From Large Scale to Small Scale Geometric Topology of Networks: Wireline, Wireless, Quantum, and Power

Edmond A. Jonckheere
Dept of Electrical Engineering- Systems
University of Southern California
Los Angeles, CA 90089-2563
(213) 740-4457
jonckhee@usc.edu

April 11, 2020

Preface

The thesis of this book is that what is probably driving the field of applied mathematics at the dawn of the 21st century is the Internet and the generalized concept of networks (wireline, wireless, power, quantum, social, percolation, etc.) that sprang out of the *size* and *complexity* of such networks. Surely, the self-similar behavior of Internet traffic signals, a manifestation of *complex* behavior outside the realm of Gaussian processes, has given a boost to the theory of nonGaussian α-stable distributions defining a landscape where *Mathematics meet the Internet [96].* Phasor Measurement Unit (PMU) signals recorded on smart grid show similar nonclassical behavior [81, 82]. However, here, our claim is that the formidable *size* of network graphs is driving yet another field. Certainly, the growth of the Internet has led to a flareup of research in random graph generators (Erdös-Renyi, Barabasi-Albert, Small-World), a research that has been fairly successful at modeling *some* network phenomena. However, it is argued that the clarification of other phenomena, typically congestion, requires a *geometric topology* approach to graphs, which is the main topic of this book. The essential point of the *geometric topology* of graphs is that, if the graph is viewed as a transport medium (transport of packets, heat diffusion, spin & entanglement transport, etc.), its relevant properties are encapsulated in the concept of curvature. The motivation for the latter approach is that the negative/positive curvature dichotomy *does not map in an obvious way* to the classical graph generator classification [68]. The first historical manifestation of geometric topology in networks is that of Gromov-hyperbolic or Gromov negatively curved graphs, that is, graphs that behave metrically like saddle-shaped surfaces where geodesic triangles are “thin.” If routing (or more generally transport) is along geodesics (minimum cost paths) the thinness of the geodesic triangles will create congestion in the areas where sides of triangles come close. The latter has been probably the most spectacular application of Gromov geometry. But next to transport along minimum length paths, nature teaches us another, very efficient kind of transport—heat diffusion, where calories move from a hot source to a cold sink driven by very local temperature gradient. This idea was put to use in the Heat Diffusion protocol (although it was already in a less obvious way involved in the Back-Pressure), where the movement of packets is driven by local queue backlogs. For such transports, the large-scale Gromov curvature is irrelevant and must be substituted by the Ollivier-Ricci curvature—a very local one.

The present book is hence organized around this large-scale versus small-scale dichotomy. The title of this book, *"... from Large-Scale to Small-Scale"* reflects this historical paradigm shift.

Parallel to this geometrical line of ideas, there is an algebraic avenue of approach, referred to as *noncommutative geometry*, since it trades trades geometrical objects for noncommutative algebras. Such approach is sometimes referred to as *coarse geometry* since it involves an algebraic coarsening of geometrical objects. It can be claimed that large scale geometry is coarse as it trades local details for a better understanding of the global properties; here, however, *"coarse"* will be reserved to the algebraic approach. This algebraic approach is a bit outside the main geometrical stream of this book; however, in our humble opinion it cannot be ignored as there is growing presence of C^*-algebra dynamics in quantum mechanics. To put it simply, Heisenberg uncertainty principle is a coarsening of the phase space. In order to have the geometrical and the algebraic approaches living together, the main stream of the book is definitely geometrical, while the algebraic coarse approach is relegated to appendices. In general, when material becomes somewhat too algebraic, we have moved it to end-of-part appendices.

An outline of the book follows. Part I deals with network graphs, disregarding their metric structures, so that the topological and algebraic aspects are dominating. We begin, after introducing fundamental network and random graph concepts, with the traditional problem of embedding a graph in a surface, so that the reader gets used, from the beginning on and in the simplest possible context, to think of graphs and surfaces as the same kind of objects. Next, still in the same spirit, we take a more modern view at the problem by embedding the graph in a simplicial complex and, if possible, embedding the latter in a (possibly stratified) manifold.

A notable omission from this Part I is noncommutative geometry. Although it is an algebraic theory that provides another formalization of the embedding of graphs in surfaces, it takes its real significance in the coarse metric context via the coarse Baum-Connes conjecture, so that it is relegated to the next Part.

Part II deals with metric structures on manifolds and graphs. The first chapter of this part deals with the general problem of shortest path computation—in both manifolds and graphs. We take a unified view on the problem by showing that Euler-Lagrange equations, geodesic equations, and even shortest path algorithms on a graph can all be derived from Bellman's principle of optimality. Another aim of this chapter is to serve as some kind of a casual introduction to the two following chapters on Riemannian geometry, where the concept of geodesic is more precisely redefined as a curve with its tangent field parallel to itself. The Gauss theory of surfaces serves as an introduction to formal Riemannian geometry. In the latter chapter, we already introduce Ollivier-Ricci curvature again to get the reader acquainted to think graphs and manifolds as the same objects. After reviewing constant curvature spaces, we introduce comparison theory, also known as CAT (Cartan-Alexandroff-Toponogov) theory, as a first attempt at bringing such concepts as angle and curvature in the metric space context. The idea is to redraw a metric space triangle isometrically in a standard constant curvature Riemannian space and declare that the angle and curvature properties of the metric triangle are those it has in the Riemannian comparison space. The resulting curvature concepts are all at the *small scale* of the triangles viewed as basic building blocks. To complete the geometric approach, we have felt compelled to develop some elements of Finsler geometry as a way to formalize the fact that, on a network graph, the communication cost from A to B need not be the same as the communication cost from B to A. The next chapter deals with "coarse geometry." We begin by introducing such elementary coarse metric concepts as coarse map and coarse structures to eventually introduce the less intuitive concept of coarsening operator. Then the same chapter develops the algebraic approach to coarse geometry, referred to as noncommutative geometry, where a space, faithfully represented by a commutative algebra, is given a noncommutative algebra representation, with an inevitable damage to, or *coarsening* of, the space. An example shows that representing a graph by a noncommutative algebra results in a surface in which the graph is embeddable, thereby establishing the connection with Part I. More formally, we present the *coarse Baum-Connes conjecture*, known to be true in δ-hyperbolic spaces, as some way of asserting that the damage done to the space by going to a noncommutative algebra representation is no more than a coarsening. Precisely, the algebraic K-theory of the noncommutative algebra is the K-homology of a coarsening of the space.

Part III deals with large scale Gromov δ-hyperbolic spaces and its manifestation in many physical and logical network graphs, where the δ-hyperbolic property can be viewed as a formalization of the well known, visually intuitive "core concentric" property. The first chapter introduces the various fatness, thinness slimness of triangles to define large scale negative curvature. The next chapter investigates what such concepts become on constant curvature Riemannian manifolds. The heart of this part is the next chapter, where the classical graph generators are examined in the light of Gromov geometry. probably the most relevant result is that the well know growth/preferential attachment model of Internet build up leads—under some conditions—to δ-hyperbolic graphs, emphasizing that the positive/negative curvature does not trivially map to classical generators.

Part III is devoted to worm propagation, with a touch of control flavor encapsulated in the concept of "worm defense." Worm propagation is approached from the unifying point of view that a worm propagates on a graph, which depending on the mode of propagation of the worm could have varying topological, random, and curvature properties. Our contention is that the speed of propagation and the ability of the defense to slow it down depend heavily on the curvature.

A connection is then established with combinatorial group theory, which produces the so-called Cayley graphs, known to be hyperbolic with probability one, and on which the propagation can be expressed in terms of such group theoretic concepts as growth functions.

Contents

I Graph Topology without Curvature 13

1 Graphs and simplicial complexes 15
1.1 combinatorial graph theory 15
1.2 topological graph theory 19

2 Classical Random Graph Generators 21
2.1 Erdős & Rényi random graphs 21
2.1.1 simple asymptotic properties 23
2.1.2 connectivity 24
2.1.3 degree 24
2.1.4 diameter 25
2.2 Watts & Strogatz Small World graphs 26
2.2.1 α-model 27
2.2.2 β-model 28
2.3 Barabási & Albert Scale Free graphs 28
2.4 historical and bibliographical remarks 30

3 Embedding graphs on surfaces 31
3.1 compact surface classification 31
3.2 surface embedding 34
3.3 Algorithm for minimum embedding 36
3.4 extension of edge flow to surface 39
3.4.1 extension to surface elements 40
3.4.2 extension around vertices 42

4 Embedding Graphs in 3-manifolds 43

5 Extending graph to the simplicial complex of a possibly stratified manifold 45
5.1 Poincaré complexes 47
5.2 (simple) homotopy equivalence 48
5.3 surgery exact sequence 50
5.4 extension of flows 52
5.4.1 extension of flows to volume elements 53

5.4.2 extension around vertices 53

II Riemannian and nonRiemannian Geometry 55

6 Shortest Distance: Concepts and Computations 59
6.1 Riemannian spaces 59
6.1.1 Bellman's principle 59
6.1.2 Euler-Lagrange equations 60
6.1.3 Geodesics 61
6.2 NonRiemannian spaces 64
6.2.1 Metric, geodesic, and length spaces 64
6.2.2 Bellman-Ford and Dijkstra algorithms 67

7 Gauss Theory of Surfaces and the *Theorema Egregium* 69
7.1 Surfaces 69
7.2 Curvature 70
7.3 Gauss-Bonnet theorem 73

8 Riemannian Geometry 75
8.1 Vector fields and Lie bracket 75
8.2 Levi-Civita model 75
8.2.1 kinematic interpretation 75
8.2.2 formal definition 76
8.2.3 parallel displacement 77
8.2.4 arc length minimization 77
8.3 covariant derivative 78
8.4 curvature 80
8.5 sectional curvature 81
8.6 Gauss curvature of surfaces 85
8.7 Ricci and scalar curvature 89
8.8 Jacobi field 89
8.9 Ollivier-Ricci curvature 91
8.9.1 Foundation of differential geometry 91
8.9.2 Connection with wireline networks and diffusion processes 92
8.9.3 Towards Ollivier-Ricci curvature 92
8.9.4 From Riemannian manifolds to graphs 94
8.10 local curvature concepts at a point 95
8.10.1 2-dimensional case 95
8.10.2 higher-dimensional case 97
8.10.3 towards medium scale curvature 97
8.11 constant sectional curvature spaces 98

9 Standard constant sectional curvature spaces 101
9.1 Jacobi field 101
9.1.1 Euclidean space 101
9.1.2 Sphere model 102
9.1.3 Hyperboloid model 103
9.2 Models of Hyperbolic spaces with constant curvature -1 104
9.2.1 n-dimensional models 104
9.2.2 2-dimensional models 105

10 Isometric embedding in constant curvature space 107
10.1 distance structure and links 107
10.2 Gram matrices 108
10.2.1 positive curvature 108
10.2.2 negative curvature 108
10.3 Cayley-Menger matrix 109
10.4 congruence invariants 110
10.5 Main results 113
10.6 isometric embedding of basic graph structures 115
10.6.1 complete graph 115
10.6.2 star 118
10.6.3 core concentric structure 120
10.6.4 tree 121
10.7 dual theory 124

11 comparison geometry 127
11.1 CAT comparison theory 127
11.2 Alexandrov angle 128
11.3 example of application of Alexandroff angles: isometric embedding on surface 129
11.4 higher dimensional comparison 129
11.5 nonpositively curvature in the sense of Buseman 130
11.6 negative curvature in the sense of Alexandrov 130

12 Finsler geometry 131

A Coarse metric geometry 133
A.1 coarse map and coarse structure 133
A.1.1 quasi-isometry versus coarse equivalence 134
A.2 coarsening 135
A.3 coarse homology 136
A.4 Noncommutative geometry 137
A.4.1 preview: trading spaces for commutative algebras 137
A.4.2 trading spaces for noncommutative algebras 139
A.4.3 example 142
A.4.4 stabilization and scaling up 145
A.4.5 other algebras 146

A.4.6 further remarks 146

III Large scale geometry 149

13 large scale metric hyperbolic geometry 151
13.1 Various definitions of large scale negatively curved spaces 151
13.1.1 fatness 152
13.1.2 slimness 154
13.1.3 thinness 154
13.1.4 4 point inequality, 4 point condition, and Ptolemaic inequality 156
13.2 quasi-isometries 158
13.3 geodesics versus quasi-geodesics 161
13.3.1 Riemannian viewpoint 161
13.3.2 metric viewpoint 166
13.4 k-local versus quasi-geodesics 166
13.5 density of quasi-geodesics 167
13.6 δ-Computation 168
13.6.1 fatness computation 168
13.6.2 Gromov product 169

14 metric hyperbolic geometry of constant curvature spaces 171
14.1 Hyperbolic trigonometry 171
14.2 Slimness computation 173
14.3 Insize computation 175
14.4 Thinness computation 177
14.5 fatness and billiard dynamics 178
14.5.1 First order conditions 185
14.5.2 Hyperbolic orthocenter construction 187
14.5.3 Second order variation 191
14.5.4 Uniqueness 199
14.5.5 Fatness formula 202
14.6 bibliographical and historical notes 204

15 random versus hyperbolic graphs 207
15.1 setting up the various graph generators 207
15.2 fatness analysis 209
15.2.1 Erdős and Rényi Random graphs 210
15.2.2 Watts-Strogatz Small world graphs 210
15.2.3 Barabási-Albert scale free graphs 210
15.2.4 Growth with uniform attachment graphs 210
15.2.5 Comparison among all graph generators 210
15.3 Gromov product analysis 210
15.3.1 Erdős and Rényi Random graphs 210
15.3.2 Watts-Strogatz Small world graphs 220

15.3.3 Barabási-Albert scale free graphs 223
15.3.4 Growth with uniform attachment graphs 223
15.3.5 Comparison among all graph generators 228
15.4 comparison between δ_F and δ_G analyses 233

IV Flows on large-scale structures 235

16 Robustness against Curvature Uncertainty 237
16.1 Introduction 237
16.2 sensitivity to Christoffel's symbols 238
16.3 sensitivity of geodesics in negative curvature space 238

17 Worm propagation 239
17.1 Introduction 239
17.2 epidemiological model 240
17.3 Analytical Active Worm Propagation 241
17.4 comparison between two models 242

18 Groups, Cayley graphs, and Cayley Complexes 243
18.1 basic definitions 243
18.2 examples of groups 245
18.2.1 polynomial matrices groups 245
18.2.2 The modular group 245
18.2.3 combinatorial groups 246
18.2.4 Braid group 247
18.2.5 the fundamental group 247
18.3 The Burnside problem 249
18.4 Commutator calculus 250
18.5 Cayley graphs 251
18.6 growth functions 252

19 worm propagation on Cayley graphs 253
19.1 growth function propagation model 253
19.2 Elementary examples of worm propagation 254
19.2.1 free group 254
19.2.2 Abelian group 254
19.3 cancellation groups 256
19.3.1 one versus two cancellation 256
19.3.2 two versus three cancellation 260

20 Worm defense 271

B Coarse Flow Control by C^*-Algebra Dynamics **273**
B.1 Topological Dynamics 273
B.1.1 Commutative Case 274
B.1.2 Noncommutative Case 277
B.2 C^*-Algebra Control 278
B.2.1 Commutative case 278
B.2.2 Noncommutative case 278
B.3 Functorial Coarsening 278

V Appendices: wireline, wireless, quantum and power networks **279**

A Wireline Physical and Logical Graphs **281**
A.1 physical graphs 281
A.1.1 nodes 282
A.1.2 links 282
A.1.3 packet switched versus circuit switched network 283
A.1.4 protocol stack 284
A.2 logical graphs 285
A.3 overlaying 285
A.4 Topological Aspects of Overlaying 286
A.4.1 Graph Theoretic Aspects of Overlaying 287
A.4.2 Combinatorial Topological Aspects of Overlaying 288
A.4.3 Geometrical Aspects of Overlaying 288

B wireless **291**

C quantum **293**
C.1 Spintronic networks 293
C.2 Entanglement networks 293

D power **295**

Part I

Graph Topology without Curvature

Chapter 1

Graphs and simplicial complexes

1.1 combinatorial graph theory

From both the concept of physical graph, consisting of vertices corresponding to bridges and/or routers and of edges corresponding to communication links and the concept of logical graphs, such as the World Wide Web (WWW) consisting of vertices corresponding to documents (web pages) and edges corresponding to hyper-links (URL's), there emerges the need for a formal definition of a graph. We first look at graphs from the combinatorial viewpoint. We begin with the most elementary, but most restrictive, formalization and then extend it to include other graph concepts.

Definition 1 *A (simple or simplicial [1]) graph G consists of a non-empty set V_G of vertices, a set E_G of edges, and $\forall e \in E_G$ a nonempty end point set $V_G(e)$ of vertices upon which e is incident subject to the following restrictions:*

1. *The endpoint set $V_G\left(e\right)$ contains two distinct vertices.*

2. *If $e \neq \acute{e}$, then $V_G\left(e\right) \neq V_G\left(\acute{e}\right)$.*

From the definition of a graph, there is a 1-1 correspondence between an edge $e \in E_G$ and its endpoint set $V_G\left(e\right)$. Therefore, en edge e is sometimes denoted by $\{v_1, v_2\} = V_G\left(e\right)$.

Definition 2 *The order of G is $|V_G|$ and the size of G is $|E_G|$. A graph G is finite if $|V_G| < \infty$ and $|E_G| < \infty$. A graph is infinite if it is not finite.*

[1] This terminology [42, p. 1] is not uniform, as "simple" is very often omitted when the graph does not have multiple edges nor self-loops. The terminology of "simplicial" is sometimes used to refer to the fact that graphs by this restricted definition are 1-dimensional simplicial complexes.

With few exceptions, the graphs considered here are finite.

We now extend the concept of graphs by removing some of the restrictions on the end point set.

Definition 3

1. *A loop is an edge e such that $V_G(e)$ contains exactly one vertex. A loop graph is a graph in which loops are allowed.*
2. *An edge e is multiple if there exists $\acute{e} \neq e$, $\acute{e} \in E_G$ such that $V_G(e) = V_G(\acute{e})$. A multigraph is a graph in which multiple edges are allowed.*
3. *A pseudograph allows for loops and multiples edges* [2]

Now, we summarize the above in the following:

Definition 4 *A pseudograph is a triple (V_G, E_G, ∂), where V_g is the vertex set, E_G is the edge set, and ∂ is the boundary operator defined such that, $\forall e \in E_G$, $\partial e = V_G(e)$.*

We now allow for edges to be directed.

Definition 5 *A direction for an edge e is a function*

$$\sigma(e) : \{BEGIN, END\} \to V_G(e).$$

A directed edge e^σ is an edge e with a direction $\sigma(e)$. A directed graph, or digraph, is a graph in which every edge is a directed edge.

Definition 6 *An underlying graph is a corresponding graph where all edge directions are deleted and any multiple edges are coalesced.*

Definition 7 *The vertices u and v are adjacent if there exists an edge $e \in E_G$ such that $V_G(e) = \{u, v\}$, in which case the vertex u and the edge e are incident upon each other. The set $I_G = \{V_G(e) \mid e \in E_G\}$ is the incident structure.*

Definition 8 *The degree (or valence) $\deg(v)$ of the vertex v is the number of edges (with each loop being counted twice) that v is incident upon, that is, $\deg(v) = |\{e \in E_G \mid v \in V_G(e)\}|$. A vertex of degree 0 is an isolated vertex. A graph G where each vertex has the same degree k is k-regular graph; 3-regular graphs are cubic.*

For every graph, the sum of the degrees of the vertices is equal to twice the number of edges, that is,

$$\sum_{v \in V_G} \deg(v) = 2 \sum_{e \in E_G} |E_G|$$

[2] Again, this terminology [43, p. 10] is not uniform, as some authors [64, p. 3] use the terminology of “multigraph” when both multiple edges and self-loops are allowed. Here we prefer to make a clear distinction between pseudgraphs and multigraphs because in noncommutative geometry self-loops play a role completely different than multiple edges.

Indeed, each edge contributes exactly 2 to the sum of the degrees.

A graph can be represented by its incident structure. In addition, a graph can be represented by its adjacency matrix or its incidence matrix.

Definition 9 *The adjacency matrix A_G for a graph G is a function*

$$V_G \times \ V_G \to \mathbb{N}$$

where the entry $A_G(i,j)$ corresponding to row i and column j is the multiplicity of the adjacency between the vertices v_i and v_j.

Definition 10 *The incidence matrix T_G of a graph G is a function*

$$V_G \times E_G \to \mathbb{N}$$

where the entry $T_G(i,j)$, corresponding to row i and column j, is defined by

$$T_G(i,j) = \begin{cases} 0, & \textit{if } v_i \notin V(e_j) \\ 1, & \textit{if } v_i \in V(e_j) \ \textit{ and } \ |V(e_j)| = 1 \\ 2, & \textit{if } v_i \in V(e_j) \ \textit{ and } \ |V(e_j)| = 2 \end{cases}$$

Representation of a graph by an incident or adjacency matrix may involve a loss of space efficiency compared with incident structure because of the additional zeros. However, the advantage of this representation by matrix form is the information recovery.

Definition 11 *A graph H is a subgraph of a graph G if $V_H \subset V_G$ and $E_H \subset E_G$. In addition, if $V_H = V_G$, a graph H is a spanning subgraph of a graph G.*

An important example of the subgraph is a walk. Intuitively, a walk in a graph is the combinatorial analog of a continuous image in a closed line segment, which may arbitrarily often cross or retrace itself, backward or forward. The formal definition of a walk is as follows:

Definition 12 *A walk W of length n from vertex u to vertex v of a graph G is an alternating sequence of vertices and directed edges,*

$$W = v_0, e_1^{\sigma_1}, v_1, \ldots, v_{n-1}, e_n^{\sigma_n}, v_n$$

where $v_0 = u$, $v_n = v$, and each edge is incident on the two vertices immediately preceding and following it, i.e.,

$$\begin{aligned} \sigma_i(\textit{BEGIN}) &= v_{i-1}, \\ \sigma_i(\textit{END}) &= v_i \end{aligned}$$

In addition, if $v_0 = v_n$, the walk is closed; otherwise, it is open. The walk is a trail, if all its edges are distinct. The walk is a path, if all its vertices are distinct. A closed walk of length $n \geq 3$ with distinct vertices (except $v_0 = v_n$) is designated as a cycle.

A walk $W = v_0, e_1^{\sigma_1}, v_1, \ldots, v_{n-1}, e_n^{\sigma_n}, v_n$ can be denoted as $(v_0, v_1) \ldots (v_{n-1}, v_n)$.

Definition 13 *A graph G is connected, if for every pair of vertices $u, v \in V(G)$, then there exists a path in G joining u to v. A component of G is a maximal connected subgraph of G.*

Two graphs that can be illustrated by the same picture are isomorphic. To be more precise,

Definition 14 *Two graphs G_1 and G_2 are isomorphic $(G_1 \cong G_2)$, if there exists a bijective map $\theta : V_{G_1} \longrightarrow V_{G_2}$ preserving adjacency, that is, $\{u, v\} \in I_{G_1}$ if and only if $\{\theta(u), \theta(v)\} \in I_{G_2}$.*

Isomorphic graphs have the same degree. Therefore, the degree is an invariant.

Given that two graphs G_1 and G_2 with $V_{G_1} \cap V_{G_2} = \emptyset$, several operations that generate new graphs from G_1 and G_2 are defined as follows:

Definition 15

1. *The union $G = G_1 \cup G_2$ is such that $V_G = V_{G_1} \cup V_{G_2}$ and $E_G = E_{G_1} \cup E_{G_2}$.*

2. *The join (suspension) $G = G_1 + G_2$ is such that $V_G = V_{G_1} \cup V_{G_2}$ and $E_G = E_{G_1} \cup E_{G_2} \cup (V_{G_1} \times V_{G_2})$, where the endpoints of an edge $e = (u, v) \in (V_{G_1} \times V_{G_2})$ are the vertices $u \in V_{G_1}$ and $v \in V_{G_2}$.*

3. *The Cartesian product $G = G_1 \times G_2$ is such that $V_G = V_{G_1} \times V_{G_2}$ and $E_G = (E_{G_1} \times V_{G_2}) \cup (V_{G_1} \times E_{G_2})$, where the endpoints of an edge $(e, v) \in (E_{G_1} \times V_{G_2})$ are (u_1, v) and (u_2, v) with $V_{G_1}(e) = \{u_1, u_2\}$, and the endpoints of an edge $(u, f) \in (V_{G_1} \times E_{G_2})$ are (u, v_1) and (u, v_2) with $V_{G_2}(f) = \{v_1, v_2\}$.*

4. *The composition (lexicographic product) $G = G_1[G_2]$ is such that $V_G = V_{G_1} \times V_{G_2}$ and $E_G = (E_{G_1} \times V_{G_2} \times V_{G_2}) \cup (V_{G_1} \times E_{G_2})$, where the endpoints of an edge $(e, v_1, v_2) \in (E_{G_1} \times V_{G_2} \times V_{G_2})$ are (u_1, v_1) and (u_2, v_2) with $V_{G_1}(e) = \{u_1, u_2\}$, and the endpoints of an edge $(u, f) \in (V_{G_1} \times E_{G_2})$ are (u, v_1) and (u, v_2) with $V_{G_2}(f) = \{v_1, v_2\}$.*

5. *The edge-complement $\bar{G}$ of a graph G is such that $V_{\bar{G}} = V_G$ and*
$$E_{\bar{G}} = \{(v_1, v_2) \in V_G \times V_G \mid (v_1, v_2) \notin V_G(e) \text{ for } e \in E_G\}.$$

Definition 16

1. *A tree is a connected graph with no cycles.*

2. *A complete graph K_n is a graph of n vertices such that every pair of vertices is adjacent (i.e., all $\binom{n}{2}$ possible edges are present).*

3. *A totally disconnected (or empty) graph $\bar{K}_n$ is a graph of n vertices such that $E_{\bar{K}_n} = \emptyset$.*

4. *A complete bipartite graph $K_{m,n}$ is a graph of order $m+n$ such that $K_{m,n} = \bar{K}_m + \bar{K}_n$.*

5. *A complete npartite graph $K_{p_1,p_2,\ldots,P_n}$ is a graph of $\sum_{i=1}^{n} p_i$ vertices such that $K_{p_1,p_2,\ldots,P_n} = \bar{K}_{p_1} + \bar{K}_{p_2} + \ldots + \bar{K}_{p_n}$.*

6. *An n-cube Q_n is a graph that is defined recursively as*

$$\begin{aligned} Q_1 &= K_2 \\ Q_n &= K_2 \times Q_{n-1}, n \geq 2. \end{aligned}$$

1.2 topological graph theory

It should be observed that the preceding definition of graph is purely combinatorial in the sense that an "edge" $e \in E_G$ is an abstract element devoid of topological structure. Of course, it is tempting to think an edge as the homeomorph of the unit interval $[0,1]$, but this need not be done. In fact, leaving an edge as an abstract element makes the definition of graph by far more general. However, there are situations where there are some benefits at topologizing a graph. Indeed, when a graph is a topological space, we could certainly consider the issue of embedding it in a surface or a manifold and under some conditions think the graph as a surface or a manifold.

Definition 17 *Let V_G be a vertex set and let $\Sigma = \{\sigma(e) : \{0,1\} \to V_G\}$ be a set of functions. A topological graph is the topological space obtained from the disjoint union*

$$V_G \sqcup (\Sigma \times [0,1])$$

after the identification

$$\begin{aligned} \sigma(e)(0) &= (\sigma(e), 0) \\ \sigma(e)(1) &= (\sigma(e), 1) \end{aligned}$$

$\Sigma \times [0,1]$ *is the edge set E_G, topologized as many copies of the unit interval.*

Chapter 2

Classical Random Graph Generators

In this chapter, we follow the chronological development of Internet modeling by reviewing random graphs, as they were the first models that were proposed to cope with the phenomenal complexity of the Internet infrastructure. In order to model the phenomenal size of the Internet, it is customary to consider graphs under the asymptotic situation of an infinitely large number of vertices. There are several such asymptotic models, which may or may not exhibit some of the desirable properties that a random graph model of the Internet should enjoy. Among such relevant properties are the connectivity, the diameter, the distribution of the degree, etc. Another relevant issue is whether a property could exhibit the thresholding phenomenon, that is, whether it could appear or disappear abruptly under some change in the asymptotic parameters.

2.1 Erdős & Rényi random graphs

The Erdős-Rényi approach, which was the startup of random graph theory, is characterized by purely probabilistic methods, without reference to the growth phenomenon which later became predominant in internet research. Intuitively, a random graph in the sense of Erdős and Rényi is defined such that, given a number of vertices, the connections among them are specified in a random way. More specifically, for a given number n of vertices, that is, $V_G = \{1, 2, \ldots, n\}$, let $G(n)$ denote the set of all graphs G of order n. The two simplest models of random graphs of order n are as follows:

1. $G(n, m)$ model: Given positive integers n and m, where $0 \leq m \leq \binom{n}{2}$, then the $G(n, m)$ model is contained in $G(n)$ and consists of all labeled graphs of order n and of size m, with uniform probability distribution.

Therefore, $G\left(n,m\right)$ has

$$\binom{\binom{n}{2}}{m} = \binom{\frac{n(n-1)}{2}}{m}$$

elements and the probability of each graph $G \in G\left(n,m\right)$ is given by

$$p\left(G\right) = \binom{\frac{n(n-1)}{2}}{m}^{-1}.$$

2. $G\left(n,p\right)$ model: Given a positive integer n and a real number p where $0 \leq p \leq 1$, then the $G\left(n,p\right)$ model is contained in $G\left(n\right)$ and consists of all labeled graphs of order n and size $m \leq \binom{n}{2}$, and the probability measure on those graphs is specified by that fact that every pair of vertices is linked, independently of the other pairs, with probability p. Therefore, the probability measure in nonuniform as it depends on m. Specifically, the probability that $G \in G\left(n,p\right)$ has m edges is given by

$$p\left(G\right) = \binom{\binom{n}{2}}{m} p^{m} \left(1-p\right)^{\binom{n}{2}-m}.$$

An important part of random graph theory is the relationship between some property Q and the random graph parameter p or m that guarantees that the property Q arises asymptotically as $n \to \infty$. We now proceed more formally.

Definition 18 *Given that Ω_n is a model of random graphs of order n, then almost every random graph in Ω_n has a property Q in the sense of Erdős-Rényi if the probability of the graphs with this property approaches* 1 *as $n \to \infty$.*

In the following two definitions, if G and H are graphs and Q is a graph property, by $G \subset H$, we mean that G is a subgraph of H and by $G \in Q$ we mean that the graph G has the property Q.

Definition 19 *A property Q is monotone increasing if for every $G \in Q$ and $G \subset H$, then $H \in Q$.*

Definition 20 *The property Q is convex if for every $F \subset G \subset H$ and $F \in Q, H \in Q$, then $G \in Q$.*

An important relationship between the $G\left(n,m\right)$ and the $G\left(n,p\right)$ models is that, under some conditions, if one model enjoys some property, so does the other model, as stated by the following theorem:

Theorem 1 *Given that $m = m\left(n\right)$ is any sequence of positive integers such that*

$$\left(1-\varepsilon\right) p \binom{n}{2} < m(n) < \left(1+\varepsilon\right) p \binom{n}{2}$$

where $\varepsilon > 0$ is fixed, then if almost every graph has a property Q in the $G(n,m)$ model, then so it has in the $G(n,p)$ model. Given that the property Q is convex and

$$m(n) = \left\lfloor p\binom{n}{2} \right\rfloor,$$

then if almost every graph has a property Q in the $G(n,p)$ model, then so it has in the $G(n,m)$ model.

Probably the most spectacular phenomenon with random graphs is that many monotone-increasing properties appear *suddenly*, i.e., graphs of a link probability or size slightly less than a certain threshold are very unlikely to have property Q, whereas graphs of a link probability or size slightly greater than the threshold almost certainly have this property. A formal definition of threshold function is given as follows:

Definition 21 *Given that Ω_n is a model of random graphs of order n and Q is a property of graphs, then $t = t(c,n)$ is a threshold function for property Q if there exists a number c_0 such that, for $G \in \Omega_n$, with either $p(n) = t(c,n)$ in $G(n,p)$ or $m(n) \sim t(c,n)$ in $G(n,m)$, the following holds:*

1. *if $c > c_0$, almost all graphs G have the property Q.*

2. *if $c < c_0$, almost no graph G has the property Q.*

In a variant of the definition of threshold function, the issue is not the value of a parameter c relative to some threshold value c_0, but the asymptotic behavior of $p(n)$ or $m(n)$ relative to some yardstick function $p_c(n)$ or $m_c(n)$.

Definition 22 *Given that Q is a monotone increasing property, then the function $p_c(n)$ is a threshold function for Q in $G(n,p(n))$ if*

$$\lim_{n\to\infty} \frac{p(n)}{p_c(n)} = \begin{cases} 0, & \text{then } P_{n,p(n)}(Q) \longrightarrow 0 \\ 1, & \text{then } P_{n,p(n)}(Q) \longrightarrow 1 \end{cases}$$

Similarly, the function $m_c(n)$ is a threshold function for Q in $G(n,m(n))$ if

$$\lim_{n\to\infty} \frac{m(n)}{m_c(n)} = \begin{cases} 0, & \text{then } P_{n,m(n)}(Q) \longrightarrow 0 \\ \infty, & \text{then } P_{n,m(n)}(Q) \longrightarrow 1 \end{cases}$$

2.1.1 simple asymptotic properties

Here we consider the properties of the $G(n,p)$ model under the simple asymptotic condition $n \to \infty$ while p is held constant.

Theorem 2 *In the $G(n,p)$ model, with p constant and $n \to \infty$,*

1. *almost every graph has diameter 2;*

2. *almost every graph is k-connected;*

3. *almost every graph contains a given subgraph of order k as an induced graph;*

4. *almost every graph is nonplanar.*

The preceding theorem already gives us the clue that the Erdös-Rényi model might not be appropriate for the AS or the physical graph modeling, because the latter definitely do not have such a small diameter as 2.

2.1.2 connectivity

The following theorem defines the threshold function for the connectivity property.

Theorem 3 *For either $G \in G(n,p)$ with $p(n) = c\frac{\log n}{n}$ or $G \in G(n,m)$ with $m(n) \sim c\frac{1}{2}n\log n$, the following holds:*

1. *if $0 < c < 1$, almost every graph G is disconnected.*

2. *if $c > 1$, almost every graph G is connected.*

2.1.3 degree

Probably the most important parameter of a graph, as it relates to the Internet, is the degree of its nodes. From the previous facts, the degree distribution of a $G(n,p)$ random graph follows the binomial distribution

$$p(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

which can be approximated by a Poisson distribution

$$p(k) \approx \exp(-np)\frac{(pn)^k}{k!}$$

for large n. Putting it another way, if X_d is the number of nodes of degree d, the probability distribution of the random variable X_d follows the Poisson distribution

$$p(X_d = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

where

$$\lambda = n\binom{n-1}{d}p^d(1-p)^{n-1-d}$$

Indeed, since the nodes are independent, $X_d = k$ is the occurrence of k times the elementary event that the node has degree d, which occurs with probability $\binom{n-1}{d}p^d(1-p)^{n-1-d}$.

In fact, a Poisson distribution of the degree of the nodes occurs under more general circumstances than the strict Poisson conditions that require $p \downarrow 0$, $n \uparrow \infty$, while $pn = \lambda$ remains constant. This is made precise in the following theorem, which applies to both the $G(n,p)$ and the $G(n,m)$ models.

Theorem 4 *Given that d is a fixed nonnegative integer and X_d is the number of vertices of degree d, then if*

1. $0 < p(n) < 1$, $p(n) = \frac{\log n}{n} + d\frac{\log\log n}{n} + \frac{x}{n} + o\left(\frac{1}{n}\right)$ *in* $G(n,p)$ *or*
2. $0 < m(n) < \binom{n}{2}$, $m(n) \sim \frac{1}{2}n\log n + \frac{d}{2}n\log\log n + \frac{x}{2}n + o(n)$ *in* $G(n,m)$,

then X_d converges in distribution to a Poisson random variable with parameter $\lambda(d) = \frac{\exp(-x)}{d!}$, that is,

$$P(X_d = k) \to \exp(-\lambda(d))\frac{\lambda(d)^k}{k!}.$$

Observe that the $G(n,p)$ case is set up such that $p(n)n$ behaves as $\log n$ as $n \to \infty$, and as such the model differs from the Poisson model where $p(n)n$ converges as $n \to \infty$. With this clarification, the degree distribution of the nodes in the $G(n,p)$ case can be evaluated as

$$\frac{X_d}{n} \approx E\left(\frac{X_d}{n}\right) = \frac{\lambda(d)}{n} = \frac{e^{-p(n)n}(\log n)^d}{d!}$$

Likewise, the degree distribution of the nodes in the $G(n,m)$ model is easily seen to be

$$\frac{X_d}{n} \approx E\left(\frac{X_d}{n}\right) = \frac{\lambda(d)}{n} = \frac{e^{-\frac{2m(n)}{n}}(\log n)^d}{d!}.$$

The above has the consequence of ruling out the Erdős-Rényi random graphs as good models of the AS and the physical graphs, because the latter have their degree characterized by a power tail law.

2.1.4 diameter

The following theorem defines the threshold function for the "diameter 2" property of a graph.

Theorem 5 *Given that $p(n)^2 n - 2\log n \to \infty$ and $n^2(1-p(n)) \to \infty$ in $G(n,p)$ or $0 < m(n) < \binom{n}{2}$ and $2\frac{m(n)^2}{n^3} - \log n \to \infty$ in $G(n,m)$, then almost every graph has diameter 2.*

As a generalization of the preceding, we define the threshold function of the "diameter d" property of a graph.

Theorem 6 *Given a function $d(n) \geq 3$,$\frac{\log n}{d(n)} - 3\log\log n \to \infty$, then if*

1. $0 < p(n) < 1$ *satisfies*

$$\begin{array}{rcl} p(n)^{d(n)} n^{d(n)-1} - 2\log n & \to & \infty \\ p(n)^{d(n)-1} n^{d(n)-2} - 2\log n & \to & -\infty \end{array}$$

in the $G(n,p)$ model or

2. $0 < m(n) < \binom{n}{2}$ *satisfies*

$$\begin{aligned} 2^{d(n)-1} m(n)^{d(n)} n^{-d(n)-1} - \log n &\to \infty \\ 2^{d(n)-2} m(n)^{d(n)-1} n^{-d} - \log n &\to -\infty \end{aligned}$$

in the $G(n,m)$ *model,*

then almost every graph has diameter d*.*

2.2 Watts & Strogatz Small World graphs

The Small World phenomenon is the coexistence of two apparently contradictory features in a social network, in which the vertices are individuals and the links represent some social acquaintances between pairs of individuals. At the small scale of an individual, the cluster of his (her) social acquaintances has extremely small size compared with the world's population (the degree is small); nevertheless, a landmark experiment demonstrated that a letter sent from friends to friends needed an average of 6 hops to go from one individual to any other individual [22, pp. 25-27] (the diameter of the social network is small). Another phenomenon in social networks is that the acquaintances of an individual in general have social acquaintances among themselves. The tendency of the friends of an individual to make social connections is measured by the clustering coefficient:

$$C(v) = \frac{\#\{\triangle vxy : \{v,x\},\{v,y\} \in E_G(v)\}}{\binom{deg(v)}{2}}$$

It is the ratio of the number of triangles with a vertex at v over the maximum possible number of triangles that that can be constructed with pairs of links at v. It is in general agreed that a social network has a high clustering coefficient ≈ 1. In fact, in anticipation of the geometric theory of graphs, the latter is closely related to local positive curvature at v. The random graphs as modeled by Erdős and Rényi clearly do not generate the Small World phenomenon, for the main reason that they have too small a diameter (2) as $n \to \infty$. Their clustering coefficient is roughly p, so that in an attempt to approach a social network we would let $p \to 1$, but this would in turn create too large a degree $\approx (n-1)p$.

This clearly motivates a new graph paradigm, the Small World paradigm, that should satisfy the following properties:

1. The order of the graph n must be large.

2. The graph must be sparse, in the sense that the degree k of the nodes must be small; more precisely, $k << n$.

3. There must be some clustering around every node in the sense that the immediate neighbors of v must be well interlinked. Precisely, the clustering coefficient must approach 1.

4. The *characteristic path length*, that is, the average distance between two points, must be small, of the order of 6...

Watts and Strogatz [92] proposed the so-called α- and β-models, which generate graphs with the Small World property. The α-model is motivated by how people actually make new acquaintances in a real social network. In contrast, the β-model is motivated by removing the reference to social network.

2.2.1 α-model

The aim of the α-model is to construct a graph that captures the nature of connections in a social network. The α-model is designed so as to generate a graph that lies between two extreme models: the caveman and the Solaria world. The caveman world represents the model with the property that everybody you know knows everybody else you know and no one else. In contrast, the Solaria world represents the model with the property that the influence of current friendships over new friendships is so slight as to be indistinguishable from random chance. Given that n is the number of vertices, $\bar{k}$ is the average degree, and $\alpha \in [0, \infty)$ is a tunable parameter, then the α-model, which is a graph of order n and of size $\lfloor \bar{k}\frac{n}{2} \rfloor$, can be constructed as follows:

1. Randomly select vertex i from n possible vertices and for every $j \neq i$, compute a measure R_{ij} of vertex i's propensity to connect to vertex j. R_{ij} is equal to 0 if vertices i and j are already connected; otherwise, R_{ij} is computed by the following formula:

$$R_{ij} = \begin{cases} p, & m_{ij} = 0 \\ \left(\frac{m_{ij}}{\bar{k}}\right)^{\alpha}(1-p) + p, & 0 < m_{ij} < \bar{k} \\ 1, & m_{ij} \geq \bar{k} \end{cases}$$

   where m_{ij} is the number of vertices that are adjacent to both i and j and $p \ll \binom{n}{2}^{-1}$ is a baseline random probability of an edge (i, j) existing.

2. The probability that vertex i will connect to vertex j is given by

$$P_{ij} = \frac{R_{ij}}{\sum_{j \neq i} R_{ij}}.$$

   Then randomly select vertex $\tilde{j} \neq i$ according to the probability P_{ij} and connect vertex i to vertex $\tilde{j}$.

3. Repeat step 1 with the restriction that, if the vertex i is chosen, it cannot be chosen again until all other vertices have been chosen. This procedure is repeated until the graph consists of $\lfloor \bar{k}\frac{n}{2} \rfloor$ edges.

2.2.2 β-model

The β-model is a one-parameter model that lies between an ordered finite dimensional lattice and a random graph. In fact, the β-model has similar properties as the α-model without the concept of social network. When β varies from 0 to 1, the corresponding graph changes from a regular lattice graph to an approximately random graph in $G(n, m)$ model with $m = \left\lfloor \bar{k}\frac{n}{2} \right\rfloor$. Given that n is the number of vertices, $\bar{k}$ is the average degree, and $\beta \in [0, 1]$ is a tunable parameter, then the β-model, which is a graph of order n and of size $\left\lfloor \bar{k}\frac{n}{2} \right\rfloor$, can be constructed as follows:

1. Starting with a ring lattice graph of order n in which every vertex is connected to its first $\bar{k}$ nearest neighbors ($\frac{\bar{k}}{2}$ on either side). In order to have a connected graph at all times, the graph should have $n \gg \bar{k} \gg \ln n \gg 1$.

2. Each vertex i is chosen in turn, along with the edge that connected it to its nearest neighbor in a clockwise sense $(i, i+1)$. Each $(i, i+1)$ edge is randomly rewired according to the probability β. If the $(i, i+1)$ edge is selected to be rewired, then the $(i, i+1)$ edge is deleted and rewired such that i is connected to another vertex j, which is chosen uniformly at random from the entire graph (excluding self-connections and repeated connections).

3. Repeat step 2 until all vertices have been considered once; then the procedure continues for next nearest neighbor $(i, i+2)$ edges, and so on. The procedure is completed until all edges in the graph have been considered for rewiring exactly once.

The shape of the degree distribution of the Watts-Strogatz model is similar to that of the random graph, which has a peak at $\bar{k}$ and decays exponentially for large k.

2.3 Barabási & Albert Scale Free graphs

Recently, there has been a fair amount of study suggesting that complex real-world data networks (for example, the AS network) are following neither the Erdős-Rényi nor the Watts-Strogatz graph model, in the sense that the degree distribution $p(k)$ of real world data networks significantly deviates from the Poisson distribution. In fact, most complex real-world data networks have their degree distributions following the so-called *power-law tail*, that is, $p(k) \sim k^{-\gamma}$, where γ is independent of the scale of the network. Such a network is called *scale free*. Barabási and Albert suggested that the scale free networks can be generated by combining two mechanisms: growth and preferential attachment. Most of the complex real-world data networks can indeed be described as open systems, where new vertices can be added throughout the life of the networks, and the new vertices are more likely to be connected to vertices of higher degrees, so that newly added vertices have easy “gateways” to the rest of the network. The algorithm for the Barabási-Albert growth and preferential attachment model is given as follows:

1. Growth: Given a graph G_0 of small order n_0 and of size m_0, then G_{t+1} is recursively obtained from G_t by adding a new vertex v with $l\,(\leq n_0)$ edges to G_t such that the new vertex connects to l different vertices in G_t.

2. Preferential Attachment: The l different vertices in G_t are chosen such that the probability $p\left((i,v) \in E_{G_{t+1}}\right)$ that that vertex $i \in G_t$ is connected to the new vertex v depends on the degree k_i of the vertex i, i.e.,

$$p\left((i,v) \in E_{G_{t+1}}\right) = \frac{k_i}{\sum\limits_j k_j}.$$

After t time steps, the resulting graph G_t is a graph with $n = t + n_0$ vertices and $lt + m_0$ edges. In addition, the degree distribution $p\,(k)$ can be computed using continuum, master-equation or rate equation approaches. All of these approaches provide the same asymptotic results.

In the continuum approach, which was proposed by Barabási-Albert, the degree distribution $p\,(k)$ is computed by considering the rate of change of the degree of a given vertex i, with the assumption that the probability that the vertex i is added at time t_i is uniformly distributed over $[0,t]$. This yields the following formula:

$$p\,(k) = \frac{2l^2\left(t + \frac{m_0}{l}\right)}{n_0 + t}\frac{1}{k^3}.$$

Hence, in the limit as $t \to \infty$,

$$p\,(k) = 2l^2k^{-3}$$

which predicts that the degree distribution follows the power law tail with the exponent $\gamma = 3$.

In the master-equation approach, which was proposed by Dorogovtsev, Mendes, and Samukhin, the asymptotic limit of degree distribution $P\,(k)$ is computed by the following formula.

$$p\,(k) = \frac{2l\,(l+1)}{k\,(k+1)\,(k+2)}.$$

In the rate equation approach, which was proposed by Krapivsky, Redner, and Leyvraz, the asymptotic limit of the degree distribution $p\,(k)$ yields the same formula as the master-equation approach.

Hence all three approaches reveal that the degree distribution $p\,(k) \sim k^{-\gamma}, \gamma = 3$ and independent of l.

The diameter of the Barabási-Albert models can be shown to be, asymptotically as $n \longrightarrow \infty$, given by the following theorem.

