal to approximately 0.59. The exact value is

$$\lim_{\varepsilon \to 0} \tilde{p}(\varepsilon) = 1 - \frac{3\sqrt{3}}{4\pi},$$

which is precisely what $\tilde{p}(\varepsilon)$ would be if the underlying space was $\mathbb{R}^2$ instead of $\mathbb{S}^2_\oplus$. It makes sense intuitively, since $\mathbb{S}^2_\oplus$ is locally ($\varepsilon \to 0$) indistinguishable from $\mathbb{R}^2$. The maximal value of $\tilde{p}(\varepsilon)$ for $\varepsilon \to \pi/2$ is approximately 0.68, and in exact form is

$$\lim_{\varepsilon \to \pi/2} \tilde{p}(\varepsilon) = \frac{\pi - 1}{\pi}.$$

The intersection of two hemispheres is a spherical lune, whose area scales linearly with the angle between the two hemispheres.

$$\begin{aligned}\tilde{p}(\varepsilon) &= \frac{1}{2\pi} \cdot \int_0^{2\pi} \int_0^{\pi/2} \sin(\psi) \cdot \frac{\pi - \psi}{\pi} \, d\psi \, d\phi \\ &= \frac{1}{\pi} \int_0^{\pi/2} \sin(\psi)(\pi - \psi) \, d\psi \\ &= \frac{1}{\pi} \int_\pi^{\pi/2} \psi \cdot \sin(-\psi) \, d\psi \\ &= \frac{\pi - 1}{\pi}.\end{aligned}$$

Values strictly within the interval $(0, \pi/2)$ are not so easily computed and take some computational time to approximate with Mathematica's numerical integration methods. Without loss of generality, we can set $r_\oplus = 1$ to improve computational time and accuracy.

B.5 Minimum distance of nodes

Let ρ be the expected great-circle distance from a node v on $\mathbb{S}^2_\oplus$ to its nearest neighbor, given that there are a total of n nodes. If the nearest neighbor to v has an angular distance of θ, then the probability density function of this θ is given by

$$f(\theta) = \frac{1}{2} \sin(\theta),$$

and the cumulative distribution function of θ is

$$F(\theta) = \int_0^\theta f(\psi) \, d\psi = \left(-\frac{1}{2} \cos(\psi) \right) \Big|_0^\theta = -\frac{1}{2} \cos(\theta) + \frac{1}{2}.$$

We can use the double-angle formula to simplify further as such:

$$\begin{aligned}F(\theta) &= -\frac{1}{2} \cos(\theta) + \frac{1}{2} \\ &= -\frac{1}{2} \left(2\cos^2\left(\frac{\theta}{2}\right) - 1 \right) + \frac{1}{2} \\ &= \sin^2\left(\frac{\theta}{2}\right).\end{aligned}$$

Given this c.d.f., we can compute the c.d.f. $\Lambda(\theta)$ of the minimum out of $n-1$ samples of $F(\theta)$ as follows.

$$\begin{aligned}\Lambda(\theta) &= \mathbb{P}\left(\exists\, v_i : \mathrm{d}(v, v_i) \le \theta\right) \\ &= 1 - \mathbb{P}\left(\forall\, v_i : \mathrm{d}(v, v_i) > \theta\right) \\ &= 1 - \left(\mathbb{P}(\mathrm{d}(v, v_i) > \theta)\right)^{n-1} \\ &= 1 - (1 - F(\theta))^{n-1} \\ &= 1 - \left(1 - \sin^2\left(\frac{\theta}{2}\right)\right)^{n-1} \\ &= 1 - \cos^{2n-2}\left(\frac{\theta}{2}\right).\end{aligned}$$

Then we differentiate to compute the corresponding p.d.f. $\lambda(\theta)$.

$$\begin{aligned}\lambda(\theta) &= \frac{d}{d\theta} \left(1 - \cos^{2n-2}\left(\frac{\theta}{2}\right) \right) \\ &= (n-1) \sin\left(\frac{\theta}{2}\right) \cos^{2n-3}\left(\frac{\theta}{2}\right).\end{aligned}$$

Now we take the expectation and multiply by $r_\oplus$ to convert angular distance to great-circle distance.