Theorem 7 *Given that m is an integer greater than 2 and $\varepsilon > 0$ is fixed, then a.e. $G_m^n \in \mathcal{G}_m^n$ is connected and has diameter satisfying*

$$(1-\varepsilon)\frac{\log n}{\log\log n} \leq diam\,(G_m^n) \leq (1+\varepsilon)\frac{\log n}{\log\log n}.$$

If $m = 1$, G_1^m is connected and has diameter satisfying

$$\left(\alpha^{-1}-\varepsilon\right)\log n \leq diam\,(G_1^n) \leq \left(\alpha^{-1}+\varepsilon\right)\log n$$

where α is a solution of $\alpha \exp(1+\alpha) = 1$.

In order to gauge the importance of the preferential attachment in a growth process, Barabasi et al devised the growth/uniform attachment construction, in which each of the l links of the new vertex v is attached with uniform probability to the pre-exiting ones. In other words,

1. Growth: Same as before.
2. Uniform Attachment: Any of the l new edges brought in addition to those of G_t attaches to vertex $i \in G_t$ with uniform probability, that is,

$$p((i,v) \in E_{G_{t+1}}) = \frac{1}{\#V_{G_t}}$$

From the above construction, the importance of the preferential attachment is easily understood after observing that the above does not yield a power tail law.

2.4 historical and bibliographical remarks

The theory of random graphs was developed by Paul Erdős and Alfréd Rényi in the 1950's. Initially, it was a purely mathematical endeavor, with no specific applications in mind, which explains why this theory remained unnoticed for many years.

The Watts & Strogatz model was developed in the 1960's and its origin is traditionally said to be the modeling of social networks. However, another preoccupation of Watts and Strogatz was the synchronization of little bursts of light emitted by fireants. The puzzling feature was that fireants could synchronize their bursts over a distance that was beyond their visual acuity. The key feature of the model of this phenomenon was the clustering around every ant v of those ants within visual contact of the nominal ant v. The clusters were of a size much smaller than the whole population of ants. To secure synchronization across long distances, Watts and Strogatz conjectured that some ants had superior visual acuity, contributing to synchronization, and modeled as the rewiring of a neighboring link to a long distance link with probability β.

None of these models, however, was able to reproduced the heavy tailed phenomenon that became the predominant feature of the physical and autonomous systems graphs. For this reason, the Barabasi et al model was developed, quite recently, in the late 1990's.

Chapter 3

Embedding graphs on surfaces

In this Chapter, we look at the somewhat simplified problem of topologically embedding a communication graph on such a simple object as a surface, with the aim of developing a continuous geometry model of the information flow. The embedding is deliberately taken to be topological and disregards metric structure, as a first step towards the concept of *coarse structure* in which the metric is imprecisely defined.

First, the topological concept of surface and the classification of compact surfaces are discussed. Next, embedding algorithms are proposed. Finally, the flow on the edges of the graph is extended to the whole surface, revealing such phenomena as stagnation points.

3.1 compact surface classification

The topological concept of a surface is a mathematical abstraction of the idea of a surface made up of pieces of paper glued together. Hence a surface is a topological space with the same local properties as the Euclidean plane or the Euclidean unit disk.

Definition 23 *A surface is a connected* 2*-dimensional manifold, i.e., a Hausdorff space in which every point has an open neighborhood homeomorphic to the open* 2*-dimensional disc* $\mathbb{D} = \left\{x \in \mathbb{R}^2 : |x| < 1\right\}$*. A surface is orientable, if every closed path is orientation preserving (i.e., the orientation is preserved by traveling once around the closed path). A surface is nonorientable, if it is not orientable.*

An example of a compact orientable surface is the 2-sphere $S^2 = \left\{x \in \mathbb{R}^3 : |x| = 1\right\}$; another important compact orientable surface is the torus, which can be described as a surface that is homeomorphic to the surface of a doughnut. An example of a compact nonorientable surface is the real projective plane N_1, which can be described as a surface that is homeomorphic to the quotient space of the 2-sphere S^2 obtained by identifying every pair of diametrically opposite points.

There are several elementary operations that can be performed on a surface and that as such create new surfaces from old ones.

1. Adding a **handle**: Remove two open triangles of disjoint closure from the surface, give them opposite orientation, and then glue the two triangles along their edges.

2. Adding a **crosscap**: Remove an open square from the surface, give opposite edges opposite orientation, and then glue opposite edges.

Another technique to create new surfaces out of old ones is to combine two surfaces through some operations. The **connected sum** of two disjoint surfaces S_a and S_b, denoted by $S_a \# S_b$, is formed by cutting a small circular hole in each surface and by gluing the two surfaces together along the boundaries of the holes. To be precise, subsets $D_a \subset S_a$ and $D_b \subset S_b$ are chosen such that D_a and D_b are homeomorphic to the unit disc D. The complements of the interior of D_a and D_b are denoted by S'_a and S'_b, respectively. Given that $h : \partial D_a \to \partial D_b$ is a homeomorphism, then the connected sum $S_a \# S_b$ is the quotient space of $S'_a \cup S'_b$ obtained by identifying the point x and $h(x)$ for all points x in the boundary of D_a. The connected sum of n tori can be viewed as a sphere with n handles and can be denoted by S_n. Similarly, the connected sum of n projective planes can be viewed as a sphere with n crosscaps and can be denoted by N_n.

The connected sum of a torus and a projective plane is homeomorphic to the connected sum of three projective planes.

Definition 24 *A triangulation of a compact surface S consists of a finite family of closed subsets $\{T_1, T_2, \ldots, T_n\}$ that cover S and a family of homeomorphisms $\varphi_i : \tilde{T}_i \to T_i, i = 1, 2, \ldots, n$ where each $\tilde{T}_i$ is a triangle in the plane $\mathbb{R}^2$.The image of the vertices under φ_i is designated as vertex set and the image of the edges under φ_i is designated as edge set. In addition, it is required that any two triangles T_i and T_j, either be disjoint, or have a single vertex in common, or have one entire edge in common.*

From the strong form of the Jordan curve theorem, there exists such a triangulation for any compact surface S. Therefore, the Euler characteristic $\chi(S)$ for a compact surface S with triangulation $\{T_1, T_2, \ldots, T_n\}$ can be defined as

$$\chi(S) = v - e + f \tag{3.1}$$

where

$$\begin{aligned} v &= \text{total number of vertices of } S, \\ e &= \text{total number of edges of } S, \\ f &= \text{total number of triangles } (= n). \end{aligned}$$

surface	Euler characteristic
Sphere	2
connected sum of n tori	$2-2n$
connected sum of n projective planes	$2-n$

Table 3.1: Euler characteristic of various compact surfaces

In fact, the Euler characteristic depends only on the surface S, not on the triangulation chosen. In addition, the subdivision of S can be allowed into arbitrary polygons such that the interior of each polygon is homeomorphic to an open disc and the closure of each edge is homeomorphic to a closed interval or a circle. Finally, the Euler characteristic is a topological invariant and does not depend on the subdivision of S into polygons. Therefore, the Euler characteristic of a surface S can be redefined as follows:

Definition 25 *For a compact surface S, the Euler characteristic $\chi(S)$ is defined as*

$$\chi(S) = v - e + f \tag{3.2}$$

where

$$\begin{aligned} v &= \text{total number of vertices} \\ e &= \text{total number of edges} \\ f &= \text{total number of regions or faces} \end{aligned}$$

and v, e, f are obtained from arbitrary subdivision of a surface S into polygons.

The Euler characteristic of the connected sum $S_a \# S_b$ can be computed by the following formula:

$$\chi(S_a \# S_b) = \chi(S_a) + \chi(S_b) - 2 \tag{3.3}$$

The Euler characteristics of particular surfaces are shown in Table 3.1.

Definition 26 *The genus $g(S)$ of a surface S is the number of handles (for an orientable surface) or crosscaps (for a nonorientable surface) that one must add to the sphere to obtain a surface that is homeomorphic to the surface S. A surface that is a connected sum of n tori or n projective planes is said to be of genus n. A sphere is of genus* 0.

Proposition 1

$$g(S) = \begin{cases} \frac{1}{2}(2-\chi(S)) & \text{in the orientable surface case} \\ 2-\chi(S) & \text{in the nonorientable surface case.} \end{cases} \tag{3.4}$$

Theorem 8 *Given that S_a and S_b are compact surfaces, then S_a and S_b are homeomorphic if and only if their Euler characteristics are equal and both are orientable or both are nonorientable.*

Proof. See [62, Chap. 1, Thm. 8.2]. ■

The classification theorem for compact surfaces can be stated as follows:

Theorem 9 (classification of compact surface) *Every compact surface is homeomorphic to a sphere, to a connected sum of tori, or to a connected sum of projective planes.*

Proof. See [62, Chap. 1, Thm. 5.1]. ■

From this theorem, every compact surface can be represented by a polygon subject to gluing selected pairs of edges. In all three cases, the (oriented) edges of the polygonal line are listed multiplicatively as we go clockwise along the line; the notation a^{-1} denotes the edge obtained from a by reversing orientation. The occurrence of twice the same edge along the polygonal line means that the two edges are glued, with consistent orientation.

1. The sphere: aa^{-1}
2. The connected sum of g tori: $a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}\dots a_gb_ga_g^{-1}b_g^{-1}$
3. The connected sum of k projective planes: $c_1c_1c_2c_2\dots c_kc_k$.

Polygon representations of sphere, torus, and projective plane are shown in Figure ???

From the canonical representation of a surface as a polygon with pairs of edges identified, it follows that the fundamental group of a surface is generated by the edges of the polygonal line subject to the relation that translates the fact that the polygonal path subject to the gluing is shrinkable to a point. Using the formal notation of presentation of groups by generators and relations of Chapter 18, it is easy to see that the fundamental groups of the surfaces are as follows:

1. The sphere
$$\pi_1(S^2) = \langle a | aa^{-1} = 1\rangle = 1$$
2. The compact orientable surface of genus g:
$$\pi_1(S_g) = \langle a_1, b_1, ...a_g, b_g | a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}\dots a_gb_ga_g^{-1}b_g^{-1} = 1\rangle$$
3. The compact nonorientable surface with k crosscaps:
$$\pi_1(N_k) = \langle c_1, ..., c_k | c_1^2...c_k^2 = 1\rangle$$

3.2 surface embedding

Since the internet has become a world wide service, it is useful to think the internet as a graph drawn on the surface S^2 of the earth. If there are satellites and their links, they can be projected orthogonally back on the surface of the earth, so that the graph is still drawn on S^2.

Recall that a graph is *planar* if it can be drawn on the sphere S^2 without links crossing. Using a stereographic projection argument, it is easy to show that a graph is planar iff it can be drawn on the plane $\mathbb{R}^2$ without links crossing. A celebrated theorem asserts that a graph is planar iff it does not contain any of the Kuratowsky graphs K_5, $K_{3,3}$ as subgraphs.

Consider a graph G, which we write on the sphere S^2, possibly with some edges crossing. For each edge crossing, "pull a handle" and draw one of the edges on the handle rather than on the sphere. After pulling a handle for every pair of crossing edges, the graph is written–without edge crossings– on a sphere with g handles, that is, the compact surface S_g of genus g. This process is of course nonunique, but among all such processes there is one that leads to a minimum number of handles, called the *genus of the graph* [95, Def. 6-9].

The computational implementation of this process relies on the so-called *Hefter-Edmonds rotation system* [64, Sec. 3.2]. Intuitively, a surface encompassing a graph induces up to cyclic permutation an ordering of the edges flowing out of each vertex. Conversely, assume that we have a *rotation system* $\{\pi_v : v \in G_0\}$, that is, $\forall v \in G_0$, we are given a permutation π_v of the edges flowing out of v. Among the great many rotation systems that are associated with a graph, some of them induces *cellular embeddings* of G, that is, $S \setminus G$ breaks as the disjoint union of *(acyclic) cells*, i.e., homeomorphs of the open disk. The cells of the embedding are given by the π *walks* [64, Sec. 3.2]: Start at a vertex v^0 and proceed along an outflowing edge e^0 until we reach another vertex v^1 from which we proceed along the edge $\pi_{v^1}(e^0)$, etc., until we come back to v^0; this closed path bounds a cell and the cells are "glued" along the edges to yield S_g. A fundamental result says that those rotation systems that yield the minimum genus embedding are cellular embeddings [64, Prop. 3.4.1]. However, it is also possible to have a cellular embedding on a surface of a genus higher than the genus of the graph. The maximum of all genera of surfaces S_g on which the graph has a cellular embedding is called the *maximum genus*, $g_M(G)$ (see [64, Sec. 4.5]).

We now proceed more formally.

Definition 27 *A graph G is embedded in a surface S, if it is drawn in S so that edges intersect only at their common vertices. The components of $S - G$ are regions (or faces) of the embedding. A region is a 2-cell, if it is homeomorphic to the open unit disk. In addition, if every region of an embedding is a 2-cell, the embedding is a 2-cell embedding.*

In addition, a region is a 2-cell, if any simple closed curve in this region can be continuously deformed or contracted into a single point.

Definition 28 *The genus $g(G)$ of a graph G is the minimum genus among the genera of all surfaces S in which G can be embedded. An embedding of a graph G in the compact surface S_k is a minimal embedding, if $g(G) = k$.*

Theorem 10 *Given that G is a connectedgraph embedded in a surface of genus g which is equal to the genus of the graph, then every region of G is a 2-cell and the embedding is a 2-cell embedding.*

Proof. See [95, Thm. 6-11]. ■

Definition 29 *The maximum genus $g_M(G)$ of a graph G is the maximum genus among the genera of all orientable surfaces in which G can be 2-cell embedded.*

Definition 30 *A graph is planar, if it can be embedded in the plane (or equivalently, in the surface S_0 by the stereographic projection).*

If G is planar, then $g(G) = 0$. If $g(G) = k, k > 0$, then G has an embedding in S_k, but not in S_h, for $h < k$.

Theorem 11 (Kuratowski) *A graph G is planar if and only if it contains no subgraph that is homeomorphic to either K_5 or $K_{3,3}$.*

Proof. See [42, Sec. 1.4.5]. ■

The complete graph K_5 and the complete bipartite graph $K_{3,3}$ are called Kuratowski's graphs.

Definition 31 *Given that G is a connected graph with a 2-cell embedding on an orientable surface S, then the Euler characteristic of a 2-cell embedding $G \to S$, $\chi(G \to S)$ is defined by*

$$\chi(G \to S) = |V| - |E| + |F| \tag{3.5}$$

where V is the set of vertices, E is the set of edges, and F is the set of regions.

Therefore, the Euler characteristic of a 2-cell embedding is equal to the Euler characteristic of surface S, i.e.,

$$\begin{aligned} \chi(G \to S) &= \chi(S) \\ &= \begin{cases} 2 - 2g, & \text{if } S = S_g \\ 2 - k, & \text{if } S = N_k \end{cases} \end{aligned} \tag{3.6}$$

Theorem 12 *For each orientable surface S_g $(g = 0, 1, 2, \ldots)$, there exists a connected graph G and a 2-cell embedding $G \to S_g$ whose Euler characteristic satisfies the equation $\chi(G) = 2 - 2g$. For each nonorientable surface N_k $(k = 0, 1, 2, \ldots)$, there is a graph G and a 2-cell embedding of G into the surface N_k such that $\chi(G) = 2 - k$.*

Proof. See [42, Thm. 3.3.1 and Thm. 3.3.2]. ■

3.3 Algorithm for minimum embedding

An important problem in graph theory is how to determine the genus of a graph. This problem can be translated into a combinatorial problem of determining a rotation system with the maximum number of regions. Intuitively, each rotation system can be considered as an algebraic description of a 2-cell embedding.

The vertex set of a connected graph G can be denoted by $V_G = \{1, 2, \ldots, n\}$. Given that $V(i) = \{k \in V_G \mid \{i, k\} \in E_G\}$ for each $i \in V_G$, then define $p_i : V(i) \to V(i)$ to be a cyclic permutation on $V(i)$, of length $n_i = |V(i)|$; p_i is designated as the rotation at i and the set $\{p_1, p_2, \ldots, p_n\}$ of rotations is designated as a rotation system, or rotation scheme.

Theorem 13 *Every rotation system $\{p_1, p_2, \ldots, p_n\}$ for a graph G determines a 2-cell embedding of G into an oriented surface S, such that the orientation on S induces a cyclic ordering of the edges $\{i, k\}$ at i in which the immediate successor to $\{i, k\}$ is $\{i, p_i(k)\}, i = 1, \ldots, n$. In fact, given $\{p_1, p_2, \ldots, p_n\}$, there is an algorithm which produces the embedding. Conversely, given a 2-cell embedding in a surface S with a given orientation, there is a corresponding $\{p_1, p_2, \ldots, p_n\}$ determining that embedding.*

Proof. See [95, Thm. 6-50]. ■

Given $E = \{(a, b) \mid \{a, b\} \in E_G\}$, and P a permutation on the set E of directed edges (where each edge of G is associated with two possible directions) such that $P(a, b) = (b, p_b(a))$, then an orbit under P is a closed walk $W = (i_0, i_1) \ldots (i_{m-1}, i_m)$ with the following properties:

1. For every $j \neq k, (i_{j-1}, i_j) \neq (i_{k-1}, i_k)$.
2. $(i_k, i_{k+1}) = P(i_{k-1}, i_k)$, $k = 1, 2, \ldots, m-1$
3. $(i_0, i_1) = P(i_{m-1}, i_m)$.

Notice that $(i_j, i_k) \neq (i_k, i_j)$ since they consists of different directions. Then each orbit under P determines a 2-cell region of the corresponding embedding. Hence the number of orbits is the number of faces of the 2-cell embedding. Finally, regions can be pasted together with (a, b) matched with (b, a) to obtain an orientable surface.

Since an embedding of G into $S_{g(G)}$ is a 2-cell embedding, there is a rotation system corresponding to this 2-cell embedding. It now follows that the genus of any connected graph can be computed by selecting among the $\prod_{i=1}^{n} (n_i - 1)!$ possible rotation systems the one that gives the maximum number of orbits, and hence determines the genus of a graph. The maximum genus of a graph can be computed from a rotation system with minimum number of orbits.

To illustrate the concept of rotation system, three rotation systems of the complete graph K_5 are considered.

Let $V_{K_5} = \{1, 2, 3, 4, 5\}$ with $V(i) = V_{K_5} - \{i\}$. Define a rotation system by

$$\begin{aligned} p_1 &: (2, 3, 4, 5) \\ p_2 &: (3, 4, 5, 1) \\ p_3 &: (4, 5, 1, 2) \\ p_4 &: (5, 1, 2, 3) \\ p_5 &: (1, 2, 3, 4) \end{aligned}$$

The orbits under this rotation system are

1. $(1,2)\,(2,3)\,(3,4)\,(4,5)\,(5,1)$
2. $(1,3)\,(3,2)\,(2,4)\,(4,3)\,(3,5)\,(5,4)\,(4,1)\,(1,5)\,(5,2)\,(2,1)$
3. $(1,4)\,(4,2)\,(2,5)\,(5,3)\,(3,1)$

From this rotation system, we have $\chi(K_5) = 5 - 10 + 3 = -2$. This implies that $2 - 2g = -2,\ g = 2$. This rotation system corresponds to an embedding of the graph K_5 into an orientable surface with genus 2. Now consider another rotation system defined by

$$\begin{aligned} p_1 &: (5,4,3,2) \\ p_2 &: (1,4,3,5) \\ p_3 &: (2,1,4,5) \\ p_4 &: (5,3,2,1) \\ p_5 &: (3,4,1,2) \end{aligned}$$

The orbits under this rotation system are

1. $(1,2)\,(2,4)\,(4,1)\,(1,3)\,(3,4)\,(4,2)\,(2,3)\,(3,1)$
2. $(1,4)\,(4,5)\,(5,1)$
3. $(1,5)\,(5,2)\,(2,1)$
4. $(2,5)\,(5,3)\,(3,2)$
5. $(3,5)\,(5,4)\,(4,3)$

Hence this rotation system yields $\chi(K_5) = 5 - 10 + 5 = 0$. This implies that $2 - 2g = 0,\ g = 1$. This rotation system corresponds to an embedding of the graph K_5 into an orientable surface with the minimum genus of the graph K_5.

Consider the other rotation defined by

$$\begin{aligned} p_1 &: (2,5,4,3) \\ p_2 &: (1,5,4,3) \\ p_3 &: (1,5,4,2) \\ p_4 &: (1,5,3,2) \\ p_5 &: (1,3,4,2) \end{aligned}$$

The orbits under this rotation system are

1. $(1,2)\,(2,5)\,(5,1)\,(1,4)\,(4,5)\,(5,2)\,(2,4)\,(4,1)\,(1,3)\,(3,5)$
   $(5,4)\,(4,3)\,(3,2)\,(2,1)\,(1,5)\,(5,3)\,(3,4)\,(4,2)\,(2,3)\,(3,1)\,.$

Graph	genus g	maximum genus g_M
Q_n	$1+2^{n-3}(n-4)\,,\ n\geq 2$	$(n-2)\,2^{n-2}, n\geq 2$
K_n	$\left\lceil \frac{(n-3)(n-4)}{12} \right\rceil,\ n\geq 3$	$\left\lfloor \frac{(n-1)(n-2)}{4} \right\rfloor$
$K_{m,n}$	$\left\lceil \frac{(m-2)(n-2)}{4} \right\rceil,\ m,n\geq 2$	$\left\lfloor \frac{(m-1)(n-1)}{2} \right\rfloor$

Table 3.2: Genus and maximum genus of some families graphs

Hence this rotation system yields $\chi(K_5) = 5-10+1 = -4$. This implies that $2-2g = -4,\ g = 3$. This is the maximum genus of the graph K_5. Hence this rotation system corresponds to an embedding of the graph K_5 into an orientable surface with maximum genus.

In fact, among all 7776 possible permutations $\left(((4-1)!)^5 = 7776\right)$, there are 462 rotation systems of genus one, 4974 rotation systems of genus two, and 2340 rotation systems of genus three.

Table 3.2 shows the formula for some well known graphs [95, Chap. 6, Sec. 6-4 and 6-5]. Note that $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x; $\lceil x \rceil$ denotes the least integer greater than or equal to x.

The problem of computing the genus of a graph is known to be NP-complete Therefore, there is no polynomial bounded algorithm for deciding the genus of graph.

What about embeddability of infinite graphs???...

3.4 extension of edge flow to surface

Recall that the objective of interpolating the internet grid with such a continuous geometric structure as a surface is to have a convenient description of the traffic as a flow on a manifold. Clearly, the flow is *objectively* defined on the edges. However, after interpolating the graph with a surface, it is necessary to define the flow on the surface elements as well. From this perspective, the most fundamental problem appears to be how to solve such a PDE as the 2-dimensional Navier-Stokes equation given the boundary conditions on the links. The problem we will be facing taking that path, besides the issue of justifying the fluid flow model, is that at the vertices the flow is not continuous, not even defined. We therefore opt for a more constructive approach, which is not subject to the apriori choice of a Navier-Stokes model and which in a sense allows for all reasonable extensions of the boundary flow to the surface element.

Assume that the embedding $G \to S$ is cellular. Take a surface element s_2, that is, a connected component of $S \setminus G$. Recall that this surface element is simply connected, hence a cell homeomorphic to $\mathbb{D}$. The boundary of this surface element is a polygonal collection of links and vertices, homeomorphic to S^1. On that boundary, the flow is objectively defined and the problem is to extend the flow in some continuous fashion to the whole surface element. One first problem is that the link traffic is time-varying, but one has the option of freezing it at a certain time or averaging it over time. As already said, a more bothersome problem is that the flow along the edges of a surface element—take a triangle $a^k a^0 a^{k+1}$ to make the notation easy—is *not* continuous at the vertices. This can be fixed by a drawing a small smooth curve $a^{0_k} a^{0_{k+1}}$ tangent to $a^k a^0$ and $a^0 a^{k+1}$ at a^{0_k}, $a^{0_{k+1}}$, respectively, with similar curves near the vertices a^{k+1} and a^k. Along $a^{k_0} a^{0_k}$, the flow v is tangent to the link and along the curve $a^{0_k} a^{0_{k+1}}$ the vector field is defined to be smoothly connecting the field on $a^{k_0} a^{0_k}$ with that on $a^{0_{k+1}} a^{k+1_0}$. Thus, instead of a flow along $a^k a^0 \cup a^0 a^{k+1} \cup a^{k+1} a^k$, we consider the flow along $a^{k_0} a^{0_k} \cup a^{0_k} a^{0_{k+1}} \cup a^{0_{k+1}} a^{k+1_0} \cup a^{k+1_0} a^{k+1_k} \cup a^{k+1_k} a^{k_{k+1}} \cup a^{k_{k+1}} a^{k_0}$. The directional number of packets $v(p)$ flowing along that path per unit time is spatially smooth along this path.

The drawback of this trick is that not only do we have to extend the flow from the boundary to the interior of such a "smoothed triangle" as $a^{k_0} a^{0_k} \cup a^{0_k} a^{0_{k+1}} \cup a^{0_{k+1}} a^{k+1_0} \cup a^{k+1_0} a^{k+1_k} \cup a^{k+1_k} a^{k_{k+1}} \cup a^{k_{k+1}} a^{k_0}$ but we have to do the same extension for the concave curvilinear polygon $a^{0_1} a^{0_2} \cup \ldots \cup a^{0_k} a^{0_{k+1}} \cup \ldots \cup a^{0_{n-1}} a^{0_n}$ around the vertex a^0. The latter is necessary in order to elucidate the singularities at the vertices.

3.4.1 extension to surface elements

The surface element is in general a polygonal cell, but to make the notation easier we take it to be a triangle. There is no conceptual difference between the case of a triangle and the general polygonal case.

The extension of the flow from the boundary of the smoothed triangle to its interior is the well defined mathematical problem of *extension of vector fields*. As is well known, there might be obstructions to extending the everywhere nonvanishing continuous boundary flow to an *everywhere nonvanishing* continuous vector field inside the curve. In other words, if there are obstructions, there are points inside the triangle where the vector field will be discontinuous, will no longer exist, or will vanish.

Precisely, invoke the Jordan curve theorem to define s_2 to be the Jordan region of

$$\partial s_2 = a^{k_0} a^{0_k} \cup a^{0_k} a^{0_{k+1}} \cup a^{0_{k+1}} a^{k+1_0} \cup a^{k+1_0} a^{k+1_k} \cup a^{k+1_k} a^{k_{k+1}} \cup a^{k_{k+1}} a^{k_0}$$

Clearly, there exists a diffeomorphism

$$h : \partial s_2 \to S^1$$

Define along S^1 the vector field

$$w(e^{\jmath\theta}) = \frac{v(h^{-1}(e^{\jmath\theta}))}{||v(h^{-1}(e^{\jmath\theta}))||}$$

The map

$$\begin{array}{rcl} f : S^1 & \to & S^1 \\ e^{\jmath\theta} & \mapsto & w(e^{\jmath\theta}) \end{array}$$

where $w(e^{\jmath\theta})$ is identified with a point on the unit circle is referred to as *Gauss map*. Clearly, there exists an extension of the everywhere nonvanishing field $v : \partial s_2 \to \mathbb{R}^2 \setminus 0$ to an everywhere nonvanishing field $v : s_2 \to \mathbb{R}^2 \setminus 0$ iff there exists an extension of the Gauss map to $f : \mathbb{D} \to S^1$.

Theorem 14 *There is no obstruction to the extension of the everywhere nonvanishing vector field $v : \partial s_2 \to \mathbb{R}^2 \setminus 0$ to an everywhere nonvanishing field $v : s_2 \to \mathbb{R}^2 \setminus 0$ iff the degree of the Gauss map*

$$\deg(f) := \frac{1}{2\pi\jmath} \oint_{S^1} d \log f(e^{\jmath\theta})$$

vanishes.

Proof. Assume the degree of the Gauss map is vanishing. Then the winding number of the boundary field along the boundary vanishes; hence the flow can be extended as a continuous flow. Now, assume the flow can be extended and let us show by contradiction that the degree must vanish. Assume it does not. Since the flow can be extended, it can be homotopically deformed to a constant flow. But since the Gauss map is a homotopy invariant, the degree of the Gauss map on the boundary vanishes. A contradiction. ■

Intuitively, the degree is the winding number of the boundary field relative to an interior point.

Now, observe that the boundary vector field along the smoothed triangle is very specific: it is constant along $a^{k_0}a^{0_k}$, $a^{0_{k+1}}$, and $a^{k+1_k}a^{k_{k+1}}$. Furthermore, if the extensions along $a^{0_k}a^{0_{k+1}}$, $a^{k+1_0}a^{k+1_k}$, and $a^{k_{k+1}}a^{k_0}$ are minimal in the sense that the angle variations are minimum, then clearly the degree is restricted.

Theorem 15 *If the field extension along $a^{0_k}a^{0_{k+1}}$ is minimum in the sense that*

$$\frac{1}{2\pi\jmath} \int_{a^{0_k}}^{a^{0_{k+1}}} d \log v$$

is minimum among all extensions subject to the boundary conditions $f(a^{0_k})$ and $f(a^{0_{k+1}})$, and if a similar statement holds along $a^{0_{k+1}}$ and $a^{k+1_k}a^{k_{k+1}}$, then

$$\deg(f) = 0, \pm 1$$

Proof. This theorem is easily proved observing that there are only three cases of flows along the edges of the triangle: i) a counterclockwise circulating flow (in which case the degree is $+1$), a clockwise circulating flow (in which case the degree is -1); or a noncirculating flow (in which case the degree is 0). ■

Now, assume that the degree does not vanish. The extension of the map $f : S^1 \to S^1$ to a map $f : \mathbb{D} \to S^1$ is not possible; in other words, the extension field will have discontinuities or will have points where it does not exists if we insist that it be nonvanishing. If we use the fluid flow metaphor and extend the field to a potential field, the extension will have such points as

1. **sources**, that is, points 0 such that the radial velocity is $v_r = \frac{m}{r}$

2. **sinks**, that is, points 0 such that the radial velocity is $v_r = -\frac{m}{r}$
3. **vortices**, that is, points of angular velocity $v_\theta = \frac{k}{2\pi r}$

Another way to proceed is to consider an extension $f : \mathbb{D} \to \mathbb{D}$ and argue that the field will have points where it vanishes. Again using the fluid flow metaphor, these points would be called stagnation points, that is, points where the fluid is motionless. We would like to make a statement as to how many and what kind of stagnation points there are in $\mathbb{D}$. We follow a purely topological approach, which is independent of any fluid flow model.

Corollary 1 *If* $\deg(f) \neq 0$*, the extension of* $v : \partial s_2 \to \mathbb{R}^2 \setminus 0$ *to* $s_2 \to \mathbb{R}^2$ *has singularities* z_i *such that*

$$\sum_i \mathrm{index}(f, z_i) = \deg(f)$$

where index *denotes the degree of the local Gauss map around the singularity*

Proof. Proceed by homotopy invariance of the degree. deform the path ∂s_2 into the interior of s_2 until it encircles the singularities; the path can then be broken down into several paths around the respective singularities, in which case the integration yeilds the degree of the local gauss map. ■

3.4.2 extension around vertices

Bibliographical and historical notes

The material in Section 3.1 follows [62], closely.

Chapter 4

Embedding Graphs in 3-manifolds

Chapter 5

Extending graph to the simplicial complex of a possibly stratified manifold

Here we propose another approach to the problem of how to approximate the internet grid with some continuous geometric structure. By its nature, the internet is a graph. From a more topological viewpoint, this graph is a 1-dimensional simplicial complex. This 1-dimensional simplicial complex has an *objective* geometric realization as a highly singular curve in the sense that every vertex of a degree of three or more is a multiple point of the curve.

In an attempt to make the interpolating geometric structure less singular, one could associate, *purely formally*, a 2-simplex with every collection of three routers interconnected with links. In such a case as the graph that is the 1-dimensional skeleton of the triangulation of a compact surface, going from a 1-dimensional to a 2-dimensional realization removes all singularities. But this is a not a general fact. In general, associating a 2-simplex with every triple of interconnected vertices will result in a singular surface, in the sense that edges common to three or more 2-simplexes are singular lines.

Observe that this more topological construction will not in general result in the same surface as that constructed from graph theory. Indeed, the graph theoretic approach leads to a smooth surface, while the new construction will in general lead to a highly singular surface.

But probably more conceptual is the issue that the archetypal topological construction of going to higher dimensions has resulted in *some* decrease of the degree of singularity. Clearly, the smoother the geometric structure the easier it is to describe the flow running on it. From this point on, we could go even further and speculate whether the higher the dimension of the geometric realization of the grid, the less singular it could be made. The hope is that the whole internet grid could be viewed as a high dimensional compact manifold, over which the flow could be described by a partial differential equation.

In view of these observations, it appears that, conceptually, the problem is to associate with the internet grid a higher-dimensional simplicial complex, properly physically motivated and properly structured so that it be underlying simplicial complex of such a smooth object as a manifold, or a stratified manifold. Another requirement is that the extension of the link flow to the whole manifold should be tractable.

The fundamental approach is to assign to every complete subgraph K_n of G a $(n-1)$-simplex. This will indeed result in the highest dimensional simplicial complex that can possibly be associated with the graph. The problem is that the resulting simplicial complex will not, in general, be homogeneous. As such, this simplicial complex is not likely to be the simplicial complex of a manifold, not even a manifold with boundary, albeit it is more likely to be the simplicial complex of a stratified manifold.

Since the problem of finding a manifold that has a given simplicial complex as triangulation–the so-called *regularization problem*–is not likely to be successful, we opt for another approach. But probably a more important motivation for this latter approach is that, if it does not completely remove, at least it mitigates the subjectivity of the assignment of n-simplexes to complete n-subgraphs. Crudely speaking, the simplicial complex is allowed to be preprocessed, to undergo a *simple homotopy equivalence*, to make it more likely to be the complex of a manifold. The simple homotopy equivalence somehow manipulates the n-simplexes are replace them by a collection of $(n-1)$-simplexes, in an attempt to make the complex closer to the complex of a manifold. Of course, this distort the local structure of the complex, but leave the global–the large scale–structure unchanged.

Whatever the simple homotopy equivalence does to the simplexes, the remaining simplexes are still made up of vertices that are connected by edges and as such, by a recursive construction starting up with the same procedure as Section 3.4.1, it is possible to extend the flow along the links to a flow throughout the closure of the simplexes.

There is another, *completely different*, motivation for the preceding construction involving some concepts borrowed from the Čech homology. The starting point is to observe that a router in its routing protocol depends heavily on, and communicate heavily with, the nearby routers, that is, those that are directly linked to the nominal one. This leads to the idea of declaring the router a^0 along with its neighbors an *open set*, $V(a^i)$, that is, a vertex in the Čech homology theory. Following up the path of the Čech homology theory, it follows that the routers $a^0, ..., a^n$ form an n-simplex iff $\cap_{i=0}^{n} V(a^i) \neq \emptyset$. With a little bit of combinatorics, it is easily seem that the routers $a^0, ..., a^n$ form an n-simplex iff for every pair of routers there is a link between them. This provides another justification of the idea of declaring a set of interlinked routers a “simplex.”

5.1 Poincaré complexes

Ever since Poincaré, it has been customary to devise simplicial complexes that were meant to be combinatorial models of simplicial decompositions of topological manifolds. The concepts of *homology manifold* (useful to address Poincaré duality) and *pseudomanifold* (useful in Brouwer degree theory) are among such examples. Both of these classes have been known to encompass topological manifolds. Unfortunately, the gap between the topological manifolds and their easy combinatorial models does not close. There are counterexamples of homology manifolds that are not manifolds (see [67, p. 376]) and counterexamples of pseudomanifolds that are not manifolds (see [61, p. 100]). The path of approach taken here is first to choose the minimum structure a complex should enjoy for it to have a chance of being the complex of a manifold and second to decide whether such complex is indeed the complex of a manifold. This most fundamental property a complex should enjoy is the Poincaré duality.

Definition 32 *The integral group ring $\mathbb{Z}\Gamma$ of a discrete group $\Gamma = (\{\gamma_1, \gamma_2, ...\}, \cdot)$ is the set of formal series*

$$\sum_{n_i \in \mathbb{Z}} n_i \gamma_i$$

along with the operations:

- *group operation: $(\sum n_i\gamma_i) + (\sum m_i\gamma_i) = \sum(n_i + m_i)\gamma_i$*
- *ring product: $(\sum n_i\gamma_i) \cdot (\sum m_j\gamma_j) = \sum(n_i m_j)\,(\gamma_i \cdot \gamma_j)$*

Definition 33 *Let K be a complex. Let $C_*(K)$ and $C^*(K)$ be the chain complex and the cochain complex, respectively, of K for any coefficients. Take a chain $c_n \in C_n(K)$ and a cochain $c^i \in C^i(K)$. By decomposing any n-simplex of the chain c_n as a product of a i-simplex and a (n-i)-simplex, write $c_n = \sum c'_i c''_{n-i}$. Then the cap product $c_n \cap c^i \in C_{n-i}(K)$ is defined as*

$$c_n \cap c^i = \sum c^i(c'_i) c''_{n-i}$$

Theorem 16 (Poincaré duality) *A necessary condition for an n-dimensional simplicial complex K to be the simplicial complex of a manifold is that it is a Poincaré complex, that is, there exists a homology class $[c_n] \in H_n(K, \mathbb{Z}\pi_1(K))$ such that the cap multiplication by $[c_n]$ map*

$$\begin{array}{rcl} [c_n] \cap \cdot : H^i(K, \mathbb{Z}\pi_1(K)) & \to & H_{n-i}(K, \mathbb{Z}\pi_1(K)) \\ [c^i] & \mapsto & [c_n] \cap [c^i] = \sum c^i(c'_i) c''_{n-i} \end{array}$$

is an isomorphism for the (co)homology with coefficients in $\mathbb{Z}[\pi_1(K)]$, the integral group ring of the fundamental group $\pi_1(K)$.

The gap between Poincaré manifold and topological manifolds is in part due to the fact that topological manifolds satisfy a Poincaré duality already at the chain level in the sense that there exists a chain duality map

$$\cap c_n : C^i(M) \longrightarrow C_{n-i}(M)$$

5.2 (simple) homotopy equivalence

In topology, it is customary to classify spaces up to homotopy equivalence, because all algebraic objects associated with spaces via their simplicial, CW, or cell complexes are homotopy invariants. The problem is that homotopy equivalence is too coarse a classification. One way to make this classification somewhat finer is to restrict ourselves to the most obvious kind of homotopy equivalences: the so-called *simple* homotopy equivalences.

A **simple homotopy equivalence** is a sequence of elementary collapses and expansions. Before proceeding to a formal definition, we give some examples which illustrate its relevance to internet problem: Consider a set of three interlinked routers. In our topological modeling, this is a closed 2-simplex: $\overline{a^0a^1a^2}$. An **elementary collapse** is an operation like

$$\overline{a^0a^1a^2} \searrow \overline{a^0a^1} \cup \overline{a^0a^2}$$

The internet significance should be obvious: An elementary collapse of a system of three interlinked routers is the disabling of a link. An **elementary expansion** is just the reverse operation. (An expansion is basically the addition of a link.) Now, consider a system of 4 interlinked routers. In our topological model, this is a closed 3-simplex, $\overline{a^0a^1a^2a^3}$. An elementary collapse is

$$\overline{a^0a^1a^2a^3} \searrow \overline{a^0a^1a^2} \cup \overline{a^0a^2a^3} \cup \overline{a^0a^1a^3}$$

Here, the elementary collapse does not entail the disabling of any router; it simply removes the simplex property of the set of routers $a^0a^1a^2a^3$ and $a^1a^2a^3$. The elementary expansion is just the reverse operation of assigning a 2-simplex with $a^1a^2a^3$ and a 3-simplex to $a^0a^1a^2a^3$. The concepts of elementary collapse and expansion and simple homotopy equivalence can be extended to polyhedra by applying the elementary collapses and expansions to the "free" simplexes–that is, those that have no faces in common with other simplexes.

As already discussed, there is some subjectivity in the assignment of simplexes to interlinked routers. However, as we shall see soon, the manifold interpolating the grid can only be defined up to (simple) homotopy equivalence, and as such this removes some of the apriori in the assignment of simplexes to set of interlinked routers. The relevance of the concept of collapse and expansion to internet problem should be obvious.

Now we proceed with more formal definitions in the more general context of cell complexes

Definition 34 *A homotopy equivalence $L \hookrightarrow K$ is an elementary collapse, written $K \searrow L$, if K is obtained from L be attaching two cells of contiguous dimensions:*

$$K = (L \cup_f D^{k-1}) \cup_g D^k$$

where f, g are the attaching maps

$$\begin{array}{rcl} f : S^{k-2} & \to & L \\ g : S^{k-1} & \to & (L \cup_f D^{k-1}) \end{array}$$

An elementary expansion $L \nearrow K$ is just the reverse operation. A homotopy equivalence is simple if it is a sequence of elementary collapses and expansions.

To better appreciate this definition, we quote the following theorems:

Theorem 17 *A map $f : K \to L$ between polyhedra is a simple homotopy equivalence if upon extending the map to regular neighborhoods in a sufficiently high dimensional Euclidean space, it is homotopic to a PL homeomorphism.*

Theorem 18 (s-cobordism) *Let $M^{n\geq 5}$ be a compact connected manifold and consider all manifolds M' such that there exists a manifold W^{n+1} such that $\partial W = M \sqcup M'$ and W retracts onto M and M'. Then $W \cong M \times [0,1]$ iff $M \hookrightarrow W$ is a simple homotopy equivalence.*

Homotopy equivalences are classified by their *Whitehead torsion.* The torsion of a homotopy equivalence $f : K \to L$, $\tau(f)$, somehow measures the extend to which the homotopy equivalence differs from a simple homotopy equivalence. The torsion of a homotopy equivalence takes values in the *Whitehead group* $\mathrm{Wh}(\pi_1(K))$. The Whitehead group is a K-theoretic concept. Define

$$GL(\mathbb{Z}\pi_1(K)) = \cup_{n=1}^{\infty} GL_n(\mathbb{Z}\pi_1(K))$$

Let $E_n(\mathbb{Z}\pi_1(K))$ be the subgroup of GL_n generated by elementary matrices of the form $1 + ge_{ij}$, $1 \leq i,j \leq n$, $g \in \pi_1(K)$, where e_{ij} is the matrix with zeros everywhere except $1_{\pi_1(K)}$ in the entry (i,j). Let $E(\mathbb{Z}\pi_1(K)) = \cup_{n=1}^{\infty} E_n(\mathbb{Z}\pi_1(K))$ and let $[GL(\mathbb{Z}\pi_1(K)), GL(\mathbb{Z}\pi_1(K))]$ be the commutator subgroup.

Lemma 1 (Whitehead) *The commutator subgroup $[GL(\mathbb{Z}\pi_1(K)), GL(\mathbb{Z}\pi_1(K))]$ is normal and*

$$[GL(\mathbb{Z}\pi_1(K)), GL(\mathbb{Z}\pi_1(K))] = E(\mathbb{Z}\pi_1(K))$$

Definition 35 *The 1st order algebraic K-group of the integral group ring $\mathbb{Z}\pi_1(K)$ is the Abelian (multiplicative) group*

$$K_1(\mathbb{Z}\pi_1(K)) = GL(\mathbb{Z}\pi_1(K))/[GL(\mathbb{Z}\pi_1(K)), GL(\mathbb{Z}\pi_1(K))]$$

In a sense, $K_1(\mathbb{Z}\pi_1(K))$ is an Abelianized version of $GL(\mathbb{Z}\pi_1(K))$.

Definition 36 *The Whitehead torsion $Wh(\pi_1(K))$ of the group $\pi_1(K)$ is defined as*

$$Wh(\pi_1(K)) = K_1(\mathbb{Z}[\pi_1(K)])/\left\{ \begin{pmatrix} \pm g_1 & 0 & \cdots \\ 0 & \pm g_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} : g_i \textit{ is a unit of } \pi_1(K) \right\}$$

We first define the torsion of a homotopy equivalence $L \hookrightarrow K$, where L is a subcomplex of K. Consider the relative chain group $C_*(K, L, \mathbb{Z}\pi_1(K))$ along with its differential $\partial_k : C_k(K, L, \mathbb{Z}\pi_1(K)) \to C_{k-1}(K, L, \mathbb{Z}\pi_1(K))$. ∂_k has a matrix representation for some basis of the module $C_*(K, L, \mathbb{Z}\pi_1(K))$. Hence,

Definition 37 *The torsion $\tau(L \hookrightarrow K)$ is the image of the matrix representation of ∂_* in $Wh(\pi_1(K))$.*

For a general homotopy equivalence $f : K \to L$, consider the mapping cylinder $C_f = K \times [0,1] \cup_f L$, where the attaching map is $f : K \times \{0\} \to L$. Clearly, $K \times \{0\} \hookrightarrow C_f$ is a homotopy equivalence. Hence,

Definition 38

$$\tau(f : K \to L) = \tau(K \times \{0\} \hookrightarrow C_f)$$

.

An equivalent definition consists in considering the chain complex $K_n \oplus L_{n+1}$ along with the boundary operator

$$\begin{pmatrix} \partial_{K_n} & 0 \\ (-1)^n f_n & \partial_{L_{n+1}} \end{pmatrix} : K_n \oplus L_{n+1} \to K_{n-1} \oplus L_n$$

and define the torsion to be the image of the matrix representation of the above boundary in $\mathrm{Wh}(\mathbb{Z}\pi_1(K))$.

Now, we can formulate the algebraic characterization of simple homtopy equivalence:

Theorem 19 *The homotopy equivalence $f : K \to L$ is simple iff $\tau(f) = 0$.*

Definition 39 *A simple Poincaré complex is a Poincaré complex with the property that its chain duality map is a simple homotopy equivalence.*

A compact manifold has a simple chain duality map. It can also be shown that a compact homology manifold is a simple Poincaré complex.

5.3 surgery exact sequence

Given a (simple) Poincaré complex K, there is a need to provide a precise definition of the set $\mathcal{S}^{(s)}(K)$ of equivalence classes of manifold structures that have K as a triangulation.

Definition 40 *A structure on a (simple) Poincaré complex K is pair (M, ϕ), where M is a compact manifold and $\phi : M \to K$ is a (simple) homotopy equivalence. The structure set $\mathcal{S}^{(s)}(K)$ is defined to be the set of equivalence classes of structures on K for the equivalence relation $(M, \phi) \sim (M', \phi')$ defined as*

- *M and M' are cobordant, that is, there exists a manifold W such that $\partial W = M \sqcup M'$.*
- *There exists a homotopy equivalence $F : W \to K \times [0,1]$ such that $F|M = f$ and $F|M' = f'$.*

We note that there are essentially two theories: one for simple Poincaré complexes and another one where the simplicity assumption is not required. Here, the primary focus is on the latter case.

Having defined $\mathcal{S}(K)$, the objective is two-fold:

1. Decide whether $\mathcal{S}(K) \neq \emptyset$.

2. If the answer to 1) is affirmative, characterize how big $\mathcal{S}(K)$ is.

A complete answer to the two above questions would take us along a very long path, so that only the main ideas are developed here.