$$\begin{aligned}\rho := \mathbb{E}[\lambda(\theta)] &= r_\oplus \cdot \int_0^\pi \theta \cdot \lambda(\theta)\, d\theta \\ &= r_\oplus (n-1) \cdot \int_0^\pi \theta \sin\left(\frac{\theta}{2}\right) \cos^{2n-3}\left(\frac{\theta}{2}\right)\, d\theta.\end{aligned}$$

We use the power-reduction formula for cosines to turn the function $\cos^{2n-3}$ into a sum of simpler terms:

$$\rho = r_\oplus (n-1) \cdot \int_0^\pi \theta \sin\left(\frac{\theta}{2}\right) \frac{2}{2^{2n-3}} \sum_{k=0}^{n-2} \binom{2n-3}{k} \cos\left((2n-3-2k)\cdot\frac{\theta}{2}\right)\, d\theta.$$

We can exchange the integral and the summation and then apply the product-to-sum trigonometric identity:

$$\begin{aligned}\rho &= \frac{r_\oplus (n-1)}{2^{2n-4}} \cdot \sum_{k=0}^{n-2} \binom{2n-3}{k} \int_0^\pi \theta \sin\left(\frac{\theta}{2}\right) \cos\left((2n-3-2k)\cdot\frac{\theta}{2}\right)\, d\theta \\ &= \frac{r_\oplus (n-1)}{2^{2n-3}} \cdot \sum_{k=0}^{n-2} \binom{2n-3}{k} \int_0^\pi \theta\left(\sin((n-k-1)\cdot\theta) - \sin((n-k-2)\cdot\theta)\right)\, d\theta.\end{aligned}$$

The final term in the sum $k = n-2$ is considered separately here. Integrating this particular term by parts gives

$$\begin{aligned}&\frac{r_\oplus (n-1)}{2^{2n-3}} \cdot \binom{2n-3}{n-2} \int_0^\pi \theta \sin(\theta)\, d\theta \\ =&\frac{r_\oplus (n-1)}{2^{2n-3}} \cdot \binom{2n-3}{n-2} \left((-\theta\cos(\theta))|_0^\pi + \int_0^\pi \cos(\theta)\, d\theta \right) \\ =&\frac{r_\oplus (n-1)}{2^{2n-3}} \cdot \frac{1}{2}\binom{2n-2}{n-1} \cdot \pi \\ =&\frac{\pi r_\oplus (n-1)}{4^{n-1}} \binom{2n-2}{n-1}. \qquad (\star)\end{aligned}$$

We reintroduce this term at a later stage. Considering the other terms from $k = 0$ to $n-3$, using integration by parts the integrals cancel and simplify to the following.

$$\begin{aligned}&\frac{r_\oplus (n-1)}{2^{2n-3}} \cdot \sum_{k=0}^{n-3} \binom{2n-3}{k} \left(\int_0^\pi \theta \sin((n-k-1)\theta)\, d\theta - \int_0^\pi \theta \sin((n-k-2)\theta)\, d\theta \right) \\ =&\frac{\pi r_\oplus (n-1)}{2^{2n-3}} \cdot \sum_{k=0}^{n-3} \binom{2n-3}{k} (-1)^{n-k} \left[\frac{1}{n-k-1} + \frac{1}{n-k-2} \right].\end{aligned}$$

We can use a telescoping argument to rearrange the terms in the sum to be grouped by the fraction $\frac{1}{n-k-2}$ from $k = 0$ to $k = n-4$. Rearranging and adding limit terms gives

$$\begin{aligned}&\frac{\pi r_\oplus (n-1)}{2^{2n-3}} \cdot \left(\binom{2n-3}{0} \frac{(-1)^n}{n-1} - \binom{2n-3}{n-3} + \sum_{k=0}^{n-4} \frac{(-1)^{n-k}}{n-k-2} \left[\binom{2n-3}{k} - \binom{2n-3}{k+1} \right] \right) \\ =&\frac{\pi r_\oplus (n-1)}{2^{2n-3}} \cdot \left(\frac{(-1)^n}{n-1} - \frac{n-2}{2n}\binom{2n-2}{n-1} - \sum_{k=0}^{n-4} \frac{(-1)^{n-k}}{n-k-2} \cdot \frac{2n-2k-4}{2n-2} \binom{2n-2}{k+1} \right) \\ =&\frac{\pi r_\oplus (n-1)}{2^{2n-3}} \cdot \left(\frac{(-1)^n}{n-1} - \frac{n-2}{2n}\binom{2n-2}{n-1} - \frac{(-1)^n}{n-1} \sum_{k=0}^{n-4} (-1)^k \binom{2n-2}{k+1} \right).\end{aligned}$$