A manifold has a normal bundle and the first question is whether the complex K can be endowed with some "normal data" compatible with that of a manifold. The normal bundle of a manifold is classified by a map $M \to BO$, where BO is the classifying space of vector bundles with orthogonal structure group. The problem is that a Poincaré complex need not have a normal bundle; it only has a (Spivak) spherical fibration. Spherical fibrations over a Poincaré complex K are classified by maps $K \to BG$, where BG is the base space of the universal sphere bundle. While a vector bundle trivially induces a spherical fibration via a map $BO \to BG$, not all spherical fibrations come from a vector bundle. Whether a spherical fibration comes from a vector bundle is formulated as to whether there exists a lifting of the classifying map of the spherical fibration, as shown by the diagram

$$\begin{tikzcd} & BO \arrow[d] \\ K \arrow[r] \arrow[ur, "?"] & BG \end{tikzcd}$$

Such a lifting is called a *normal invariant.* It can be shown that there exists a normal invariant iff the composite map $K \to BG \to B(G/O)$ is null homotopic. The homotopy class of this composite is the first obstruction to the Poincaré complex being homotopic to a manifold. It can also be shown that the set of normal invariants $\mathcal{N}(K)$ can be identified with $[K, G/O]$. Next, we investigate how a manifold can fit within the above diagram. Define a *degree one normal map* $f : M \to K$ to be a degree one map that makes the following diagram commute:

$$\begin{tikzcd} M \arrow[r] \arrow[d, "f"'] & BO \arrow[d] \\ K \arrow[r] \arrow[ur] & BG \end{tikzcd}$$

The top row is the classifying map of the normal bundle of the manifold M, the bottom row is the classifying map of the spherical fibration over K, and the diagonal arrow is a lifting of the spherical fibration classifying map. The map $f : M \to K$ is said to be of *degree one* if, given a fundamental homology class $[z_n^M] \in H_n(M)$ and a fundamental homology class $[z_n^K] \in H_n(K)$, we have $f_*([z_n^M]) = [z_n^K]$. It is tacitly assumed that the normal map f comes equipped with the classifying map $M \to BO$ of the normal bundle of the manifold M. It is easy to see by a pullback argument that the classifying map $M \to BO$ induces a classifying map $K \to BO$. Conversely, it can be shown that every normal invariant comes from some degree one normal map. By the Browder-Novikov theorem, the vector bundle reduction of the spherical fibrations are in a one-to-one correspondence with the bordism classes of degree one normal maps. Clearly, a homotopy equivalence is a degree one normal map and this implies existence of a map $\mathcal{S}(M) \to \mathcal{N}(M)$, which will appear in the surgery exact sequence.

The next step is to determine whether the degree one normal map $f : M \to K$ is a homotopy equivalence. For this, it is necessary that $\pi_*(f) = 0$. Set $n = 2m$ or $n = 2m + 1$. It turns out that it is always possible to "kill" $\pi_i(f)$ for $i \leq m$. Whether it is possible to "kill" the remaining part of $\pi_*(f)$ is the secondary, surgery theoretic obstruction to making f a homotopy equivalence. The surgery theoretic obstruction takes value in the surgery obstruction group $L_n(\pi_1(K))$. The latter implies existence of a map $\theta : \mathcal{N}(M) \to L_n(\pi_1(M))$.

The main result is formulated in the following theorem:

Theorem 20 (surgery exact sequence) *Let K be a Poincaré complex of formal dimension n.*

1. *$\mathcal{S}(K) \neq \emptyset$ iff there exists a degree one normal map $f : M \to K$ such that its surgery obstruction*

$$\theta_*(f) \in L_n(\mathbb{Z}\pi_1(K))$$

vanishes.

2. *Given a homotopy equivalence $f : M \to K$, the structure set fits within the surgery exact sequence*

$$\mathcal{S}(M) \xrightarrow{\eta} \mathcal{N}(M) \xrightarrow{\theta} L_n(\pi_1(M))$$

where $\mathcal{N}(M)$ is the normal data and L_n the surgery obstruction group.

If this test does not pass, then we will have to first investigate whether K is the simplicial complex of a manifold with boundary. The way to go is to consider a (simple) Poincaré pair $(K, \partial K)$ and redevelop a relative approach to the above.

If K cannot be realized as a manifold with boundary, then the next logical step would be to determine whether K can be realized as a *stratified manifold.* Here we are reaching the forefront of current mathematical research [94].

5.4 extension of flows

This higher-dimensional extension problem follows pretty much the 2-dimensional case of section 3.4.1, with the same difficulty that at the vertices the link flow is ill-defined.

The approach is essentially recursive: the 2-simplexes have their vertices rounded; in a first step the flow is continuously extended to the curves near the vertices and in a second step the flow is extended the the interior of the rounded simplex.

Consider then a tetrahedron. By the previous construction its faces are approximated by rounded triangles. The tetrahedron is hence rounded by a piece of surface near its vertices and the flow is extended to this piece of surface. As such the flow is defined on ∂s_3 and extended to s_3. In another operation the flow on the surface around the apexes is extended as well.

5.4.1 extension of flows to volume elements

Consider the problem of extending a vector field v defined on ∂s_n, where s_n denotes the "rounded" simplex. Clearly, there exists a diffeomorphism

$$h : \partial s_n \to S^{n-1}$$

Define on S^{n-1} the vector field

$$w(x) = \frac{v(h^{-1}(x))}{||v(h^{-1}(x))||}$$

The map

$$f : S^{n-1} \to S^{n-1}$$

is referred to as Gauss map and as such has a degree, defined homologically as

$$\deg(f) = \frac{f_*([c_{n-1}])}{[c_{n-1}]}$$

where $[c_{n-1}]$ is the fundamental class of $H_{n-1}(S^{n-1})$.

Proposition 2 *There is an extension iff the degree of the Gauss map vanishes.*

Proposition 3 *Any extension of the vector field v defined on ∂s_n to s_n has zeros z_i such that*

$$\sum_i \mathit{index}(f, z_i) = \deg(f)$$

where index *denotes the degree of the local Gauss map.*

Proof. See [20, Proposition 12.12] ■

5.4.2 extension around vertices

Part II

Riemannian and nonRiemannian Geometry

This part is the heart of the book. It will probably appear the most difficult one to the novice, as it attempts to unify the concepts of Riemannian manifolds and weighted graphs, an apparently problematic operation.

Naturally, this starts with a review of classical Riemannian geometry, in which we emphasize the fact that those behavioral issues of the geodesics potentially relevant to routing are encapsulated in that single parameter: the sectional curvature. Thereafter, motivated by the fact that a curvature-like concept might explain such aberrant routing behavior as fluttering, the developments take a less traditional form, as we endeavor to define curvature for general metric spaces. The first step along that line only slightly departs from the Riemannian tradition by attempting an isometric embedding of a graph in constant curvature space. Indeed, if this can be done, it could be argued that the graph has the curvature of the space it is embedded in. A first problem with this approach is that a graph may be embeddable in spaces of different curvature. Furthermore, it turns out that, for the tree-like graphs, the embedding can only be done in a space of "infinite negative curvature," a potentially problematic concept.

Despite its shortfalls, this embedding approach gives us the clue that probably a better idea would be to attempt a constant curvature space embedding only at a small scale. The latter is, in a sense, the idea of comparison geometry. In this approach, a geodesic triangle of a graph is isometrically embedded in a constant curvature space to yield the comparison triangle; then the graph is declared to have locally the curvature of the comparison space if the triangle in the graph has the same metric properties as the comparison triangle. Using this comparison concept, we next endeavor to define local curvature concepts for graphs. Positive curvature appears the most relevant at a low scale and is in fact relevant to the WWW graph. The problem is that the details of the local curvature can be very nastily heterogeneous and in fact irrelevant if the graph is viewed from the large scale. This is essentially the concept of large scale geometry, in which negative curvature is more relevant. To circumvent the difficulty that trees are only embeddable in "infinite negative curvature spaces," we use a comparison argument at the idealized scale of the infinitely large triangles. This yields the concept of Gromov (δ)-hyperbolic spaces. More specifically, a Gromov δ hyperbolic metric space is defined such that any geodesic triangle, no matter its size, has an inscribed triangle of finite perimeter δ. The latter is a property shared by *all* Riemannian manifolds of an arbitrary curvature $\kappa < 0$. As such, this δ-hyperbolic space concept gets out of the straitjacket of constant curvature. However, the latter is not directly applicable to Engineering problems because, no matter how large Internet graphs are, they are finite, while the Gromov δ-hyperbolic property only make sense for infinite graphs. This problem is resolved by properly scaling δ relative to the diameter of the graph and by defining a finite graph to be hyperbolic if δ/diam is less that a certain threshold. The relevance of the latter to networking problems is proved by showing that scale free graphs are indeed δ-hyperbolic.

Chapter 6

Shortest Distance: Concepts and Computations

As a prelude to Riemannian geometry in a broad sense to include nonRiemamannian geometry and its networking applications, here, we develop length minimizing paths computation from the point of view of optimal control–more specifically, from the point of view of Bellman's Principle of Optimality. The general Euler-Lagrange equations of the calculus of variations and its specialization to length minimizing paths, usually referred to as equations of geodesics, are derived from the unified standpoint of Bellman's principle. To further folster the common mathematical structure shared by manifolds and graphs, we derive the Bellman-Ford algorithm for finding a cost minimizing path on a graph from the same Bellman Principle of Optimality. This common approach to length minimizing paths in manifolds and graphs serves as a spring board towards the more formal unification of graphs and manifolds insofar as their geodesics are concerned under the concepts of geodesic space and length space via the Hopf-Rinow theorem.

6.1 Riemannian spaces

6.1.1 Bellman's principle

We start with the general optimal control problem:

$$\inf_u \int_{t_0}^{t_1} L(x(\tau), u(\tau)) d\tau$$

subject to

$$\dot{x} = f(x, u), \quad x(0) = x_0$$

In the above, x is the state, u is the control, and L stands for either "loss" or, as we shall see later, "Lagrangian" function. At the heart of the Principle of Optimality is the concept of the "cost to go,"

$$V(x,t) = \inf_{u,x(t)=x} \int_t^{t_1} L(x(\tau),u(\tau))d\tau$$

By the principle of optimality, we have

$$V(x,t) = \inf_{u(\tau)} \left(\int_t^{t+\epsilon} L(x(\tau),u(\tau))d\tau + V(x(t+\epsilon),t+\epsilon) \right)$$

Equivalently,

$$\inf_u \left(\int_t^{t+\epsilon} L(x(\tau),u(\tau))d\tau + V(x(t+\epsilon),t+\epsilon) - V(x(t),t) \right) = 0$$

So far, we have been arguing in a global, coordinate independent fashion, but now it is most convenient to give ourselves a local coordinate system $x^1,...,x^n$. Letting $\epsilon \downarrow 0$ in the above equation yields

$$\min_u \left(L(x,u) + \sum_i \frac{\partial V}{\partial x^i} f^i(x,u) + \frac{\partial V}{\partial t} \right) = 0$$

The above is sometimes referred to as *Hamilton-Jacobi-Bellman's equation.* If the control u is unconstrained, the above infimization yields an implicit relation involving the partial derivatives of L and f^i relative to the various components of u. If the conditions of the implicit function theorem are satisfied, the implicit relation can be rewritten in an explicit form as $u = P(V)$, for some partial differential operator P. Hence the whole problem resides in the solution to the partial differential equation for $V(x,t)$,

$$L(x,P(V)) + \sum_i \frac{\partial V}{\partial x^i} f^i(x,P(V)) + \frac{\partial V}{\partial t} = 0$$

This equation can be solved in the well-known linear quadratic problem. In this case, and for that matter in all other cases where it can be solved, u becomes an explicit function of x,t, say, $u = K(x,t)$. Plugging the latter in the state equation yields $\dot{x}(t) = f(x(t),K(x(t),t))$. Existence of a unique solution is guaranteed, at least locally, once an initial condition $x(t_0)$ is provided.

6.1.2 Euler-Lagrange equations

To derive Euler's equations, it suffices to choose the control to be $\dot{x}$, in which case the preceding infimization yields,

$$\frac{\partial L}{\partial \dot{x}^i} + \frac{\partial V}{\partial x^i} = 0 \tag{6.1}$$

Next, we differentiate the above relative to t, and we observe that

$$\begin{aligned}\frac{d}{dt}\frac{\partial V}{\partial x^i} &= \frac{\partial^2 V}{\partial x^i \partial x^j}\dot{x}^j + \frac{\partial^2 V}{\partial x^i \partial t} \\ &= \frac{\partial}{\partial x^i}\left(\frac{\partial V}{\partial x^j}\dot{x}^j + \frac{\partial V}{\partial t}\right) \\ &= \frac{\partial}{\partial x^i}\frac{dV}{dt} \\ &= -\frac{\partial L}{\partial x^i}\end{aligned}$$

Therefore, the differentiation of (6.1) relative to t yields the celebrated Euler equations:

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}^i} - \frac{\partial L}{\partial x^i} = 0$$

The problem is that the differentiation relative to t has made the resulting Euler equation nonuniquely solvable for a given initial condition x_0. In fact, it should already be clear from the above that Euler's equation is a second order ordinary differential equation, which has a guaranteed unique solution once $x(t_0)$ and $\dot{x}(t_0)$ are given.

6.1.3 Geodesics

Here we derive the geodesic equation from the Euler-Lagrange equation applied to the arc length and the arc energy.

Given that $\gamma : [a, b] \to M$ is a differentiable curve in a Riemannian manifold M, then the length of γ is given by

$$\begin{aligned}\ell(\gamma) &= \int_a^b \left\| \frac{d\gamma}{dt}(t) \right\| dt \\ &= \int_a^b \left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle^{\frac{1}{2}} dt \qquad (6.2)\end{aligned}$$

and the energy of γ is given by

$$E(\gamma) = \frac{1}{2}\int_a^b \left\| \frac{d\gamma}{dt}(t) \right\|^2 dt. \qquad (6.3)$$

The length and energy of γ can be expressed in local coordinates $x =$

$\left(x^1, x^2, \ldots, x^n\right)$ by the following formula:

$$\ell\left(\gamma\right)=\int\limits_a^b \sqrt{g_{ij}\left(x\left(\gamma\left(t\right)\right)\right)\dot{x}^i\left(t\right)\dot{x}^j\left(t\right)}dt, \tag{6.4}$$

$$E\left(\gamma\right)=\frac{1}{2}\int\limits_a^b g_{ij}\left(x\left(\gamma\left(t\right)\right)\right)\dot{x}^i\left(t\right)\dot{x}^j\left(t\right)dt. \tag{6.5}$$

Lemma 2 *For each differentiable curve $\gamma : [a,b] \to M$*

$$\ell\left(\gamma\right)^2 \leq 2\left(b-a\right)E\left(\gamma\right), \tag{6.6}$$

and equality holds if and only if $\left\|\frac{d\gamma}{dt}\right\| =$ constant.

Proof. This is a direct consequence of the Cauchy-Schwarz inequality:

$$\int_a^b \|\dot{\gamma}\|^2 \cdot 1dt \leq \left(\int_a^b \|\dot{\gamma}\|^2 dt\right)^{1/2}\left(\int_a^b 1dt\right)^{1/2}$$

(See [51, Chap. 1, Lemma 1.4.2].) ■

Definition 41 *A curve γ is parameterized proportionally to arc-length if $\left\|\frac{d\gamma}{dt}\right\| =$ constant almost everywhere. In addition, if $\left\|\frac{d\gamma}{dt}\right\| = 1$, the curve γ is parameterized by arc-length.*

Hence a geodesic is always parameterized proportionally to arc-length. In fact, there exists a diffeomorphism $\psi : [0, \ell\left(\gamma\right)] \to [a,b]$ such that

$$\left\|\frac{d\gamma \circ \psi}{dt}\right\| = 1$$

for almost every t. Hence a geodesic γ can be reparameterized such that $\gamma : [0, \ell\left(\gamma\right)] \to M$ is parameterized by arc-length.

Lemma 3 *If $\gamma : [a,b] \to M$ is a differentiable curve, and $\psi : [\alpha, \beta] \to [a,b]$ is a change of parameter, then*

$$\ell\left(\gamma \circ \psi\right) = \ell\left(\gamma\right). \tag{6.7}$$

Proof. See [51, Chap. 1, Lemma 1.4.3]. ■

From lemma 3, the length of a differentiable curve is invariant under a change of parameter.

Arc length minimization

By definition, a geodesic is (locally) arc length minimizing, so that the relevant Lagrangian is

$$L(x,\dot{x}) = \left(g_{ij}\dot{x}^i\dot{x}^j\right)^{1/2}$$

Hence, Lagrange's equations yield, successively,

$$\begin{aligned} \frac{d}{dt}\frac{\partial}{\partial \dot{x}^k}\left(g_{ij}\dot{x}^i\dot{x}^j\right)^{1/2} - \frac{\partial}{\partial x^k}\left(g_{ij}\dot{x}^i\dot{x}^j\right)^{1/2} &= 0 \\ \frac{d}{dt}\frac{1}{L}\left(g_{kj}\dot{x}^j + g_{ik}\dot{x}^i\right) - \frac{1}{L}\frac{\partial g_{ij}}{\partial x^k}\left(\dot{x}^i\dot{x}^j\right) &= 0 \end{aligned}$$

At this stage, it is convenient to reparameterize the problem in terms of the arc length s. The transformation $t \mapsto s$ is certainly invertible because its Jacobian $\frac{ds}{dt} = L$ is nonvanishing. After the reparameterization, we get

$$\frac{d}{ds}\left(g_{kj}\dot{x}^j + g_{ij}\dot{x}^i\right) - \frac{\partial g_{ij}}{\partial x^k}\left(\dot{x}^i\dot{x}^j\right) = 0$$

where now $\dot{x}$ is a short for $\frac{dx}{ds}$. Computing the partial derivatives yields,

$$g_{kj}\ddot{x}^j + g_{ik}\ddot{x}^i + \left(\frac{\partial g_{kj}}{\partial x^l}\dot{x}^l\dot{x}^j + \frac{\partial g_{ik}}{\partial x^l}\dot{x}^l\dot{x}^i - \frac{\partial g_{kj}}{\partial x^k}\dot{x}^i\dot{x}^j\right) = 0$$

Manipulating the indices yields

$$g_{kj}\ddot{x}^j + \frac{1}{2}\left(\frac{\partial g_{kj}}{\partial x^i} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k}\right)\dot{x}^i\dot{x}^j = 0$$

Premultiplying by g^{lk} and contracting the index k yields the celebrated geodesic equation

$$\frac{d^2x^k}{ds^2} + \Gamma^k_{ij}\frac{dx^i}{ds}\frac{dx^j}{ds} = 0$$

where

$$\Gamma^k_{ij} = \frac{1}{2}g^{kl}\left(\frac{\partial g_{lj}}{\partial x^i} + \frac{\partial g_{il}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^l}\right)$$

are the so-called Christoffel symbols of the Levi-Civita connection.

Energy minimization

The Euler-Lagrange equations of energy minimization are given by the following equation

$$\frac{d}{dt}\frac{\partial E}{\partial \dot{x}^i} - \frac{\partial E}{\partial x^i} = 0,\ i = 1,2,\ldots,n, \tag{6.8}$$

where

$$\begin{aligned} E\left(x,\dot{x}\right) &= \frac{1}{2}g_{jk}\left(x\right)\dot{x}^j\dot{x}^k, \\ x^i\left(t\right) &= x^i\left(\gamma\left(t\right)\right). \end{aligned}$$

Then

$$\begin{aligned}\frac{d}{dt}\left(g_{ik}\left(x\left(t\right)\right)\dot{x}^{k}\left(t\right)+g_{ji}\left(x\left(t\right)\right)\dot{x}^{j}\left(t\right)\right)&=g_{jk,i}\left(x\left(t\right)\right)\dot{x}^{j}\left(t\right)\dot{x}^{k}\left(t\right)\\ g_{ik}\ddot{x}^{k}+g_{ji}\ddot{x}^{j}+g_{ik,l}\dot{x}^{l}\dot{x}^{k}+g_{ji,l}\dot{x}^{l}\dot{x}^{j}&=g_{jk,i}\dot{x}^{j}\dot{x}^{k}\\ g_{lm}\ddot{x}^{m}+\frac{1}{2}\left(g_{lk,j}+g_{jl,k}-g_{jk,l}\right)\dot{x}^{j}\dot{x}^{k}&=0\\ g^{il}g_{lm}\ddot{x}^{m}+\frac{1}{2}g^{il}\left(g_{lk,j}+g_{jl,k}-g_{jk,l}\right)\dot{x}^{j}\dot{x}^{k}&=0.\end{aligned}$$

Hence

$$\ddot{x}^{i}\left(t\right)+\Gamma^{i}_{jk}\left(x\left(t\right)\right)\dot{x}^{j}\left(t\right)\dot{x}^{k}\left(t\right)=0,\ i=1,2,\ldots,n. \tag{6.9}$$

That is, the geodesic is the solution of the Euler-Lagrange equations for energy E.

Theorem 21 *A geodesic γ on a Riemannian manifold M satisfies the Euler-Lagrange equations for the energy E.*

Proof. See [51, Chap. 1, Lemma 1.4.4]. ■

6.2 NonRiemannian spaces

6.2.1 Metric, geodesic, and length spaces

Definition 42 *A metric space (X,d) is a set X endowed with a symmetric, positive definite, nondegenerate form satisfying the triangle inequality.*

Definition 43 *Given that (X,d) is a metric space, then a curve or a path γ in X is a continuous mapping from a compact interval $[a,b]\subset\mathbb{R}$ into X.*

Definition 44 *Given that (X,d) is a metric space, then the length of a curve $\gamma:[a,b]\to X$ is*

$$l\left(\gamma\right)=\sup_{a=t_0\leq t_1\leq\cdots\leq t_n=b}\sum_{i=0}^{n-1}d\left(\gamma\left(t_i\right),\gamma\left(t_{i+1}\right)\right),$$

where the supremum is taken over all possible partitions (no bound on n) with $a=t_0\leq t_1\leq\cdots\leq t_n=b$. A curve γ is said to be rectifiable if it has finite length.

Definition 45 *Given that (X,d) is a metric space, then a geodesic path γ is an isometric embedding $\gamma:[a,b]\to X$, that is, a map such that*

$$d\left(\gamma\left(t\right),\gamma\left(t'\right)\right)=\left|t-t'\right|,\quad\forall t,t'\in\left[a,b\right].$$

It is trivial to observe that the parameter t of the geodesic is the arc length, which we shall from here on denote as s. Next, if A, B are the end points of the geodesic, that is, $\gamma(a) = A$, $\gamma(b) = B$, then the above implies that

$$\ell(\gamma) = \sup_{a=s_0 \leq \ldots s_i \leq s_{i+1} \ldots \leq s_n = b} \sum_{i=0}^{n-1} d(\gamma(s_i), \gamma(s_{i+1})) = \sum_{i=0}^{n-1} |s_{i+1} - s_i| = b - a$$

Furthermore, the isometric embedding property also implies that $d(A, B) = b - a$, so that $\ell(\gamma) = d(A, B)$. It is easily see from the triangle inequality that any path joining A to B has length at least $d(A, B)$. Hence it follows that the geodesic γ joining A to B is a shortest length arc joining A to B. As a word of caution, this definition of a geodesic as an isometric embedding implies that the geodesic is *globally* arc length minimizing, whereas the Riemannian definition of geodesic implies that it is *only locally* arc length minimizing.

Definition 46 *A metric space (X, d) is a geodesic space if for every two points in X there exists a geodesic path joining them. A metric space (X, d) is a uniquely geodesic space if there is exactly one geodesic path joining two points in X. A metric space (X, d) is an r-geodesic space if for every two points in X of which the distance between them is less than r, there is a geodesic path joining them.*

An example of a geodesic space is a complete metric space (X, d) with the property that $\forall x, y \in X$, there exists a point $m \in X$ such that

$$d(x, m) = d(y, m) = \frac{1}{2} d(x, y)$$

m is called the midpoint between x and y (see [23, p. 41]). To prove the preceding, find the midpoint between x and m, the midpoint between m and y, and iterate until the sequence of midpoints converges to an arc, easily seen to be the geodesic from x to y.

The Euclidean space $\mathbb{E}^n$ is an example of a uniquely geodesic space. The punctured Euclidean space, $\mathbb{E}^n \setminus \{0\}$, is not a geodesic space, because, for example, the points $(-1, 0)$ and $(+1, 0)$ cannot be joined by a geodesic. The issue as to whether a Riemannian manifold is a geodesic space is dealt with by the celebrated Hopf-Rinow theorem:

Theorem 22 (Hopf-Rinow) *For a Riemannian manifold M, the following statements are equivalent:*

1. *M is metrically complete*

2. *M is geodesically complete*

Furthermore, under any of these conditions, M is a geodesic space.

Proof. See [51, Th. 1.4.8], [21, Corollary 3.20]. ■

A finite graph in which every edge is assigned a "weight," as it is practiced in communication networks, is a geodesic space; the distance between two vertices is defined as the infimum of the weights of all paths joining A and B, where the weight of a path is the sum of the weights of the constituting edges of the path. An infinite graph need not be a geodesic space; for example, the non locally finite graph on two vertices A, B joined by countably infinitely many edges with weights $1 + 1/n,\ n = 1, 2, ...$ is not a geodesic space, because $d(A, B) = 1$, yet there does not exist a geodesic of length 1 joining A, B.

Observe that, for a weighted graph, the distance is defined from the link weights. This can be formalized under the concept of length space.

Definition 47 *A length structure defined over a topological space X is a subset A of arcs in X along with a function $L : A \to \mathbb{R}^+ \cup \infty$, and subject to the following restrictions:*

1. *closure under restriction*
2. *closure under concatenation*
3. *closure under reparameterization.*

Clearly, a topological space together with a length structure can be made a metric space as follows:

$$d_L(A, B) = \inf\{\ell(L) : L \text{ joins A to B}\}$$

Such a distance emanating from a length structure is called length distance. More formally,

Definition 48 *A metric space (X, d) whose metric is a length metric is called length space.*

Clearly, a graph along with the metric d induced by the link weights is a length space.

Another example of length space is a Riemannian manifold endowed with its usual distance, since this distance emanates from the length element ds^2.

In general, a metric space need not be a length space and a length space need not be a geodesic space. For example, $\mathbb{R} \setminus \{0\}$ with the metric $d\,(x, y) = |x - y|$ is a metric space, but it is not a length space. In addition, $\mathbb{E}^2 \setminus \{0\}$ is a length space, but it is not a geodesic space, as already observed.

However, from a generalized version of the Hopf-Rinow theorem, sometimes referred to as Hopf-Rinow-Cohn-Vossen theorem (see [23, Th. 2.5.28]), a complete locally compact length space is a geodesic space.

Theorem 23 (Hopf-Rinow-Cohn-Vossen) *Given that X is a complete locally compact length space, then*

1. *every closed bounded subset of X is compact,*

2. X is a geodesic space.

From the above, we recover the already known result that a complete Riemannian manifold is a geodesic space.

Therefore, finitely connected graphs, Cayley graphs, and complete Riemannian manifolds are unified under the concept of geodesic metric spaces. Clearly, we have found a common structure for the geometric objects on which hydrodynamic flows and information flows propagate. We have fallen, however, a bit short of encompassing those networks where the cost of communicating from A to B is not the same as the cost of communicating from B to A, also referred to as *digraphs*. This could be accomplished by extending the approach of Section A.4 to nonsymmetric adjacency matrices.

6.2.2 Bellman-Ford and Dijkstra algorithms

Here we look at arc length minimization on a graph. The two well known algorithms—the Dijkstra and the Bellman-Ford algorithms—both involve some version of the Principle of Optimality. The only difference between the two algortihms is that the Bellman-Ford algorithm uses time parameterization, that is, parameterization by the number of hops, while the Dijkstra algorithm uses arc length parameterization.

To understand what version of the Principle of Optimality is specifically involved in arc length minimization, we temporarily remains within the confines of Riemannian geometry. Arc length minimization specifically involves mixed conditions: an initial condition x_0 and a terminal condition x_1. The generic problem is to join x_0 to x_1 along an arc that minimizes some cost functional, not necessarily the energy functional as is the case with the reachability gramian. The problem is that, if we define

$$W(x_0, x_1, t_1 - t_0) = \inf \int_{t_0}^{t_1} L d\tau$$

subject to $x(t_0) = x_0$ and $x(t_1) = x_1$ for *fixed* $t_1 - t_0$, the above may not have a finite solution. To increase the chances of existence of a finite solution, we rather define

$$W(x_0, x_1, T) = \inf_{\substack{u([t_0,t_1]) \\ t_1 - t_0 \leq T}} \int_{t_0}^{t_1} L d\tau$$

that is, the cost it takes to go from x_0 to x_1 in T *or less than* T units of time. The principle of optimality in this newer context does not involve the cost-to-go, but the "cost-to-backstep," $W(x_0, x, T)$; specifically,

$$W(x_0, x_1, T + \Delta) = \inf_{x(t)=x, t_1 - t \leq \Delta} \left\{ W(x_0, x, T) + \int_t^{t_1} L d\tau \right\}$$

Bellman-Ford algorithm

Let the weighted graph (G, w) be given, where $w(v_i, v_j)$ is the weight of the edge joining v_i to v_j. The Bellman-Ford algorithm recursively constructs the tree $T(v_0, h)$ of all geodesic paths from v_0 having no more than h hops. For $v \in T(v_0, h)$, let $W(v_0, v, h) < \infty$ be the minimum cost of going from the vertex v_0 to the vertex v in no more than h hops. For $v \notin T(v_0, h)$, $W(v_0, v, h) = \infty$.

The algorithm is initialized as $T(v_0, 0) = \{v_0\}$, $W(v_0, v_0, 0) = 0$ and $W(v_0, v \neq v_0, h > 0) = \infty$. The basic recursive step is the following version of the Principle of Optimality

$$W(v_0, v, h+1) = \min_{v_i} \left(W(v_0, v_i, h) + w(v_i, v)\right)$$

and the tree is updated as

$$T(v_0, h+1) = T(v_0) \cup \{v_i\}$$

Dijkstra algorithm

Given a weighted graph (G, w), the Dijkstra algorithm recursively constructs the tree $T(v_0, s)$ consisting of all vertices within a distance from v_0 not exceeding s. For any $v \in T(v_0, s)$, the distance $d(v_0, v)$ is defined. For s sufficiently large, $T(v_0, s)$ becomes the spanning tree of all geodesics emanating from v_0.

The algorithm is initialized as $T(v_0, 0) = \{v_0\}$, along with $d(v_0, v) = w(v_0, v)$, $\forall v \in E(v_0)$. Assume the tree $T(v_0, s)$ is known. The basic recursive step is is

$$s + \Delta s = \min_{v_i \notin T(v_0,s)} \left(\min_{v \in T(v_0,s)} \left(d(v_0, v) + w(v, v_i)\right) \right)$$

and the tree is updated as

$$T(v_0, s + \Delta s) = T(v_0, s) \cup \{v_i\}$$

Chapter 7

Gauss Theory of Surfaces and the *Theorema Egregium*

7.1 Surfaces

Let $\mathbb{E}^3$ be charted with coordinates (x, y, z) relative to the basis (f_x, f_y, f_z). A surface can be defined locally by the chart

$$\begin{array}{rcl} D & \to & \mathbb{E}^3 \\ (u,v) & \mapsto & \xi(u,v) \end{array}$$

where D is the diffeomorph of the open unit disk of $\mathbb{R}^2$. As such (u, v) are local coordinates of the surface. The atlas of a surface is the collection of charts (D_i, ξ_i) properly glued.

It is convenient to associate with the surface a moving frame defined as

$$E_u = \frac{\partial \xi}{\partial u}, \quad E_v = \frac{\partial \xi}{\partial v}, \quad E_w = E_u \times E_v$$

The span of the vectors E_u, E_v defines the tangent space $T_\xi S$. Clearly, E_w is the normal to the surface.

The inner product $\langle , \rangle$ in $\mathbb{E}^3$ induces a metric on the surface as follows:

$$\begin{aligned} ||d\xi||^2 &= \left\| \frac{\partial \xi}{\partial u} du + \frac{\partial \xi}{\partial v} dv \right\|^2 \\ &= \left\langle \frac{\partial \xi}{\partial u} du + \frac{\partial \xi}{\partial v}, \frac{\partial \xi}{\partial u} du + \frac{\partial \xi}{\partial v} \right\rangle \end{aligned}$$

The above leads to the so-called first quadratic form:

$$ds^2 = \begin{pmatrix} du & dv \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix} \begin{pmatrix} du \\ dv \end{pmatrix}$$

where,

$$E = \left\| \frac{\partial \xi}{\partial u} \right\|^2, \quad F = \left\langle \frac{\partial \xi}{\partial u}, \frac{\partial \xi}{\partial v} \right\rangle, \quad G = \left\| \frac{\partial \xi}{\partial v} \right\|^2$$

It is noted that E, F, G are functions of (u, v). As such, the first quadratic form is relevant to the intrinsic geometry of the surface, since it can be expressed entirely in terms of the local coordinates (u, v), which can be viewed as abstract coordinates, without reference to the encompassing space $\mathbb{E}^3$.

Another important metric quantity is the area. Consider in D a small parallelogram constructed on the vectors (du_1, dv_1), (du_2, dv_2). This in turn maps to a parallelogram constructed on the vectors

$$\begin{aligned} d\xi_1 &= \frac{\partial \xi}{\partial u} du_1 + \frac{\partial \xi}{\partial v} dv_1, \\ d\xi_2 &= \frac{\partial \xi}{\partial u} du_2 + \frac{\partial \xi}{\partial v} dv_2 \end{aligned}$$

in the tangent plane to the surface S. It is well known that in $\mathbb{E}^3$, the area of a parallelogram constructed with the above vectors is given as

$$dA^2 = \det \begin{pmatrix} \langle d\xi_1, d\xi_1 \rangle & \langle d\xi_1, d\xi_2 \rangle \\ \langle d\xi_2, d\xi_1 \rangle & \langle d\xi_2, d\xi_2 \rangle \end{pmatrix}$$

After a few manipulations, this yields,

$$dA = \sqrt{EG - F^2}(du_1 dv_2 - du_2 dv_1)$$

7.2 Curvature

Positive curvature versus negative curvature at a point on a surface is basically the issue as to whether the surface is on one side or on both sides of the tangent plane at that point.

Consider the point $\xi(0,0) \in S$ and the tangent plane $T_{\xi(0,0)}S$. The "height" of the tangent plane relative to the surface is

$$\left\langle (\xi(u,v) - \xi(0,0)), \frac{E_w(0,0)}{\|E_w(0,0)\|} \right\rangle$$

Clearly, the first order term of the expansion of the height function vanishes, so that the next term is the second order one:

$$\left\langle (\xi(u,v) - \xi(0,0)), \frac{E_w(0,0)}{\|E_w(0,0)\|} \right\rangle = \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} L & M \\ M & N \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + ...$$

In the above, L, M, N are the second order derivatives of $\langle \xi, E_w(0,0)/\|E_w(0,0)\| \rangle$. Recall that the norm of $E_w = E_u \times E_v$ is the area of the parallelogram constructed on E_u, E_v, that is, $\sqrt{EG - F^2}$. In anticipation of some tedious manipulation, L, M, N are defined in simplified notation as

$$L\sqrt{EG - F^2} = \langle \xi_{uu}, E_w \rangle, \quad M\sqrt{EG - F^2} = \langle \xi_{uv}, E_w \rangle, \quad N\sqrt{EG - F^2} = \langle \xi_{vv}, E_w \rangle$$

where the subscripts of ξ denote partial derivatives.

The second order term is called *second quadratic form.* Clearly, the sign of the eigenvalues of $\begin{pmatrix} L & M \\ M & N \end{pmatrix}$ dictates whether the surface is above, below, or on both sides of the tangent plane. Since the product of the eigenvalues is the determinant, it is clear that positive curvature means $LN - M^2 > 0$ while negative curvature means $LN - M^2 < 0$. We introduce the area as a normalization factor and define the Gauss (or sectional) curvature to be

$$\kappa = \frac{LN - M^2}{EG - F^2}$$

The above normalization also has the effect of making κ invariant under change of parameterization.

The famous *Theorema Egregium* of Gauss asserts that the curvature is intrinsic to the geometry of the surface and hence is expressible in terms of the E, F, G only. We proceed from

$$\kappa(EG - F^2)^2 = (\langle \xi_{uu}, (E_u \times E_v) \rangle)(\langle \xi_{vv}, (E_u \times E_v) \rangle) - (\langle \xi_{uv}, (E_u \times E_v) \rangle)^2$$

By elementary vector geometry, all of the inner products appearing in the right hand side of the above are expressible in term of determinants constructed on the x, y, z coordinates of the various vectors. Precisely,

$$(\langle \xi_{uu}, (E_u \times E_v) \rangle) = \det \begin{pmatrix} \xi_{uu,x} & \xi_{uu,y} & \xi_{uu,z} \\ \xi_{u,x} & \xi_{u,y} & \xi_{u,z} \\ \xi_{v,x} & \xi_{v,y} & \xi_{v,z} \end{pmatrix}$$

$$(\langle \xi_{vv}, (E_u \times E_v) \rangle) = \det \begin{pmatrix} \xi_{vv,x} & \xi_{vv,y} & \xi_{vv,z} \\ \xi_{u,x} & \xi_{u,y} & \xi_{u,z} \\ \xi_{v,x} & \xi_{v,y} & \xi_{v,z} \end{pmatrix}$$

$$(\langle \xi_{uv}, (E_u \times E_v) \rangle) = \det \begin{pmatrix} \xi_{uv,x} & \xi_{uv,y} & \xi_{uv,z} \\ \xi_{u,x} & \xi_{u,y} & \xi_{u,z} \\ \xi_{v,x} & \xi_{v,y} & \xi_{v,z} \end{pmatrix}$$

Clearly, the first term in the right hand side of $\kappa(EG - F^2)^2$ is the product of two determinants, hence the determinant of the product of the respective matrices. Next, since the determinant is not affected by matrix transposition, we have

$$\begin{aligned} & (\langle \xi_{uu}, (E_u \times E_v) \rangle)(\langle \xi_{vv}, (E_u \times E_v) \rangle) \\ = \quad & \det \left(\begin{pmatrix} \xi_{uu,x} & \xi_{uu,y} & \xi_{uu,z} \\ \xi_{u,x} & \xi_{u,y} & \xi_{u,z} \\ \xi_{v,x} & \xi_{v,y} & \xi_{v,z} \end{pmatrix} \begin{pmatrix} \xi_{vv,x} & \xi_{u,x} & \xi_{v,x} \\ \xi_{vv,y} & \xi_{u,y} & \xi_{v,y} \\ \xi_{vv,z} & \xi_{u,z} & \xi_{v,z} \end{pmatrix} \right) \\ = \quad & \det \begin{pmatrix} \langle \xi_{uu}, \xi_{vv} \rangle & \langle \xi_{uu}, \xi_u \rangle & \langle \xi_{uu}, \xi_v \rangle \\ \langle \xi_u, \xi_{vv} \rangle & E & F \\ \langle \xi_v, \xi_{vv} \rangle & F & G \end{pmatrix} \end{aligned}$$

Using exactly the same manipulation, it is found that

$$\begin{aligned}
&(\langle \xi_{uv}, (E_u \times E_v) \rangle)^2 \\
&= \det\left(\begin{pmatrix} \xi_{uv,x} & \xi_{uv,y} & \xi_{uv,z} \\ \xi_{u,x} & \xi_{u,y} & \xi_{u,z} \\ \xi_{v,x} & \xi_{v,y} & \xi_{v,z} \end{pmatrix} \begin{pmatrix} \xi_{uv,x} & \xi_{u,x} & \xi_{v,x} \\ \xi_{uv,y} & \xi_{u,y} & \xi_{v,y} \\ \xi_{uv,z} & \xi_{u,z} & \xi_{v,z} \end{pmatrix} \right) \\
&= \det \begin{pmatrix} \langle \xi_{uv}, \xi_{uv} \rangle & \langle \xi_{uv}, \xi_u \rangle & \langle \xi_{uv}, \xi_v \rangle \\ \langle \xi_u, \xi_{uv} \rangle & E & F \\ \langle \xi_v, \xi_{uv} \rangle & F & G \end{pmatrix}
\end{aligned}$$

Putting all of the pieces together, and after some further elementary matrix manipulation, it is found that the crucial curvature-related term becomes

$$\begin{aligned}
&\kappa(EG - F^2)^2 = \\
&\quad (\langle \xi_{uu}, \xi_{vv} \rangle - \langle \xi_{uv}, \xi_{uv} \rangle)(EG - F^2) \\
&+ \det \begin{pmatrix} 0 & \langle \xi_{uu}, \xi_u \rangle & \langle \xi_{uu}, \xi_v \rangle \\ \langle \xi_u, \xi_{vv} \rangle & E & F \\ \langle \xi_v, \xi_{vv} \rangle & F & G \end{pmatrix} \\
&- \det \begin{pmatrix} 0 & \langle \xi_{uv}, \xi_u \rangle & \langle \xi_{uv}, \xi_v \rangle \\ \langle \xi_u, \xi_{uv} \rangle & E & F \\ \langle \xi_v, \xi_{uv} \rangle & F & G \end{pmatrix}
\end{aligned}$$

The final twist is to show that all of the partial derivatives appearing in the right hand side of the above are easily expressible in terms of the first quadratic form. First, from $\langle \xi_u, \xi_u \rangle = E$, $\langle \xi_u, \xi_v \rangle = F$, $\langle \xi_v, \xi_v \rangle = G$, we obtain

$$\begin{aligned}
\langle \xi_{uu}, \xi_u \rangle &= \frac{1}{2}\frac{\partial E}{\partial u} \\
\langle \xi_{uu}, \xi_v \rangle &= \frac{\partial F}{\partial u} - \frac{1}{2}\frac{\partial E}{\partial v} \\
\langle \xi_u, \xi_{vv} \rangle &= \frac{\partial F}{\partial v} - \frac{1}{2}\frac{\partial G}{\partial u} \\
\langle \xi_v, \xi_{vv} \rangle &= \frac{1}{2}\frac{\partial G}{\partial v} \\
\langle \xi_{uv}, \xi_u \rangle &= \frac{1}{2}\frac{\partial E}{\partial v} \\
\langle \xi_{uv}, \xi_v \rangle &= \frac{1}{2}\frac{\partial G}{\partial u}
\end{aligned}$$

Differentiating the 2nd equation relative to v and the last equation relative to u yields, respectively,

$$\begin{aligned}
\langle \xi_{uuv}, \xi_v \rangle + \langle \xi_{uu}, \xi_{vv} \rangle &= \frac{\partial^2 F}{\partial v \partial u} - \frac{1}{2}\frac{\partial^2 E}{\partial v^2} \\
\langle \xi_{uuv}, \xi_v \rangle + \langle \xi_{uv}, \xi_{uv} \rangle &= \frac{1}{2}\frac{\partial^2 G}{\partial u^2}
\end{aligned}$$

Subtracting the second from the first yields

$$\langle \xi_{uu}, \xi_{vv} \rangle - \langle \xi_{uv}, \xi_{uv} \rangle = -\frac{1}{2}\frac{\partial^2 G}{\partial u^2} + \frac{\partial^2 F}{\partial u \partial v} - \frac{1}{2}\frac{\partial^2 E}{\partial v^2}$$

Clearly, our objective of writing κ as a function of the matrix of the first quadratic form has been achieved.

In order to find a compact expression of κ in terms of E, F, G it is convenient to assume that the moving frame E_u, E_v is orthogonal. In this case, we obtain

$$\kappa = -\frac{1}{2}\frac{1}{\sqrt{EG}}\left(\frac{\partial}{\partial v}\frac{\partial E/\partial v}{\sqrt{EG}} + \frac{\partial}{\partial u}\frac{\partial G/\partial u}{\sqrt{EG}}\right)$$

7.3 Gauss-Bonnet theorem

Let $C : [0, 2\pi) \ni t \mapsto c(t) \in S$ be a piecewise differentiable curve on S. Let V be a unit vector tangent to S and displaced parallel to itself along the curve. The Gauss-Bonnet theorem asserts that the parallel displacement of $V(0)$ along C results in a $V(2\pi)$ not aligned with the initial vector; more precisely, the mismatch angle $\angle(V(0), V(2\pi))$ is equal to the integral of the curvature along the surface bounded by C.

As a simple illustration, consider a vector tangent to an arc of great circle of the unit sphere. In its parallel displacement along the entire arc of great circle, the vector remains tangent to the arc of great circle and hence $\angle(V(0), V(2\pi)) = 2\pi$. Since for a unit sphere $\kappa = 1$, the integral of the curvature along the half-sphere is clearly the area of the half-sphere, that is 2π, hence illustrating the Gauss-Bonnet theorem.

To prove the theorem, we attach a moving frame E_u, E_v to the curve C, so that the frame has vanishing index along C. The latter means that neither E_u nor E_v makes a complete 2π rotation along the curve. Let the normal vector be the cross product $E_u \times E_v$. Let $\theta = \angle(E_u(c(t)), V(c(t)))$ be the angle by which E_u has to rotate to align itself with V. Since the frame has vanishing index, it follows that

$$\angle(V(0), V(2\pi)) = \oint d\theta$$

The problem is to show that the right hand side integral is related to the curvature on the surface bounded by C.

From here on, it is assumed that $E_u \perp E_v$. The fact that V has unit norm yields the parameterized expression:

$$V = \frac{\cos\theta}{\sqrt{E}}E_u + \frac{\sin\theta}{\sqrt{G}}E_v$$

Next,

$$\begin{aligned}\dot{V} &= -\frac{\sin\theta}{\sqrt{E}}\dot{\theta}E_u + \frac{\cos\theta}{\sqrt{G}}\dot{\theta}E_v \\ &\quad +\frac{\cos\theta}{\sqrt{E}}\dot{E}_u + \frac{\sin\theta}{\sqrt{G}}\dot{E}_v \\ &\quad -\frac{1}{2}\frac{\cos\theta}{E^{3/2}}\dot{E}E_u - \frac{1}{2}\frac{\sin\theta}{G^{3/2}}\dot{G}E_v\end{aligned}$$

Next, we require $V(t)$ to be a parallel field, that is, we require $\dot{V}$ to have an E_w component only, that is, $\langle \dot{V}, E_u\rangle = 0$ and $\langle \dot{V}, E_v\rangle = 0$. Taking the inner product of the above with E_u and observing that $\dot{E} = 2\langle \dot{E}_u, E_u\rangle$ yields

$$\sin\theta\left(-\sqrt{E}\dot{\theta} + \frac{1}{\sqrt{G}}\langle \dot{E}_v, E_u\rangle\right) = 0$$

Likewise, taking the scalar product with E_v and observing that $\dot{G} = 2\langle E_v, \dot{E}_v\rangle$ yields

$$\cos\theta\left(\sqrt{G}\dot{\theta} + \frac{1}{\sqrt{E}}\langle \dot{E}_u, E_v\rangle\right) = 0$$

If we further observe that $\langle \dot{E}_v, E_u\rangle + \langle E_v, \dot{E}_u\rangle = 0$, the above two equations are consistent with solution

$$\dot{\theta} = \frac{1}{\sqrt{EG}}\langle E_u, \dot{E}_v\rangle$$

Next, a bit of vector calculus shows that

$$\begin{aligned}\langle E_u, \dot{E}_v\rangle &= \langle E_u, \frac{\partial E_v}{\partial u}\rangle\dot{u} + \langle E_u, \frac{\partial E_v}{\partial v}\rangle\dot{v} \\ &= \langle E_u, \frac{\partial E_v}{\partial u}\rangle\dot{u} - \langle\frac{\partial E_u}{\partial v}, E_v\rangle\dot{v} \\ &= \langle E_u, \frac{\partial E_u}{\partial v}\rangle\dot{u} - \langle\frac{\partial E_v}{\partial u}, E_v\rangle\dot{v} \\ &= \frac{1}{2}\left(\frac{\partial E}{\partial v}\dot{u} - \frac{\partial G}{\partial u}\dot{v}\right)\end{aligned}$$

Finally, we apply Stokes' theorem to get

$$\begin{aligned}\oint \dot{\theta}dt &= \oint \frac{1}{2\sqrt{EG}}\left(\frac{\partial E}{\partial v}du - \frac{\partial G}{\partial u}dv\right) \\ &= -\frac{1}{2}\int\int\left(\frac{\partial}{\partial u}\frac{G'_u}{\sqrt{EG}} + \frac{\partial}{\partial v}\frac{E'_v}{\sqrt{EG}}\right)dudv \\ &= \int\int K\sqrt{EG}dudv \\ &= \int\int KdA\end{aligned}$$

Chapter 8

Riemannian Geometry

This chapter develops formal Riemannian geometry. In particular, geodesics, which were defined to be length minimizing paths in the previous chapter, are here redefined more formally as curves that have their tangent fields parallel to themselves.

8.1 Vector fields and Lie bracket

8.2 Levi-Civita model

8.2.1 kinematic interpretation

Consider a particle moving along the C^1 motion $x(t)$ on a C^1 surface S embedded in $\mathbb{E}^3$. By most elementary kinematics, the velocity, viz.,

$$V(x(t)) = \frac{dx(t)}{dt}$$

is in the tangent plane $T_{x(t)}S$. The acceleration $\frac{dV(x(t))}{dt}$, however, will not in general lie in the tangent plane. While this is not a matter of significant importance as long as the surface is embedded in $\mathbb{E}^3$, it becomes a conceptual problem if we want to define an "acceleration" in the context of the intrinsic geometry of the surface. To define an acceleration that remains within the tangent plane, an acceleration referred to as *tangential* in elementary kinematics, we define the *covariant*, rather than absolute, derivative as

$$\frac{DV(x(t))}{dt} := P_{T_{x(t)}S} \frac{dV(x(t))}{dt}$$

where $P_{T_{x(t)}S}$ denotes the orthogonal projection onto the tangent plane $T_{x(t)}S$.

8.2.2 formal definition

We now proceed a bit more formally by removing the reference to kinematics. In this context, we are given a tangent field V, that is, an assignment $S \ni x \mapsto V(x) \in T_x S$, along which we want to differentiate the C^1 vector field W. The covariant, or Levi-Civita, derivative of W along V is the vector field defined as

$$\nabla_V W(x_0) = P_{T_{x(t_0)}S} \left. \frac{dW(x(t))}{dt} \right|_{t=t_0}$$

where $x(t)$ is the integral curve of V passing through x_0, that is, the solution to $\dot{x}(t) = V(x(t))$ subject to $x(t_0) = x_0$. $\nabla_V W$ is sometimes rewritten, more informally, as $\frac{DW}{dt}$, when the tangent field along with the differential is taken is obvious from the context.

The Levi-Civita derivative enjoys the following arch typical properties:

Theorem 24 *The Levi-Civita derivative satisfies the following properties:*

1. *linearity:*
$$\nabla_{\alpha_1 V_1 + \alpha_2 V_2} W = \alpha_1 \nabla_{V_1} W + \alpha_2 \nabla_{V_2} W$$

2. *product rule:*
$$\nabla_V \alpha W = \alpha \nabla_V W + \frac{d\alpha(x(t))}{dt} W$$

3. *symmetry: If E_1, E_2 is a tangent frame,*
$$\nabla_{E_i} E_j = \nabla_{E_j} E_i$$

4. *compatibility with the metric:*
$$\frac{d}{dt}\langle W, Z\rangle = \langle \nabla_V W, Z\rangle + \langle W, \nabla_V Z\rangle$$

Proof. Only the symmetry property is nontrivial. Let (F_1, F_2, F_3) be an orthonormal reference frame of $\mathbb{E}^3$. Clearly, there exists a C^0 coefficient matrix A such that

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = A \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}$$

By linearity of the covariant derivative, we have

$$\begin{pmatrix} \nabla_{E_1} \\ \nabla_{E_2} \end{pmatrix} = A \begin{pmatrix} \nabla_{F_1} \\ \nabla_{F_2} \\ \nabla_{F_3} \end{pmatrix}$$

Combining the above two, we get

$$\begin{aligned}\begin{pmatrix} \nabla_{E_1} \\ \nabla_{E_2} \end{pmatrix} \begin{pmatrix} E_1 & E_2 \end{pmatrix} &= A \begin{pmatrix} \nabla_{F_1} \\ \nabla_{F_2} \\ \nabla_{F_3} \end{pmatrix} \begin{pmatrix} F_1 & F_2 & F_3 \end{pmatrix} A^T \\ &= AIA^T \\ &= AA^T\end{aligned}$$

and the connection is symmetric. ■

8.2.3 parallel displacement

If $W_0 \in T_{x(t_0)}S$, the parallel displacement of W along $x(t)$ in the ambient space $\mathbb{R}^3$ is the solution of the differential equation $\frac{dW(x(t))}{dt} = 0$ subject to the initial condition $W(x(t_0)) = W_0$. This solution will in general take W outside the tangent space. To remain within the confines of the intrinsic geometry of the surface, we define the Levi-Civita parallel displacement of W along $x(t)$ via the differential equation

$$\frac{DW(x(t)}{dt} = \nabla_{\dot{x}} W(x(t)) = 0$$

subject to the intial condition

$$W(x(t_0)) = W_0$$

This time, the parallel displacement will result in tangent vectors. Given $W_0 \in T_{x(t_0)}S$, the parallel displacement will provide a $W_1 \in T_{x(t_1)}S$. This yields a unique assignment $T_{x(t_0)}S \ni W_0 \mapsto W_1 \in T_{x(t_1)}S$, the *Levi-Civita connection.*

With this concept the Levi-Civita derivative can be redefined as

$$\nabla_V W(x_0) = \lim_{\tau \downarrow 0} \frac{W_0(t_0 + \tau) - W(t_0)}{\tau}$$

where $W_0(t_0 + \tau)$ is the parallel displacement of $W(t_0 + \tau)$ to $T_{x(t_0)}S$.

8.2.4 arc length minimization

Probably the most important motivation of the concept of covariant derivative is that it is a convenient way to formulate arc length optimality. Define inn $\mathbb{R}^3$ the surface $S = \{x \in \mathbb{R}^3 : \langle x|x\rangle = c\}$, where $\langle\cdot|\cdot\rangle$ is a possibly nonsign definite quadratic form and c is a constant. It is well known that the metric $\langle dx|dx\rangle$ on $\mathbb{R}^3$ induces a metric ds^2 on S. As such, given two points $a, b \in S$, there arises the problem of finding a rectifiable curve along which $\int_a^b ds$ is minimum.

Theorem 25 *The arc-length parameterized curve*

$$\begin{aligned}\gamma : [0, d(a,b)] &\to S \\ s &\mapsto \gamma(s)\end{aligned}$$

is a geodesic iff

$$\nabla_{\dot\gamma}\dot\gamma := \frac{D\dot\gamma(s)}{ds} = P_{T_{\gamma(s)}S}\ddot\gamma = 0$$

Proof. We treat this problem as a constraint optimization with Lagrange multiplier:

$$L = \sqrt{\sum_i \dot\gamma_i^2} + \lambda\langle\gamma|\gamma\rangle$$

Writing the Euler-Lagrange equations relative to the generalized coordinates γ_i yields,

$$\begin{aligned}\ddot\gamma_i &= \lambda\langle\frac{\partial\gamma}{\partial\gamma_i}|\gamma\rangle\\ &= \lambda\langle E_i|\gamma\rangle\end{aligned}$$

It follows that $\sum_i \ddot\gamma_i\dot\gamma_i = \lambda\langle\dot\gamma|\gamma\rangle = 0$, where the last equality stems from the fact that $\frac{d}{ds}\langle\gamma|\gamma\rangle = \frac{d}{ds}c = 0$. But $\dot\gamma \in T_{\gamma(s)}S$. Hence $P_{T_{\gamma(s)}S}\ddot\gamma = 0$. ■

8.3 covariant derivative

Here we redefine $\nabla_V W$ in a manner totally independent of any embedding of the manifold in a Euclidean space. The idea is to develop an intrinsic definition that enforces linearity and product rule:

$$\begin{aligned}\nabla_V W &= \nabla_{v^iE_i} w^jE_j\\ &= v^i\nabla_{E_i} w^jE_j\\ &= v^i\left(w^j\nabla_{E_i}E_j + \frac{\partial w^j}{\partial\xi^i}E_j\right)\end{aligned}$$

It follows from the above that linearity and product rule will be satisfied no matter how $\nabla_{E_i}E_j$ is defined. Clearly, the latter is a vector field and as such must be expressible in terms of basis vectors,

$$\nabla_{E_i}E_j = \Gamma_{ij}^k E_k$$

so that the covariant derivative is uniquely defined by the coefficients Γ_{ij}^k, called Christoffel symbols.

In principle, the Christoffel symbols could be chosen arbitrarily. However, this could lead to strange geometries. It is therefore desirable to mimic the Levi-Civita model. The connection is said to be symmetric if $\Gamma_{ij}^k = \Gamma_{ji}^k$; it is said to be compatible with the metric iff the same relation as for the Levi-Civita model holds.

Theorem 26 *There exists a unique connection, called Levi-Civita connection, that is symmetric and compatible with the metric and it is given by*

$$\Gamma_{ij}^k = \frac{1}{2}\left(\frac{\partial g_{jl}}{\partial\xi^i} + \frac{\partial g_{il}}{\partial\xi^j} - \frac{\partial g_{ij}}{\partial\xi^l}\right)g^{lk}$$

Proof. Using the compatibility between the covariant derivative and the metric,

$$\nabla_{E_k}\langle E_i, E_j\rangle = \langle \nabla_{E_k} E_i, E_j\rangle + \langle E_i, \nabla_{E_k} E_j\rangle$$

along with the symmetry of the fundamental tensor, we get

$$\frac{\partial g_{ij}}{\partial \xi^k} = \Gamma^l_{ki} g_{jl} + \Gamma^l_{kj} g_{il} \tag{8.1}$$

By permutation of indices, we find

$$\frac{\partial g_{ki}}{\partial \xi^j} = \Gamma^l_{jk} g_{il} + \Gamma^l_{ji} g_{kl} \tag{8.2}$$

$$\frac{\partial g_{kj}}{\partial \xi^i} = \Gamma^l_{ik} g_{jl} + \Gamma^l_{ij} g_{kl} \tag{8.3}$$

Adding 8.2, 8.3 and subtracting 8.1, and using the symmetry of the Γ's, we get

$$\frac{1}{2}\left(\frac{\partial g_{kj}}{\partial \xi^i} + \frac{\partial g_{ki}}{\partial \xi^j} - \frac{\partial g_{ij}}{\partial \xi^k}\right) = \Gamma^l_{ij} g_{kl}$$

Multiplying both sides by g^{kn}, summing over k, and after some elementary indices manipulation, we obtain the result. ■

Closely associated with the concept of covariant derivative is the concept of parallel displacement of a vector in the tangent space. Let $m(t)$ be the integral curve of a vector field V passing through m_0. Let $V_0 \in T_{m_0}M$. Then V is said to be the parallel displacement of V_0 along the curve m iff $\nabla_V W = 0$.

Finally, consider the case of a curve such that its velocity along the curve is the parallel displacement of itself, that is,

$$\nabla_{\dot{c}}\dot{c} = 0$$

Consider the above in a local coordinate patch ξ^k charted with moving frame E_i, that is, $c = c^i E_i$, in which case we obtain

$$\begin{aligned}
\nabla_{v^j E_j} v^i E_i &= v^j \nabla_{E_j} v^i E_i \\
&= v^j (v^i \nabla_{E_j} E_i + \frac{\partial v^i}{\partial \xi^j} E_i) \\
&= v^j (v^i \Gamma^k_{ji} E_k + \frac{\partial v^k}{\partial \xi^j} E_k) \\
&= (\frac{dm^j}{ds}\frac{dm^i}{ds}\Gamma^k_{ji} + \frac{dv^k}{\partial \xi^j}\frac{dm^j}{ds})E_k \\
&= (\frac{dm^j}{ds}\frac{dm^i}{ds}\Gamma^k_{ji} + \frac{dv^k}{ds})E_k \\
&= (\frac{dm^j}{ds}\frac{dm^i}{ds}\Gamma^k_{ji} + \frac{d^2 m^k}{ds^2})E_k \\
&= 0
\end{aligned}$$

It follows that

$$\frac{d^2m^k}{ds^2} + \frac{dm^j}{ds}\frac{dm^i}{ds}\Gamma^k_{ji} = 0$$

which is exactly the equation of length minimization. Therefore, a curve that has its tangent vector parallel to itself is a geodesic.

Given that $\gamma : [a, b] \to M$ is a geodesic and that the connection is compatible with the metric, then

$$\begin{aligned} \frac{d}{dt}\left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle &= 2\left\langle \frac{D}{dt}\frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle = 0, \\ \left| \frac{d\gamma}{dt} \right| &= c \neq 0, \\ L\left(\gamma\right) &= c\left(b - a\right). \end{aligned} \tag{8.4}$$

8.4 curvature

Given two vector fields X, Y, themselves defining a section through the tangent bundle, the curvature of the Riemannian manifold, as seen from the section defined by X, Y, is defined as the operator

$$R(X, Y) = \nabla_Y\nabla_X - \nabla_X\nabla_Y + \nabla_{[X,Y]}$$

acting on vector arbitrary vectors fields. In a certain sense, it is the commutator of the covariant derivatives along X and Y, and the Lie bracket term ensures linearity of $R(X, Y)$ as a differential operator and linearity of $R(X, Y)$ relative to the arguments X, Y.

More precisely, observe the following:

$$\begin{aligned} R(X,Y)fZ &= (\nabla_Y\nabla_X - \nabla_X\nabla_Y + \nabla_{[X,Y]})fZ \\ &= \nabla_Y\nabla_X fZ - \nabla_X\nabla_Y fZ + \nabla_{[X,Y]}fZ \\ &= \nabla_Y(f\nabla_X Z + X(f)Z) - \nabla_X(f\nabla_Y Z + Y(f)Z) + ([X,Y](f)Z + f\nabla_{[X,Y]}Z) \\ &= Y(f)\nabla_X Z + f\nabla_Y\nabla_X Z + X(f)\nabla_Y Z + YX(f)Z \\ &\quad -X(f)\nabla_Y Z - f\nabla_X\nabla_Y Z - XY(f)Z - Y(f)\nabla_X Z \\ &\quad +[X,Y](f)Z + f\nabla_{[X,Y]}Z \\ &= f(\nabla_Y\nabla_X Z - \nabla_X\nabla_Y Z + \nabla_{[X,Y]})Z \\ &= fR(X,Y)Z \end{aligned}$$

Regarding the second linearity claim, consider the following string:

$$
\begin{aligned}
R(fX_1+gX_2,Y) &= \nabla_Y\nabla_{fX_1+gX_2} - \nabla_{fX_1+gX_2}\nabla_Y + \nabla_{[fX_1+gX_2,Y]} \\
&= \nabla_Y f\nabla_{X_1} + \nabla_Y g\nabla_{X_2} - f\nabla_{X_1}\nabla_Y - g\nabla_{X_2}\nabla_Y + \nabla_{[fX_1,Y]} + \nabla_{[gX_2,Y]} \\
&= Y(f)\nabla_{X_1} + f\nabla_Y\nabla_{X_1} + Y(g)\nabla_{X_2} + g\nabla_Y\nabla_{X_2} \\
&\quad -f\nabla_{X_1}\nabla_Y - g\nabla_{X_2}\nabla_Y \\
&\quad +f\nabla_{[X,Y]} - Y(f)\nabla_{X_1} + g\nabla_{[X_2,Y]} - Y(g)\nabla_{X_2} \\
&= f(\nabla_Y\nabla_{X_1} - \nabla_{X_1}\nabla_Y + \nabla_{[X_1,Y]}) + g(\nabla_Y\nabla_{X_2} - \nabla_{X_2}\nabla_Y + \nabla_{[X_2,Y]}) \\
&= fR(X_1,Y) + gR(X_2,Y)
\end{aligned}
$$

Since $R(X,Y)Z$ is linearity relative to X, Y, and Z, it can be expressed in terms of the contravariant components of X, Y and Z in some mobile reference frame, say $\{E_i\}$. Write

$$X = x^iE_i, \quad Y = y^jE_j, \quad Z = z^kE_k$$

then, by linearity,

$$R(X,Y)Z = R(x^iE_i, y^jE_j)z^kE_k = x^iy^jz^kR(E_i,E_j)E_k$$

By definition, $R(E_i,E_j)E_k$ is a vector field and as such it must be expressible in terms of the mobile frame,

$$R(E_i,E_j)E_k = R^l_{ijk}E_l$$

The system of quantities R^l_{ijk} is called curvature tensor. The curvature tensor can itself be expressed in terms of the covariant derivative:

$$
\begin{aligned}
R(E_i,E_j)E_k &= \nabla_{E_j}\nabla_{E_i}E_k - \nabla_{E_i}\nabla_{E_j}E_k \\
&= \nabla_{E_j}\Gamma^l_{ik}E_l - \nabla_{E_i}\Gamma^l + jkE_l \\
&= \Gamma^l_{ik}\Gamma^s_{jl}E_s + \frac{\partial\Gamma^l_{ik}}{\partial x^j}E_l - \Gamma^l_{jk}\Gamma^s_{il}E_s - \frac{\partial\Gamma^l_{jk}}{\partial x^i}E_l \\
&= (\Gamma^s_{ik}\Gamma^l_{jk} - \Gamma^s_{jk}\Gamma^l_{is} + \frac{\partial\Gamma^l_{ik}}{\partial x^j} - \frac{\partial\Gamma^l_{jk}}{\partial x^i})E_l
\end{aligned}
$$

It follows that

$$R^l_{ijk} = \Gamma^s_{ik}\Gamma^l_{jk} - \Gamma^s_{jk}\Gamma^l_{is} + \frac{\partial\Gamma^l_{ik}}{\partial x^j} - \frac{\partial\Gamma^l_{jk}}{\partial x^i}$$

8.5 sectional curvature

It is through the concept of *sectional* curvature that it is possible to give the abstract concept of curvature some intuitive geometric interpretation. Given two tangent vectors $X, Y \in T_{m_0}M$, the sectional curvature relative to the surface element induced by X, Y is defined as

$$K(X,Y) = \frac{\langle R(X,Y)X, Y\rangle}{||X||^2||Y||^2 - \langle X,Y\rangle^2}$$

Observe that, by the Cauchy-Schwarz inequality, the denominator is positive; in fact, this denominator is, up to the fourth order, the square of the area of the parallelogram constructed on X, Y. The following is a key result:

Theorem 27 *Let X, Y be two vector fields defined in a neighborhood of $p \in M$ and consider a small parallelogram constructed on the vectors $Xdt, Yd\tau$ at the point $p \in M$. Let $W \in T_pM$ be a tangent vector, which is displaced parallel to itself counterclockwise along the parallelogram to come back as $\tilde{W} \in T_pM$. Then, up to the second order,*

$$\tilde{W} = W + R(Xdt, Yd\tau)W$$

Proof. The result of the parallel displacement of the vector W along the integral curve of X for an amount of time dt results in

$$w^k - \Gamma^k_{ij} w^j x^i dt$$

The vector field Y, displaced parallel to itself along the integral curve of X for an amount of time dt results in

$$y^k - \Gamma^k_{ij} y^j x^i dt$$

The parallel displacement of the new W along the new Y results in

$$\left(w^k - \Gamma^k_{ij} w^j x^i dt\right) - \Gamma^k_{ij}(p + Xdt)\left(w^j - \Gamma^j_{ml} w^l x^m dt\right)\left(y^i - \Gamma^i_{pq} y^p x^q dt\right) d\tau$$

After a few manipulation, the above reduces to

$$\tilde{W}_{YX} = w^s - \Gamma^s_{ij} w^j x^i dt - \Gamma^s_{ij} w^j y^i dt + \left(\Gamma^l_{ik}\Gamma^s_{jl} + \Gamma^l_{ji}\Gamma^s_{kl} - \frac{\partial \Gamma^s_{kj}}{\partial \xi^i}\right) w^k x^i y^j dt d\tau$$

Now, we take the vector W and first displace it parallel to itself along $Yd\tau$ and then displace the resulting vector along the result of the parallel displacement of X. Instead of using the same procedure as above, we simply interchange the role of x, y and we obtain

$$\tilde{W}_{XY} = w^s - \Gamma^s_{ij} w^j y^i d\tau - \Gamma^s_{ij} w^j x^i dt + \left(\Gamma^s_{jk}\Gamma^s_{il} + \Gamma^l_{ij}\Gamma^s_{kl} - \frac{\partial \Gamma^s_{ki}}{\partial \xi^j}\right) w^k x^i y^j dt d\tau$$

Clearly,

$$\begin{aligned} \tilde{W} - W &= W_{YX} - W_{XY} \\ &= \left(\Gamma^l_{ik}\Gamma^s_{jl} - \Gamma^l_{jk}\Gamma^s_{il} + \frac{\partial \Gamma^s_{ik}}{\partial \xi^j} - \frac{\partial \Gamma^s_{jk}}{\partial \xi^i}\right) w^k x^i dt y^j d\tau e_s \\ &= R(Xdt, Yd\tau)W \end{aligned}$$

■

Theorem 28 *Let $X, Y \in T_{m_0}M$ and let S be the surface element consisting of all geodesics γ such that $\gamma(0) = m_0$ and such that $\dot{\gamma}$ lies in the linear span of X, Y in $T_{m_0}M$. Let c be a closed curve in S, let W be a tangent vector at a point $c_0 \in c$, and let $\tilde{W}$ be the tangent vector obtained by parallel displacement of W along the curve from c_0 back to c_0. Then*

$$\tilde{W} = W + \frac{area(c)}{area(parallelogram)} R(X,Y)W$$

Proof. See [13, p. 233]. ■

From the above, we can derive the Gauss-Bonnet theorem. Consider the section generated by X, Y and let $\widetilde{X}$ be the result of the parallel displacement of X along a closed curve embedded in the surface element. From the above,

$$\widetilde{X} = X + A(c)\frac{R(X,Y)X}{\sqrt{||X||^2||Y||^2 - \langle X, Y\rangle^2}}$$

and taking the inner product with Y yields

$$\langle \widetilde{X}, Y\rangle = \langle X, Y\rangle + A(c)\frac{\langle R(X,Y)X, Y\rangle}{\sqrt{||X||^2||Y||^2 - \langle X, Y\rangle^2}}$$

If $\cos(\widetilde{X}, Y)$ denotes the cosine of the angle between the vectors $\widetilde{X}$ and X, with a similar definition for $\cos(X, Y)$, we get

$$||\widetilde{X}||||Y|| \cos(\widetilde{X}, Y) = ||X||||Y|| \cos(X, Y) + A(c)\frac{\langle R(X,Y)X, Y\rangle}{\sqrt{||X||^2||Y||^2 - \langle X, Y\rangle^2}}$$

Agreeing that $||\widetilde{X}|| \approx ||X||$, we get

$$\frac{||X||||Y||}{\sqrt{||X||^2||Y||^2 - \langle X, Y\rangle^2}} \Big(\cos(\widetilde{X}, Y) - \cos(X, Y)\Big) = A(c)K(X,Y)$$

and upon some elementary inner product manipulation we get

$$\frac{\cos(\widetilde{X}, Y) - \cos(X, Y)}{\sin(X, Y)} = A(c)K(X,Y)$$

If we agree that the angle (X, Y) is much bigger than the angle $(\widetilde{X}, X)$, we have

$$\begin{aligned} &\cos(\widetilde{X}, X) - \cos(X, Y) \\ &= \quad 2\sin\frac{(\widetilde{X}, Y) + (X, Y)}{2} \sin\frac{(\widetilde{X}, Y) - (X, Y)}{2} \\ &\approx \quad 2\sin(X, Y)\sin\frac{(\widetilde{X}, X)}{2} \\ &\approx \quad \sin(X, Y)(\widetilde{X}, X) \end{aligned}$$

Therefore, we get,

$$(\widetilde{X}, X) = A(c)K(X,Y)$$

The above is a formula "in the small." "In the large," we get the celebrated Gauss-Bonnet theorem:

$$\int\int_S KdS = (\widetilde{X}, X)$$

where S is a surface bounded by a closed curve c X is a tangent vector in $T_{c_0}S$ and $\widetilde{X}$ is the result of the parallel displacement of X along the closed curve c, from c_0 back to c_0.

Definition 49 *Given that $\gamma : [0,l] \to M$ is a curve parameterized by arc-length in M, then the covariant derivative*

$$\frac{D}{dt}\left(\frac{d\gamma}{dt}\right) = \kappa_g \tag{8.5}$$

of $\dot{\gamma}$ along the curve γ at p is called the geodesic curvature of γ at p.

Theorem 29 *Given that M is an oriented 2-dimensional Riemannian manifold with sectional curvature κ and volume element dA, and $N \subset M$ is a polygon which is diffeomorphic to a subset of $\mathbb{R}^2$ such that ∂N has vertices at $t_1, \ldots, t_n$ with discontinuity $\theta_1, \ldots, \theta_n$ and κ_g is its geodesic curvature with arc length ds, then*

$$\int\limits_N \kappa dA + \int_{\partial N} \kappa_g ds + \sum_{i=1}^{n} \theta_i = 2\pi. \tag{8.6}$$

Proof. See [31, Sec. 4-5]. ■

Theorem 30 (Global Gauss-Bonnet) *Given that M is an oriented 2-dimensional Riemannian manifold with sectional curvature κ and volume element dA, and $R \subset M$ is a regular region of M such that ∂R has vertices at $t_1, \ldots, t_n$ with discontinuity $\theta_1, \ldots, \theta_n$ and κ_g is its geodesic curvature with volume element ds, then*

$$\int\limits_R \kappa dA + \int_{\partial R} \kappa_g ds + \sum_{i=1}^{n} \theta_i = 2\pi\chi(R). \tag{8.7}$$

Proof. See [31, Sec. 4-5]. ■

Corollary 2 *Given that M is an orientable compact surface, then*

$$\int_M \kappa dA = 2\pi\chi(M). \tag{8.8}$$

Proof. See [31, Sec. 4-5, Cor. 2]. ■

Corollary 3 *Given a geodesic triangle $\triangle(A,B,C)$ with interior angle α, β, γ in a Riemannian manifold with constant sectional curvature κ, then*

1. if $\kappa < 0$, then area of $\triangle(A, B, C) = \frac{\pi-(\alpha+\beta+\gamma)}{-\kappa}$;

2. if $\kappa = 0$, then $\alpha + \beta + \gamma = \pi$;

3. if $\kappa > 0$, then area of $\triangle(A, B, C) = \frac{(\alpha+\beta+\gamma)-\pi}{\kappa}$.

Proof. This follows from the Local Gauss-Bonnet Theorem with $\theta_1 = \pi - \alpha$, $\theta_2 = \pi - \beta$, $\theta_3 = \pi - \gamma$. Thus $\int \kappa dA = -\pi + (\alpha + \beta + \gamma)$. The result follows easily. ■

8.6 Gauss curvature of surfaces

Here we take a little pause in the development of Riemannian geometry and take a step backward to the Gauss theory of surfaces. The first objective in doing so is purely illustrative: to make the concept of sectional curvature more palatable. More important, however, is the fact that, since any graph can be embedded in a surface, a thorough understanding of the curvature of surfaces is warranted.

Let $\mathbb{E}^3$ be the usual Euclidean space endowed with the inner product $\langle , \rangle$. A surface can be defined locally by the chart

$$\begin{array}{ccc} D & \to & \mathbb{E}^3 \\ (u,v) & \mapsto & \xi(u,v) \end{array}$$

where D is the diffeomorph of the open unit disk of $\mathbb{R}^2$. As such (u, v) are local coordinates of the surface. The atlas of the surface is the collection of charts (D_i, ξ_i) properly glued. The vectors $E_u = \frac{\partial \xi}{\partial u}$, $E_v = \frac{\partial \xi}{\partial v}$ define a basis of the tangent space to the surface. The inner product of $\mathbb{E}^3$ induces a metric $ds^2 = g_{uu}du^2 + 2g_{uv}dudv + g_{vv}dv^2$ where $g_{uu} = \|E_u\|^2$, $g_{vv} = \|E_v\|^2$ and $g_{uv} = \langle E_u, E_v \rangle$. The cross product $E_u \times E_v$ defines the normal vector. Consider the point $\xi(0, 0) \in S$ and the tangent plane $T_{\xi(0,0)}S$. The "height" of the surface relative to the tangent plane is

$$\left\langle (\xi(u,v) - \xi(0,0)), \underbrace{\frac{E_u(0,0) \times E_v(0,0)}{\|E_u(0,0) \times E_v(0,0)\|}}_{N(0,0)} \right\rangle$$

Clearly, the first order term of the expansion of the height function vanishes, so that the next term is the second order one:

$$\langle (\xi(u,v) - \xi(0,0)), N(0,0) \rangle = \frac{1}{2} \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} L & M \\ M & N \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + o\left(\left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|^2 \right)$$

In the above, L, M, N are the second order derivatives of $\langle \xi(u,v) - \xi(0,0), N(0,0) \rangle$ evaluated at $(0, 0)$. The second order term is called *second quadratic form.* We introduce the square of the area of the parallelogram constructed on E_u, E_v as a normalization factor and define the Gauss curvature of the surface at the point p to be

$$\kappa = \frac{LN - M^2}{\|E_u\|^2\|E_v\|^2 - \langle E_u, E_v\rangle^2} \tag{8.9}$$

The above normalization has the effect of making κ invariant under change of parameterization.

Clearly, the sign definiteness of $\begin{pmatrix} L & M \\ M & N \end{pmatrix}$ dictates whether the surface is above, below, or on both sides of the tangent plane. This leads to probably the most intuitive interpretation of the Gauss curvature: $\kappa > 0$ means that the surface is locally on one side of the tangent plane (e.g., sphere); $\kappa < 0$ means that the surface is locally on both sides of the tangent plane (e.g., saddle); and $\kappa = 0$ means that the surface along the image of the eigenvector corresponding to $\lambda = 0$ coincides up to second order with the tangent plane (e.g., cylinder).

To provide a quantitative interpretation of the Gauss curvature, it is convenient to introduce the shape operator $S : TS \to TS$ defined by linear extension from

$$S(E_u) = -\frac{\partial}{\partial u}N, \quad S(E_v) = -\frac{\partial}{\partial v}N$$

where N is the *unit* normal vector. Since $\frac{\partial}{\partial u}\langle N, N\rangle = -2\langle S(E_u), N\rangle = 0$, it follows that $S(E_u)$, and $S(E_v)$ by the same reasoning, are in the tangent plane. The connection between the shape operator and the second quadratic form is easily seen to be

$$\begin{pmatrix} \langle S(E_u), E_u\rangle & \langle S(E_u), E_v\rangle \\ \langle S(E_v), E_u\rangle & \langle S(E_v), E_v\rangle \end{pmatrix} = \begin{pmatrix} L & M \\ M & N \end{pmatrix} \tag{8.10}$$

Now, consider in the tangent plane the extreme values of

$$\frac{\langle S(uE_u + vE_v), uE_u + vE_v\rangle}{\|uE_u + vE_v\|^2}$$

A classical Lagrange multiplier argument reveals that the extreme values are given by the generalized eigenvalue problem:

$$\begin{pmatrix} L & M \\ M & N \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \lambda \begin{pmatrix} g_{uu} & g_{uv} \\ g_{vu} & g_{vv} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}$$

Write the generalized eigenvalues as $\lambda_1 = \frac{\text{sign}\lambda_1}{R_1}$, $\lambda_2 = \frac{\text{sign}\lambda_2}{R_2}$. Let X_1, X_2 be the generalized eigenvectors. Since $S(X_i) = \lambda_i X_i$ is (minus) the directional derivative of N along the integral curve of X_i, it is easily seen that R_1, R_2 are the curvature radii of the integral curves of the generalized eigenvectors associated with λ_1, λ_2. By the extremal property of the eigenvalues, λ_1, λ_2 may be called **principal curvatures**. Since $\lambda_1\lambda_2 = \det\left(\begin{pmatrix} g_{uu} & g_{uv} \\ g_{vu} & g_{vv} \end{pmatrix}^{-1}\begin{pmatrix} L & M \\ M & N \end{pmatrix}\right)$, the Gauss curvature as defined by Eq. 8.9 then appears to be $\kappa = \lambda_1\lambda_2 = \frac{\text{sign}(\lambda_1\lambda_2)}{R_1R_2}$.

Clearly, the formula 8.9 bears some resemblance with the definition of the sectional curvature, and this prompt us to show that

$$\langle R(E_u, E_v)E_u, E_v\rangle = LN - M^2$$

The above, however, requires some formalization of what was done above.

Let $\bar{\nabla}$ be the covariant derivative in $T\mathbb{R}^3$, while ∇ denotes the covariant derivative in the tangent space to S. Clearly, by the Levi Civita model, $\bar{\nabla}_X Y - \nabla_X Y$ is the normal component of $\bar{\nabla}_X Y$. Define

$$B(X,Y) = \bar{\nabla}_X Y - \nabla_X Y = \left(\bar{\nabla}_X Y\right)^N$$

where $(.)^N$ denotes the normal component. It is easily seen that $B(X,Y)$ is a bilinear form, but less trivial is the fact that it is symmetric:

$$\begin{aligned} B(Y,X) &= \bar{\nabla}_Y X - \nabla_Y X \\ &= \bar{\nabla}_X Y - \bar{[X,Y]} - \nabla_X Y + [X,Y] \\ &= \bar{\nabla}_X Y - \nabla_X Y \\ &= B(X,Y) \end{aligned}$$

In the above $\bar{[X,Y]}$ denotes the Lie bracket of X, Y in $T\mathbb{R}^3$, which is easily seen to be the same as the Lie bracket $[X,Y]$ of X,Y in the tangent space to S. Since $B(X,Y)$ is in the normal space, it is convenient to write it as

$$B(X,Y) = h(X,Y)N$$

where N is a unit normal vector. This bilinear form turns up to be a formalization of the second quadratic form, as it is easily seen that

$$\begin{aligned} L &= h(E_u, E_u) \\ M &= h(E_u, E_v) \\ N &= h(E_v, E_v) \end{aligned}$$

Indeed,

$$h(E_u, E_u) = \langle B(E_u, E_u), N\rangle = \langle \bar{\nabla}_{E_u} E_u, N\rangle = L$$

with a similar argument for the other component of the second quadratic form.

Next, we provide a formalization of the shape operator $S : TS \to TS$, which in this new context is defined as

$$\langle S(X), Y\rangle = \langle B(X,Y), N\rangle$$

Clearly,

$$\langle S(X), Y\rangle = \langle \bar{\nabla}_X Y - \nabla_X Y, N\rangle = \langle \bar{\nabla}_X Y, N\rangle = -\langle \bar{\nabla}_X N, Y\rangle$$

where the third equality stems from $\bar{\nabla}_X \langle Y, N\rangle = 0$. Therefore,

$$S(X) = -\bar{\nabla}_X N$$

In fact, $-\bar{\nabla}_X N$ is easily seen to be in the tangent space, because indeed, $\langle \bar{\nabla}_X N, N\rangle = \frac{1}{2}\bar{\nabla}_X \langle N, N\rangle = 0$. Therefore, the shape operator can also be defined as

$$S(X) = -\nabla_X N$$

There is a big difference between what has been done previously and this formalization. Previously, we proceeded from *coordinate* vector fields E_u, E_v, which almost trivially led to the symmetric property of the the matrix 8.10. Here, the development is more general, in the sense that the vector fields X, Y are not necessarily coordinate fields, and thus the proof of the symmetry of $B(X, Y)$ requires a bit of a less trivial argument.

Now, with this formalization, we can prove the following:

Theorem 31

$$\langle R(X, Y)X, Y\rangle = \langle B(X, X), B(Y, Y)\rangle - B(X, Y)^2$$

Proof. Since the curvature of Euclidean space is vanishing, we will prove that

$$\langle R(X, Y)X, Y\rangle - \langle \bar{R}(X, Y)X, Y\rangle = B(X, X)B(Y, Y) - B(X, Y)^2$$

Since the Lie brackets of X, Y are the same in TS and in $T\mathbb{R}^3$, it suffices to show that

$$\langle (\nabla_Y \nabla_X - \nabla_X \nabla Y)\, X, Y\rangle - \left\langle \left(\bar{\nabla}_Y \bar{\nabla}_X - \bar{\nabla}_X \bar{\nabla}_Y\right) X, Y\right\rangle = B(X, X)B(Y, Y) - B(X, Y)^2$$

Consider the following:

$$\begin{aligned} \bar{\nabla}_Y \bar{\nabla}_X X &= \bar{\nabla}_Y (B(X, X) + \nabla_Y X) \\ &= \bar{\nabla}_Y (h(X, X)N + \nabla_Y X) \\ &= h(X, X)\bar{\nabla}_Y N + (\bar{\nabla}_Y h)N + \bar{\nabla}_Y \nabla_Y X) \end{aligned}$$

Next, take the inner product with X and observe that $X \perp N$,

$$\begin{aligned} \left\langle \bar{\nabla}_Y \bar{\nabla}_X, X\right\rangle &= h(X, X)\langle \bar{\nabla}_Y N, Y\rangle + \langle \bar{\nabla}_Y \nabla_Y X, Y\rangle \\ &= h(X, X)\langle \bar{\nabla}_Y N, Y\rangle + \langle \nabla_Y \nabla_Y X, Y\rangle \end{aligned}$$

Since $\langle N, Y\rangle = 0$, it follows that

$$\begin{aligned} \langle \bar{\nabla}_Y N, Y\rangle &= -\langle N, \bar{\nabla}_Y\rangle \\ &= -\langle B(Y, Y), N\rangle \\ &= -h(Y, Y) \end{aligned}$$

Combining the previous two expressions we get

$$\left\langle \bar{\nabla}_Y \bar{\nabla}_X, X\right\rangle = -h(X, X)h(Y, Y) + \langle \nabla_Y \nabla_Y X, Y\rangle$$

Working out the other components of the curvature the same way, and putting everything together, the result is obtained. ■

8.7 Ricci and scalar curvature

Given that $p \in M$, that $\{z_1, z_2, \ldots, z_n\}$ is an orthonormal basis of T_pM, then the Ricci curvature in the direction z_i is

$$Ric_p\left(z_i\right) = \frac{1}{n-1} \sum_j \left\langle R\left(z_i, z_j\right) z_i, z_j \right\rangle, \tag{8.11}$$

where $j = 1, \ldots, n$ and $j \neq i$. In addition, the scalar curvature at p is

$$\begin{aligned}\kappa\left(p\right) &= \frac{1}{n} \sum_{i=1}^{n} Ric_p\left(z_i\right) \\ &= \frac{1}{n\left(n-1\right)} \sum_{i,j} \left\langle R\left(z_i, z_j\right) z_i, z_j \right\rangle \end{aligned} \tag{8.12}$$

and it is independent of a choice of orthonormal basis of T_pM.

8.8 Jacobi field

In a nutshell, the Jacobi field is a vector field along a nominal geodesic that indicates how the geodesic is perturbed as we perturb its end points or the angles at its end points. Formally, let $\{\gamma_t : t \in (-\epsilon, +\epsilon)\}$ be a smooth family of geodesics, all geodesics in the family being parameterized by the arc length. γ_0 is referred to as the “nominal” geodesic. The Jacobi field along the nominal geodesics is defined as

$$J(s) = \left. \frac{d\gamma_t(s)}{ds} \right|_{t=0}$$

Clearly, it is the infinitesimal sensitivity of the geodesic in the family.

The Jacobi field satisfies a differential equation of the Riccati type, which is derived as follows: By definition, any geodesic in the family satisfies $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$, where from here on we drop the subscript t for convenience. Taking the covariant derivative relative to the field J yields $\nabla_J \nabla_{\dot{\gamma}}\dot{\gamma} = 0$, which further yields

$$\nabla_{\dot{\gamma}} \nabla_J \dot{\gamma} + (\nabla_J \nabla_{\dot{\gamma}} - \nabla_{\dot{\gamma}} \nabla_J)\dot{\gamma} = 0$$

Now, we use the relationship between the commutator and the curvature. To this end, we observe that $[J, \dot{\gamma}] = 0$. The latter is a corollary of the way the problem is parameterized. Precisely, observe that $\gamma_{\cdot}(s) : (-\epsilon, +\epsilon) \to M$ is an integral curve of the Jacobi field $J(s)$. Now, start at a point $m_0 = \gamma_0(s_0)$ on the nominal geodesic, proceed for a lenth Δs along the nominal geodesic to reach $\gamma_0(s_0 + \Delta s)$, at which point we switch to the integral curve of the Jacobi field and follow it for Δt until we reach the final point $\gamma_{\Delta t}(s_0 + \Delta s)$. Clearly, the same point can be reached by starting from $m_0 = \gamma_0(s_0)$, by following the integral curve of the Jacobi field for Δt to reach $\gamma_{\Delta t}(s_0)$, and then switching to the geodesic and following it for a length Δs to reach $\gamma_{\Delta t}(s_0 + \Delta s)$. In other words, the parallelogram closes and the Lie bracket vanishes. With $[J, \dot{\gamma}] = 0$, we get

$$\nabla_{\dot{\gamma}} \nabla_J \dot{\gamma} + R(\dot{\gamma}, J)\dot{\gamma} = 0$$

Next observe the following string:

$$\begin{aligned}
\nabla_{\dot{\gamma}} \nabla_J \dot{\gamma} &= \nabla_{\dot{\gamma}} \nabla_J \nabla_{\dot{\gamma}} \gamma \\
&= \nabla_{\dot{\gamma}} \nabla_{\dot{\gamma}} \nabla_J \gamma \\
&= \nabla_{\dot{\gamma}} \nabla_{\dot{\gamma}} J
\end{aligned}$$

Hence we obtain the celebrated Riccati equation of the Jacobi field:

$$\nabla_{\dot{\gamma}} \nabla_{\dot{\gamma}} J + R(\dot{\gamma}, J)\dot{\gamma} = 0$$

The following lemma is useful:

Lemma 4 *Let J be a Jacobi field relative to the geodesic γ for a covariant derivative compatible with the metric. Then*

$$\langle J(s), \gamma'(s) \rangle = \langle J'(0), \gamma'(0) \rangle s + \langle J(0), \gamma'(0) \rangle \tag{8.13}$$

Proof. Using the compatibility between the metric and the covariant derivative, the definition of the geodesic γ, and the Jacobi equation, respectively, we get

$$\langle J', \gamma' \rangle' = \langle J'', \gamma' \rangle + \langle J', \gamma'' \rangle = \langle J'', \gamma' \rangle = \langle R(J, \gamma')\gamma', \gamma' \rangle$$

Furthermore,

$$\begin{aligned}
\langle R(J, \gamma')\gamma', \gamma' \rangle &= \left\langle \left(\nabla_{\gamma'} \nabla_J - \nabla_J \nabla_{\gamma'} - \nabla_{[J,\gamma']} \right) \gamma', \gamma' \right\rangle \\
&= \langle \left(\nabla_{\gamma'} \nabla_J - \nabla_J \nabla_{\gamma'} \right) \gamma', \gamma' \rangle \\
&= \langle \nabla_{\gamma'} \nabla_J \gamma', \gamma' \rangle \\
&= \nabla_{\gamma'} \langle \nabla_J \gamma', \gamma' \rangle \\
&= \frac{1}{2} \nabla_{\gamma'} \nabla_J \langle \gamma', \gamma' \rangle \\
&= 0
\end{aligned}$$

Hence, $\langle J', \gamma' \rangle$ is a constant. Invoking again the definition of the geodesic γ and using the latter result, we get

$$\langle J, \gamma' \rangle' = \langle J', \gamma' \rangle + \langle J, \gamma'' \rangle = \langle J', \gamma' \rangle = \langle J'(0), \gamma'(0) \rangle$$

Integrating the above, one obtains (8.13). ■

It is well known in Euclidean geometry that the shortest distance between a point x and a curve c is the length of the line segment from x abutting c perpendicularly. As an application of the concept of Jacobi field, we extend this basic fact to Riemannian geometry.

Let $c : [a, b] \to M$ be a C^1 curve and let $x \notin c$ be a point. Let $\gamma : [0, \ell] \to M$ be the normalized geodesic such that $\gamma(0) = x$, and $\gamma(\ell) = p \in c$ with

$\langle \gamma'(\ell), c'(c^{-1}(p)) \rangle = 0$. The proof relies on considering the field $\tilde{\gamma} : [0,\ell] \times (-\epsilon, +\epsilon) \to M$ of perturbed geodesics such that $\tilde{\gamma}(0,t) = \gamma(0)$, $\gamma(\cdot) = \tilde{\gamma}(\cdot, 0)$, and $\tilde{\gamma}(\cdot, t)$ is also normalized. Define the Jacobi field $J(s) = \frac{d\tilde{\gamma}(s,t)}{dt}\Big|_{t=0}$. Clearly $J(0) = 0$. Furthermore, because $\gamma(\cdot)$ and $\tilde{\gamma}(\cdot, t)$ are normalized, it follows that $\langle J'(0), \gamma'(0) \rangle = 0$.

Now applying the lemma, it follows that $J(\ell) \perp \gamma'(\ell)$. Therefore, a perturbed geodesic of length ℓ, $\tilde{\gamma}(\cdot, \epsilon)$, for ϵ very small will not reach the curve. Therefore, γ is the shortest geodesic.

8.9 Ollivier-Ricci curvature

8.9.1 Foundation of differential geometry

Fundamental in the process of extending geometry in the Euclidean plane to geometry on a surface $\mathcal{S} \subset \mathbb{R}^3$ is the intuitive idea of projecting the ordinary derivative $\frac{d}{dt}X(c(t))$ of a tangent vector field X, defined along a curve c, on the tangent space to the surface, leading to the concept of Levi-Civita connection

$$\nabla_{\dot{c}} X := P_{T_{c(t)}\mathcal{S}} \left(\frac{d}{dt} X(c(t)) \right) \in T_{c(t)}S.$$

The covariant derivative $\nabla_C X$ of the vector field X along the vector field C (not necessarily the tangent to a curve) in a Riemannian manifold $\mathcal{M}$ is a formalization of the intuitive geometric concept of restricting the differential to the tangent space, along with symmetry, $\nabla_C X = \nabla_X C$, linearity relative to C, the product rule relative to scalar multiplication of X and compatibility with the Riemannian metric, viz., $\frac{d}{dt}\langle X(c(t)), Y(c(t)) \rangle = \langle \nabla_{\dot{c}} X, Y \rangle + \langle X, \nabla_{\dot{c}} Y \rangle$

A vector field X is said to be *parallel to itself* along the curve $c : [0,1] \to \mathcal{M}$, if it satisfies the partial differential equation $\nabla_{\dot{c}} X = 0$. Under such conditions, $X(c(1))$ is said to be a *parallel displacement* of $X(c(0))$. This formal definition calls into question by how much this parallel displacement differs from the ordinary Euclidean one. A nonvanishing curvature is precisely symptomatic of such discrepancy. But the immediate problem is that $X(c(0))$ and $X(c(1))$ lives in different tangent spaces and are difficult to compare. One way to go around this difficulty—challenged by the Ollivier [71] concept of curvature—is to bring $X(c(1))$ back to $T_{c(0)}\mathcal{M}$ by another parallel displacement along an extension of c to a closed curve. To somewhat simplify the problem without sacrificing generality in our Ollivier-Ricci curvature objective, assume the curve c and the vector field X live in a 2-dimensional tangent bundle span$\{X, Y\}$. Then

$$\angle(X(c(1)), X(c(0)) = \text{Area}(c) K(X,Y), \tag{8.14}$$

where $K(X,Y)$ is the sectional curvature, a curvature where the parallel displacement is restricted to a 2-dimensional facet. Precisely,

$$K(X,Y) = \frac{\langle R(X,Y)X, Y \rangle}{\|X\|^2 \|Y\|^2 - \langle X, Y \rangle^2},$$

where

$$R(X,Y) = \nabla_Y \nabla_X - \nabla_X \nabla_Y + \nabla_{[X,Y]}$$

is the fundamental curvature operator.