Notice how the term $\frac{(-1)^n}{n-1}$ corresponds to the term for $k = -1$ in the summation. Hence we can shift the index to $j = k + 1$ and get

$$\begin{aligned}
&\frac{\pi r_{\oplus}(n-1)}{2^{2n-3}} \cdot \left(-\frac{n-2}{2n}\binom{2n-2}{n-1} - \frac{(-1)^n}{n-1}\sum_{k=-1}^{n-4}(-1)^k\binom{2n-2}{k+1} \right) \\
=&\frac{\pi r_{\oplus}(n-1)}{2^{2n-3}} \cdot \left(-\frac{n-2}{2n}\binom{2n-2}{n-1} + \frac{(-1)^n}{n-1}\sum_{j=0}^{n-3}(-1)^j\binom{2n-2}{j} \right).
\end{aligned}$$

The partial alternating sum of binomial coefficients can be simplified to

$$\begin{aligned}
&\frac{\pi r_{\oplus}(n-1)}{2^{2n-3}} \cdot \left(-\frac{n-2}{2n}\binom{2n-2}{n-1} + \frac{(-1)^n}{n-1}\cdot(-1)^{n-3}\binom{2n-3}{n-3} \right) \\
=&\frac{\pi r_{\oplus}(n-1)}{2^{2n-3}} \cdot \left(-\frac{n-2}{2n}\binom{2n-2}{n-1} - \frac{1}{n-1}\cdot\frac{n-2}{2n}\binom{2n-2}{n-1} \right) \\
=&\frac{\pi r_{\oplus}(n-1)}{4^{n-1}} \cdot \left(-\frac{n-2}{n}\binom{2n-2}{n-1} - \frac{n-2}{n(n-1)}\binom{2n-2}{n-1} \right).
\end{aligned}$$

Reintroducing the term for $k = n - 2$ derived separately in ($\star$) eventually simplifies to

$$\rho = \frac{\pi r_{\oplus}}{4^{n-1}}\binom{2n-2}{n-1}.$$

As $n \to \infty$, we can use *Stirling's formula* to show that ρ converges to zero at a rate of order $n^{-1/2}$:

$$\rho = \frac{\pi r_{\oplus}}{4^{n-1}}\binom{2n-2}{n-1} \sim \frac{\sqrt{\pi} r_{\oplus}}{\sqrt{n}}.$$

Appendix C

List of countries and territories

The column labeled **CC** indicates what connected component of the worldgraph the entry is part of: 1 indicates the landmass comprising Europe, Asia, Africa, and part of Oceania; 2 indicates the Americas; entries with no value in **CC** are part of much smaller components.