8.9.2 Connection with wireline networks and diffusion processes

Wireline networks in general send packets along optimal paths, along *geodesics* in Riemannian language. Note that a geodesic is only locally length $\ell(\gamma) = \int_\gamma ds$ optimal, as formally the geodesic is defined such that its tangent is parallel to itself, $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$, where the geodesics is parameterized by arc length and $\dot{\gamma} := \frac{d\gamma(s)}{ds}$. Motivated by network outage where optimal paths have to be quickly recomputed, the geodesic nominal γ is embedded in a family of geodesics, γ_p, $p \in (-\epsilon, +\epsilon)$ with $\gamma_0 = \gamma$. The *Jacobi field* $J(s) := \frac{d}{dp}\gamma_p(s)\Big|_{p=0}$ quantifying the variation of geodesic satisfies the equation

$$\nabla_{\dot{\gamma}}\nabla_{\dot{\gamma}} J + K(J,\dot{\gamma})J = 0. \tag{8.15}$$

It is convenient to search a solution of the form $J(s) = j(s)W(s)$ where $W(s)$ is orthogonal to $\gamma(s)$ under uniform curvature K, in which case

$$\frac{d^2}{ds^2} j(s) + Kj(s) = 0. \tag{8.16}$$

Clearly, if $K < 0$, geodesics are diverging, an observation that lies at the foundation of congestion in wireline Gromov hyperbolic networks [49].

Other processes of the diffusion type, that is, such processes as heat diffusion and Heat Diffusion wireless networking [6, 8, 90, 9, 7, 91] involving the Laplace operator, do not “diffuse” along geodesics, but rather follow some thermodynamical-like processes, where the heat kernel exposes the curvature in its Ricci format. The Ricci curvature $\mathrm{Ric}(X)$ is the average of $K(X,Y)$ over all facets $\mathrm{span}\{X,Y\}$ containing X.

Note the fundamental difference between wireline-like networking and diffusion. Wireline networking involve large scale optimal paths, whereas wireless networking in both its backpressure and Heat Diffusion implementation is driven by strictly local queue backlogs, in the same way as heat diffusion is driven by a strictly local temperature gradient.

8.9.3 Towards Ollivier-Ricci curvature

Contrary to what is usually done, here, we attempt to define curvature by reference to different tangent spaces, one centered at $\gamma(0)$, the other at $\gamma(\epsilon)$. Consider two δ-radius balls $B_{\gamma(0)}$, $B_{\gamma(\epsilon)}$. We establish a correspondence between the two balls as follows: Consider $x \in B_{\gamma(0)}$ along with $X = \exp^{-1}_{\gamma(0)}(x)$. Displace X parallel to itself along γ from $\gamma(0)$ to $\gamma(\epsilon)$ to obtain Y. Define $y = \exp_{\gamma(\epsilon)}(Y)$.

This establishes the correspondence $T : x \mapsto y$. To introduce a *transport* idea, the ball $B_{\gamma(0)}$ is endowed with a probability measure μ_0 and $d\mu_0(x)$ is transported to $y = T(x)$ along a geodesic arc $[x, y]$ of length equal to the distance $d(x, y)$.

Invoking the Jacobi field (8.15)-(8.16), the distance $d(x, T(x))$ along the "perturbed" geodesic $[x, y]$ and how its relates to the distance $d(\gamma(0), \gamma(\epsilon)) = \epsilon$ along the "nominal" geodesic depends on the sectional curvature $K(X, \dot{\gamma})$. Therefore, the cost of the transport

$$C(T) = \int_{B_{\gamma(0)}} d(x, T(x)) d\mu_0(x), \tag{8.17}$$

since it involves an integral over all $x \in B_{\gamma(0)}$, tacitly involves an integral over all tangent vectors $X \in T_{\gamma(0)}B_{\gamma(0)}$ and as such averages $K(X, \dot{\gamma})$ over all X to yields the Ricci curvature $\mathrm{Ric}(\dot{\gamma}(0))$.

In 0-curvature, the distance $d(x, T(x))$ is independent of x and therefore the transport cost is $d(\gamma(0), \gamma(\epsilon) = \epsilon$. It remains to see how this distance is affected by the curvature. Define $d\theta(s)$ to be the elementary angle swept by the normal $W(s)$ to the geodesic under an elementary move ds along such geodesic. Then

$$d(x, T(x)) = \epsilon + \int_0^\epsilon j(s) d\theta(s)$$

$j(s)$ is the distance between the nominal and perturbed geodesics measured along the normal to the nominal geodesic and using (8.16) is evaluated as

$$\begin{aligned} j(s) &= \delta \cosh(\sqrt{-K}s) - \frac{\epsilon\delta}{2}\sqrt{-K}\sinh(\sqrt{-K}s) \\ &\approx \delta \cosh(\sqrt{-K}s) \end{aligned}$$

Next, we apply (8.14) to the closed path made up with $\dot{\gamma}_0(s)ds$, $j(s+ds)W(s+ds)$, $-\dot{\gamma}_\delta(s+ds)ds$ and $-j(s)W(s)$. Noting that the left-hand side of (8.14) is the full discrepancy angle around the closed path while we only need the discrepancy along the nominal geodesic, we get

$$\begin{aligned} d\theta &= \frac{1}{2} d\mathrm{Area}(j, ds)\sqrt{-K} \qquad (8.18) \\ &= \frac{1}{2} j(s) ds \sqrt{-K} \qquad (8.19) \end{aligned}$$

Putting everything together and after an elementary integration, it is found that

$$d(x, T(x)) \approx \epsilon(1 - \frac{1}{2}K\delta^2),$$

an estimate consistent with that of [71, Prop. 6, Sec. 8].

The above estimate was derived nominally in a negatively curved manifold, but redeveloping the same argument with ordinary trigonometry rather than hyperbolic trigonometry would validate it in positively curved spaces.

The above clearly indicates that in negative curvature, the transportation cost from x to $T(x)$ is larger than along the nominal geodesic. In positive curvature, the x to $T(x)$ cost is smaller than along γ.

To summarize:

$$\begin{array}{lcl}
\text{Ric} < 0 & \Leftrightarrow & \int_{B_{\gamma(0)}} d(x,T(x))d\mu_0(x) > d(\gamma(0),\gamma(\epsilon) \\
\text{Ric} = 0 & \Leftrightarrow & \int_{B_{\gamma(0)}} d(x,T(x))d\mu_0(x) = d(\gamma(0),\gamma(\epsilon) \\
\text{Ric} > 0 & \Leftrightarrow & \int_{B_{\gamma(0)}} d(x,T(x))d\mu_0(x) < d(\gamma(0),\gamma(\epsilon)
\end{array}$$

8.9.4 From Riemannian manifolds to graphs

On a graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ endowed with a distance $d(\cdot,\cdot)$, we need to emulate the Riemannian manifold environment. We identify an edge ij of the graph with the geodesic $\gamma([0,\epsilon])$ and the graph theoretic neighborhoods $\mathcal{N}_i$, $\mathcal{N}_j$ of i and j with the balls $B_{\gamma(0)}$, $B_{\gamma(\epsilon)}$ centered at $\gamma(0)$, $\gamma(\epsilon)$. Discrete probabilities μ_i, μ_j on $\mathcal{N}_i$, $\mathcal{N}_j$ are obvious substitutes for the measures μ_0, μ_ϵ on the balls $B_{\gamma(0)}$, $B_{\gamma(\epsilon)}$.

The difficulty is to emulate the Riemannian connection resorting only to the graph theoretic distance, or at the very least redefine the cost $C(T)$ in (8.17) in a way that does not involve parallel displacement. It is easily observed that

$$C = \inf_{T:B_{\gamma(0)}\to B_{\gamma(\epsilon)}} \int_{B_{\gamma(0)}} d(x,T(x))d\mu_0(x)$$

where T is restricted to be one-to-one. The graph theoretic emulation of the above is

$$\vec{C}_{\mathcal{G}} = \min_{\mathcal{N}_i \ni k \to \ell \in \mathcal{N}_j} \sum_{k\in\mathcal{N}_i} d(k,\ell)\mu_i(k)$$

In this case, because the cardinalities of $\mathcal{N}_i$ and $\mathcal{N}_j$ might not be the same, the mapping $k \mapsto \ell$, while one-to-many, could be many-to-one. As such, the formula lacks symmetry and cannot be used as a Wasserstein-like distance. To remedy this situation, we introduce a *transference plan* $\xi^{ij}(k,\ell)$ as a substitute for the many-to-many mapping $k \mapsto \ell$, with the added generality that only a piece $\xi^{ij}(k,\ell)$ of $\mu_i(k)$ is transferred to ℓ. The above formula hence becomes

$$C_{\mathcal{G}} = \min_{\xi^{ij}(k,\ell)} d(k,\ell)\xi^{ij}(k,\ell)$$

with of course the consistency conditions

$$\sum_{\ell} \xi^{ij}(k,\ell) = \mu_i(k), \quad \sum_{k} \xi^{ij}(k,\ell) = \mu_j(\ell)$$

The curvature concept that emanates from this cost ($C_{\mathcal{G}} > (<)\epsilon \Leftrightarrow \text{Ric} < (>)0$) is very local, around an edge, in contradiction with the global Gromov concept. This explains why such concept appears the correct one to anticipate performance of backpressure and Heat Diffusion protocols on wireless networks [90, 91].

8.10 local curvature concepts at a point

8.10.1 2-dimensional case

Assume for the sake of the argument that the graph G has been topologically embedded in S_g using the rotation system $\{\pi_x : x \in V_G\}$. Consider the vertex A, choose an edge e incident upon A, and define vertices $B^1, ..., B^n$ such that $AB^1 = e$, $AB^2 = \pi_A(e)$, $AB^3 = \pi_A^2(e)$, $AB^n = \pi_A^{n-1}(e)$, and finally $\pi_A(AB^n) = \pi_A(AB^1)$. Observe that B^i, B^{i+1} need not be connected by an edge, but that there is a distance $d(B^i, B^{i+1})$ associated with them.

Assume that the system of points $A, B^1, ..., B^n$ and distances $d(A, B^i)$, $d(B^i, B^{i+1})$ is embeddable in $\mathbb{E}^3$, using the techniques of Sec. 10.5. In this case, we can connect B^i, B^{i+1} with an edge of length $d(B^i, B^{i+1})$.

The cone with A as apex and the polygonal line $B^1B^2...B^n$ as base is a singular (pyramidal) surface and we strive to define the curvature at its apex. To make this pyramidal surface differentiable, we "round" the edges AB^i to make them cylinders with their symmetry axes parallel to the original edges and tangent to the original faces of the pyramid. It is unclear what is happening at the apex, but, after rounding the edges, the surface away from the apex becomes differentiable, so that the curvature can be obtained via the Gauss-Bonnet theorem by integration along some closed path on the differentiable part of the surface.

We choose as integration path the broken geodesic consisting of straight lines within the "flat" part of the faces and arcs of circles orthogonal to the cylinders. On the other hand, the Gauss curvature vanishes on the flat portion of the faces AB^iB^{i+1} and on the cylinders smoothing over the edges, so that the only contribution to the surface integral is that provided by what remains of the apex after rounding the edges. Let S_A denote a neighborhood of the apex so that its complement is differentiable.

In the limiting case of cylinders of arbitrarily small radii, a turn angle is typically the angle between a straight line $[B^{i-1}B^i]$ and the perpendicular to the edge AB^i, and then the angle $\bar{\beta}^i$ between the perpendicular to AB^i and B^iB^{i+1}, etc., as shown in Figure **??**. Let C^i be the intersection of the perpendiculars to AB^i and AB^{i+1}. Let $\gamma^i = \angle(B^iC^i, C^iB^{i+1})$. Let $\alpha^i = (\angle B^iA, AB^{i+1})$ Since the sum of the internal angles of the quadrilateral $AB^iC^iB^{i+1}$ is 2π, we have

$$\alpha^i + \pi + \gamma^i = 2\pi$$

But in $\triangle B^iC^iB^{i+1}$ we have

$$\gamma^i + \bar{\beta}^i + \bar{\beta}^{i+1} = \pi$$

It follows that

$$\alpha^i = \bar{\beta}^i + \bar{\beta}^{i+1}$$

and the sum of the turn angles is

$$\sum_i (\bar{\beta}^i + \bar{\beta}^{i+1}) = \sum_i \alpha^i$$

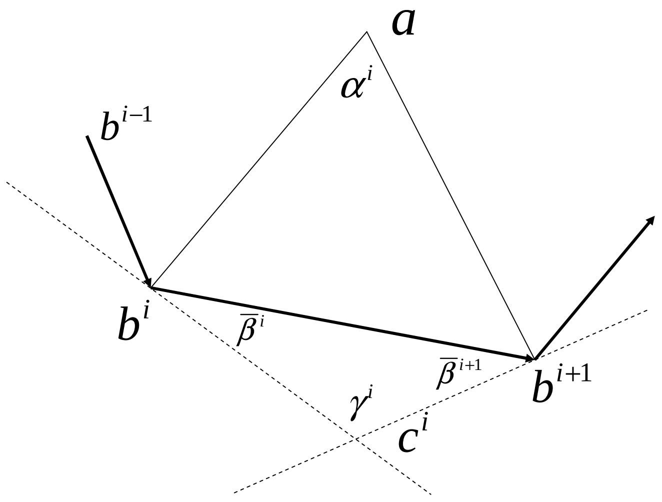


Figure 8.1: The face ab^ib^{i+1} of the pyramid with apex a. The dotted lines are orthogonal to the edges and may be thought as containing the projections of those parts of the integration path following geodesics of the cylinders smoothing the edges. In case of cylinders of vanishingly small radii, the integration path around b^i becomes $b^{i-1}b^ib^{i+1}$.

Hence the Gauss-Bonnet theorem yields

$$\int\int_{S_A} \kappa dA = 2\pi - \sum_i \alpha_i$$

In other words, the neighborhood of the apex has hyperbolic, Euclidean, or spherical geometry depending on whether the sum of the Alexandroff angles at the apex is larger than, equal to, or greater than 2π, respectively.

If $\sum_i \angle B^i A B^{i+1} > 2\pi$, then there is a *pleat* singularity at A and the surface is said to have a *singular hyperbolic metric* (see [52, Chap. 14]). If, on the other hand, $\sum_i \angle B^i A B^{i+1} < 2\pi$, then there is a *conical singularity* at A (see [52, Sec. 61, 6.2]), for indeed the sum of the angles at the apex of a cone is $< 2\pi$.

8.10.2 higher-dimensional case

Now take a graph, let its complete n-subgraphs be the n-simplexes, $n \leq 3$, of a 3-D complex, and *assume* that this simplicial complex is the triangulation of a 3-manifold M^3. Then the (sectional) curvature (of the complex and hence the graph) can be defined as the function $\kappa(x^i x^j) = 2\pi - \sum_{kl} \alpha_{ij}^{kl}$, where α_{ij}^{kl} is the dihedral angle around $x^i x^j$ in the tetrahedron $x^i x^j x^k x^l$ (see [58]). This definition extends trivially to $n > 3$. A confirmation of the validity of this definition is provided by the 3-manifold fact that, if $\kappa(x^i x^j) > 0$, then M^3 has a spherical metric [57]. A scalar curvature can also be defined as $S(x^i) = \sum_{jkl} \kappa(x^i x^j) \text{vol}(x^i x^j x^k x^l)$ (see [58]). The Yamabe flow problem asserts that the scalar curvature of a Riemannian manifold can be deformed to a constant one [97, 89, 5, 80]. However, this is not in general true for the *combinatorial* Yamabe problem [58], a fact that has unfortunate consequences for network congestion.

A disturbing observation is that the combinatorial curvature depends on the triangulation of the manifold. Here, the nagging question is whether there is consistency among the combinatorial curvatures of the various simplicial complexes that can be constructed from a graph.

8.10.3 towards medium scale curvature

The next issue is to define a larger scale curvature. Quite unfortunately in dimension 2, it is not possible to define the curvature of a triangle $\triangle ABC$ where AB, BC and CA are links, because this metric triangle is embeddable in a space of any curvature. If however, we take a geodesic triangle $\triangle ABC$ the edges of which contain many vertices, viz., $AB = AB'...A''B$, $BC = BC''...B'''C$, $AC = AC'...A'''C$, then in the triangle $AB'C'$ we could define the Alexandrov angle $\angle B'AC'$ and likewise, $\angle A''BC''$ and $\angle A'''CB'''$. The curvature would be deemed negative, vanishing, or positive depending on whether $\angle B'AC' + \angle A''BC'' + \angle A'''CB'''$ is less than, equal to, or greater than π (see [23, 4.1.5]).

8.11 constant sectional curvature spaces

Now we look more specifically at the Jacobi field of a constant sectional curvature space. In this case, the relevant sectional curvature is

$$K(J,\dot\gamma) = \frac{< R(J,\dot\gamma)J,\dot\gamma >}{||\dot\gamma||^2||J||^2 - < \dot\gamma, J >^2}$$

Now, recall that $||\dot\gamma|| = 1$ because of the arc length parameterization and let us find a Jacobi field J orthogonal to the geodesic, viz., $< \dot\gamma, J >= 0$. We get

$$K = \frac{< R(J,\dot\gamma)J,\dot\gamma >}{||J||^2}$$

For the above to be constant, we must have

$$R(\dot\gamma, J)\dot\gamma = KJ$$

Hence the equation of the Jacobi field becomes

$$\nabla_{\dot\gamma}\nabla_{\dot\gamma}J + KJ = 0$$

Now, we try a solution like

$$J = jW$$

where W is a unit vector orthogonal to the geodesic. Clearly,

$$\begin{aligned}\nabla_{\dot\gamma}(jW) &= W\frac{dj}{ds} + j\nabla_{\dot\gamma}W \\ &= W\frac{dj}{ds}\end{aligned}$$

Next,

$$\begin{aligned}\nabla_{\dot\gamma}\nabla_{\dot\gamma}(jW) &= \nabla_{\dot\gamma}(W\frac{dj}{ds}) \\ &= \nabla_{\dot\gamma}(\frac{dj}{ds})W + \frac{dj}{ds}\nabla_{\dot\gamma}W \\ &= \frac{d^2j}{ds^2}\end{aligned}$$

Finally, the equation for a Jacobi field orthogonal to the geodesic in a constant curvature space reads,

$$\frac{d^2j}{ds^2} + Kj = 0$$

We look at the solution to this equation in the three different geometries:

- Hyperbolic Geometry: This is the case $K < 0$, with a solution of the form

  $$j(s) = j(0)\cosh(\sqrt{-K}s) + j'(0)\sinh(\sqrt{-K}s)$$

  It is easily verified that $j'(0) = \tan\theta$, where θ is the angle between $\gamma'(0)$ and $\dot\gamma_t(0)$, where $\gamma_t(0) = \exp_{\gamma(0)}(j(0)W)$. (Probably the easiest way to comprehend the latter is to set $j(0) = 0$.) The crucial point is to observe that the geodesics *diverge exponentially.*

- Euclidean geometry: This is the case $K = 0$, with a solution of the form

  $$j(s) = j(0) + j'(0)s$$

  In other words the geodesic *diverge linearly.*

- Spherical geometry: This is the case $K > 0$, with a solution of the form

  $$j(s) = j(0)\cos(\sqrt{K}s) + j'(0)\sin(\sqrt{K}s)$$

  In other words, the geodesics are *oscillatory.*

Chapter 9

Standard constant sectional curvature spaces

9.1 Jacobi field

Here we introduce the standard model spaces $\mathbb{M}_\kappa^n$ carrying n-dimensional Riemannian manifold structures with constant sectional curvature κ. They are also referred to as *"comparison"* spaces, because, as will be done in Chapter 11, the comparison spaces are used as some kinds of yardsticks in the sense that, if a space X behaves metrically in a manner *comparable* with $\mathbb{M}_\kappa^n$ in the sense of the so-called CAT-inequality, then X will be said to have its curvature bounded by κ.

In what follows, $\mathbb{E}^n$ is the standard Euclidean space with inner product $\langle x, y\rangle = \sum_i x^i y_i$.

9.1.1 Euclidean space

It is intuitively clear that the standard Euclidean space $\mathbb{E}^n$ has vanishing curvature, so that it will be the n-dimensional model of vanishing curvature, $\mathbb{M}_0^n$. The metric

$$d : \mathbb{E}^n \times \mathbb{E}^n \to \mathbb{R}^+$$

is given by

$$d(x, y) = \sqrt{\langle x - y, x - y\rangle}$$

where $\langle \cdot, \cdot\rangle$ is the standard $\mathbb{E}^n$ inner product. If $x \neq y$, the geodesic joining x to y is given by

$$\gamma\,(t) = x + \frac{y - x}{\|y - x\|} t$$

To prove, formally, that the above is indeed the geodesic, it suffices to observe that $\|d\gamma\| = \|dt\|$.

Let $\triangle ABC$ be a geodesic triangle in such space. The cosine law is easily derived as

$$\begin{aligned} d(a,b)^2 &= \langle ca-cb, ca-cb\rangle \\ &= d(c,a)^2+d(c,b)^2-2d(c,a)d(c,b)\cos\gamma \end{aligned}$$

9.1.2 Sphere model

Consider in $\mathbb{E}^{n+1}$ the sphere

$$\mathbb{S}^n_R=\{(x_1,...,x_{n+1}):x_1^2+...+x_{n+1}^2=R^2\}$$

Invoking most elementary argument, we define the curvature to be

$$\kappa=\frac{1}{R^2}$$

Since the geometry of the sphere is so well understood, let us agree to choose $\mathbb{S}^n_R$ to be the n-dimensional model $\mathbb{M}^n_\kappa$ of spaces of constant positive curvature $\kappa>0$. Define

$$d:\mathbb{S}^n_R\times\mathbb{S}^n_R\to\left[0,\frac{\pi}{\sqrt{\kappa}}\right]$$

by

$$\cos\sqrt{\kappa}d\left(x,y\right)=\kappa\left\langle x,y\right\rangle \tag{9.1}$$

where $\langle x,y\rangle$ is the usual inner product of $x,y\in\mathbb{E}^{n+1}$. $d(\cdot,\cdot)$ is a metric, as it is induced on $\mathbb{S}^n_R$ by the usual metric of $\mathbb{E}^{n+1}$. Given that $x\neq y\in\mathbb{S}^n_R$, the geodesic path joining x to y is the arc of great circle

$$\begin{aligned} \gamma:[0,d(x,y)] &\to \mathbb{S}^n_R \\ t &\mapsto \gamma\left(t\right)=\left(\cos\sqrt{\kappa}t\right)x+\left(\sin\sqrt{\kappa}t\right)u \end{aligned}$$

where

$$u=\frac{y-\kappa\left\langle x,y\right\rangle x}{\sqrt{\kappa}\left\|y-\kappa\left\langle x,y\right\rangle x\right\|}$$

is a vector of norm R orthogonal to x in the (x,y)-plane. That γ is the geodesic is easily verified by observing that $\gamma(0)=x$ and $\gamma(d(x,y))=y$ and that γ is an isometry, viz., $\langle d\gamma,d\gamma\rangle=dt^2$.

Consider now a geodesic triangle $\triangle abc$. The angle between two geodesic paths $[ca]$ and $[cb]$ issued from a point c of $\mathbb{S}^n_R$ with initial unit tangent vectors v and w, resp., is the unique number $\gamma\in[0,\pi]$ such that $\cos\gamma=\langle v,w\rangle$. From this fact, we derive the cosine law of spherical trigonometry:

$$\begin{aligned} \cos\sqrt{\kappa}d(a,b) &= \kappa\langle a,b\rangle=\langle\bar{a},\bar{b}\rangle \\ &= \langle\bar{c}\langle\bar{a},\bar{c}\rangle+v\langle\bar{a},v\rangle,\bar{c}\langle\bar{b},\bar{c}\rangle+w\langle\bar{b},w\rangle\rangle \\ &= \cos\sqrt{\kappa}d(c,a)\cos\sqrt{\kappa}d(c,b)+\sin\sqrt{\kappa}d(c,a)\sin\sqrt{\kappa}d(c,b)\cos\gamma \end{aligned}$$

where $\bar{a}=a/R$, $\bar{b}=b/R$, and $\bar{c}=c/R$.

9.1.3 Hyperboloid model

Consider in $\mathbb{E}^{n+1}$ the hyperboloid upper sheet

$$\mathbb{H}_R^n = \left\{ (x_1, ..., x_{n+1}) : \begin{array}{l} x_{n+1} \geq 0 \\ x_1^2 + ... + x_n^2 - x_{n+1}^2 = -R^2 \end{array} \right\}$$

The above can be, a bit more formally, redefined as the sphere of radius $\jmath R$

$$\mathbb{H}_R^n = \left\{ (x_1, ..., x_{n+1}) : \begin{array}{l} x_{n+1} \geq 0 \\ \langle x|x \rangle = (\jmath R)^2 \end{array} \right\}$$

in the quadratic space $\mathbb{E}^{n,1}$ defined to be the vector space $\mathbb{R}^{n+1}$ endowed with the bilinear form

$$\langle x \mid y \rangle = \left(\sum_{i=1}^{n} x^i y^i \right) - x^{n+1} y^{n+1}$$

for all $x, y \in \mathbb{R}^{n+1}$ (see [72, Chap. 4]). The fact that the curvature is negative can be seen by computing the osculating surface and observing that it has mixed signature. Let us agree to call this space the model space $\mathbb{M}_\kappa^n$ for negative curvature $\kappa = -\frac{1}{R^2} < 0$. The metric

$$d : \mathbb{H}_R^n \times \mathbb{H}_R^n \to \mathbb{R}^+$$

is defined by

$$\cosh \sqrt{-\kappa} d\,(x, y) = \kappa \, \langle x \mid y \rangle \tag{9.2}$$

The above is in fact the metric induced on $\mathbb{H}_R^n$ by the metric $\langle \cdot, \cdot \rangle$ of $\mathbb{E}^{n,1}$. Given that $x \neq y \in \mathbb{H}_R^n$, then the geodesic path joining x to y is

$$\begin{array}{rcl} \gamma : [0, d(x,y)] & \to & \mathbb{H}_R^n \\ t & \mapsto & \gamma\,(t) = \left(\cosh \sqrt{-\kappa} t \right) x + \left(\sinh \sqrt{-\kappa} t \right) u \end{array}$$

where

$$u = \frac{y - \kappa \langle x \mid y \rangle x}{\sqrt{-\kappa} \sqrt{\langle y - \kappa \langle x \mid y \rangle x | y - \kappa \langle x \mid y \rangle x \rangle}}$$

is a vector of norm R orthogonal to x, that is,

$$\langle u \mid u \rangle = R^2, \quad \langle u \mid x \rangle = 0$$

in the (x, y)-plane. To verify that the above is indeed a geodesic, observe first that $\langle \gamma | \gamma \rangle = -R^2$, so that $\gamma(t)$ remains in the hyperboloid. Next, it is easily verified, using a little bit of hyperbolic trigonometry, that $\gamma(0) = x$ and $\gamma(d(x, y)) = y$. Finally, observe that γ is an isometry, as $\langle d\gamma | d\gamma \rangle = dt^2$.

The hyperbolic angle between two geodesic paths $[ca]$ and $[cb]$ issued from a point $c \in \mathbb{H}_R^n$ with initial unit tangent vectors v and w, resp., is the unique number $\gamma \in [0, \pi]$ such that $\cos\gamma = \langle u \mid v\rangle$. From this, we derive the hyperbolic cosine law:

$$\begin{aligned}
\cosh\sqrt{-\kappa}d(a,b) &= \kappa\langle a,b\rangle = -\langle\bar{a}|\bar{b}\rangle \\
&= -\langle\bar{c}\langle\bar{a}|\bar{c}\rangle + v\langle\bar{a}|v\rangle|\bar{c}\langle\bar{b}|\bar{c}\rangle + w\langle\bar{b}|w\rangle\rangle \\
&= \cosh\sqrt{-\kappa}d(c,a)\cosh\sqrt{-\kappa}d(c,b) - \sinh\sqrt{-\kappa}d(c,a)\sinh\sqrt{-\kappa}d(c,b)\cos\gamma
\end{aligned}$$

where $\bar{a} = a/R$, $\bar{b} = b/R$, and $\bar{c} = c/R$.

9.2 Models of Hyperbolic spaces with constant curvature -1

9.2.1 n-dimensional models

In this section, five analytic models for hyperbolic space are introduced. Each model is a complete Riemannian manifold with an associated Riemannian metric. In addition, each model has a constant sectional curvature -1. The analytic model for hyperbolic spaces are given as follows:

Half-space model H^n :

$$H^n = \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_n > 0\}$$

with the associated Riemannian metric

$$g_{ij}(x_1, \dots, x_n) = \frac{\delta_{ij}}{x_n^2}.$$

Poincaré ball model D^n:

$$D^n = \left\{(x_1, \dots, x_n) \in \mathbb{R}^n : x_1^2 + \cdots + x_n^2 < 1\right\}$$

with the associated Riemannian metric

$$g_{ij}(x_1, \dots, x_n) = \frac{4\delta_{ij}}{\left(1 - (x_1^2 + \cdots + x_n^2)\right)^2}.$$

Hyperboloid model $\mathbb{H}^n$:

$$\mathbb{H}^n = \left\{(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1} : x_1^2 + \cdots + x_n^2 - x_{n+1}^2 = -1, x_{n+1} > 0\right\}$$

with the associated Riemannian metric

$$g_{ij}(x_1, \dots, x_{n+1}) = \begin{cases} 0 & i \neq j \\ 1 & i = j \neq n+1 \\ -1 & i = j = n+1 \end{cases}$$

Jemisphere model J^n:

$$J^n = \left\{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} : x_1^2 + \cdots + x_n^2 + x_{n+1}^2 = 1, x_{n+1} > 0\right\}$$

with the associated Riemannian metric

$$g_{ij}\left(x_1, \ldots, x_n\right) = \frac{\delta_{ij}}{x_{n+1}^2}$$

Klein model K^n:

$$\left\{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1^2 + \cdots + x_n^2 < 1\right\}$$

with the associated Riemannian metric

$$g_{ij}\left(x_1, \ldots, x_n\right) = \frac{\delta_{ij}}{1 - (x_1^2 + \cdots + x_n^2)} + \frac{x_i x_j}{\left(1 - (x_1^2 + \cdots + x_n^2)\right)^2}$$

9.2.2 2-dimensional models

Poincaré disk model

Given the unit disk $D^2 = \{z \in \mathbb{C} : |z| < 1\}$ with Riemannian metric

$$ds = \frac{2\,|dz|}{1 - |z|^2},$$

the distance d_D is given by

$$d_D\left(z, w\right) = \ln \frac{|1 - \bar{z}w| + |w - z|}{|1 - \bar{z}w| - |w - z|} = \tanh^{-1} | \frac{z - w}{1 - z\bar{w}} |$$

and the geodesic for each pair of z, w in D^2 with $z \neq w$ is the unique Euclidean circle C or line L which contains z, w and orthogonal to the unit circle.

Poincaré half plane model

Given the half space $H^2 = \{z \in \mathbb{C} : \Im\left(z\right) > 0\}$ with Riemannian metric

$$ds = \frac{|dz|}{\Im\left(z\right)},$$

the distance d_H is given by

$$d_H\left(z, w\right) = \ln\left(\frac{|z - \bar{w}| + |z - w|}{|z - \bar{w}| - |z - w|}\right) = 2\tanh^{-1}\frac{|z - w|}{|z - \bar{w}|}$$

and the geodesic for each pair of z, w in H^2 with $z \neq w$ is the unique Euclidean circle C or line L which contains z, w and orthogonal to the real axis.

Chapter 10

Isometric embedding in constant curvature space

The premise of the previous chapter was that any triangle in a metric space can be isometrically embedded in any of the standard constant curvature spaces. The problem is that for such more complicated objects as graphs, this is not always true. The purpose of this chapter is to derive conditions for metric graphs to be embeddable in a constant curvature space. If this can be accomplished, then the graph can be given a curvature, rather than a curvature bound as in the previous chapter.

10.1 distance structure and links

Given a weighted graph (G, w), it is easy to construct a metric structure (X, d) on its vertex set $X = \{x^1, ..., x^n\}$. Conversely, there arises the question as to whether, given a metric space (X, d) on a finite or countably infinite set of points, there exists a weighted graph (G, w) on the vertex set X that induces the distance $d : X \times X \to \mathbb{R}$. The answer is given by the following:

Theorem 32 *Given a set of points $x^1, ..., x^n$ and a symmetric distance matrix $d(x^i, x^j)$, there exists a unique weighted graph (G, w) that induces the metric $d : X \times X \to \mathbb{R}$.*

Proof. Two arbitrary nodes x^i, x^j are linked if and only if, for an arbitrary node x^k, $k \neq i, j$, we have

$$d(x^i, x^j) < d(x^i, x^k) + d(x^k, x^j), \tag{10.1}$$

in which case $w(x^i x^j) = d(x^i, x^j)$. This yields a weighted graph structure. To prove that it is unique, let (G_1, w_1), (G_2, w_2) be two such structures. Take two points x^i, x^j and assume they are linked in (G_1, w_1) and not linked in (G_2, w_2). Therefore, in (G_2, w_2), there must exist a path $x^i = x^{k_0}, x^{k_1}, ..., x^{k_{M-1}}, x^{k_M} = x^j$

such that $d(x^i,x^j) = \sum_m d(x^{k_m},x^k_{m+1})$. Using the triangle inequality, this implies that, for any intermediate point of this path, $d(x^i,x^j) < d(x^i,x^k_m) + d(x^k_m,x^j)$, which contradicts (10.1). ■

If we start with a weighted graph (G,w), construct the metric space (X,d) on its vertex set, and them construct the graph (G_d,w_d) using the previous theorem, it is not always true that, topologically and consequently metrically, $G_d = G$. A discrepancy would typically happen when G has a link $x^m x^{m+1}$ of such a large weight that it does not contribute to the metric d; precisely, for no pairs of vertices x^i, x^j do we have $d(x^i,x^j) = \ell(... \cup x^m x^{m+1} \cup ...)$.

If, to gauge the curvature properties of a graph, we attempt an isometric embedding $(G,d) \to \mathbb{M}^r_\kappa$ in one of the standard constant curvature spaces, those links too heavily weighted to contribute to the metric will be problematic, and for this reason we prefer to look at the isometric embedding $(X,d) \to \mathbb{M}^r_\kappa$.

10.2 Gram matrices

10.2.1 positive curvature

A set of n points $x^1,...,x^n$ on $\mathbb{S}^r_R$ can be considered as a set of n $(r+1)$-dimensional vectors in the underlying space $\mathbb{E}^{r+1}$. Write those vectors as linear combination of the basis $\{E_i : i = 1,...,r+1\}$ of $\mathbb{E}^{r+1}$: $x_i = x_i^j E_j$. Then it is easily verified that the Gram matrix $\{\langle x^i, x^j\rangle\}$ can be written as

$$\begin{pmatrix} \langle x_1,x_1\rangle & \dots & \langle x_1,x_n\rangle \\ \vdots & \ddots & \vdots \\ \langle x_n,x_1\rangle & \dots & \langle x_n,x_n\rangle \end{pmatrix} =$$

$$\begin{pmatrix} x_1^1 & \dots & x_1^{r+1} \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^{r+1} \end{pmatrix} \begin{pmatrix} \langle E_1,E_1\rangle & \dots & \langle E_1,E_{r+1}\rangle \\ \vdots & \ddots & \vdots \\ \langle E_{r+1},E_1\rangle & \dots & \langle E_{r+1},E_{r+1}\rangle \end{pmatrix} \begin{pmatrix} x_1^1 & \dots & x_n^1 \\ \vdots & \ddots & \vdots \\ x_1^{r+1} & \dots & x_n^{r+1} \end{pmatrix}$$

The Gram matrix of the basis vectors $\{\langle E_i, E_j\rangle\}$ is positive definite for the usual inner product. Next, in the $\mathbb{M}^r_\kappa$ model, $\langle x_i, x_j\rangle = \frac{1}{\kappa}\cos\sqrt{\kappa}d(x_i,x_j)$. It therefore appears that the embeddability of n points in r-dimensional manifold of constant positive curvature κ is related to whether the matrix $\{\cos\sqrt{\kappa}d(x^i,x^j)\}$ is positive (semi)definite of rank $(r+1)$.

10.2.2 negative curvature

A similar argument holds for negative curvature embedding, with the difference that we use the $\mathbb{H}^r_R$ model embedded in the space $\mathbb{E}^{r+1,1}$ with nonsign definite inner product $\langle\cdot|\cdot\rangle$. Gain, it is easily checked that

$$\begin{pmatrix} \langle x_1|x_1\rangle & \dots & \langle x_1|x_n\rangle \\ \vdots & \ddots & \vdots \\ \langle x_n|x_1\rangle & \dots & \langle x_n|x_n\rangle \end{pmatrix} =$$

$$\begin{pmatrix} x_1^1 & \dots & x_1^{r+1} \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^{r+1} \end{pmatrix} \begin{pmatrix} \langle E_1|E_1\rangle & \dots & \langle E_1|E_{r+1}\rangle \\ \vdots & \ddots & \vdots \\ \langle E_{r+1}|E_1\rangle & \dots & \langle E_{r+1}|E_{r+1}\rangle \end{pmatrix} \begin{pmatrix} x_1^1 & \dots & x_n^1 \\ \vdots & \ddots & \vdots \\ x_1^{r+1} & \dots & x_n^{r+1} \end{pmatrix}$$

The Gram matrix of the basis vectors $\{\langle E_i|E_j\rangle\}$ now has signature $+1,+1,...,+1,-1$. Furthermore, in this model, $\langle x_i|x_j\rangle = \frac{1}{\kappa}\cosh\sqrt{-\kappa}d(x_i,x_j)$. Since $\kappa < 0$, the matrix $\{\cosh\sqrt{-\kappa}d(x_i,x_j)\} =: \Delta$ can be written as $X'EX$, where E has signature $-1,-1,...,-1,+1$. By elementary linear algebra, Δ is congruent to the direct sum $\Delta_{(r+1)\times(r+1)} + 0_{(n-r-1)\times(n-r-1)}$ and $\Delta_{(r+1)\times(r+1)}$ is itself congruent to diag$(-1,-1,...,-1,+1)$. Therefore, embeddability is related to whether $\{\cosh\sqrt{-\kappa}d(x_i,x_j)\}$ has a sequence of nested principal minors of alternate sign, up to an including order $(r+1)\times(r+1)$, and vanishing thereafter.

10.3 Cayley-Menger matrix

Consider a set of vectors $x^1,...,x^r$ in Euclidean space $\mathbb{E}^r$ with inner product $\langle\cdot,\cdot\rangle$. It is well known that the volume $V(x^1,...,x^k)$ of the parallelepiped constructed on the vectors $x^1,...,x^r$ is given by

$$V^2(x^1,...,x^r) = \det\begin{pmatrix} \langle x^1,x^1\rangle & \dots & \langle x^1,x^r\rangle \\ \vdots & \ddots & \vdots \\ \langle x^r,x^1\rangle & \dots & \langle x^r,x^r\rangle \end{pmatrix}$$

Using a classical Schur complement argument, it is easily found that the above can be rewritten as

$$V^2(x^1,...,x^r) = -\det\begin{pmatrix} \langle x^1,x^1\rangle & \dots & \langle x^1,x^r\rangle & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \langle x^r,x^1\rangle & \dots & \langle x^r,x^r\rangle & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix}$$

In Euclidean space, we have the relationship

$$\langle x^i,x^j\rangle = \frac{1}{2}\left(||x^i||^2 + ||x^j||^2 - d(x^i,x^j)^2\right)$$

Replace the inner product in the volume expression by the more primitive concepts of norm and distance; next subtract the last row (column) multiplied by

$\frac{1}{2}||x^i\|^2$ from the ith row (column) and it is easily found that

$$V^2(x^1,...,x^r) = (-1)^{r+1} \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & d(x^1,x^2) & \dots & d(x^1,x^r) \\ 1 & d(x^2,x^1) & 0 & \dots & d(x^2,x^r) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d(x^r,x^1) & d(x^r,x^2) & \dots & 0 \end{pmatrix}$$

10.4 congruence invariants

The isometric embedding of a set of points in the standard constant curvature spaces is a traditional problem, which has received a practical solution in terms of the signs of a nested sequence of principal minors of some matrices. For embedding in constant curvature $\kappa > 0$ space, the relevant matrix associated with a set of points $x^1, ..., x^n$ is

$$\Delta_\kappa(x^1,...,x^n) = \left\{\cos\left(d(x^i,x^j)\sqrt{\kappa}\right)\right\}_{1\leq i,j\leq n}$$

For embedding in the Euclidean space, the relevant matrix is the so-called Cayley-Menger matrix:

$$D(x^1,...,x^n) = \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & d(x^1,x^2)^2 & \dots & d(x^1,x^n)^2 \\ 1 & d(x^2,x^1)^2 & 0 & \dots & d(x^2,x^n)^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d(x_n,x^1)^2 & d(x^n,x^2)^2 & \dots & 0 \end{pmatrix}$$

For embedding in the standard constant curvature space $\kappa < 0$, the relevant matrix is

$$\Delta_\kappa(x^1,...,x^n) = \left\{\cosh\left(d(x^i,x^j)\sqrt{-\kappa}\right)\right\}_{1\leq i,j\leq n}$$

The embeddability condition in positive curvature space is equivalent to the sequence of top left hand corner principal minors of the relevant matrix to be positive, which is equivalent to positive definiteness of the same matrix. For embeddability in Euclidean and negative curvature spaces, the same sequence but for the other relevant matrices must have alternating signs; precisely, the $k \times k$ principal minor must have sign $-(-1)^k$. The following reformulation will turn out to be helpful:

Lemma 5 *Let Δ_k be the top left hand corner principal $k \times k$ submatrix of the $n \times n$ real symmetric matrix Δ. Assume $\det \Delta_k \neq 0$, $k = 1, ..., n$. Then sign $\det \Delta_k = -(-1)^k$ if and only if Δ has $(n-1)$ negative eigenvalues, $\lambda_1 \leq ... \leq \lambda_{n-1} < 0$, and one positive eigenvalue $\lambda_n > 0$.*

Proof. The proof is by induction on the size n of the matrix. Clearly, the statement of the theorem is completely trivial for $n = 1$. Suppose now that the result is valid up to order $n-1$ and let us show that it is valid up to order n. Partition Δ as

$$\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}$$

where Δ_{11} is $(n-1) \times (n-1)$ and nonsingular. As is well known, a Schur complement argument yields the congruence relation

$$\begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix} \sim \begin{pmatrix} \Delta_{11} & 0 \\ 0 & \Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12} \end{pmatrix} \tag{10.2}$$

Hence the above two matrices have the same signs for their eigenvalues. Clearly, $\operatorname{sign}\det\Delta_k = -(-1)^k$, $k \leq n$, if and only if $\operatorname{sign}\det\Delta_k = -(-1)^k$, $k \leq n-1$ and $\operatorname{sign}\left(\Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right) = -1$. By induction hypothesis, $\operatorname{sign}\det\Delta_k = -(-1)^k$, $k \leq n-1$, iff Δ_{11} has $(n-1)$ negative and one positive eigenvalue. On the other hand, in view of (10.2), the condition $\operatorname{sign}\left(\Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right) = -1$ means that Δ has one negative eigenvalue in addition to those of Δ_{11}. Hence the statement of the lemma holds up to order n. ■

For the matrix D of Euclidean embedding, the 1×1 top left hand corner principal minor vanishes, so that the previous result need a slight revision:

Lemma 6 *Let D_k be the top left hand corner principal $k \times k$ submatrix of the $(n+1) \times (n+1)$ real symmetric matrix D, where $n \geq 2$. Assume $\det D_k \neq 0$, $k = 2, ..., n+1$. Then sign $\det D_k = -(-1)^k$, $k = 2, ..., n+1$, if and only if D has n negative eigenvalues, $\lambda_1 \leq ... \leq \lambda_n < 0$, and one positive eigenvalue $\lambda_{n+1} > 0$.*

Proof. The proof is the same as that of Lemma 5, except for the fact that the startup of the induction argument is $n = 2$ with the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ rather than $n = 1$ with the "matrix" 1. ■

When some top left hand corner minors (of D) vanish (for $k > 2$), some caution must be exercised in order to be able to infer something about the eigenvalues of Δ. Indeed, once a top left hand corner minor (of D) vanishes (for $k > 2$), so must all of those containing it in order to be able to conclude that some eigenvalues vanish. Contrary to Lemma 5 where a relabeling would not affect the sign of the sequence of nested principal minors, here, as a general rule, the required behavior of the principal minors as an eigenvalue test can only be achieved after a relabeling of the rows and the columns of Δ.

Lemma 7 *Let Δ have nonvanishing diagonal elements. Then there exists a relabeling of the rows and columns of Δ such that sign $\det\Delta_k = -(-1)^k$, $k \leq n_1$ and $\det\Delta_k = 0$ for $n_1 + 1 \leq k \leq n$ iff Δ has $n_1 - 1$ negative eigenvalues, $n - n_1$ vanishing eigenvalues, and 1 positive eigenvalue.*

Proof. Since Δ has nonvanishing diagonal elements, there exists at least one top left hand corner principal minor that does not vanish. Let Δ_{11} be the maximum nonsingular top left hand corner principal submatrix of Δ. As before, we have the congruence relationship

$$\left(\begin{array}{cc} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{array}\right) \sim \left(\begin{array}{cc} \Delta_{11} & 0 \\ 0 & \Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12} \end{array}\right)$$

If $\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}=0$, the theorem is proved. In the opposite case, there exists a principal submatrix of $\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}$ that is nonsingular. (Indeed, if all such minors vanish, the characteristic polynomial of $\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}$ is s^m, all eigenvalues vanish, and so does the matrix.) Relabel the rows and the columns of $\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}$, which amounts to relabel the rows and the columns of Δ of indices from n_1+1 to n, such that the nonsingular principal submatrix is brought to the top left hand corner position in of $\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}$, that is,

$$P^T\left(\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right)P = \left(\begin{array}{cc} \Delta^c_{\alpha\alpha} & \Delta^c_{\alpha\beta} \\ \Delta^c_{\beta\alpha} & \Delta^c_{\beta\beta} \end{array}\right)$$

where $\Delta^c_{\alpha\alpha}$ is nonsingular, α denotes those row and column indices of Δ corresponding to the nonsingular principal submatrix of $\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}$, P denotes the column permutation matrix, and the superscript "c" denotes the Schur complement. From these considerations, it follows that

$$P^T\left(\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right)P = \left(\begin{array}{cc} \Delta_{\alpha\alpha} & \Delta_{\alpha\beta} \\ \Delta_{\beta\alpha} & \Delta_{\beta\beta} \end{array}\right) - \left(\begin{array}{c} \Delta_{\alpha 1} \\ \Delta_{\beta 1} \end{array}\right)\Delta_{11}^{-1}\left(\begin{array}{cc} \Delta_{1\alpha} & \Delta_{1\beta} \end{array}\right)$$

Clearly, we have the congruence relations

$$\Delta \sim \left(\begin{array}{ccc} \Delta_{11} & 0 & 0 \\ 0 & \Delta_{\alpha\alpha}-\Delta_{\alpha 1}\Delta_{11}^{-1}\Delta_{1\alpha} & 0 \\ 0 & 0 & X \end{array}\right) \sim \left(\begin{array}{ccc} \Delta_{11} & \Delta_{1\alpha} & 0 \\ \Delta_{\alpha 1} & \Delta_{\alpha\alpha} & 0 \\ 0 & 0 & X \end{array}\right)$$

where X denotes the Schur complement of $\Delta_{\alpha\alpha}-\Delta_{\alpha 1}\Delta_{11}^{-1}\Delta_{1\alpha}$ in $P^T\left(\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right)P$. Clearly,

$$\left(\begin{array}{cc} \Delta_{11} & \Delta_{1\alpha} \\ \Delta_{\alpha 1} & \Delta_{\alpha\alpha} \end{array}\right)$$

is nonsingular. From here on, we iterate until we reach the congruence relation

$$\Delta \sim \left(\begin{array}{cc} \Delta_{\gamma\gamma} & 0 \\ 0 & 0 \end{array}\right)$$

where $\Delta_{\gamma\gamma}$ is nonsingular. The latter along with Lemma 5 proves the theorem. ■

Again, the preceding result needs a slight amendment to be applicable to D.

Lemma 8 *Let the $(n+1)\times(n+1)$ real symmetric matrix D have nonsingular top left hand corner principal submatrix D_2. Then there exists a relabeling of the rows and columns of Δ of indices $3, ..., n+1$ such that sign $\det D_k = -(-1)^k$, $k \leq n_1$ and $\det D_k = 0$ for $n_1 + 1 \leq k \leq n+1$ iff D has $n_1 - 1$ negative eigenvalues, $n+1-n_1$ vanishing eigenvalues, and 1 positive eigenvalue.*

Proof. Essentially the same as the preceding. ■

10.5 Main results

With these preliminaries, we can state the fundamental embedding theorems:

Theorem 33 *The metric space (X,d) can be isometrically embedded in the standard constant curvature $\kappa > 0$ space of dimension r iff the diameter of the set of points does not exceed $\frac{\pi}{\sqrt{\kappa}}$ and, possibly after some relabeling, $\Delta^\kappa(x^1, ..., x^{r+1}) > 0$ and $\det \Delta^\kappa(x^1, ..., x^{r+1}, x^{r+2}, ..., x^k) = 0$ for $r+2 \leq k \leq n$.*

Proof. Necessity is obvious. Hence we only prove sufficiency. Clearly, the matrix $\Delta^k(x^1, ..., x^{r+1}, x^{r+2}, ..., x^n)$ is congruent to diag$(+1, +1, ..., +1, 0, ..., 0)$. Write this congruence as

$$\Delta^\kappa(x^1, ..., x^{r+1}, x^{r+2}, ..., x^n) = \frac{1}{R^2} X' \mathrm{diag}(+1, +1, ..., +1) X$$

where the columns of X have norm R. Using an orthonormal basis in the standard space $\mathbb{E}^{r+1}$, the above can be rewritten

$$\Delta^\kappa(x^1, ..., x^{r+1}, x^{r+2}, ..., x^n) = \frac{1}{R^2} X' \begin{pmatrix} \langle E_1, E_1 \rangle & \dots & \langle E_1, E^{r+1} \rangle \\ \vdots & \ddots & \vdots \\ \langle E^{r+1}, E^1 \rangle & \dots & \langle E^{r+1}, E^{r+1} \rangle \end{pmatrix} X$$

With X we define the embedding $x_i \mapsto x^i_k E_k$. The congruence relation yields

$$\cos \sqrt{\kappa} d(x^i, x^j) = \kappa \langle x^i_k E_k, x^j_l E_l \rangle$$

From the above, $d(x^i, x^j)$ is the distance between the points $x^i_k E_k$, $x^j_l E_l$ on the sphere $\mathbb{S}^r_R$. Hence the embedding $(X,d) \hookrightarrow \mathbb{M}^r_\kappa$ is isometric. ■

Corollary 4 *There exists an isometric embedding $(X,d) \to (\mathbb{M}^r_{\kappa>0}, \bar{d})$ iff the diameter of the set of points does not exceed $\frac{\pi}{\sqrt{\kappa}}$ and the eigenvalues of $\Delta(x^1, ..., x^n)$ are as*

$$0 = \lambda_1 = ... = \lambda_{n-r-1} < \lambda_{n-r} \leq ... \leq \lambda_n$$

Theorem 34 *The metric space (X,d) can be isometrically embedded in the standard constant curvature $\kappa > 0$ space of dimension r iff the diameter of the set of points does not exceed $\frac{\pi}{\sqrt{\kappa}}$ and*

1. *There exists a sequence of points, possibly after relabeling,* $x^1, ..., x^{r+1}$ *such that*
$$\det \Delta^\kappa(x^1, ..., x^k) > 0, \quad k \le r+1$$

2. *For any two points* $x, x' \in X$*, we have*
$$\det \Delta^\kappa(x^1, ..., x^{r+1}, x) = 0$$
and
$$\det \Delta^\kappa(x^1, ..., x^{r+1}, x, x') = 0$$

Proof. See [15, Th. 63.1]. ■

Theorem 35 *There exists an embedding* $(X, d) \to \mathbb{E}^r$ *iff*

1. *There exists, possibly after relabeling,* $x^1, ..., x^{r+1}$*, such that*
$$\textit{sign} \det D(x^1, ..., x^k) = (-1)^k, \quad k \le r+1$$

2. *For any two points* $x, x' \in X$*, we have*
$$\det D(x^1, ..., x^{r+1}, x) = 0$$
$$\det D(x^1, ..., x^{r+1}, x, x') = 0$$

Proof. See [15, Th. 41.1, 42.1]. ■

Corollary 5 *There exists an embedding* $(X, d) \to \mathbb{E}^r$ *iff the matrix* $D(x^1, ...x^n)$ *has eigenvalues*
$$\lambda_1 \le ... \le \lambda_{r+1} < 0 = \lambda_{r+2} = ... = \lambda_n < \lambda_{n+1}$$

Theorem 36 *The metric space* (X, d) *can be isometrically embedded in the standard constant curvature* $\kappa < 0$ *space of dimension* r *iff, possibly after some relabeling,* $\det \Delta^\kappa(x^1, ..., x^k) = (-1)^{k+1}$*, for* $k \le r+1$*, and* $\det \Delta^k(x^1, ..., x^{r+1}, x^{r+2}, ..., x^k) = 0$ *for* $r+2 \le k \le n$*.*

Proof. Necessity is obvious. Hence we only prove sufficiency. Clearly, the matrix $\Delta^k(x^1, ..., x^{r+1}, x^{r+2}, ..., x^n)$ is congruent to $\text{diag}(-1, -1, ..., -1, +1, 0, ..., 0)$. Write this congruence as
$$\Delta^k(x^1, ..., x^{r+1}, x^{r+2}, ..., x^n) = \frac{1}{R^2} X' \text{diag}(-1, -1, ..., -1, +1) X$$

Choosing the basis $\{E_i\}$ of the quadratic space $\mathbb{E}^{r+1,1}$, this congruence can be rewritten
$$\Delta^k(x^1, ..., x^{r+1}, x^{r+2}, ..., x^n) = -\frac{1}{R^2} X' \begin{pmatrix} \langle E_1 | E_1 \rangle & \dots & \langle E_1 | E^{r+1} \rangle \\ \vdots & \ddots & \vdots \\ \langle E^{r+1} | E^1 \rangle & \dots & E^{r+1} | E^{r+1} \rangle \end{pmatrix} X$$

The above yields

$$\cosh\sqrt{-\kappa}d(x^i,x^j)=\kappa\langle x^i_k E_k, x^j_l E_l\rangle$$

Hence $d(x^i,x^j)$ is the distance between the points $x^i_k E_k$, $x^j_l E_l$ on the hyperboloid $\mathbb{H}^r_R$. Hence the embedding is isometric. ■

Theorem 37 *There exists an isometric embedding $(X,d)\to\mathbb{M}^r_{\kappa<0}$ iff*

1. *There exists a sequence of points $x^1,...,x^{r+1}$, possibly after relabeling, such that*
$$sign\det\Delta^\kappa(x^1,...,x^k)=-(-1)^k,\quad k\leq r+1$$

2. *For any two points $x,x'\in X$, we have*
$$\det\Delta^\kappa(x^1,...,x^{r+1},x)=0$$
$$\det\Delta^\kappa(x^1,...,x^{r+1},x,x')=0$$

Proof. See [15, Th. 106.1 and Cor.]. ■

Corollary 6 *There exists an isometric embedding $(X,d)\to\mathbb{M}^r_{\kappa<0}$ iff the eigenvalues of $\Delta(x^1,...,x^n)$ are as*

$$\lambda_1\leq...\leq\lambda_r<0=\lambda_{r+1}=...=\lambda_{n-1}<\lambda_n$$

Proof. From the preceding and Lemma 5. ■

10.6 isometric embedding of basic graph structures

10.6.1 complete graph

Embeddability of the complete graph with uniform link weight in both the constant curvature hyperbolic space and the constant curvature spherical space involves the special $k\times k$ Toeplitz structure

$$T_k=\begin{pmatrix}1 & c & \dots & c\\ c & 1 & \dots & c\\ \vdots & \vdots & \ddots & \vdots\\ c & c & \dots & 1\end{pmatrix},\quad k\geq 1$$

where $c=\cosh\left(d(x^i,x^j)\sqrt{-\kappa}\right)$ in the hyperbolic case and $c=\cos\left(d(x^i,x^j)\sqrt{\kappa}\right)$ in the spherical case. The issue is the sequence of principal minors of such a Toeplitz-structured matrix. Set

$$t_k=\det T_{k\times k}$$

and we have the following lemma:

Lemma 9 *The recursion on the principal minors of the Toeplitz-structured matrix T_k is*

$$t_{k+1} = (1-c)t_k + (1-c)^2 t_{k-1} - (1-c)^3 t_{k-2}$$

subject to the initial conditions

$$\begin{aligned} t_1 &= 1 \\ t_2 &= 1-c^2 \\ t_3 &= (1-c)^2(2c+1) \end{aligned}$$

Furthermore, the solution to the above recursion is given by

$$t_k = (1-c)^{k-1}\left((k-1)c+1\right), \quad k \geq 1$$

Proof. By subtracting the first column from the last column of T_k, we get

$$\det T_k = (-1)^{k+1}(c-1)\det T + (1-c)\det T_{k-1}$$

where T is the Toeplitz matrix with 1's on the superdiagonal and c's everywhere else. Again, by subtracting the first row from the last row of T, we get

$$\det T = (-1)^{k+1}(c-1)\det\begin{pmatrix} c & ce_{k-3}^T \\ ce_{k-3} & T_{k-3} \end{pmatrix}$$

where e_k is the k-dimensional column made up of 1's. Observing that

$$\begin{pmatrix} c & ce_{k-3}^T \\ ce_{k-3} & T_{k-3} \end{pmatrix} = \begin{pmatrix} (c-1) & 0 \\ 0 & 0 \end{pmatrix} + T_{k-2}$$

and remembering that the determinant of the sum of two matrices equals the sum of the determinants of all matrices constructed with some columns of the first matrix and the complementary columns of the second matrix, we get

$$\begin{aligned} \det\begin{pmatrix} c & ce_{k-3}^T \\ ce_{k-3} & T_{k-3} \end{pmatrix} &= \det T_{k-2} + \det\begin{pmatrix} (c-1) & ce_{k-3} \\ 0 & T_{k-3} \end{pmatrix} \\ &= \det T_{k-2} + (c-1)\det T_{k-3} \end{aligned}$$

Combining all of the above yields the recursion. The initial conditions on the recursion are trivial to verify. The explicit solution is easily seen by direct verification to satisfy the recursion and its initial conditions. ■

From the above, it is possible to say something about the eigenvalues of T_n.

Corollary 7

$$\det(sI - T_n) = (s-(1-c))^{n-1}(s-((n-1)c+1))$$

Proof. Recall that the coefficient of s^{n-k} in $\det(sI - T_n)$ is $(-1)^k$ times the sum of all principal minors of order k of T_n. There are $\binom{n}{k}$ such principal minors, all equal to t_k. Hence,

$$\begin{aligned}
\det(sI - T_n) &= \sum_{k=0}^{n}(-1)^k\binom{n}{k}t_k s^{n-k} \\
&= \sum_{k=0}^{n}(-1)^k\binom{n}{k}(1-c)^{k-1}\left((k-1)c+1\right)s^{n-k} \\
&= \left(\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}(1-c)^k s^{n-1-k}\right)(s-((n-1)c+1)) \\
&= (s-(1-c))^{n-1}(s-((n-1)c+1))
\end{aligned}$$

■

Theorem 38 *The complete graph $K_{n\geq 2}$ with uniform link weight $d(x^i,x^j) > 0$ is irreducibly isometrically embeddable in $\mathbb{E}^{n-1}$ and in $M^{n-1}_{\kappa<0}$. The same graph is isometrically embeddable in $M^{n-1}_{\kappa>0}$ if and only if*

$$\kappa \leq \left(\frac{\cos^{-1}\left(-\frac{1}{n-1}\right)}{d(x^i,x^j)}\right)^2 \tag{10.3}$$

Furthermore, it is irreducibly isometrically embeddable in $M^{n-2}_{\kappa>0}$ for

$$\kappa = \left(\frac{\cos^{-1}\left(-\frac{1}{n-1}\right)}{d(x^i,x^j)}\right)^2 \tag{10.4}$$

Proof. Since Euclidean embbedding is scalable, it can be assumed that the graph has unit link weight. By subtracting the first column from the last column of $D(x^1,...,x^k)$, we get $\det D(x^1,...,x^k) = (-1)\det D(x^1,...,x^{k-1})$. Since, obviously, $\det D(x^1,x^2) = 1$ and $\det D(x^1,x^2,x^3) = -1$, the proof follows by induction.

For embeddability in hyperbolic space, set $c = \cosh\left(d(x^i,x^j)\sqrt{-\kappa}\right) > 1$, and then the lemma yields $\det\Delta_\kappa(x^1,...,x^k) = (1-c)^{k-1}((k-1)c+1)$. The principal minors of $\Delta_\kappa(x^1,...,x^n)$ clearly never vanish and their signs have the required alternating property, from which irreducible isometric embedding follows.

For embeddability in spherical space, set $c = \cos\left(d(x^i,x^j)\sqrt{\kappa}\right) \leq 1$ and then the lemma yields $\Delta_\kappa(x^1,...,x^k) = (1-c)^{k-1}((k-1)c+1)$. Isometric embeddability is hence equivalent to the sequence $(k-1)c+1$, $k = 1,...,n$ being positive, with possibly a vanishing tail. $(k-1)c+1 \geq 0$ is clearly equivalent to

$$\kappa \leq \left(\frac{\cos^{-1}\left(-\frac{1}{k-1}\right)}{d(x^i,x^j)}\right)^2 \tag{10.5}$$

and since

$$\cos^{-1}\left(-\frac{1}{n-1}\right) < \cos^{-1}\left(-\frac{1}{k-1}\right), \quad k < n$$

it follows that $\det \Delta_\kappa(x^1,...,x^k) = (1-c)^{k-1}((k-1)c+1)$ could only possibly vanish for $k=n$ and is positive for $k<n$. Hence the graph is isometrically embeddable iff (10.5) is satisfied $\forall k \leq n$, which is equivalent to (10.3). The irreducible isometric embedding in M_κ^{n-2} requires, in addition, that $(n-1)c+1 = 0$, which is equivalent to (10.4). ■

Remark 1 *Observe that an alternate proof of embeddability for all three cases is provided by a combination of Lemma 5 and Corollary 7.*

It is instructive to provide some "visual geometry" interpretation of the embedding of the uniformly weighted complete graph in a sphere. Consider first three points v^1, v^2, v^3 with uniform distance d. Clearly, this structure is embeddable in any S^2 sphere of radius $r \geq \frac{3d}{2\pi}$, that is, of curvature $\kappa \leq \left(\frac{2\pi}{3d}\right)^2$, consistently with (10.3). Furthermore, the same structure is irreducibly isometrically embeddable in the sphere S^1 of radius $r = \frac{3d}{2\pi}$, that is, of curvature $\kappa = \left(\frac{2\pi}{3d}\right)^2$, consistently with (10.4).

Consider the embedding of four points v^1, v^2, v^3, v^4, with uniform distance $d = d(v^i, v^j)$, $i \neq j$, in the sphere $S_r^2 \subseteq \mathbb{R}^3$ of radius r, hence of curvature $\kappa = \frac{1}{r^2}$. In $\mathbb{R}^3$, these four points form the vertices of a regular tetrahedron with edges $[v^iv^j]_{\mathbb{R}^3}$, written more simply as v^iv^j. Geometrically, the embedding problem in S_r^2 consists in finding the unique sphere that passes through the vertices of the regular tetrahedron and such that its arcs of great circles $[v^iv^j]_{S^2}$, $i \neq j$, have length d, that is, $d = r\angle v^iOv^j$, where O denotes the center of the sphere circumscribed to the tetrahedron. Since O is the point equidistant to v^1, v^2, v^3, v^4, the center O of the sphere lies on the segment $v^1v^1_\perp$ perpendicular to the plane $v^2v^3v^4$. By the same argument, O lies on the segment $v^2v^2_\perp$ orthogonal to $v^1v^3v^4$. Hence $O = v^1v^1_\perp \cap v^2v^2_\perp$. By elementary geometry, $v^1v^1_\perp$ and $v^2v^2_\perp$ are both in the plane v^1v^2w orthogonal to the segment v^3v^4 and intersecting this segment at $w \in v^3v^4$. Let $\alpha = \angle v^1wv^2$ be the dihedral angle of the tetrahedron. Recall that by elementary (Euclidean) geometry $\ell_{\mathbb{R}^3}(v^2v^1_\perp) = 2\ell_{\mathbb{R}^3}(v^1_\perp w)$; next, a little bit of elementary (rectilinear) trigonometry in the triangle v^1v^2w yields $\alpha = \cos^{-1}\frac{1}{3}$; finally a little bit more of (Euclidean) geometry in v^1v^2w yields $\angle v^1Ov^2 = \pi - \alpha$. Thus the relation between d and r reads $d = r(\pi - \cos^{-1}\frac{1}{3})$, that is, $\frac{1}{r} = \frac{\pi - \cos^{-1}\frac{1}{3}}{d} = \frac{\cos^{-1}(-\frac{1}{3})}{d}$, which is fully consistent with (10.4).

10.6.2 star

We now look at embeddability in another basic graph structure, that of a star. More precisely, we have a central vertex v^{n+1} with n peripheral vertices $v^1,...,v^n$ all connected to v^{n+1} with uniformly weighted links. If we observe that $d(v^i,v^j) = 2d(v^i,v^{n+1})$ for $i,j \leq n$, the ordering $v^1,...,v^n,v^{n+1}$ of the vertices is justified by the fact that the first n vertices form, as far as the distance structure is concerned, a complete graph structure with uniform link weight. Let $T_n(d,c)$ denote the $n \times n$ Toeplitz matrix with d's on the diagonal and c's off the diagonal. Embeddability of the $v^1, ..., v^{n+1}$ system relies on the matrix

$$T = \begin{pmatrix} T_n(1, \cosh(2d(v^i, v^j)\sqrt{-\kappa})) & e_n^T \cosh(d(v^i, v^j)\sqrt{-\kappa})) \\ e_n \cosh(d(v^i, v^j)\sqrt{-\kappa})) & 1 \end{pmatrix}, \quad i, j \leq n$$

By the previous theorem, the principal minors of $T_n(1, \cosh(2d(v^i, v^j)\sqrt{-\kappa}))$ have the correct sign, so that it remains to check the sign of the determinant of the above matrix.

Theorem 39 *The star structure $v^1, ..., v^{n+1}$ is embeddable in a hyperbolic space of sufficiently negative curvature κ.*

Proof. It suffices to check that $\det T$ has opposite sign to $\det T_n(1, \cosh(2d(v^i, v^j)\sqrt{-\kappa}))$. Remember that the Schur complement theorem says that

$$\det T = \\ \det T_n(1, \cosh(2d(v^i, v^j)\sqrt{-\kappa})) \\ \big(1 - e_n^T \cosh(d(v^i, v^j)\sqrt{-\kappa}))T_n^{-1}(1, \cosh(2d(v^i, v^j)\sqrt{-\kappa}))e_n \cosh(d(v^i, v^j)\sqrt{-\kappa}))\big)$$

Hence it suffices to check that

$$\big(1 - e_n^T \cosh(d(v^i, v^j)\sqrt{-\kappa}))T_n^{-1}(1, \cosh(2d(v^i, v^j)\sqrt{-\kappa}))e_n \cosh(d(v^i, v^j)\sqrt{-\kappa}))\big) < 0$$

At the $\kappa \to -\infty$ limit, the sign of the above becomes that of

$$\big(1 - e_n^T T_n^{-1}(0,1)e_n\big)$$

It is easily seen by direct verification that

$$T_n^{-1}(0,1) = \frac{1}{n-1} T_n(-(n-2), 1)$$

so that

$$\big(1 - e_n^T T_n^{-1}(0,1)e_n\big) = 1 - \frac{1}{n-1} < 0$$

and therefore the proof of embeddability in a sufficiently negatively curved space follows from an induction argument.

■

Theorem 40 *The above defined star with $n \geq 3$ is not embeddable in any positively curved space.*

Proof. Embeddability in constant positive curvature space is equivalent to

$$T = \begin{pmatrix} T_n(1, \cos(2d(v^i, v^j)\sqrt{\kappa})) & e_n^T \cos(d(v^i, v^j)\sqrt{\kappa})) \\ e_n \cos(d(v^i, v^j)\sqrt{\kappa})) & 1 \end{pmatrix} \geq 0$$

By the embeddability theorem for complete graph, we have, for some specific curvatures,

$$T_n(1, \cos(2d(v^i, v^j)\sqrt{\kappa})) \geq 0$$

Invoking again the Schur complement theorem, $T \geq 0$ only if

$$T_n(1, \cos(2d(v^i, v^j)\sqrt{\kappa})) - e_n en^T \cos^2(d(v^i, v^j)\sqrt{\kappa})) \geq 0$$

After some elementary manipulation, the above reduces to

$$T_n(\sin^2\left(d(v^i, v^j)\sqrt{\kappa}\right), -\sin^2\left(d(v^i, v^j)\sqrt{\kappa}\right)) = \sin^2\left(d(v^i, v^j)\sqrt{\kappa}\right) T_n(1, -1) \geq 0$$

By the lemma, the above is clearly impossible for $n \geq 3$. ■

10.6.3 core concentric structure

To look at the embeddability properties of such a core concentric structure as the ISP graph of Figure 13.1, consider a complete graph with vertices $v^1, ..., v^n$ with uniform link weight $d(v^i, v^j) = 1$ and to each vertex v^i let us attach a tendril of length ℓ ending up in the vertex v^{n+i}. We have the following theorem:

Theorem 41 *The above defined core concentric structure is embeddable in M_κ^{2n-1} provided $\ell\sqrt{-\kappa}$ is sufficiently large.*

Proof. Embeddability of this structure in a constant curvature $\kappa < 0$ space relies on the matrix

$$\left(\begin{array}{ccc|ccc} 1 & \dots & c1 & c_\ell & \dots & c_{\ell+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_1 & \dots & 1 & c_{\ell+1} & \dots & c_\ell \\ \hline c_\ell & \dots & c_{\ell+1} & 1 & \dots & c_{2\ell+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_{\ell+1} & \dots & c_\ell & c_{2\ell+1} & \dots & 1 \end{array}\right) = \left(\begin{array}{cc} T(1, c_1) & T(c_\ell, c_{\ell+1}) \\ T(c_\ell, c_{\ell+1}) & T(1, c_{2\ell+1}) \end{array}\right)$$

where $c_1 = \cos(\sqrt{-\kappa})$, $c_\ell = \cosh(\ell\sqrt{-\kappa})$, $c_{\ell+1} = \cos((\ell+1)\sqrt{-\kappa})$, and $c_{2\ell+1} = \cosh((2\ell+1)\sqrt{-\kappa})$. The congruence relation

$$\begin{aligned} &\left(\begin{array}{cc} T(1, c_1) & T(c_\ell, c_{\ell+1}) \\ T(c_\ell, c_{\ell+1}) & T(1, c_{2\ell+1}) \end{array}\right) \\ &\quad \sim \left(\begin{array}{cc} T(1, c_1) & 0 \\ 0 & T(1, c_{2\ell+1}) - T(c_\ell, c_{\ell+1})T^{-1}(1, c_1)T(c_\ell, c_{\ell+1}) \end{array}\right) \end{aligned}$$

indicates that it suffices to argue on the sign of the eigenvalues of the block diagonal matrix. Since the complete graph core is embeddable, $T(1, c_1)$ has $(n-1)$ negative eigenvalues and one positive eigenvalue. Hence it suffices to show that the Schur complement $T(1, c_{2\ell+1}) - T(c_\ell, c_{\ell+1})T^{-1}(1, c_1)T(c_\ell, c_{\ell+1})$

has n negative eigenvalues as $\ell\sqrt{-\kappa} \to \infty$. Elementary computation shows that the limit of the $n \times n$ Schur complement is

$$\begin{pmatrix} 1-c_\ell^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1-c_\ell^2 \end{pmatrix}$$

Since $c_\ell \to \infty$, the theorem is proved. ■

10.6.4 tree

As elementary counterexamples prove, a tree is not in general embeddable in a standard *finitely* negatively curved space. However, the fact that a tree is Gromov hyperbolic for $\delta = 0$ provides the clue that a tree may be embeddable, in some limiting sense, in a infinitely negatively curved space.

If we are given a distance system $d(v^i, v^j)$ and if

$$\max_{i,j} d(v^i, v^j) = d(v^{i^*}, v^{j^*}) = d^*$$

is unique, then the dominant terms in the matrix $\Delta_\kappa(v^1, v^2, ...)$ are $\cosh(d^*\sqrt{-\kappa})$ in the (i^*, j^*) and (j^*, i^*) positions. More precisely,

$$\lim_{\kappa\to-\infty} \frac{1}{\cosh(d^*\sqrt{-\kappa})}\Delta_\kappa(v^1, v^2, ...) = \begin{pmatrix} 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1 & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 0 \end{pmatrix} \sim \left(\begin{array}{cc|c} 0 & 1 & 0 \\ 1 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right)$$

Clearly, the limiting matrix, which is the embeddability matrix of the line segment $v^{i^*}v^{j^*}$ of distance $d(v^{i^*}v^{j^*})$, has the correct eigenvalue structure so that the tree structure is embeddable is an infinitely negatively curved space.

Consider now the case of a uniform tree of degree k; typically, the Cayley graph of a free group on k generators. Let every link be assigned the weight ℓ and consider the subtree consisting of all nodes at a distance less than or equal to ℓl from the root v^0. Clearly, the longest distance that can be achieved on this tree occurs for a pair of nodes at maximal distance ℓl from the root and such that their shortest connecting path goes thorough the root. This maximum achievable distance is $2\ell l$. The set of k^l nodes at a maximal distance from the root decomposes into k subsets $S_1, ..., S_k$ of k^{l-1} modes such that the distance between two nodes in the same subset is less than $2\ell l$ and the distance between two nodes in different subsets is exactly $2\ell l$. Clearly, each such subset S_i has its shortest path to the root crossing one and exactly one of the k nodes that spring off the root. Clearly, the submatrix of $\Delta_\kappa(v^0, v^1, v^2, ..., v^{k^l})$ dominant as

$\kappa \to -\infty$ corresponds to indices in the subsets $S_1, ..., S_k$ and as such is congruent to the $k \times k$ block matrix

$$\tilde{\Delta}(k, C_{l-1}) = \begin{pmatrix} 0 & C_{l-1} & \dots & C_{l-1} \\ C_{l-1} & 0 & \dots & C_{l-1} \\ \vdots & \vdots & \ddots & \vdots \\ C_{l-1} & C_{l-1} & \dots & 0 \end{pmatrix}$$

where C_{l-1} is the $(l-1) \times (l-1)$ matrix fully populated with $\cosh(2\ell l\sqrt{-\kappa})$:

$$C_{l-1} = \begin{pmatrix} \cosh(2\ell l\sqrt{-\kappa}) & \dots & \cosh(2\ell l\sqrt{-\kappa}) \\ \vdots & \ddots & \vdots \\ \cosh(2\ell l\sqrt{-\kappa}) & \dots & \cosh(2\ell l\sqrt{-\kappa}) \end{pmatrix}$$

It remains to show that this matrix has the correct signature for its eigenvalues. To this effect, we slightly generalize the above structure to the $k \times k$ block structure in which each block C_m has size $m \times m$,

$$\tilde{\Delta}(k, C_m) = \begin{pmatrix} 0 & C_m & \dots & C_m \\ C_m & 0 & \dots & C_m \\ \vdots & \vdots & \ddots & \vdots \\ C_m & C_m & \dots & 0 \end{pmatrix}$$

where C_m is the $m \times m$ matrix

$$C_m = \begin{pmatrix} c & c & \dots & c \\ c & c & \dots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \dots & c \end{pmatrix}$$

We need some preliminary results. Define the $k \times k$ Toeplitz matrix

$$T_k(0, c) = \begin{pmatrix} 0 & c & \dots & c \\ c & 0 & \dots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \dots & 0 \end{pmatrix}$$

and let

$$\tau_k = \det T_k(0, c)$$

Lemma 10 *The recursion on the principal minors of the Toeplitz-structured matrix $T_k(0, c)$ is*

$$\tau_{k+1} = -c\tau_k + c^2\tau_{k-1} + c^3\tau_{k-2}$$

subject to the initial conditions

$$\begin{aligned} \tau_1 &= 0 \\ \tau_2 &= -c^2 \\ \tau_3 &= 2c^3 \end{aligned}$$

Furthermore, the solution to the above recursion is given by

$$\tau_k = (-1)^{k-1}(k-1)c^k, \quad k \geq 1$$

Proof. The proof follows the same lines as that of Lemma 9 and is omitted. ■

Proposition 4 *The characteristic polynomial of the matrix* $\tilde{\Delta}(k, C_m)$ *is*

$$s^{k(m-1)}\left(s+mc\right)^{k-1}\left(s-m(k-1)c\right)$$

Proof. First, observe that $\tilde{\Delta}(k, C_m)$ has rank k. Next, remember that the coefficient of s^{km-i}, $i \geq 1$, in the characteristic polynomial of $\tilde{\Delta}(k, C_m)$ is $(-1)^i$ times the sum of all its principal minors of order i. Because the rank of $\tilde{\Delta}(k, C_m)$ is k, only those principal minors up to order k are nonvanishing. To construct a nonvanishing minor of order i, only one row or column index can be chosen in every block row or column. For each size i, there are obviously $\binom{k}{i}m^i$ such nonvanishing principal minors. By lemma 10, all such minors equal $(-1)^{i+1}(i-1)c^i$. Therefore, the characteristic polynomial is

$$s^{km} - \sum_{i=2}^{k}\binom{k}{i}m^i(i-1)c^i s^{km-i}$$

Now observe that

$$\begin{aligned}
&-\binom{k}{i}(i-1) \\
&= \frac{(k-1)!(k-i-ki+i)}{i!(k-i)!} \\
&= \frac{(k-1)!(k-i)-(k-1)(k-1)!i}{i!(k-i)!} \\
&= \binom{k-1}{i} - (k-1)\binom{k-1}{i-1}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&s^{km} - \sum_{i=2}^{k}\binom{k}{i}m^i(i-1)c^i s^{km-i} \\
&= s^{km} - \sum_{i=2}^{k}\left(\binom{k-1}{i} - (k-1)\binom{k-1}{i-1}\right)m^i c^i s^{km-i} \\
&= s^{k(m-1)}\left(s^k - \sum_{i=2}^{k}\left(\binom{k-1}{i} - (k-1)\binom{k-1}{i-1}\right)m^i c^i s^{k-i}\right) \\
&= s^{k(m-1)}\left(\sum_{i=0}^{k-1}\binom{k-1}{i}s^{k-1-i}m^i c^i\right)\left(s-m(k-1)c\right) \\
&= s^{k(m-1)}\left(s+mc\right)^{k-1}\left(s-m(k-1)c\right)
\end{aligned}$$

■

Remark 2 *The same result can also be proved by observing that*

$$\tilde{\Delta}(k, C_m) = \begin{pmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 0 \end{pmatrix} \otimes C_m$$

and by remembering that $\left\{\lambda_l(\tilde{\Delta}(k, C_m))\right\} = \{\lambda_i(T_k(0,1))\lambda_j(C_m)\}$. *The eigenvalues of* C_m *are clearly* $\{0, mc\}$ *and those of* $T_k(0,1)$ *are found to be* $\{-1, k-1\}$.

Corollary 8 *The uniform tree of degree* k *is embeddable in an infinitely negatively curved space.*

10.7 dual theory

The preceding embeddability theory is based on *distances* among a set of vertices. For every of the three curvature cases–negative, vanishing, positive– the theory provides a matrix such that, if some sign definiteness conditions are satified, the vertices are embeddable in the corresponding constant curvature space. One already perceives the problem that it might take up to the construction of 3 matrices before being able to determine in what space the set of vertices is embeddable. The dual theory proceed from the *dihedral angles*, and has the advantage that, given a system of dihedral angles, *one single* Gram matrix allows us to infer whether the geometry if hyperbolic, Euclidean, or spherical. Its disadvantage is that it applies to simplexes only and as such is a rather local theory.

In the ordinary $\mathbb{E}^3$ space, the concept of dihedral angles along an edge x^ix^j of a simplex $x^1x^2x^3x^4$ is pretty trivial: it is the angle between the faces incident upon x^ix^j. Already observe that it can also be defined as the complement relative to π of the usual angle between the lines orthogonal to the faces adjacent to x^ix^j. We now proceed more formally:

Definition 50 *Consider an* n*-simplex* $x^1...x^{n+1}$ *in the constant curvature space* M_κ. *Let* σ_{n-1} *be a codimension 2 subsimplex, and let* x^k, x^l *be the complement of the set of vertices of* σ_{n-1} *relative to* $x^1, ..., x^{n+1}$. *Then the dihedral angle* $\alpha(\sigma_{n-2})$ *is defined as* $\pi - \angle((x^k\sigma_{n-2})^\perp, (x^l\sigma_{n-2})^\perp)$, *where* $(x^k\sigma_{n-2})^\perp$ *denotes the geodesic orthogonal to all edges in* $x^k\sigma_{n-1}$, *with a similar definition for* $(x^l\sigma_{n-2})^\perp$.

If α_{ij}, $i \neq j$, is the dihedral angle of the edge x^ix^j, and if we agree that $\alpha_{ii} = \pi$, then what dictates the geometry is the Gram matrix $G = \{-\cos\alpha_{ij}\}$ (see [75, 57]).

$$G = \begin{pmatrix} 1 & -\cos\alpha_{12} & -\cos(\alpha_{13}) & -\cos\alpha_{14}) \\ -\cos\alpha_{12} & 1 & -\cos\alpha_{23} & -\cos\alpha_{24} \\ -\cos\alpha_{13} & -\cos\alpha_{32} & 1 & -\cos\alpha_{34} \\ -\cos\alpha_{14} & -\cos\alpha_{24} & -\cos\alpha_{34} & 1 \end{pmatrix}$$

Precisely, the geometry if hyperbolic iff $\lambda_1(G) < 0 < \lambda_2(G) \leq \lambda_3(G) \leq \lambda_4(G)$ and all (i, j) cofactors are positive; the geometry is Euclidean iff $\lambda_1(G) = 0 < \lambda_2(G) \leq \lambda_3(G) \leq \lambda_4(G)$ with the same condition on the cofactors; finally, the geometry is spherical iff $G > 0$.

Chapter 11

comparison geometry

“Comparison geometry” refers to the simple, but far reaching, idea that, if we isometrically embed an arbitrary triangle of some metric space in a standard Riemannian space, the resulting comparison triangle can be given such inner product concepts as angles and curvature. Because in comparison geometry the fundamental object is, not unlike traditional Greek geometry, the triangle, it is a local theory, especially as far as the angles are concerned. As far as curvature is concerned, one way to go from local to global would be to find a curvature bound for all triangles of the metric space, with the limitation that the overall metric space can only be given a curvature bound.

11.1 CAT comparison theory

The comparison theory was historically the first attempt at defining curvature of metric spaces and is usually credited to Alexandrov [50, Sec. 2.2], although it has roots tracing back to Menger (see [50, Sec. 2.2], [15, Sec. 40], and Sec. 10.5).

Definition 51 *Given a geodesic triangle ABC in the geodesic space (X, d), the* **comparison triangle** *[12, p. 19] in the standard constant curvature space $(M_\kappa, \bar{d})$ is a triangle $\bar{A}\bar{B}\bar{C}$ such that*

$$\begin{array}{rcl} \bar{d}(\bar{A},\bar{B}) & = & d(A,B) \\ \bar{d}(\bar{B},\bar{C}) & = & d(B,C) \\ \bar{d}(\bar{C},\bar{A}) & = & d(C,A) \end{array}$$

The space (X, d) is said to be of curvature bounded by κ if the metric properties of the triangle in (X, d) are bounded by those of the comparison triangle in M_κ. There are several precise materializations of this idea. Probably the most popular one is the CAT-theory, where “CAT” stands for Cartan-Alexandrov-Toponogov [52, Sec. 3.2] [12, p. 19], [54, Th. VIII.4.1].

Definition 52 *Take points $Z \in [AB]$, $Y \in [AC]$ along with their comparison points in M_κ. If $\kappa < 0$, then (X, d) is said to be a $CAT(\kappa)$-space, i.e., has curvature bounded from above by $\kappa < 0$ iff*

$$d(X,Y) \leq \bar{d}(\bar{Z},\bar{Y})$$

If $\kappa > 0$, then (X, d) is said to be a $CAT(\kappa)$-space, i.e., has curvature bounded from below by $\kappa > 0$ iff

$$d(X,Y) \geq \bar{d}(\bar{Z},\bar{Y})$$

(see [23, Sec. 4.1.4]).

With the above concepts, we can define a geodesic space to be negatively curved at low scale if the sum of the Alexandrov angles at the vertices of sufficiently small geodesic triangles ABC is less than π. With some caution, a similar statement can be made for positively curved spaces [23, 4.1.5].

(More generally, the Rauch-Toponogov comparison theorems [12, p. 19], [54, Th. VIII.4.1] compare metric properties of Riemannian spaces of different curvatures.)

11.2 Alexandrov angle

The comparison triangle also allows a concept of angle to be defined *solely in terms of the distance, independently of the concept of inner product.* The *Alexandrov angle* [21, Def. 1.12], [23, 4.3] at the vertex A of a geodesic triangle $\triangle ABC$ is the $\overline{\lim}_{\epsilon \to 0}$ ($\underline{\lim}_{\epsilon \to 0}$) of the angle at the vertex $\bar{A}$ of the comparison triangle $\bar{A}\bar{B}_\epsilon\bar{C}_\epsilon$ of $[AB_\epsilon C_\epsilon]$ in the standard negative (positive) curvature space M_κ, where $B_\epsilon \in [AB]$, $C_\epsilon \in [AC]$, and $d(A, B_\epsilon) = \epsilon d(A, B)$, $d(A, C_\epsilon) = \epsilon d(A, C)$. The angle $\angle \bar{B}_\epsilon \bar{A} \bar{C}_\epsilon$ depends on the metric of M_κ, but remarkably the limit as $\epsilon \downarrow 0$ does not depend on what comparison space is chosen [21, Prop. 2.9].

The Alexandrov angle is an *extremely small scale* concept, and as such it does not directly apply to graphs. Indeed if A is an arbitrary vertex in a graph, the Alexandrov angle between the links $[AB]$ and $[AC]$ would be 180 deg. At such a small scale, the geometry of the graph is completely lost so that this result is not surprising. For a graph, the lowest possible scale that can possibly provide some geometric insight in the neighborhood of a vertex A is the scale of vertices directly linked to A. At that scale, a nontrivial Alexandrov angle can be defined.

The Alexandrov notion of angle will play a crucial role in assessing under what conditions a hyperbolic graph can be *isometrically* embedded in a compact surface of genus $g > 2$ carrying a negative curvature.

11.3 example of application of Alexandroff angles: isometric embedding on surface

If $g > 2$, then it is well-known that S_g carries a hyperbolic metric g_{ij} of constant sectional curvature. If the graph G is hyperbolic, the question is whether the graph G can be *isometrically* embedded into (S_g, g_{ij}). To assess the "obstructions" in doing so, consider a situation where a vertex of the graph A is connected to vertices B_i, $i = 1, ..., n$, B_i is connected to B_{i+1}, and B_n is connected to B_1, with weights $d(A, B_i) = 1$, $d(B_i, B_{i+1}) = 1$, $d(B_n, B_1) = 1$. Let α_i be the "angle" at the vertex A of the (geodesic) triangle AB_iB_{i+1} and let α_n be the "angle" at the vertex A of the triangle AB_nB_1. Since the graph is to be embedded into a surface of constant sectional curvature $\kappa < 0$, the angles at the vertex A of the *graph* are most naturally defined following the procedure of Alexandrov in a comparison triangle in the model space of constant sectional curvature κ, M_κ [50], [21, Definition 2.15]. (In fact, a deeper result [21, Prop. 2.9] shows that, no matter what comparison space M_κ we choose, the angle is the same and hence equal to 60 degrees.) If $\sum_i \alpha_i = 2\pi$, then the subgraph with vertices A, B_i can be isometrically embedded into S_g. If, however, $\sum_i \alpha_i > 2\pi$, then the isometric embedding of the subgraph yields a *pleat* singularity at A and the surface is said to have a *singular hyperbolic metric* (see [52, Chap. 14]). If, on the other hand, $\sum_i \alpha_i < 2\pi$, then the isometric embedding of the subgraph must have a *conical singularity* at A (see [52, Sec. 61, 6.2]), for indeed the sum of the angles at the apex of a cone is $< 2\pi$. By the Gauss-Bonnet theorem, it follows that the cone has positive sectional curvature locally around its apex, so that the isometric embedding cannot be done on a surface of uniformly negative curvature.

11.4 higher dimensional comparison

A problem with the previous analysis, along with its definition of positively and negatively curved graphs, is that it clings on the assumption that the graph has been topologically embedded in a surface. Here we indicate the way to carry over this analysis to graphs embedded in 3-manifolds.

There is a dramatic transition as we go from 2-D to 3-D, because indeed, while a metric triangle, that is, a triple of points with their distances satisfying the triangle inequality, is embeddable in any space of any curvature, the basic 3-D building block, the metric tetrahedron, that is, a quadruple of points such that every of its triple satisfies the triangle inequality, is *not always* isometrically embeddable in an arbitrarily curved space (see Sec. 10.5). In addition, the 2-dimensional geometry fact that the sum of the internal angles of a triangle dictates the curvature is more complicated in 3 dimensions.

11.5 nonpositively curvature in the sense of Buseman

The concept of nonpositive curvature in the sense of Buseman is a very primitive one that does not even resort to the concept of comparison triangle.

Definition 53 *Let A, B be two points that can be joined by a geodesic $\gamma : [0, d(A,B)] \to X$ parameterized by arc length. The midpoint between A, B is defined as*

$$m(A,B) = \gamma\left(\frac{d(A,B)}{2}\right)$$

Definition 54 *A geodesic space (X,d) is said to be nonpositvely curved in the sense of Buseman if $\forall A \in X$, there exists a δ_A such that, $\forall B, C \in B(A,\delta_A)$,*

$$d(m(A,B), m(A,C)) \leq \frac{1}{2} d(B,C)$$

Corollary 9 *Let $\triangle ABC$ be an arbitrary geodesic triangle in the space (X,d). Take $B_\lambda \in [AB]$, $C_\lambda \in [AC]$ such that $d(A,B_\lambda) = \lambda d(A,B)$, $d(A,C_\lambda) = \lambda d(A,C)$. Assume that there exists a shortest geodesics $[B_\lambda C_\lambda]$ continuously depending on λ. Then, if the geodesic space (X,d) is nonpositively curved in the sense of Buseman, we have*

$$d(B_\lambda, C_\lambda) \leq \lambda d(B,C).$$

Proof. See [50, Cor. 2.2.1]. ■

11.6 negative curvature in the sense of Alexandrov

The concept of nonpositive curvature in the sense of Alexandrov refers to the behavior of the distance function between a point and a geodesic.

Chapter 12

Finsler geometry

Appendix A

Coarse metric geometry

In this chapter, we allow the metric structure of a graph, or any topological space for that matter, to be time-varying, or uncertain. This is motivated by the fact that the link costs of a communications graph are periodically readjusted to take into consideration recently observed congestion, delay, outages, etc. Conceptually, we are dealing with a graph that has a fixed underlying topological structure, but an uncertain metric structure, a concept which is here formalized as a coarse structure. The latter concept leads to coarse homology, a generalized homology that picks up only the coarse structure of a space irrespective of the "details."

Next to the above purely metric approach, there is an algebraic approach which aims at coarsening a space by trading it with a noncommutative algebra. The connection between the metric and the algebraic approaches is formulated in the celebrated coarse Baum-Connes and Novikov conjectures.

A.1 coarse map and coarse structure

Definition 55 *A not necessarily continuous function $f : X \to Y$ between metric spaces X, Y is said to be a* **coarse map** *if the following two properties hold:*

1. *Bounded Expansiveness: For any $c_x > 0$ there exists a $c_y > 0$ such that*

$$d_x(x_1, x_2) \leq c_x \Rightarrow d_y(f(x_1), f(x_2)) \leq c_y$$

2. *Metric Properness: For each bounded subset $B \subseteq Y$, $f^{-1}(B)$ is bounded.*

Definition 56 *Two coarse maps $f, f' : X \to Y$ between metric spaces X, Y are said to be coarsely equivalent if there exists a constant c such that $d_y(f(x), f'(x)) \leq c$. Two spaces X, Y are said to be coarsely equivalent if there exist coarse maps $f : X \to Y$ and $f^\dagger : Y \to X$ such that $ff^\dagger$ and $f^\dagger f$ are coarsely equivalent to 1_X, 1_Y, respectively, in which case $f^\dagger$ is called coarse inverse of f.*

Clearly, the property that two coarse functions are "coarsely equivalent" is an equivalence relation. The same applies to two spaces being "coarsely equivalent."

Observe that any two finite diameter spaces are coarsely equivalent, so that the definition really makes sense only for infinite diameter spaces.