country name	**ISO-2 code**	**dependency**	**CC**	**country name**	**ISO-2 code**	**dependency**	**CC**
Andorra	AD	independent	1	Laos	LA	independent	1
United Arab Emirates	AE	independent	1	Lebanon	LB	independent	1
Afghanistan	AF	independent	1	Saint Lucia	LC	independent	
Antigua and Barbuda	AG	independent		Liechtenstein	LI	independent	1
Anguilla	AI	United Kingdom		Sri Lanka	LK	independent	1
Albania	AL	independent	1	Liberia	LR	independent	1
Armenia	AM	independent	1	Lesotho	LS	independent	1
Angola	AO	independent	1	Lithuania	LT	independent	1
Antarctica	AQ	several claims		Luxembourg	LU	independent	1
Argentina	AR	independent	2	Latvia	LV	independent	1
American Samoa	AS	United States of America		Libya	LY	independent	1
Austria	AT	independent	1	Morocco	MA	independent	1
Australia	AU	independent	1	Monaco	MC	independent	1
Aruba	AW	Netherlands		Moldova	MD	independent	1
Aland Islands	AX	Finland		Montenegro	ME	independent	1
Azerbaijan	AZ	independent	1	Saint Martin (French part)	MF	France	
Bosnia and Herz.	BA	independent	1	Madagascar	MG	independent	
Barbados	BB	independent		Marshall Islands	MH	independent	
Bangladesh	BD	independent	1	North Macedonia	MK	independent	1
Belgium	BE	independent	1	Mali	ML	independent	1
Burkina Faso	BF	independent	1	Myanmar	MM	independent	1
Bulgaria	BG	independent	1	Mongolia	MN	independent	1
Bahrain	BH	independent		Macao	MO	China	1
Burundi	BI	independent	1	Northern Mariana Islands	MP	United States	
Benin	BJ	independent	1	Martinique	MQ	France	
Saint Barthelemy	BL	France		Mauritania	MR	independent	1
Bermuda	BM	United Kingdom		Montserrat	MS	United Kingdom	
Brunei	BN	independent	1	Malta	MT	independent	
Bolivia	BO	independent	2	Mauritius	MU	independent	
Bonaire, Sint Eustatius and Saba	BQ	Netherlands		Maldives	MV	independent	
Brazil	BR	independent	2	Malawi	MW	independent	1
Bahamas	BS	independent		Mexico	MX	independent	2
Bhutan	BT	independent	1	Malaysia	MY	independent	1
Bouvet Island	BV	Norway		Mozambique	MZ	independent	1
Botswana	BW	independent	1	Namibia	NA	independent	1
Belarus	BY	independent	1	New Caledonia	NC	France	
Belize	BZ	independent	2	Niger	NE	independent	1
Canada	CA	independent	2	Norfolk Island	NF	Australia	
Cocos (Keeling) Islands	CC	Australia		Nigeria	NG	independent	1
Dem. Rep. Congo	CD	independent	1	Nicaragua	NI	independent	2
Central African Rep.	CF	independent	1	Netherlands	NL	independent	1
Congo	CG	independent	1	Norway	NO	independent	1
Switzerland	CH	independent	1	Nepal	NP	independent	1
Côte d'Ivoire	CI	independent	1	Nauru	NR	independent	
Cook Islands	CK	New Zealand		Niue	NU	New Zealand	
Chile	CL	independent	2	New Zealand	NZ	independent	
Cameroon	CM	independent	1	Oman	OM	independent	1
China	CN	independent	1	Panama	PA	independent	2
Colombia	CO	independent	2	Peru	PE	independent	2
Costa Rica	CR	independent	2	French Polynesia	PF	France	
country name	**ISO-2 code**	**dependency**	**CC**	**country name**	**ISO-2 code**	**dependency**	**CC**

Table C.1: List of countries and territories.