As an illustration, it is easily seen that $\mathbb{Z}^n$ and $\mathbb{R}^n$ are coarsely equivalent. Observe, however, that $\mathbb{N} \times \mathbb{Z}$ and $\mathbb{R}^2$ are not coarsely equivalent.

Consider now a topological space X that can be endowed with several different metrics. Probably the best illustrative example most closely related to networking is a graph G, specified topologically by its vertices, its edges and and interconnection matrix, and for which different link costs can be assigned. Given two metrics d_1, d_2, the metric spaces (X, d_1), (X, d_2) may or may not be coarsely equivalent. However, one might be able to find a collection of metrics such that the resulting metric spaces are all coarsely equivalent. This leads to the following:

Definition 57 *A coarse structure on X is an equivalence class of distances d_i such that, for any two such distances d_i, d_j, the metric spaces (X, d_i), (X, d_j) are coarsely equivalent.*

A.1.1 quasi-isometry versus coarse equivalence

It is necessary to stress that the "coarse geometry" school of thoughts have concepts subtlety different than those of the "quasi-isometry" school of thoughts. Certainly, a quasi-isometric embedding is a coarse map, but the converse is not true. Indeed, consider the subset $\mathbb{L} := \{\pm 2^k : k \in \mathbb{N}^*\} \cup \{0\} \subseteq \mathbb{Z}$ along with the function $f : \mathbb{L}^* \ni \pm 2^k \mapsto \pm k \in \mathbb{Z}^*$, $0 \mapsto 0$. $\mathbb{L}$ is endowed with the metric it inherits from $\mathbb{Z}$ [1]. Clearly, f is a coarse map, f^{-1} is obviously a coarse inverse, and $|ff^{-1}(x) - x| = |f^{-1}f(x) - x| = 0$, so that f is a coarse equivalence. But it is not even a quasi-isometric embedding. Indeed, while f clearly has bounded distortion, that is, $|f(x_1) - f(x_2)| \leq \frac{1}{2}|x_1 - x_2|$, the reverse inequality $\frac{1}{\lambda}|x_1 - x_2| - \epsilon \leq |f(x_1) - f(x_2)|$ cannot hold for any $\epsilon, \lambda < \infty$.

A more spectacular counterexample is provided by the embedding of the lattice $\mathbb{N}$ as a "square spiral" in $\mathbb{Z}^2$, as shown in Fig. A.1. Again, the embedding $f : \mathbb{N} \to \mathbb{Z}^2$ is a coarse equivalence, but it is not a quasi-isometric embedding. Indeed, as before, the upper bound $d(f(x_1), f(x_2)) \leq |x_1 - x_2|$ holds. To see this, let $y^k : k = 0, 1, ..., n$ be the successive $\mathbb{Z}^2$ lattice points joining $f(x_1) = y^0$ to $f(x_2) = y^n$ by following the spiral. Using the triangle inequality,

$$d(f(x_1), f(x_2)) \leq \sum_{k=0}^{n-1} d(y^k, y^{k+1}) = n = |x_1 - x_2|$$

The problem is the upper bound; indeed, asymptotically,

$$0 < \lim_{|x_1 - x_2| \to \infty} \frac{|x_1 - x_2|}{d^2(f(x_1), f(x_2))} < \infty$$

[1] This robuster counterexample was developed in cooperation with Prof. Hespanha and Mr. Barooah.

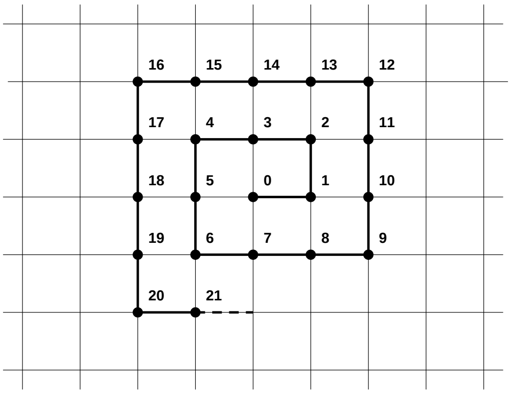


Figure A.1: The embedding $f : \mathbb{N} \to \mathbb{Z}^2$.

so that the lower bound $\frac{1}{\lambda}|x_1 - x_2| - \epsilon \leq d(f(x_1), f(x_2))$ could not hold.

Nevertheless, a coarse equivalence can be written in a way that resembles the quasi-isometric embedding.

Theorem 42 *Let X, Y be metric spaces. Then $f : X \to Y$ is a coarse embedding iff there exist functions $\underline{\rho}, \overline{\rho} : \mathbb{R}^+ \to \mathbb{R}^+$ with $\underline{\rho}(s) \to \infty$ as $s \to \infty$ such that*

$$\underline{\rho}(d(x_1, x_2)) \leq d(f(x_1), f(x_2)) \leq \overline{\rho}(d(x_1, x_2))$$

Proof. See Roe, transparencies. ■

In the one-dimensional lattice example $f : \mathbb{L} \to \mathbb{Z}$, clearly $\underline{\rho}$ could be taken as $\log_2$, but $\underline{\rho}$ cannot be taken linear as required by the quasi-isometric embedding concept.

In the other example $f : \mathbb{N} \to \mathbb{Z}^2$, the generalized quasi-isometric formulation holds; it indeed suffices to take $\underline{\rho}(s)$ asymptotically $\sqrt{s}$.

A.2 coarsening

Let G be some kind of complicated discrete space, like a large graph, that is difficult to analyze by combinatorial methods for the very reason of the curse of dimensionality. It is tempting to do the analysis on some kind of “faithful” continuous geometry model X of G. This is the concept of coarsening of the space G.

Definition 58 *Let G be a metric space. Then a coarsening of G, EG, is a metric space such that the following facts hold:*

1. *EG is a metric simplicial complex, that is, a simplicial complex equipped with a distance which on the simplexes coincides with the usual distance.*

2. *EG has bounded geometry, that is, there exists a coarse equivalence $f : EG \to Z$, where Z is a discrete metric space such that, $\forall r > 0$, $\# B_r(z) \leq b(r)$, $\forall z \in Z$, where $b(r) < B$.*

3. *EG is uniformly contractible, that is, $\forall x \in EG$, $\forall r > 0$, there exists an $R > r$ such that the inclusion $B_r(x) \to B_R(x)$ is nullhomotopic.*

4. *G and EG are coarsely equivalent.*

It is easily seen that $\mathbb{R}^n$ together with its prismatic triangulation [48] is a bounded geometry, uniformly contractible, metric simplicial complex. Since $\mathbb{R}^n$ and $\mathbb{Z}^n$ are coarsely equivalent, it follows that $E\mathbb{Z}^n = \mathbb{R}^n$.

Let Γ be a discrete group with a finite classifying space $B\Gamma$. Then the universal cover of $B\Gamma$ is coarsening of Γ.

Definition 59 *Two (continuous) maps $f, g : X \to Y$ are said to be* **properly homotopic** *if there exists a proper map $F : X \times [0, 1] \to Y$ such that $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$. Two spaces X, Y are said to be* **properly homotopically equivalent** *if there exist maps $f : X \to Y$ and $f^\dagger : Y \to X$ such that $f^\dagger f$, $f f^\dagger$ are properly homotopic to $1_X, 1_Y$, respectively. In this case, $f^\dagger$ is called proper homotopic inverse.*

Theorem 43 *EX is unique up to proper homotopy equivalence. Any coarse map $f : X \to Y$ between coarse space induces a unique proper homotopy class $[f] : EX \to EY$. Thus, E is a functor from the coarse category to the proper homotopy category, as illustrated by the following diagram:*

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d] & Y \arrow[r, "g"] \arrow[d] & Z \arrow[d] \\
EX \arrow[r, "{[f]}"] & EY \arrow[r, "{[g]}"] & EZ
\end{tikzcd}$$

Proof. Let E_1X and E_2X be two coarsenings. Since both of them are infinite simplicial complexes, their vertex sets have the same cardinality and hence there exists a bijection between the vertex sets. This vertex map induces a simplicial map between the two complexes. Finally, it is easily seen that this simplicial map has a proper homotopic inverse. Hence E_1X, E_2X are properly homotopically equivalent. The proof of the second claim proceeds along the same line and is omitted. ■

A.3 coarse homology

We would like to define a homology theory for space viewed at a distance, or space viewed through blurring lenses. The idea is that the homology theory should pick up the large scale features of the space, without the possibly cumbersome and irrelevant details of its local structure.

Definition 60 *Let K_* be a generalized homology theory [2]. Then the coarse homology of X, $KE_*(X)$, is defined as $K_*(EX)$.*

$K_*(EX)$ is indeed uniquely defined, since EX is defined up to homotopy equivalence and homology is defined up to homotopy equivalence. Furthermore, coarse homology is functorial, as the composition of the two functors E and K_*, as illustrated by the following diagram:

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "E"'] & Y \arrow[r, "g"] \arrow[d, "E"'] & Z \arrow[d, "E"'] \\
EX \arrow[r, "{[f]}"] \arrow[d, "K_*"'] & EY \arrow[r, "{[g]}"] \arrow[d, "K_*"'] & EZ \arrow[d, "K_*"'] \\
KE_*X \arrow[r, "{[f]_*}"] & KE_*Y \arrow[r, "{[g]_*}"] & KE_*Z
\end{tikzcd}$$

In other words, coarse homology is a homology theory.

A.4 Noncommutative geometry

It is well-known that a compact topological space X can be analyzed from the commutative algebra $C(X)$ of continuous complex-valued functions defined on it. Conversely, the Gelfand-Naimark theorem 45 asserts that any commutative algebra A is isometrically isomorphic to some $C(X)$. Still along the same line of ideas, the Swan-Serre theorem asserts that the algebraic (or the C^*-algebra) K-theory of the commutative algebra $C(X)$ is isomorphic to the K-groups of the vector bundles defined over X.

As a warm up exercise, we begin by reviewing the above commutative features, and then we show that replacing a commutative algebra by a noncommutastivbe one is equivalent to coarsening tyher space.

A.4.1 preview: trading spaces for commutative algebras

Gelfand-Naimark theorem

In the commutative case, an essential tool is that of the *Gelfand transform.* Let A^* denote the Banach space of continuous linear functionals defined on A. A **character** is a multiplicative bounded linear functional defined on A. Let $M(A) \subseteq A^*$ denote the set of characters of A. (The reason for the notation $M(A)$ will become clearer later.) The **Gelfand transform** is defined as

$$\begin{array}{rcl} \mathcal{G} : A & \to & C(M(A)) \\ a & \mapsto & \hat{a} \end{array}$$

where

$$\hat{a}(\mu) = \mu(a)$$

We have the following theorem:

Theorem 44 (Gelfand-Naimark Representation Theorem) *Let A be a commutative (unital) C^*-algebra. Then the Gelfand transform is an isometric $*$-isomorphism $\mathcal{G} : A \xrightarrow{\cong} C_0(M(A))$ ($\mathcal{G} : A \xrightarrow{\cong} C(M(A))$).*

Proof. See [35, Th. 2.2.2] or [39, Th. 1.4]. ■

We provide a simple illustration of this theorem:

Corollary 10 *If A is a unital C^* algebra generated by one single element, a, then $A \cong C(spec(a))$.*

Proof. For any character μ, $\mu(a) \neq 0$, and the multiplicative property of the character yields $\mu(ae) = \mu(a) = \mu(a)\mu(e)$, which implies that $\mu(e) = 1$. Since a generates the algebra, the character μ is completely defined from its value λ at a. Since $\mu(a) = \lambda$ and $\mu(\lambda e) = \lambda$, it follows that $\mu(\lambda e - a) = 0$, from which it follows that $\lambda e - a$ could not be invertible; hence $\lambda \in \text{spec}(a)$. ■

The proposition that follows justifies the notation $M(A)$ for the set of characters. Before proving that proposition, we need a lemma.

Lemma 11 *Let A be a commutative algebra and I a maximal ideal. Then $A|I \cong \mathbb{C}$.*

Proof. That $A|I$ is a field is a standard algebraic result. It is claimed that there exists a $\lambda_{[a]}$ such that $[a] = \lambda[e]$. Assume not. Then $[a - \lambda e]$ is invertible $\forall\lambda$. Define $f(\lambda) = \mu((a - \lambda e)^{-1})$. It is clearly an analytic function of λ, which has no singularities; hence it is constant. Since $f(\infty) = 0$, $f(\lambda) = 0$. Thus $(a - \lambda e)^{-1} = 0$, a contradiction. Hence $a \mapsto \lambda_a$ provides the isomorphism $A|I \to \mathbb{C}$. ■

Proposition 5 *For a commutative unital algebra A, $M(A) \cong \mathcal{M}(A)$, the set of maximal ideals of A.*

Proof. Let μ be a character and observe that, for a unital algebra, $\mu(e) = 1$. Next, take $\mu(a) = 0$. Clearly, $\forall b \in A$, $\mu(ab) = \mu(a)\mu(b) = 0$, so that whenever $a \in \ker(\mu)$, it follows that $ab \in \ker(\mu)$, $\forall b \in A$; hence $\ker(\mu)$ is an ideal. To prove that it is maximal, let I be an ideal of which $\ker(\mu)$ is a proper subset. Take $i \in I \setminus \ker(\mu)$. It is claimed that i has an inverse, for otherwise, $\mu(ai) \neq 1$, $\forall a$, which means that $\mu(ia) = 0$, so that $ia \in \ker(\mu)$ and furthermore $i \in \ker(\mu)$, a contradiction. Choose j such that $\mu(ij) = 1$, so that $ij - e \in \ker(\mu) \subseteq I$. But $ij \in I$. Hence $e \in I$, and $I = A$, the full algebra, a contradiction. Conversely, let I be a maximal ideal. Clearly, $A|I$ is a field; in fact $A|I \cong \mathbb{C}$. Define a character μ such that $\mu(a)$ is the equivalent class of a in $A|I$. Clearly, $I = \ker(\mu)$. ■

Combining the previous two results, we obtain the following:

Theorem 45 (Gelfand-Naimark Representation Theorem) *Let A be a commutative unital C^*-algebra. Then there exists an isometric *-isomorphism $A \to C(\mathcal{M}(A))$, where $\mathcal{M}(A)$ is the set of maximal ideals of A.*

Proof. For a direct proof, see [88, Th 2.2.1]. ■

$C(X)$, with $X = \mathcal{M}(A)$, is called *Gelfand representation* of A.

Recall that an algebra A is said to be *local* if it has a *unique* maximal ideal. This terminology stems from the fact that in this case the Gelfand representation of the algebra is the set of continuous functions defined on a *point*.

Swan-Serre theorem

A purely algebraic way to codify the structure of $C(X)$ is via the algebraic K-group $K_0^{alg}(C(X))$ of isomorphism classes of projective modules over the ring $C(X)$; equivalently, the Grothendieck group of conjugacy classes of *idempotents* ($e = e^2$) in $M_\infty(C(X))$, the set of arbitrarily large matrices over $C(X)$ with finitely many nonvanishing elements [77, Sec. 1.2], [93, 5.B], [35, p. 185]. Another algebraic codification of $C(X)$ exploits the fact that, in addition to being a ring, $C(X)$ is a C^*-algebra, leading to the C^*-algebra K-group $K_0(C(X))$ of equivalent classes of *projections* ($p = p^2 = p^*$) over $M_\infty(C(X))$. Because every conjugacy class of idempotents contains a projection [93, 5.B.(d)], the two algebraic codifications are the same, viz., $K_0^{alg}(C(X)) = K_0(C(X))$. The Swan-Serre theorem [48, Th. 20.51] asserts that the algebraic or C^*-algebra K-group in fact coincides with the more traditional topological K-group of complex bundles over X. In addition, the higher K-groups also coincide. Precisely,

Theorem 46 *For a compact space X,*

$$K^{-i}(X) = K_i(C(X)); \quad i = 0, 1$$

A.4.2 trading spaces for noncommutative algebras

The basic point of this subsection is that the "amount of damage" done to a space X if we attempt to describe it by means a *noncommutative* algebra $A(X)$ of operators is no more than some coarsening as defined in Section A.2.

Assume for the sake of the argument that $X = \{x^1, ..., x^n\}$ is finite. The set of functions defined over X can be identified with the set of $n \times n$ diagonal matrices. The latter can be viewed as acting on $\ell^2(\{1, ..., n\})$ and forms a *commutative* algebra of operators $\ell^2(\{1, ..., n\}) \to \ell^2(\{1, ..., n\})$. From here on, we could define a *noncommutative* algebra of nondiagonal matrices over X. The chief difference between the two cases is that, in the noncommutative case, the $(i, j), i\ /j$ element of a matrix somehow blurs the difference between x^i and x^j, while in the commutative case this blurring does not occur. The former is referred to as *noncommutative geometry* [26] in which a space is represented by some noncommutative algebra.

second Gelfand-Naimark theorem

Theorem 47 ((Second) Gelfand-Naimark representation Theorem) *Any C^*-algebra has an isometric representation as a closed subalgebra A of the algebra $B(\mathcal{H})$ of bounded linear operators defined over some Hilbert space $\mathcal{H}$.*

Proof. See, e.g., [39, Th. 1.17]. ■

It is possible to derive more specific results, but at the expense of digging much more deeply in the structure of the algebra [35, Chapter 6] than in the commutative case.

Definition 61 *An algebra A is said to be n-homogeneous if all irreducible representations have the same dimension n.*

The spirit of the results is that, under some conditions, a noncommutative algebra has a representation as a matrix-valued function.

Theorem 48 *If A is a unital, n-homogeneous C^*-algebra, then there exists an $N \geq n$ and a projection valued function $P \in C(\hat{A}, M_n(\mathbb{C}))$ such that $A \cong PC(\hat{A}, M_n(\mathbb{C}))P$.*

Proof. See [35, Th. 6.3.1]. ■

Novikov and Baum-Connes conjectures

We begin with the Novikov conjecture, since it will make the overall ideas to be developed later more palatable. Let Γ be a discrete group. Define the action of Γ on $\ell^2(\Gamma)$ as $\alpha_\gamma(f)(x) = f(\gamma x)$. Clearly $f \mapsto \alpha_\gamma(f)$ can be viewed as an operator $\ell^2(\Gamma) \to \ell^2(\Gamma)$. The set of such operators forms an algebra under the composition law $\alpha_{\gamma_1}\alpha_{\gamma_2} = \alpha_{\gamma_1\gamma_2}$, this algebra is clearly unital as $\alpha_{1_\Gamma} = 1_{B(\ell^2)}$, and it is clearly noncommutative unless the group Γ is Abelian. Furthermore, $||\alpha_\gamma(f)||_{\ell^2} = ||f||_{\ell^2}$ so that the operator α_γ is isometric and since it is invertible as $(\alpha_\gamma)^{-1} = \alpha_{\gamma^{-1}}$ it is unitary. The integral group ring $\mathbb{C}\Gamma$ clearly embeds in $B(\ell^2(\Gamma))$ as a *-subalgebra with obvious involution and the $\ell^2(\Gamma)$-closure of this *-subalgebra is the reduced algebra $A_r(\Gamma)$. Next, as already said, the coarsening of Γ is the universal cover of its classifying space, $\widetilde{B\Gamma}$. The spirit of the results we are aiming at is that the algebraic K-theory of the noncommutative algebra defined over Γ is related to the K-homology of the coarsening of Γ. Precisely,

Theorem 49 *Let Γ have no torsion. Then*

$$K_*(\widetilde{B\Gamma}) = K_*(A_r(\Gamma))$$

We now proceed to the coarse Baum-Connes conjecture, which applies to a locally compact Hausdorff space X. We proceed as in the preceding by first defining an algebra over X.

Definition 62 *An X-module is a Hilbert space H_X endowed with a C^*-homomorphism $C_0(X) \to B(H_X)$.*

Definition 63 *An operator $T \in B(H_X)$ is locally compact if, for any $\phi \in C_0(X)$, ϕT and $T\phi$ are compact.*

Definition 64 *An operator $T \in B(H_X)$ is said to have finite propagation if there exists a constant $R > 0$ such that whenever*

$$\inf\{d(x,y) : x \in \mathrm{Supp}(\phi), y \in \mathrm{Supp}(\psi)\} > R$$

we have $\phi T \psi = 0$.

Now, we choose as noncommutative algebra the $B(H_X)$-closure of the subalgebra of locally compact bounded propagation operators.

On the topological side, we need to define a generalized homology theory.

Definition 65 *An operator $T \in B(H_X)$ is pseudolocal if the commutator $[T, \phi]$ is compact whenever $\phi \in C_0(X)$.*

Pseudolocal operators form and algebra which is denoted as $\Psi^0(X)$ and the locally compact pseudolocal operators form an ideal $\Psi^{-1}(X)$. If we define $B(X)$ to be the subalgebra of $\Psi^0(X)$ of bounded propagation operators, then $\Psi^0(X)/\Psi^{-1}(X) = B(X)/A(X)$. Then the generalized homology, referred to as K-homology, is defined as

Definition 66

$$\begin{aligned} K_i(X) &= K_{i+1}(\Psi^0(X)/\Psi^{-1}(X)) \\ &= K_{i+1}(B(X)/A(X)) \end{aligned}$$

The 6 term (long) cyclic exact sequence typical of complex K-theory [93] of quotient algebras yields the so-called *assembly map* $A : K_*(X) \to K_*(A(X))$, as shown in the following diagram:

$$\begin{tikzcd}
K_0(X) \arrow[r, "A"] \arrow[d] & K_0(A(X)) \arrow[r] & K_0(B(X)) \arrow[d] \\
K_1(B(X) & K_1(A(X)) \arrow[l] & K_1(X) \arrow[l, "A"']
\end{tikzcd}$$

If $A(X)$ is commutative then this map is the K-homology/K-theory duality [14, Sec. 16.3]. The deeper stable homotopical duality between K^* and K_* was developed by Adams [2, Part III, p. 204]. Here, a more concrete realization of the duality is the bilinear pairing $K^1(X) \times \mathrm{Ext}(C^0(X)) \to \mathbb{Z}$, where Ext fits within the short exact sequence $0 \to \mathbb{C} \to \mathrm{Ext} \to C^0(X) \to 0$ where $\mathrm{Ext}(C^0(X))$ can be taken to be the definition of $K_1(X)$, and the passage between the 0th and the 1st groups is done by suspension using Bott periodicity (see [14, 16.3]). The grand unification of K-homology and K-theory of C^*-algebras is the Kasparov KK-theory [47], in which the preceding is rewritten as $KK(\mathbb{C}, A) = K_0(A)$ and $K_1(X) = \mathrm{Ext}(C^0(X)) = \mathrm{Ext}(C^0(X), \mathbb{C}) = KK^1(C^0(X), \mathbb{C})$.

Whether the space X survives the coarsening is a matter of whether the assembly map $A : K_*(X) \to K_*(A(X))$ is an isomorphism, which is not likely to occur. However, the so-called *coarse assembly map* $A_\infty : K_*(EX) \to K_*(A(X))$, defined by commutativity of the diagram

$$\begin{tikzcd}
K_*(X) \arrow[r, "A"] \arrow[d] & K_*(A(X)) \\
K_*(EX) \arrow[ur, "A_\infty"] &
\end{tikzcd}$$

is much more likely to be an isomorphism [76, Prop. 8.1, Conjecture 8.2]. This is the *coarse Baum-Connes conjecture.* It turns out that for a space with bounded fatness of the geodesic triangles, the coarse Baum-Connes conjecture holds [76, Prop 9.17].

Theorem 50 *For a δ-hyperbolic space X,*

$$K_i(EX) = K_i(A(X))$$

To sum up, the fine details as to how to coarsen a set of points consistently with a graph on the given set of points and what space results from this coarsening are left out. However, the overall philosophy should be clear at this stage: Proceed from an adjacency matrix a and construct an algebra A.

The space $EX =?$ resulting from the coarsening can be guessed from the coarse Baum-Connes conjecture as $K_*(EX =?) = K_*(A(X))$. But probably we should not hang on too tight to a concrete geometric realization $X =?$ of $K_0(A)$. Indeed, as beautifully articulated by Blackadar [14, 2.1],

> *"One of the motivations for developing the theory of noncommutative topology (...) is that in many instances in ordinary topology the natural object of study is a "singular space" [like those of Chap.* **??***] which cannot be defined and studied in purely topological terms. ... Although the singular space X may not really exist topologically, there is often a noncommutative C^*-algebra which plays the role of $C_0(X)$ in an appropriate sense."*

Other authors [39, p. 8] prefer to refer to $K_0(A)$ as a "virtual" space. In fact, as we shall see soon, the theory of C^*-dynamical systems completely abandon any concrete realization of the geometric object on which the system evolutes and completely relies on the algebra A.

A.4.3 example

To illustrate the above concepts more concretely, we show how a finite set of n points, $x^1, ..., x^n$, can be coarsened into a less trivial geometrical object by means of a noncommutative algebra associated with a graph G on the n points. A popular algebra that can be associated with a graph G with adjacency matrix a is the *digraph algebra* $A(G)$ of the graph G defined as the C^*-algebra generated by the elementary matrices $\{e_{ij} \in M_m : a_{ij} = 1\}$; see [74, 4.9,6.7]. In particular, consider the complete graph on n vertices, G_n. The digraph algebra $A(G_n)$ in this case is the full algebra M_n, from which it is easily found that $K_0(M_n) = \mathbb{Z}$. Since the C^*-algebra K-group is matched by the topological K-group $K^0(\{point\}) = \mathbb{Z}$, it follows that the complete graph has been coarsened to a point.

The problem with the digraph algebra is that it is too big an algebra (it does too much blurring) because of its transitive nature: if x^0x^1 and x^1x^2 are edges then x^0x^2 is an edge. (Transitivity is precisely what has to be abandoned to define a "tolerance relation;" see [83] for a modern exposition, although the concept can be traced to much earlier work [32].)

Here we propose to use the (noncommutative) algebra generated by the adjacency matrix [55, Sec. 2.3] of the graph. The rationale is as follows: Think a as acting on $\ell^2(\{x^1, ..., x^n\})$. As such, any element $a_{ij} \neq 0$ blurs the difference between x^i and x^j. Furthermore, it is easily seen that a^2 is the "two hop" adjacency matrix, meaning that any pair of vertices (i,j) such that $(a^2)_{ij}=1$ can be connected with two edges of the original graph. Again, $a^2:\ell^2\to\ell^2$ blurs the difference between x^i, x^j. It appears therefore that we are setting the stage for another transitivity problem. However, here, a closer look reveals that transitivity is limited by the fact that in general a has multiple eigenvalues at 0. To be specific, if m is the degree of the minimal polynomial of the adjacency matrix a, we are blurring only those pairs of points that can be connected by at most $m-1$ edges of the original graph.

Another contentious point is whether the diagonal terms of the adjacency matrix should be 0's or 1's. To understand the correct interpretation of the diagonal elements, consider the adjacency matrix $a=\begin{pmatrix}0&1\\1&0\end{pmatrix}$. The 2-hop adjacency matrix is $a^2=1$, meaning that there is a 2-hop loop on each vertex. Therefore, in this context, a diagonal element in the adjacency matrix a has to be interpreted as a single loop on the vertex, not collapsible to a point. For this reason, we prefer to set the diagonal terms of the adjacency matrix to zero. Having done so, the only projection in the algebra A is the trivial projection 1, so that $K_0(A)=\mathbb{Z}$, so that the two vertices linked by an edge have been coarsened to a point.

To be yet more specific, we show how a set of 9 points can be coarsened to a more complicated geometric object via the adjacency matrix of a graph G on the 9 points. The adjacency matrix acts on $\ell^2(\{1,...,9\})$ and "blurs" two points linked by an edge of the graph. We consider the adjacency matrix of the 1-skeleton of a torus, as shown in Fig. A.2:

$$a=\begin{pmatrix}0&1&0&0&1&0&0&1&0\\0&0&1&0&0&0&0&1&1\\1&0&0&0&1&0&0&0&1\\1&1&0&0&0&1&0&0&0\\0&0&0&1&0&1&0&1&0\\0&1&1&0&0&0&1&0&0\\1&0&1&1&0&0&0&0&0\\0&0&0&0&0&1&1&0&1\\0&0&0&1&1&0&1&0&0\end{pmatrix}+\begin{pmatrix}0&1&0&0&1&0&0&1&0\\0&0&1&0&0&0&0&1&1\\1&0&0&0&1&0&0&0&1\\1&1&0&0&0&1&0&0&0\\0&0&0&1&0&1&0&1&0\\0&1&1&0&0&0&1&0&0\\1&0&1&1&0&0&0&0&0\\0&0&0&0&0&1&1&0&1\\0&0&0&1&1&0&1&0&0\end{pmatrix}'$$

Observing that the algebra A is subject to the relation $a^3=3a^2+18a$, it is readily found that the projections are

$$p_1=-\frac{1}{6}a+\frac{1}{18}a^2,\quad p_2=-\frac{2}{9}a+\frac{1}{27}a^2,\quad p_3=\frac{1}{18}a+\frac{1}{54}a^2$$

Next, it is readily verified that $p_2+p_3=p_1$, so that the algebra is generated by p_2,p_3. Next, we check whether the set of generators can be further reduced by embedding the finite projections into the algebra $M_\infty(A)$ of arbitrarily large matrices over A and by introducing the equivalence relation $p\sim q$ iff $p=uu^*$ and $q=u^*u$ for some $u\in M_\infty(A)$. The only such relation is, possibly,

$$p_3\oplus p_3\oplus 0_\infty\sim p_2\oplus 0_\infty$$

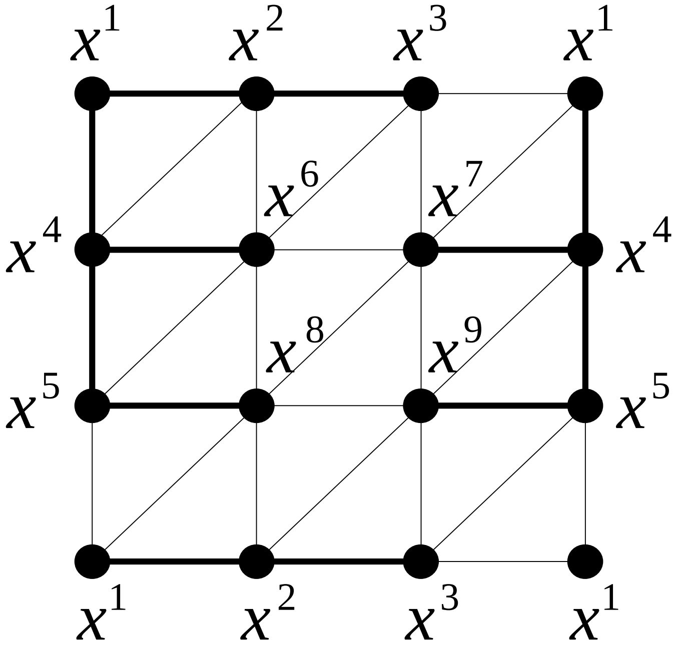


Figure A.2: The graph of the simplicial decomposition of a torus. The thick edges are those making up the maximal tree T.

since this is the only way to have the ranks match. To be more specific, if

$$p_2 \oplus 0_\infty = \begin{pmatrix} a & b & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} a & b & 0 \\ 0 & 0 & 0 \end{pmatrix}^* ; \quad a, b \in A \tag{A.1}$$

then we have to check whether

$$p_3 \oplus p_3 \oplus 0_\infty = \begin{pmatrix} a & b & 0 \\ 0 & 0 & 0 \end{pmatrix}^* \begin{pmatrix} a & b & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{A.2}$$

From (A.1), it would follow that $a \perp b$. From (A.2), it would follow that we could factor the rank one projection p_3 in two different ways, a^*a, b^*b, with $a^*b = 0$, which is clearly impossible. Therefore, the relevant projections in the algebra A are p_2, p_3. Let $V(A)$ be the commutative monoid generated by p_2, p_3 under addition. Since the algebra A is nonunital, the group $K_0(A)$ cannot be computed as the Grothendieck group of $V(A)$ (see [93, Sec. 6.2]). Let A^+ be the algebra resulting from adjoining 1 to A. Clearly, A is an ideal in A^+ and $A^+/A \simeq \mathbb{C}$, so that A^+ is a *unitization* of A [93, Sec. 2.1]. Then $K_0(A)$ can be defined from Groth$(V(A^+)) = K_0(A) \oplus \mathbb{Z}$ (see [93, Prop. 6.2.2]). Since Groth$(V(A^+)) = \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}$, it follows that the C^*-algebra K-group is $K_0(X) = \mathbb{Z} \oplus \mathbb{Z}$. This group matches the topological K-group of the torus, $K^0(\mathbb{T}^2) = \mathbb{Z} \oplus \mathbb{Z}$, and inspiring ourselves from the Swan-Serre theorem, we are led to the conclusion that the coarsening of the 9 points results in a 2-torus.

Since the degree of the minimal polynomial of a is 3, one might be under the impression that we are blurring pairs of points linked by 2 hop paths. The reason why this does not cause too much blurring given the rather coarse triangulation of $\mathbb{T}^2$ is that closer inspection of the graph reveals that any pair of vertices linked by a 2 hop path is actually linked by an edge of the graph.

Observe that, surprisingly, the above procedure *does not* blur the set of points to the mere 1-dimensional object that constitutes the graph. Indeed, if it were so, we would obtain the following obvious contradiction: Let T be a maximal tree of G; the maximal tree has 8 edges and the graph G has 27 edges. It can be shown [63, Chap. 3, Sec. 4] that the natural projection $G \to G/T$ is a homotopy equivalence and that G/T is the wedge of $27 - 8 = 19$ circles, $\vee_{i=1}^{19} S^1$, one circle for every edge of G not in T. From [46, Chap. 10, Prop 3.2], it would then follow that $\tilde{K}^0(\vee_{i=1}^{19} S^1) = \bigoplus_{i=1}^{19} \tilde{K}^0(S^1)$, where $\tilde{K}^0$ denotes the reduced K-group. Since $\tilde{K}^0(S^1) = 0$ [46, Chap. 9, Cor. 5.2], it follows that the K-group of the graph is, in the complex case, $K^0(\vee_{i=1}^{19} S^1) = \mathbb{Z}$, too small compared with the group of projections. A similar conclusion would hold in the real case. Since $KO(S^1) = \mathbb{Z} \bmod 2$ [46, Chap. 9, Cor. 5.2], we would get some torsion in $KO(\vee_{i=1}^{19} S^1)$.

A.4.4 stabilization and scaling up

In the examples we have worked out, the algebras A did not require stabilization. Some algebras–for example, purely infinite simple C^*-algebras–are known not to require stabilization [29, Th. 1.4; p. 188]. The algebra of bounded operators on an *infinite-dimensional* Hilbert space does not require stabilization either [93, 6.1.4]. However, here, we deal with algebras of operators over the finite-dimensional space $\ell^2(\{1, ..., n\})$ and the issue of whether or not these algebras need stabilization remains open. In case stabilization is needed, then the K-theoretic technique would yield a property not of the graph itself, but of some "scaled up" version of the graph. In fact, the K-theoretic tool of stabilization would give a rigorous approach to the concept of scaling. For example, consider the adjacency matrix of a triangle, homeomorphic to the circle S^1,

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

If stabilization to M_2 is needed, then the scaled up version of the graph would have adjacency matrix,

$$\begin{pmatrix} A & A \\ A & A \end{pmatrix}$$

and it is readily verified that this is the adjacency matrix of the graph of the boundary of an octahedron, one of the 4 Platonic solids, which is homeomorphic to the sphere S^2. Therefore, scaling up the circle S^1 yields the sphere S^2.

A.4.5 other algebras

We note that the Cuntz-Krieger algebra $\mathcal{O}_A$, where A is the edge matrix associated with a directed graph [37], is a popular way to associate an algebra with a graph, mainly because its K-theory is well understood [14, 10.11.9]. The connection, if any, with the previous algebra is unclear.

A.4.6 further remarks

The connection between this C^*-algebra technique and the "algebraic surgery" approach of Section 5 is probably along the lines of "bounded surgery," a particular case of "controlled topology," (see [76, pp. 53-54], [94, Chap. 9]) and some connections can already be perceived. Namely, both procedures are up to simple homotopy equivalence on the simplicial complex. Indeed, in this approach a complete graph is shrunken to a point; in surgery theory, one would say that a complete graph is simply homotopically equivalent to a point. It is as yet unclear where, if anywhere, the Poincaré complex property would be involved. Note, however, that the complex of the torus example is a Poincaré complex!

Specific Research Question: As discussed earlier, the existence of a finite δ in an infinite graph simplifies geodesic computation. Of course, the only infinite graphs that can be considered in this engineering context are those infinite graphs obtained from a finite graph pattern by a "scaling" operation [74, Example 2.1]. (This scaling operation is probably the only rigorous way to formulate the intuitively motivated idea of repeating at infinitum a graph pattern to work on an infinite graph to avoid some boundary problems.) The question we would like to address is whether some C^*-algebras of infinite graphs [37] would reveal through their K-theory the existence of a finite δ. As far as we know, this is an widely open problem.

Part III

Large scale geometry

Chapter 13

large scale metric hyperbolic geometry

In this chapter, we pursue the investigation of coarse metric hyperbolic geometry, with the difference now that the coarseness stems from the large scale point of view of the geometry.

13.1 Various definitions of large scale negatively curved spaces

In classical Riemannian geometry, "hyperbolic" or "negative sectional curvature" is a well-defined concept and it endows the space with well-behaved geodesics. The geodesic are robust in the sense that a small perturbation of the end points results in a small perturbation of the geodesic arc. This is a corollary of the Jacobi field concept [50, Section 2.1]. In contrast, a positively curved space, e.g., a sphere, does not have well behaved geodesics; think, for example, of the geodesic joining the South pole of a sphere to a point drifting around the North pole!

Inspiring ourselves from Riemannian geometry, the question arises as to whether in, say, a communication network the geodesics would be not too sensitive to perturbation of the end points. One would be tempted to say that this might be the case if the graph is "hyperbolic." The problem is that in classical Riemannian geometry, sectional curvature is a differential concept and a graph is certainly not a differential object! The approach–due to Cartan, Alexandrov, Rauch, and Toponogov [52, Sec. 3.2]–is to single out a property of a hyperbolic manifold that can be reformulated solely in terms of the distance function and geodesic.

13.1.1 fatness

One such property is the bounded fatness of the geodesic triangles. In any metric space, the *fatness* of a geodesic triangle $\triangle ABC$ is defined as

$$\delta_F(\triangle ABC) = \inf \left\{ d(X,Y) + d(Y,Z) + d(Z,X) : \begin{array}{r} x \in [BC] \\ y \in [AC] \\ z \in [AB] \end{array} \right\} \qquad (13.1)$$

The remarkable feature [76, pp. 84-85] in a Riemannian manifold of negative sectional curvature is that, no matter how big a triangle is, its fatness remains bounded by a constant depending only on the curvature. Before proving this result we need two lemmas. We first prove that the area of a Euclidean circle is smaller than the area of a hyperbolic circle; precisely,

Lemma 12 *Let C_R denote a circle of radius R in either Euclidean or hyperbolic space with sectional curvature bounded from above as $\kappa \leq \kappa_m < 0$. Then*

$$\int\int_{C_r} dS_E \leq \int\int_{C_r} dS_H$$

where dS_E, dS_H denote the Euclidean, hyperbolic, resp., surface elements.

Proof. Consider in hyperbolic space two infinitesimally close geodesic rays of length R emanating from the center O of the circle C_R with a gap angle $d\theta$. By the Jacobi field, the hyperbolic surface element between the two rays and between distances r and $r + dr$ is

$$dS_H = \frac{1}{\sqrt{-\kappa(r,\theta)}} \sinh\left(r\sqrt{-\kappa(r,\theta)}\right) dr d\theta$$

Therefore,

$$\begin{aligned}
\int\int_{C_R} dS_H &= \int_0^{2\pi}\int_0^R \frac{1}{\sqrt{-\kappa(r,\theta)}} \sinh\left(r\sqrt{-\kappa(r,\theta)}\right) dr d\theta \\
&\geq \frac{2\pi}{\sqrt{-\kappa_m}} \int_0^R \sinh\left(r\sqrt{-\kappa_m}\right) dr \\
&= \frac{2\pi}{\kappa_m}\left(\cosh\left(R\sqrt{-\kappa_m}\right) - 1\right) \\
&= \pi R^2 + \frac{2\pi}{-\kappa_m}\left(\sum_{k=2}^{\infty} \frac{R^{2k}(-\kappa_m)^{2k}}{(2k)!}\right) \\
&\geq \int\int_{C_R} dS_E
\end{aligned}$$

■

Before proceeding further, we take a very short pause to observe the following corollary:

Corollary 11 *Under the same conditions as the above, and if the sectional curvature κ is constant,*

$$\int\int_{C_H} dS_H = \frac{4\pi}{-\kappa}\left(\sinh\left(\frac{R\sqrt{-\kappa}}{2}\right)\right)^2$$

Proof. Just work out exactly the first integral of the preceding (see also [44]). ■

Next, we need a bound on the area of a hyperbolic geodesic triangle:

Lemma 13 *In a hyperbolic space with sectional curvature bounded from above as $\kappa \leq \kappa_m < 0$, we have*

$$\int\int_{\triangle ABC} dS_H \leq \frac{\pi}{(-\kappa_m)}$$

Proof. The proof follows from the curvature bound together with the Gauss-Bonnet theorem,

$$\kappa_m \int\int_{\triangle ABC} dS \geq \int\int_{\triangle ABC} \kappa dS = -\pi + (\alpha + \beta + \gamma) \geq -\pi$$

■

Now, we can formulate the key theorem:

Theorem 51 *Let M be a Riemannian manifold of sectional curvature bounded from above as $\kappa \leq \kappa_m < 0$. Then*

$$\delta_F = \sup\{\delta_F(\triangle ABC) : A, B, C, \in M\} < \frac{6}{\sqrt{-\kappa}} \tag{13.2}$$

Proof. Let r be the radius of the circle inscribed to the triangle $\triangle ABC$. Clearly,

$$\delta_F(\triangle ABC) \leq 6r$$

Furthermore, combining Lemmas 12 and 13, it follows that

$$\pi r^2 \leq \frac{\pi}{-\kappa_m}$$

Therefore,

$$r \leq \frac{1}{\sqrt{-\kappa_m}} \tag{13.3}$$

and

$$\delta_F(\triangle ABC) \leq \frac{6}{\sqrt{-\kappa_m}}$$

as claimed. ■

Here we are at the crucial point. The classical Riemannian concept of negative sectional curvature has been reformulated in terms of distance and geodesics. As such it can be extended to the concept of geodesic space, in particular to graphs. Therefore, we will say that a graph is δ_F-negatively curved if the above holds. We could go one step further and say that a graph has curvature bounded above by $\kappa < 0$ if $\delta_F < \frac{6}{\sqrt{-\kappa}}$.

13.1.2 slimness

There are many of alternative characterizations of negative curvature in terms of distance and geodesics, although not all of them are uniformly equivalent. One such characterization we will make ample use in this text is that of slimness of a geodesic triangle, which is defined as the least δ such that every edge is contained within the union of the δ neighborhoods of the other edges; formally,

$$\delta_S(\triangle ABC) = \inf \left\{ \delta : \begin{array}{c} [AB] \subseteq N_{[AC]}(\delta) \cup N_{[CB]}(\delta) \\ [BC] \subseteq N_{[BA]}(\delta) \cup N_{[AC]}(\delta) \\ [CA] \subseteq N_{[CB]}(\delta) \cup N_{[BA]}(\delta) \end{array} \right\}$$

The slimness of the geodesic triangles in a negative curvature space is the following property:

Theorem 52 *In a Riemannian manifold with its sectional curvature bounded from above as $\kappa \leq \kappa_m < 0$ the slimness of the geodesic triangles is bounded as*

$$\delta_S = \sup\{\delta_S(\triangle ABC) : A, B, C \in M\} < \frac{2}{\sqrt{-\kappa}} \tag{13.4}$$

Proof. Let r be the radius of the circle inscribed to the geodesic triangle $\triangle ABC$. Clearly, $\delta_S(\triangle ABC) < 2r$. This together with Equation 13.3 yields the result. ■

It can be shown that $\delta_F < \infty \Leftrightarrow \delta_S < \infty$. However, bounds are harder to come by. The reason why some bounds are needed is that most of our results pertaining to good behavior of the geodesics on communication graphs involve δ_S, while computationally we found that δ_F is easier to come by (see 13.6). From (13.2),(13.4), we would be tempted to assert that $\delta_F = 3\delta_S$, which appears to be confirmed by numerical exploration. However, exact bounds are harder to come by. The way to go about the problem is to use the radius r of the circle inscribed to the triangle (Gromov [40, p. 162] rather speaks of the inscribed triangle), from which it follows that $r < \delta_S < 2r$ and $2r < \delta_F < 6r$, and finally $1 < \delta_F \delta_S^{-1} < 6$. (See also [40, Sec. 6.5, 6.6] for similar bounds.)

13.1.3 thinness

Yet another approach is the "thinness" δ_T of a geodesic triangle, in which a geodesic triangle is viewed as a δ_T-fattened version of a tripod (see [40, Sec. 6.3] for detail).

Besides fatness, slimness, and thinness of geodesic triangles, we mention the approach of Alexandrov [50, Sec. 2.3] and Buseman [50, Sec. 2.2]. The approach of Alexandrov relies on the distance between a point and a parameterized point on a geodesic arc.

As an illustration, consider Figure 13.1, showing the ISP graph. Each node is an Internet Service Provider (ISP), that is, a cluster of hosts and servers, managed by the same organization, with nearly matching Internet Protocol (IP) addresses. This set of nodes is a slight clustering of the Autonomous Systems

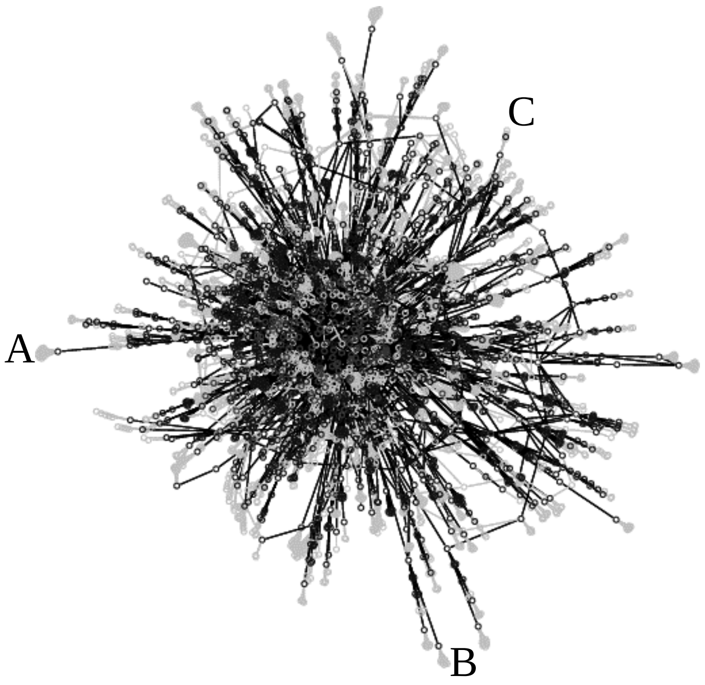


Figure 13.1: The ISP graph consisting of a highly connective core and long tendrils. Observe that the geodesic triangle $\triangle ABC$ is “slim.”