country name	ISO-2 code	dependency	CC	country name	ISO-2 code	dependency	CC
Cuba	CU	independent		Papua New Guinea	PG	independent	1
Cabo Verde	CV	independent		Philippines	PH	independent	1
Curacao	CW	Netherlands		Pakistan	PK	independent	1
Christmas Island	CX	Australia		Poland	PL	independent	1
Cyprus	CY	independent	1	Saint Pierre and Miquelon	PM	France	
Czechia	CZ	independent	1	Pitcairn	PN	United Kingdom	
Germany	DE	independent	1	Puerto Rico	PR	United States of America	
Djibouti	DJ	independent	1	Palestine	PS	Israel	1
Denmark	DK	independent	1	Portugal	PT	independent	1
Dominica	DM	independent		Palau	PW	independent	
Dominican Rep.	DO	independent		Paraguay	PY	independent	2
Algeria	DZ	independent	1	Qatar	QA	independent	1
Ecuador	EC	independent	2	Reunion	RE	France	
Estonia	EE	independent	1	Romania	RO	independent	1
Egypt	EG	independent	1	Serbia	RS	independent	1
W. Sahara	EH	contested	1	Russia	RU	independent	1
Eritrea	ER	independent	1	Rwanda	RW	independent	1
Spain	ES	independent	1	Saudi Arabia	SA	independent	1
Ethiopia	ET	independent	1	Solomon Is.	SB	independent	
Finland	FI	independent	1	Seychelles	SC	independent	
Fiji	FJ	independent		Sudan	SD	independent	1
Falkland Is.	FK	United Kingdom		Sweden	SE	independent	1
Micronesia (Federated States of)	FM	independent		Singapore	SG	independent	
Faroe Islands	FO	Denmark		Saint Helena, Ascension and Tristan da Cunha	SH	United Kingdom	
France	FR	independent	1	Slovenia	SI	independent	1
Gabon	GA	independent	1	Svalbard and Jan Mayen	SJ	Norway	
United Kingdom	GB	independent	1	Slovakia	SK	independent	1
Grenada	GD	independent		Sierra Leone	SL	independent	1
Georgia	GE	independent	1	San Marino	SM	independent	1
French Guiana	GF	France	2	Senegal	SN	independent	1
Guernsey	GG	United Kingdom		Somalia	SO	independent	1
Ghana	GH	independent	1	Suriname	SR	independent	2
Gibraltar	GI	United Kingdom	1	S. Sudan	SS	independent	1
Greenland	GL	Denmark	2	Sao Tome and Principe	ST	independent	
Gambia	GM	independent	1	El Salvador	SV	independent	2
Guinea	GN	independent	1	Sint Maarten (Dutch part)	SX	Netherlands	
Guadeloupe	GP	France		Syria	SY	independent	1
Eq. Guinea	GQ	independent	1	eSwatini	SZ	independent	1
Greece	GR	independent	1	Turks and Caicos Islands	TC	United Kingdom	
South Georgia and the South Sandwich Islands	GS	United Kingdom		Chad	TD	independent	1
Guatemala	GT	independent	2	French Southern Territories	TF	France	
Guam	GU	United States of America		Togo	TG	independent	1
Guinea-Bissau	GW	independent	1	Thailand	TH	independent	1
Guyana	GY	independent	2	Tajikistan	TJ	independent	1
Hong Kong	HK	China	1	Tokelau	TK	New Zealand	
Heard Island and McDonald Islands	HM	Australia		Timor-Leste	TL	independent	1
Honduras	HN	independent	2	Turkmenistan	TM	independent	1
Croatia	HR	independent	1	Tunisia	TN	independent	1
Haiti	HT	independent		Tonga	TO	independent	
Hungary	HU	independent	1	Turkey	TR	independent	1
Indonesia	ID	independent	1	Trinidad and Tobago	TT	independent	
Ireland	IE	independent	1	Tuvalu	TV	independent	
Israel	IL	independent	1	Taiwan	TW	China	1
Isle of Man	IM	United Kingdom		Tanzania	TZ	independent	1
India	IN	independent	1	Ukraine	UA	independent	1
British Indian Ocean Territory	IO	United Kingdom		Uganda	UG	independent	1
Iraq	IQ	independent	1	United States Minor Outlying Islands	UM	United States of America	
Iran	IR	independent	1	United States of America	US	independent	2
Iceland	IS	independent		Uruguay	UY	independent	2
Italy	IT	independent	1	Uzbekistan	UZ	independent	1
Jersey	JE	United Kingdom		Holy See	VA	independent	1
Jamaica	JM	independent		Saint Vincent and the Grenadines	VC	independent	
Jordan	JO	independent	1	Venezuela	VE	independent	2
Japan	JP	independent	1	Virgin Islands (British)	VG	United Kingdom	
Kenya	KE	independent	1	Virgin Islands (U.S.)	VI	United States of America	
Kyrgyzstan	KG	independent	1	Vietnam	VN	independent	1
Cambodia	KH	independent	1	Vanuatu	VU	independent	
Kiribati	KI	independent		Wallis and Futuna	WF	France	
Comoros	KM	independent		Samoa	WS	independent	
Saint Kitts and Nevis	KN	independent		Kosovo	XK	partially recognized	1
North Korea	KP	independent	1	Yemen	YE	independent	1
South Korea	KR	independent	1	Mayotte	YT	France	
Kuwait	KW	independent	1	South Africa	ZA	independent	1
Cayman Islands	KY	United Kingdom		Zambia	ZM	independent	1
Kazakhstan	KZ	independent	1	Zimbabwe	ZW	independent	1
country name	**ISO-2 code**	**dependency**	**CC**	**country name**	**ISO-2 code**	**dependency**	**CC**