(AS’s), because one single ISP can manage several AS’s, although this is not typical. Two ISP’s are linked if one is the next-hop of the other in the Border Gateway Protocol (BGP). Each link is assigned a weight equal to the number of paths observed between the two ISP’s by a traceroute-like routine. The weight of a link is color coded; light gray represents low weight while dark black represents high weight. High degree nodes are at the center while low degree nodes are at the periphery. This figure clearly illustrates the “core-concentric” property of the ISP graph. Further, the CAIDA project revealed that many other graphs, like the Autonomous Systems graphs, enjoy the same property. As with any graph consisting of a highly connective core and long tendrils, the geodesic lines joining three points A, B, C at the ends of the tendrils are forced to transit via the highly connective core, contributing to the slimness of the geodesic triangle ABC.

It should be stressed that core-concentricity is *not* the only property through which a graph could become hyperbolic. The popular heavy tail graphs [10], which can be thought of as consisting of many “cores,” do show negatively curved properties (see Sec. 15). Yet the hyperbolic property is even more subtle, because there are graphs that have constant degree, and as such they constitute the opposite concept of "heavy tail," yet they are hyperbolic. The classical example is the Cayley graph of the presentation $\langle g_1, g_2, ..., g_n | R_1, R_2, ..., R_p \rangle$ of a group by generators and relators. The Cayley graph of the presentation is rooted at 1, with branches corresponding to right multiplication by g_i, g_j^{-1}. The distance between two words $d(w_1, w_2)$ is the minimum number of generators g_i, g_j^{-1} needed to construct $w_1^{-1}w_2$. It is easily seen that this distance is symmetric and that the Cayley graph is a geodesic space (see Section 18.5 for detail). It is also easily seen that all nodes have degree $2n-1$, except 1 which has degree $2n$. Yet, despite this constancy of the degree, the Cayley graph is hyperbolic with probability one [40, p. 78]. The intuition behind this is that, if it weren't for the relators, the Cayley graph would be a tree; however, the tree structure is destroyed by the relators R_i which create loops; nevertheless, because of the finite number of relators and their finite lengths, the loops are "small" compared with the infinite diameter of the graph, so that from far away the graph looks like a tree and is hence hyperbolic.

13.1.4 4 point inequality, 4 point condition, and Ptolemaic inequality

The large scale negative curvature concepts proposed thus far involve three points, the three vertices of a triangle, and the geodesics joining them. A different but equivalent concept involves 4 points, but does not require the computation of the geodesics joining them. As such, this concept applies to metric spaces that need not be geodesic.

Assume we are given four points a, b, c, d in a metric space. It is instructive, but not necessary, to consider the 4 points as embedded in a geodesic space, in which case the four points can be visualized as the vertices of a quadrilateral. To simplify the notation, define,

$$\begin{aligned} x &= d(a,b), & y &= d(c,d) \\ z &= d(a,c), & w &= d(b,d) \\ u &= d(a,d), & v &= d(b,c) \end{aligned}$$

If the metric space is geodesic, $([ab],[cd])$, $([ac],[b,d])$, $([ad],[bc])$ are pairs of opposite diagonals of the quadrilateral $\Box abcd$ and $((x,y),(z,w),(u,v))$ are pairs of lengths of opposite diagonals. This quadruple of points is said to satisfy the **4 point inequality** if

$$\begin{aligned} u+v &\leq \max\{x+y, z+w\} \\ z+w &\leq \max\{x+y, u+v\} \\ x+y &\leq \max\{u+v, z+w\} \end{aligned}$$

Observe that the last two inequalities are obtained from the first one by permutation of pairs of opposite diagonals. If we define $S \leq M \leq L$ to be the smallest, medium, and largest, resp., sums of distances between end points of opposite diagonals, this condition is equivalent to $M = L$. It is easily seen that such a condition cannot hold in Euclidean space. Indeed, as a counterexample, take $\Box abcd$ to be a "fat" rectangle in $\mathbb{E}^3$. In fact, as we shall see soon, the 4 point inequality is a ($\delta = 0$)-hyperbolic condition.

A closely related condition is the so-called **Ptolemaic inequality**:

$$\begin{aligned} uv &\leq xy + zw \\ zw &\leq xy + uv \\ xy &\leq uv + zw \end{aligned}$$

Again, observe that the last two inequalities are obtained from the first one by permutation of $((x, y), (z, w), (u, v))$. If, without loss of generality, we assume that $uv \geq xy \geq zw$, the last two inequalities are automatically satisfied, so that only the first one is relevant. The following is easily proved:

Theorem 53 *If four points a, b, c, d are isometrically embeddable in Euclidean space $\mathbb{E}^3$, then they satisfy the Ptolemaic inequality.*

Proof. The proof follows from the identity

$$\det \begin{pmatrix} 0 & x^2 & z^2 & u^2 \\ x^2 & 0 & v^2 & w^2 \\ z^2 & v^2 & 0 & y^2 \\ u^2 & w^2 & y^2 & 0 \end{pmatrix} =$$

$$-(xy + zw - uv)(xy - zw + uv)(-xy + zw + uv)(xy + zw + uv)$$

From the Cayley-Menger determinant theory, if the four points are isometrically $\mathbb{E}^3$ embeddable, then the left hand side is negative. Hence the product of the first 3 factors of the right hand side must be positive. But under the (nonrestrictive) condition $uv \geq xy \geq zw$, the second and third factors are positive, so that the first factor is positive as well, that is, the Ptolemaic inequality holds. ■

Observe that the converse is not true; a quadruple of points satisfying the Ptolemaic inequality need not be $\mathbb{E}^3$ embeddable. As a counterexample, it suffices to consider $u = 2$, $v = x = y = z = w = 1$.

In fact, the Ptolemaic inequality also holds in the Poincaré disk model. Therefore, the Ptolemaic inequality is indicative of nonpositive curvature.

The connection between the 4 point inequality and the Ptolemaic inequality is more subtle[1] and will not be pursued any further here.

The four point inequality is meant to be a condition that should hold at all scales. For it to be a meaningful large scale condition, Gromov relaxed $L - M = 0$ to $L - M \leq 2\delta_G$, to hold for all quadruples of points, but for some finite δ_G.

The finite fatness and the Gromov 4 point conditions are equivalent, except that the former does not require geodesics, and as such has some computational advantages. The only problem is that the Gromov 4 point condition is by far less geometrically intuitive than the fatness condition.

[1] In [30], it is asserted that the 4 points inequality implies the Ptolemaic inequality, but this is obviously wrong as shown by the following counterexample: $u = v = 1.1$, $z = w = 1$, $x = 2.5$, and $y = 0.05$.

13.2 quasi-isometries

In general, a geometry has its "isometries." The isometries of Euclidean geometry are orthogonal matrices; the isometries of symplectic geometry are symplectic matrices; etc. In a geometry like that defined in Chapter A, in which spaces are defined up to coarse equivalence, we need a concept of isometry up to some tolerances.

Definition 67 *A* (λ, ϵ) *quasi-isometric embedding is a (not necessarily continuous) function* $f : X \to Y$ *such that such that*

$$\frac{1}{\lambda} d_X(x, x') - \epsilon \leq d_Y(f(x), f(x')) \leq \lambda d_X(x, x') + \epsilon \tag{13.5}$$

f *is said to be a quasi-isometry if, in addition, there exists a* $c > 0$ *such that*

$$\sup\{d_Y(y, f(X)) : y \in Y\} \leq c \tag{13.6}$$

In the latter case, the two metric spaces $(X, d_X), (Y, d_Y)$ *are said to be quasi-isometric.*

Observe that the second inequality of (13.5) implies that f has bounded expansiveness, whereas the first inequality guarantees metric properness, so that a quasi-isometric embedding is a coarse map.

For $\epsilon = 0$, the quasi-isometric embedding f becomes continuous. More specifically, the right-hand side of (13.5) means that the dilation of f is bounded as $\mathrm{dil}(f) := \sup_{x,x'} \frac{d(f(x), f(x'))}{d(x,x')} \leq \lambda$, that is, f is Lipschitz (see [41, Sec. 1.1]). If in addition f is homeomorphic, then the left-hand side of (13.5) means that $\mathrm{dil}(f^{-1}) \leq \lambda$, which along with $\mathrm{dil}(f) \leq \lambda$ means that f is bi-Lipschitz (see [23, Sec. 7.2]). If f is not homeomorphic, then the left-hand side of (13.5) implies that $\frac{1}{\lambda} d(u, u') \leq d(y, y'), \forall u \in f^{-1}(y), \forall u' \in f^{-1}(y')$. The latter in turn implies that $\mathrm{codil}(f) = \sup_{y,y'} \frac{d_H(f^{-1}(y), f^{-1}(y'))}{d(y,y')} \leq \lambda$, which means that f is co-Lipschitz with codilation bounded by λ (see [41, p. 24]).

Theorem 54 *A* (λ, ϵ) *quasi-isometric embedding* f *is a quasi-isometry iff there exists a quasi-inverse, that is, a* (λ', ϵ') *quasi-isometric embedding* [2]

$$f^\dagger : Y \to X$$

such that

$$d(ff^\dagger(y), y) < c_y, \forall y \in Y; \quad d(f^\dagger f(x), x) < c_x, \forall x \in X \tag{13.7}$$

Furthermore, the quasi-inverse is a quasi-isometry.

[2] There is discrepancy in the literature as to whether the quasi-inverse should be *defined* as a quasi-isometric embedding or a quasi-isometry; compare [21, I.8.16(1)] and [52, Definition 3.33]. Here, we adopt the weaker definition, since the stronger property of quasi-isometry is implied by (13.7).

Proof. First, assuming that f has a quasi-inverse $f^\dagger$, we have by definition $d(ff^\dagger(y), y) < c_y$. It follows that $\forall y \in Y$, $\exists y^* \in ff^\dagger(Y) \subseteq f(X)$ such that $d(y, y^*) < c$. This in turn implies that $d(y, f(X)) < c_y$. Hence (13.6).

Conversely, assuming that $d(y, f(X)) < c$, $\forall y \in Y$, the quasi-inverse is constructed as follows: Define $p(y)$ to be the point of $f(X)$ closest to y; clearly, $d(y, p(y)) < c$. Then, invoking the Axiom of Choice, $f^\dagger(y)$ is chosen anywhere in $f^{-1}(p(y))$. It is because $f^\dagger$ need not be continuous that we have this freedom. It remains to show that $f^\dagger$ is indeed a quasi-inverse. First, setting $f(x) = y$, we get

$$d(f^\dagger f(x), x) = d(f^\dagger(y), x) \leq \text{diam}(f^{-1}(y)) \tag{13.8}$$

Clearly, because the quasi-isometry f is a coarse map, the pre-image of a bounded set is bounded and therefore $f^{-1}(y)$ is bounded. It remains to show that $f^{-1}(y)$ is uniformly bounded. Assuming it is not. Then, one would be able to find a sequence $\{y_k\}$ such that $\text{diam}(f^{-1}(y_k)) \to \infty$ as $k \to \infty$ and one would further be able to pick two points $x_k, x'_k \in f^{-1}(y)$ such that $d(x_k, x'_k) \to \infty$. This clearly would be a contradiction to the quasi-isometric property of f,

$$\frac{1}{\lambda} d(x_k, x'_k) - \epsilon \leq d(f(x_k), f(x'_k)) = d(y, y) = 0$$

Let $c_x := \sup \text{diam}(f^{-1}(y))$. The latter along with(13.8) yields $d(f^\dagger f(x), x) < c_x$, $\forall x \in X$. Next, we have

$$\begin{aligned} d(ff^\dagger(y), y) &\leq d(ff^\dagger(y), p(y)) + d(y, p(y)) \\ &= d(ff^\dagger(p(y)), p(y)) + d(y, p(y)) \\ &= d(y, p(y)) < c \end{aligned}$$

It remains to show that $f^\dagger$ is a quasi-isometry. From the quasi-isometric property of f, we get

$$d(x, x') \leq \lambda d(f(x), f(x')) + \epsilon\lambda$$

Now, if x, x' are meant to be $f^\dagger(y), f^\dagger(y')$, respectively, then $f(x) = p(y)$ and $f(x') = p(y')$ and the quasi-isometry property yields

$$\begin{aligned} d(f^\dagger(y), f^\dagger(y')) &\leq \lambda d(p(y), p(y')) + \epsilon\lambda \\ &\leq \lambda d(y, y') + \lambda d(y, p(y)) + \lambda d(y', p(y')) + \epsilon\lambda \\ &\leq \lambda d(y, y') + 2\lambda c + \epsilon\lambda \end{aligned}$$

To prove the other inequality for $f^\dagger$, we proceed from the other inequality of the quasi-isometric property of f,

$$d(f(x), f(x')) \leq \lambda d(x, x') + \epsilon$$

As before, we set $f(x) = p(y)$, $f(x') = p(y')$, $x = f^\dagger(y)$, $x' = f^\dagger(y')$. Next, observe that $d(y, y') \leq d(p(y), p(y')) + d(y, p(y)) + d(y', p(y'))$ implies that $d(p(y), p(y') \geq d(y, y') - 2c$. Hence we get

$$\frac{1}{\lambda} d(y, y') - \frac{2c + \epsilon}{\lambda} \leq d(f^\dagger(y), f^\dagger(y'))$$

It follows that $f^\dagger$ is a (λ', ϵ') quasi-isometric embedding for $\lambda' = \lambda$ and $\epsilon' = \lambda(2c + \epsilon)$. Finally, to prove that $f^\dagger$ is a quasi-isometry, it remains to prove that every point of X is within a bounded distance from the image of $f^\dagger$, which is trivial from $d(f^\dagger f(x), x) < c_x$. ■

Observe that the quasi-inverse is a coarse inverse in the sense of A.1.

To understand the difference between the concepts of quasi-isometric embedding and quasi-isometry, observe that the inclusion $\mathbb{N} \times \mathbb{Z} \to \mathbb{R}^2$ is a quasi-isometric embedding, but not a quasi-isometry, whereas $\mathbb{Z}^2 \to \mathbb{R}^2$ is a quasi-isometry.

It is easily seen that the lattice $\mathbb{Z}^2$ is quasi-isometric to the plane $\mathbb{R}^2$, the intuition behind it being that if one looks at the lattice $\mathbb{Z}^2$ from far away, it appears as $\mathbb{R}^2$. In other words, $\mathbb{R}^2$ is a "blurring" of $\mathbb{Z}^2$. Conversely, $\mathbb{Z}^2$ is a "coarsening" of $\mathbb{R}^2$.

Beyond this trivial example, it is amazing that some geometry can be done at all up to quasi-isometries. Probably the most spectacular example comes from group theory (see Chapter 18). Let $\langle g_1, ..., g_r | R_1, ..., R_m \rangle$ be a presentation of a group Γ by generators and relators (see Section 18.1). The Cayley graph of this presentation is the graph rooted at the identity element 1, with branches corresponding to right multiplication by g_i, g_j^{-1}. The distance between two words $d(w_1, w_2)$ is the minimum number of generators g_i, g_j^{-1} needed to construct $w_1^{-1} w_2$. It is easily seen that this distance is symmetric and that the Cayley graph is a geodesic space (see Section 18.5 for detail). Let $\langle g_1', ..., g_r' | R_1', ..., R_m' \rangle$ be *another* presentation of the *same* group. It turns out that the Caley graphs of the two presentations of the same group are quasi-isometric.

Another popular result is that the fundamental group of a manifold is quasi-isometric to its covering space.

Among the concepts associated with a geometry up to quasi-isometry is the concept of *quasi-geodesic.*

Definition 68 *A (λ, ϵ) quasi-geodesic in a metric space (X, d) is a (λ, ϵ) quasi-isometric embedding*

$$\begin{array}{rcl} \tilde{\gamma} : [0, \tilde{\ell}] & \to & X \\ t & \mapsto & \tilde{\gamma}(t) \end{array}$$

that is, a function $\tilde{\gamma}$ such that

$$\frac{1}{\lambda}|t - t'| - \epsilon \leq d(\tilde{\gamma}(t), \tilde{\gamma}(t')) \leq \lambda |t - t'| + \epsilon$$

Observe that, while the parameter t of the quasi-geodesic was not specified to be the arc length, it is nevertheless closely related to the arc length. Indeed, from the definitions of quasi-geodesic and arc length,

$$\frac{1}{\lambda}|t_2 - t_1| - \epsilon \leq d(\tilde{\gamma}(t_1), \tilde{\gamma}(t_2)) \leq \ell(\tilde{\gamma}([t_1, t_2]))$$

from which it follows that the parameter is related to the arc length as

$$|t_2 - t_1| \leq \lambda(\ell(\tilde{\gamma}([t_1, t_2])) + \epsilon)$$

When $\epsilon = 0$, there is an even tighter relationship between the parameter and the arc length. Indeed, from the definition of the quasi-geodesic, we get

$$\frac{1}{\lambda}dt \leq d\tilde{\gamma} \leq \lambda dt, \tag{13.9}$$

which upon integration yields the relationship:

$$\frac{1}{\lambda}|t_2 - t_1| \leq \ell(\tilde{\gamma}([t_1, t_2])) \leq \lambda |t_2 - t_1|$$

It is sometimes convenient to refer a quasi-geodesic to the geodesic with the same end points. Let $d(\tilde{\gamma}(0), \tilde{\gamma}(\tilde{\ell})) = \ell$ and let γ be the geodesic from $\tilde{\gamma}(0)$ to $\tilde{\gamma}(\tilde{\ell})$. Set $s = \frac{t\ell}{\tilde{\ell}}$, $\bar{\lambda} = \frac{\lambda \ell}{\tilde{\ell}}$, and reparameterize the quasi-geodesic in terms of the arc length s on the geodesic, that is, $\bar{\gamma}(s) = \tilde{\gamma}(\frac{s\tilde{\ell}}{\ell})$. Then the quasi-geodesic can be rewritten

$$\frac{1}{\bar{\lambda}}d(\gamma(s_1), \gamma(s_2)) - \epsilon \leq d(\bar{\gamma}(s_1), \bar{\gamma}(s_2)) \leq \bar{\lambda} d(\gamma(s_1), \gamma(s_2)) + \epsilon$$

In Engineering, we would certainly trade a geodesic for a quasi-geodesic. The only problem is whether a quasi-geodesic is guaranteed to be close to the geodesic. This is the case of the geodesic space is hyperbolic.

13.3 geodesics versus quasi-geodesics

A crucial feature in hyperbolic space is that quasi-geodesics remain close to geodesics. To acquire a good feeling for this phenomenon, we first consider the Riemannian geometry case.

13.3.1 Riemannian viewpoint

Let $\gamma : [0, \ell] \to M$ be a geodesic and let $\tilde{\gamma} : [0, \ell] \to M$ be a *smooth* λ quasi-geodesic with the same end points, viz., $\gamma(0) = \tilde{\gamma}(0)$, $\gamma(\ell) = \tilde{\gamma}(\ell)$. The geodesic is parameterized by its arc length s as usual; however, the quasi-geodesic cannot, in general, be so parameterized, for the obvious reason that its length will not, in general, be equal to ℓ.

If we are interested in how far the quasi-geodesic can depart from the geodesic, that is, $\sup\{d(\tilde{\gamma}(t), \gamma) : t \in [0, \ell]\}$, then it is convenient to assume that the quasi-geodesic is parameterized as follows: Define the projection $p : \tilde{\gamma} \to \gamma$ such that, for $\tilde{x} \in \tilde{\gamma}$, the geodesic $[\tilde{x}, p(\tilde{x})]$ is orthogonal to γ. In other words, $p(\tilde{x})$ is the point on γ closest to $\tilde{x} \in \tilde{\gamma}$. Therefore, the convenient parameterization of the quasi-geodesic is defined as follows:

$$\begin{array}{rcl} \tilde{\gamma}^{-1} : \tilde{\gamma} & \to & [0, \ell] \\ \tilde{x} & \mapsto & \gamma^{-1}(p(\tilde{x})) \end{array}$$

By the same token, it is assumed that λ is the quasi-geodesic tolerance *for this particular parameterization.* Let $D(s) := d(\tilde{\gamma}(s), \gamma) = d(\tilde{\gamma}(s), \gamma(s))$. Let

$[\tilde{\gamma}(s_1), \gamma(s_1)]$ and $[\tilde{\gamma}(s_2), \gamma(s_2)]$ be two arbitrarily close geodesics; precisely, $s_2 - s_1 = ds$. Let q be the projection of $\tilde{\gamma}(s_1)$ on $[\tilde{\gamma}(s_2), \gamma(s_2)]$. Take $[\tilde{\gamma}(s_2), \gamma(s_2)]$ to be the nominal geodesic and consider the Jacobi field orthogonal to it, as done in Section 8.8. Clearly, $[\tilde{\gamma}(s_1), q]$ is in the Jacobi field. Since the geodesics are parameterized by the arc length, it follows that $d(\tilde{\gamma}(s_1), \gamma(s_1)) = d(q, \gamma(s_2))$. By the Jacobi field theory,

$$d(\tilde{\gamma}(s_1), q) = (s_2 - s_1)\cosh(\sqrt{-K}d(\gamma(s_2), q))$$

and for $s_2 - s_1 = ds$, we get

$$d(\tilde{\gamma}(s_1), q) = ds \cosh(\sqrt{-K}D(s))$$

The above, together with the obvious fact that

$$d(\tilde{\gamma}(s_1), \tilde{\gamma}(s_2) \geq d(\tilde{\gamma}(s_1), q)$$

and the quasi-geodesic property

$$\lambda ds \geq d(\tilde{\gamma}(s_1), \tilde{\gamma}(s_2)$$

yields

$$\lambda \geq \cosh(\sqrt{-K}D(s))$$

This above yields the bound

$$\sup\{d(\tilde{\gamma}(t), \gamma) : t \in [0, \ell]\} = \sup\{D(s) : s \in [0, \ell]\} = \frac{1}{\sqrt{-\kappa}}\cosh^{-1}\lambda \approx \frac{1}{\sqrt{-\kappa}}\log_e 2\lambda \tag{13.10}$$

In fact, we can be somewhat more accurate than the above analysis. We first need a couple of technical lemmas.

Lemma 14 *Let $\triangle ABC$ be a right-angled geodesic triangle with angles $\alpha = \frac{\pi}{2}$, β, γ and opposite edges a, b, c, respectively, in a negatively curved Riemannian manifold. Then, up to the second order, $c^2 = a^2 + b^2$.*

Proof. The hyperbolic Pythagoras theorem [11, Th. 7.11.1] says that $\cosh c = \cosh a \cosh b$, which up to the second order yields the result. ■

Lemma 15 *Let $\tilde{\gamma}$ be a smooth quasi-geodesic. Then $\tilde{\gamma}([s_1, s_2]) \to [\tilde{\gamma}(s_1), \tilde{\gamma}(s_2)]$ as $s_2 \downarrow s_1$.*

Proof. Let s_2^k, $k \in \mathbb{N}$ be a sequence decreasing to s_1 as $k \to \infty$. Draw the geodesics $[\tilde{\gamma}(s_1), \tilde{\gamma}(s_2^k)]$ and consider the situation in the tangent space via the reverse exponential map $\exp^{-1}_{\tilde{\gamma}(s_1)}$. In the tangent space, $\exp^{-1}_{\tilde{\gamma}(s_1)}([\tilde{\gamma}(s_1), \tilde{\gamma}(s_2^k)])$, $k \in \mathbb{N}$ is a pencil of lines, all passing through $\tilde{\gamma}(s_1)$, with the projection of the quasi-geodesic passing through $\exp^{-1}_{\tilde{\gamma}(s_1)}(\tilde{\gamma}(s_2^k))$ for $k \in \mathbb{N}$. The smoothness of the (projection of the) quasi-geodesic yields the result in the tangent space, and the continuity of the exponential map yields the result on the manifold. ■

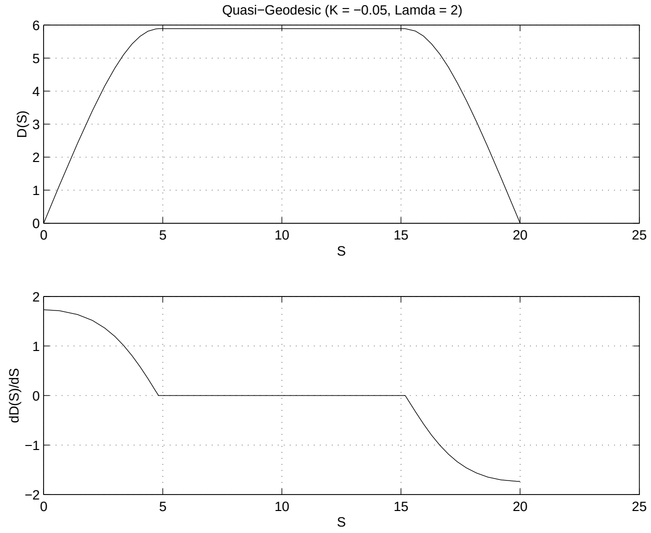


Figure 13.2: Envelope of all distance plots between λ quasi-geodesics and a nominal geodesic of length 20. The curvature is adjusted so that the solution reaches the bound, resulting in a continuously differentiable curve.

Application of the preceding two lemmas, along with the quasi-geodesic property, yields

$$\begin{aligned} d^2(\tilde{\gamma}(s_1), \tilde{\gamma}(s_2)) &= ds^2 \left(\cosh^2(\sqrt{-\kappa}D(s)) + (D'(s))^2\right) \\ &\leq \lambda^2 ds^2 \end{aligned}$$

Therefore, the λ quasi-geodesic that departs with a maximum speed from the geodesic is given by the differential equation

$$(D'(s))^2 = \lambda^2 - \cosh^2\left(\sqrt{-\kappa}D(s)\right), \tag{13.11}$$

subject to the mixed boundary condition

$$D(0) = D(\ell) = 0$$

and

$$\lambda^2 - \cosh^2\left(\sqrt{-\kappa}D(s)\right) \geq 0$$

From the latter, the bound (13.10) is easily rederived.

Figures 13.2 and 13.3 show the results of two simulation runs of Equation (13.11). In the first case, the quasi-geodesic reaches the bound (13.10) and remains at that constant distance away from the geodesic, while in the second case, the quasi-geodesic does not reach its bound.

Now, we can basically redo the above with the objective of bounding by how much the geodesic departs from the quasi-geodesic, viz., $\sup\{d(\gamma(s), \tilde{\gamma}) : s \in \gamma\}$.

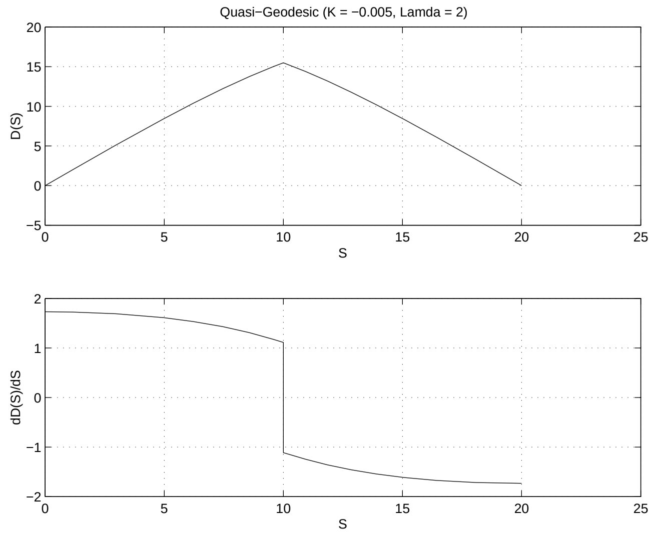


Figure 13.3: Envelope of all distance plots between λ quasi-geodesics and a nominal geodesic of length 20. The curvature is adjusted so that the solution does not reach the bound, resulting in a nondifferentiable curve.

It is necessary, however, to reparameterize the quasi-geodesic differently. Take $\tilde{x} \in \tilde{\gamma}$, draw the geodesic passing through $\tilde{x}$ and orthogonal to $\tilde{\gamma}$, until the latter geodesic intersects γ at the point $r(\tilde{x})$. Define the new parameterization of $\tilde{\gamma}$ as

$$\begin{array}{rcl} \bar{\gamma}^{-1} : \tilde{\gamma} & \to & [0, \ell] \\ \tilde{x} & \mapsto & \gamma^{-1}(r(\tilde{x})) \end{array}$$

Relative to this new parameterization, $\bar{\gamma}$ is still a quasi-geodesic, but with a tolerance of $\bar{\lambda}$, in general different from λ. Let $\bar{D}(s) = d(\bar{\gamma}(s), \gamma(s))$. Using essentially the same Jacobi field argument as in the preceding case, we derive, in the infinitesimal case $s_2 - s_1 = ds$,

$$d(\gamma(s_1), \gamma(s_2)) \geq d(\bar{\gamma}(s_1), \bar{\gamma}(s_2)) \cosh(\sqrt{-K}\bar{D}(s))$$

Now if we consider the $\bar{\lambda}$ quasi-geodesic property,

$$\frac{1}{\bar{\lambda}} d(\gamma(s_1), \gamma(s_2)) \leq d(\bar{\gamma}(s_1), \bar{\gamma}(s_2))$$

in the differential case, viz.,

$$\frac{1}{\bar{\lambda}} ds \leq d\bar{\gamma}$$

the Jacobi field argument yields

$$\bar{\lambda} \geq \cosh(\sqrt{-K}\bar{D}(s))$$

and finally,

$$\bar{D} \leq \frac{1}{\sqrt{-K}} \cosh^{-1} \bar{\lambda}$$

that is, the same bound as the previous one, except that λ is replaced with $\bar{\lambda}$.

Still as in the preceding case, observe that, as a consequence of the Pythagorean theorem,

$$\begin{array}{rcl} d^2(\gamma(s_1), \gamma(s_2)) & = & d\bar{s}^2 \left(\cosh^2(\sqrt{-\kappa} D(\bar{s})) + (D'(\bar{s}))^2\right) \\ & \leq & \lambda^2 d\bar{s}^2 \end{array}$$

where $d\bar{s}$ is the arc length on the quasi-geodesic and D' is the derivative relative to $\bar{s}$. From the above, the quasi-geodesic such that the geodesic departs at maximum speed from the quasi-geodesic is easily found and the bound is recovered.

It remains to find some bound for $\bar{\lambda}$. Clearly, $\tilde{\gamma}(s)$ and $\bar{\gamma}(s)$ are on the same quasi-geodesic, but in general at different points. In view of (13.9), $\frac{1}{\lambda}s \leq \ell(\tilde{\gamma}([0,s])) \leq \lambda s$ and $\frac{1}{\lambda}s \leq \ell(\bar{\gamma}([0,s])) \leq \lambda s$. Assume first that $\bar{\gamma}(s)$ is between $\gamma(0)$ and $\tilde{\gamma}(s)$. Then it follows that the length of the arc $\bar{\gamma}(s), \tilde{\gamma}(s)$ is bounded by $\left(\lambda - \frac{1}{\lambda}\right) s$, so that $\frac{|d(\bar{\gamma}-\tilde{\gamma})|}{ds} = \lambda - \frac{1}{\lambda}$. Then

$$\left|\frac{d\bar{\gamma}}{ds}\right| \leq \left|\frac{d\tilde{\gamma}}{ds}\right| + \left|\frac{d(\bar{\gamma} - \tilde{\gamma})}{ds}\right| \leq \lambda + \left|\lambda - \frac{1}{\bar{\lambda}}\right| = \bar{\lambda}$$

It follows that

$$\bar{\lambda} = \lambda + \sqrt{\lambda^2 - 1}$$

Now, assume that $\bar{\gamma}(s)$ is between $\tilde{\gamma}(s)$ and $\gamma(\ell)$. In this case, the length of the arc $\tilde{\gamma}(s), \bar{\gamma}(s)$ is bounded by $\left(\bar{\lambda} - \frac{1}{\lambda}\right) s$ and

$$\left|\frac{d\tilde{\gamma}}{ds}\right| \leq \left|\frac{d\bar{\gamma}}{ds}\right| + \left|\frac{d(\tilde{\gamma} - \bar{\gamma})}{ds}\right| \leq \bar{\lambda} + \left|\bar{\lambda} - \frac{1}{\lambda}\right| = \lambda$$

and it follows that

$$\bar{\lambda} = \frac{\lambda^2 + 1}{2\lambda}$$

Now, we combine all cases to get the Hausdorff distance between γ and $\tilde{\gamma}$. Define

$$\lambda_m = \max\{\lambda, \lambda + \sqrt{\lambda^2 - 1}, \frac{\lambda^2 + 1}{2\lambda}\}$$

Then the Hausdorff distance between the geodesic and the λ quasi-geodesic is bounded as

$$d_H(\gamma, \tilde{\gamma}) \leq \frac{1}{\sqrt{-K}} \cosh^{-1} \lambda_m$$

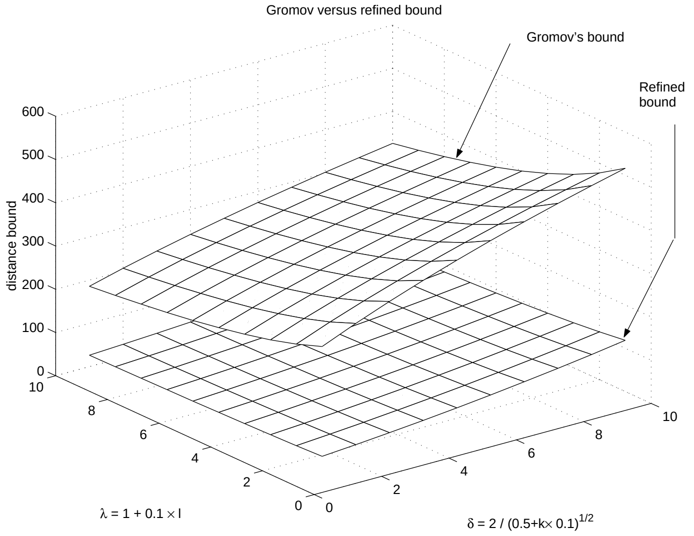


Figure 13.4: Gromov's bound versus refined bound.

13.3.2 metric viewpoint

We now look at tight bounds in arbitrary geodesic spaces. Using the material of [21], it can be shown that a bound is given by

$$D_{max} = D_0(\lambda^2 + 1) + \frac{\lambda}{2}(2\lambda^2 + 3)$$

where D_0 is the maximum solution to

$$\begin{aligned} D_0 \quad \leq \quad & \delta_S \log_2 \left(\frac{1}{\delta_S} \left(D_0 \left(6\lambda^2 + 2 \right) + \lambda \left(2\lambda^2 + 3 \right) \right) \right) \\ & + \frac{\delta_S}{2} \end{aligned}$$

Using some equation solver, this bound is easily computed and compared with the traditional Gromov bound in Figure 13.4, showing a substantial improvement by more careful analysis of the bounds.

13.4 k-local versus quasi-geodesics

An important concept in geodesic spaces is that of a k-local geodesic, defined to be a continuous map $a : [0, l] \to X$ such that the restriction $a|_{[t_1,t_2]}$ is an isometry for $|t_2 - t_1| < k$. Is a k-local geodesic a geodesic? As we will see soon, we conjecture that the answer is "yes," in some hyperbolic spaces.

An interesting fact in δ negatively curved spaces is that k-local geodesics, quasi-geodesics, and k-local quasi-geodesics are "close" to geodesics. This offers the possibility of computing geodesics by piecing together k-local geodesics, but some caution needs to be exercised before that statement can be made rigorous... Specifically, a k-local, $k > 8\delta_S$, geodesic is a (λ, ϵ) quasi-geodesic for $\lambda = \frac{k+4\delta_S}{k-4\delta_S}$ and $\epsilon = 2\delta_S$ (see [21, Th. III.1.13]). Next, there exists a universal tolerance constant $r(\delta, \lambda, \epsilon)$ such that if γ is a geodesic and $\tilde{\gamma}$ a (λ, ϵ) quasi-geodesic with the same end points, we have $d_H(\tilde{\gamma}([0,1]), \gamma([0,1])) < r(\delta, \lambda, \epsilon)$ (see [21, Th. III.1.7]). Combining the above two results, it follows that, if γ, $\tilde{\gamma}_k$ are the geodesic and a k-local geodesic, respectively, we have

$$d_H(\tilde{\gamma}_k, \gamma) < r(\delta, \frac{k+4\delta_S}{k-4\delta_S}, 2\delta_S)$$

Here, we see the problem–even if we take $k \to \infty$, we get $d_H(\tilde{\gamma}_k, \gamma) < r(\delta, 1, 2\delta) \neq 0$, so that we cannot guarantee that the k-local geodesic has been forced to be a geodesic. This observation is symptomatic of the general fact that, even if we take k arbitrarily large, the k-local geodesic cannot in general be forced to a geodesic, as shown by the following counterexample [3]: Consider two copies L_a, L_b of the real line, parallel to each other at a unit distance and such that any common perpendicular crosses the two lines at the same real number. Let the vertices be the integers, denoted as a_k, b^k, $k \in \mathbb{Z}$, on L_a, L_b, respectively. Draw the edges $[a_k, b_k]$, $k \in \mathbb{Z}$. Clearly, this "bi-infinite ladder" is Gromov hyperbolic for some finite δ, because it is quasi-isometric to the real line. Clearly, the path $a_0b_0b_1...b_ka_k$ is a k-local geodesic, but not a geodesic since its length is $k+2$, and the same fact remains true as $k \to \infty$.

The impact of this "local-to-global" property on the routing problem is unclear, but it is probably to be found along the following lines: Each router s would keep a dictionary of the k-local geodesic rooted at s and the routing would consist in piecing them together. Occasionally, this tree would have to be updated in case of link failure, etc., but this creates the inescapable problem of figuring out how robust δ_S is against minor topology changes.

13.5 density of quasi-geodesics

For reliability purposes, it would be interesting to have a great many quasi-geodesics with costs within a small bound away from the geodesic cost. The number of quasi-geodesics is quantified by their distances to the geodesic, arguing that if they go far away there is plenty of space to pack them. This distance is normalized by the length of the λ quasi-geodesic, bounded as $\lambda\ell(\gamma)$ (see [40, Sec. 7.2, Prop 7.2.A]). For a differentiable hyperbolic structure, we could define the *"differential density of quasi-geodesics per unit geodesic length"* as

$$\frac{\partial d_H(\gamma, \tilde{\gamma})}{\partial \lambda} = \frac{100\delta_S \log_2 e}{\lambda}$$

[3] This counterexample was brought to our attention by Prof. Misha Kapovich, University of Utah.

while, for discrete hyperbolic structures, the *"average density of quasi-geodesics per unit geodesic length"* is defined as

$$\lim_{\lambda\to\infty} \frac{100\delta_S(1+\log_2\lambda)}{\lambda} = 0$$

(Another approach would be based on the isoperimetric inequality [21, III.2].) The latter observation points to the fact that the drawback of the good behavior of the geodesics in a hyperbolic space is that the density of quasi-geodesics is small. To have a good density of quasi-geodesics, we would have to go to the opposite spaces of positive curvature. Besides, positive curvature is more appropriate for small diameter graphs.

A positive curvature graph can be defined via the angles. Consider a situation where a vertex of the graph A is connected to vertices B_i, $i = 1,...,n$, B_i is connected to B_{i+1}, and B_n is connected to B_1, with weights $d(A, B_i) = 1$, $d(B_i, B_{i+1}) = 1$, $d(B_n, B_1) = 1$. Let α_i be the angle at the vertex A of the (geodesic) triangle AB_iB_{i+1} and let α_n be the angle at the vertex A of the triangle AB_nB_1. The angles at the vertex A of the *graph* are most naturally defined following the procedure of Alexandrov in a comparison triangle in the model space of constant sectional curvature κ, M_κ [50], [21, Definition 2.15]. (In fact, a deeper result [21, Prop. 2.9] shows that, no matter what comparison space M_κ we choose, the angle is the same and hence equal to 60 degrees.) Then the graph is locally positively curved at A if $\sum_i \alpha_i < 2\pi$. The graph is positively curved if it is locally positively curved at every vertex. Such graphs enjoy some of the properties of manifolds of positive sectional curvature.

Unfortunately, positive curvature is topologically much more constraining than negative curvature (because, for example, the Lichnerowicz-Weitzenböck formula [39, Th. 9.16], [76, p. 6] implies that the Dirac operator has vanishing index) and except for some scarce discrete geometry results (e.g., [40, Prop. 7.2.E]) and of course the Gromov-Lawson theory of manifolds with positive *scalar* curvature, the theory is on shaky mathematical ground.

13.6 δ-Computation

In this section, we look at the computation of the δ in the specific case of metric graphs. Since in the realm of computation all graphs no matter how large are finite, such graphs can only be claimed to exhibit some δ-hyperbolic properties if δ is small compared with the diameter of the graph. As such, the issue becomes δ/diam. Wit this understanding, we specifically look at two different measures of δ-hyperbolicity: the fatness δ_F and the δ_G issued from the Gromov product.

13.6.1 fatness computation

The distance on a graph is initially defined over the vertex set, but it can easily be affinely extended to the edges. Then, a key result in the computation of δ_F for graphs is that the infimum in (13.1) can be attained on the vertices; precisely,

Theorem 55 *For a graph G and with the distance $d(\cdot,\cdot)$ affinely extended to the edges, there exists a solution X,Y,Z to (13.1) on the vertices of $[AB],[BC],[CA]$, respectively; that is,*

$$\inf\left\{ d(X,Y)+d(Y,Z)+d(Z,X) : \begin{array}{r} x\in[BC] \\ y\in[AC] \\ z\in[AB] \end{array} \right\}$$

$$= \quad \inf\left\{ d(X,Y)+d(Y,Z)+d(Z,X) : \begin{array}{r} x\in[BC]_0 \\ y\in[AC]_0 \\ z\in[AB]_0 \end{array} \right\}$$

where $[AB]_0$ denotes the 0th skeleton of $[AB]$ viewed as a simplicial complex, that is, the set of vertices.

Proof. Assume first that the optimum points X,Y,Z are in links $[B_iC_j],[C_jA_i],[A_iB_j]$ that are pairwise nonintersecting. Assume by contradiction that the infimum is reached for $X\in(B_i,C_j)$ where B_i and C_j are vertices of $[BC]$ connected by a direct link. Since the graph is a geodesic space, there exist geodesics $[YX],[ZX]\subseteq G$ such that $\ell([YX])=d(Y,X)$ and $\ell([ZX])=d(Z,X)$. Clearly, in this case, $[YX]\ni B_i$ or $[YX]\ni C_j$, since B_i,C_j are the only "gateways" from X to Y, which by assumption lies in another link. Then either both geodesics pass through the same end vertex of $[B_i,C_j]$ or they pass through different vertices. The first case where the two geodesics $[YX],[ZX]$ pass through the same vertex, say B_i, is impossible, because taking $X=B_i$ would result in a lower cost. The remaining possibility is both geodesics passing through different gateways, say, $[ZX]\ni B_i$ and $[YX]\ni C_j$. In the latter case, the cost is independent on the position of $X\in[B_iC_j]$. Hence we have the freedom to choose , say, $X=B_i$ so that the optimum cost (13.1) can be achieved for X on a vertex. By a similar argument, the optimum cost can be achieved for Y,Z on vertices as well. Again, a similar argument takes care of the case where any pair of vertices X,Y,Z are in intersecting links. ■

Clearly, the domain in which (X,Y,Z) runs is $[AB]\times[BC]\times[CA]$ and the simplicial decomposition of the factors endows the product with the structure of a polyhedron.

13.6.2 Gromov product

The Gromov product δ_G is by far more trivial to compute than δ_F and its computation will not be discussed any further here. We note, however, that while the δ_G appears an attractive tool to gauge the negative curvature properties of massive graphs, it is by far more difficult to interpret, as we shall see soon.

Chapter 14

metric hyperbolic geometry of constant curvature spaces

The present chapter is some kind of a buffer between Riemannian geometry and pure metric geometry, in the sense that we take a geodesic triangle in a constant negative curvature space, which is uniquely defined by its internal angles, and derive formulas for the the various δ measures explicitly as functions of the internal angles. This relies a lot on intensive symbolic manipulations on hyperbolic trigonometry formulas. In addtion, the fatness measure is provided a billiard dynamics interpretation.

14.1 Hyperbolic trigonometry

With the development of δ-hyperbolic spaces, the hyperbolic conditions (δ_S-slim, δ_T-thin, δ_I-insize,and δ_F-fatness) for geodesic triangles have come to play a crucial role in the sense that they provide a substitute for the differential concept of curvature that has traditionally been derived from the Riemannian connection. A manifestation of negative curvature of a Riemannian manifold is the fact that it satisfies the δ hyperbolic conditions. In this section, the hyperbolic conditions δ_S, δ_T, and δ_I are computed for a hyperbolic metric space (M, d) with constant negative curvature κ. Given that $A, B, C \in M$ with $[AB]$, $[BC]$, $[CA]$ the shortest length geodesic arcs joining A to B, B to C, and C to A, respectively, then the geodesic triangles $\triangle(A, B, C)$ is $[AB] \cup [BC] \cup [CA]$. This $\triangle ABC$ in Figure ??? is uniquely specified up to isometry by the three internal angles α, β, γ at the vertices A, B, C, respectively, provided that $\alpha + \beta + \gamma < \pi$. Given that a, b, c are the the lengths of the sides opposite to the angles α, β, γ, respectively, with the assumption that $\alpha, \beta, \gamma > 0$, then the Sine and Cosine Rules for the geodesic triangle $\triangle ABC$ can be summarized as follows:

1. The Sine Rule:

$$\frac{\sinh\left(\sqrt{-\kappa}a\right)}{\sin\alpha}=\frac{\sinh\left(\sqrt{-\kappa}b\right)}{\sin\beta}=\frac{\sinh\left(\sqrt{-\kappa}c\right)}{\sin\gamma}$$

2. The Cosine Rule I:

$$\begin{aligned}\cosh\left(\sqrt{-\kappa}a\right) &= \cosh\left(\sqrt{-\kappa}b\right)\cosh\left(\sqrt{-\kappa}c\right)-\sinh\left(\sqrt{-\kappa}b\right)\sinh\left(\sqrt{-\kappa}c\right)\cos\left(\alpha\right)\\ \cosh\left(\sqrt{-\kappa}b\right) &= \cosh\left(\sqrt{-\kappa}c\right)\cosh\left(\sqrt{-\kappa}a\right)-\sinh\left(\sqrt{-\kappa}c\right)\sinh\left(\sqrt{-\kappa}a\right)\cos\left(\beta\right)\\ \cosh\left(\sqrt{-\kappa}c\right) &= \cosh\left(\sqrt{-\kappa}a\right)\cosh\left(\sqrt{-\kappa}b\right)-\sinh\left(\sqrt{-\kappa}a\right)\sinh\left(\sqrt{-\kappa}b\right)\cos\left(\gamma\right)\end{aligned}$$

3. The Cosine Rule II:

$$\begin{aligned}\cosh\left(\sqrt{-\kappa}a\right) &= \frac{\cos\beta\cos\gamma+\cos\alpha}{\sin\beta\sin\gamma}\\ \cosh\left(\sqrt{-\kappa}b\right) &= \frac{\cos\gamma\cos\alpha+\cos\beta}{\sin\gamma\sin\alpha}\\ \cosh\left(\sqrt{-\kappa}c\right) &= \frac{\cos\alpha\cos\beta+\cos\gamma}{\sin\alpha\sin\beta}\end{aligned}$$

In addition, the Pythagoras' theorem for a geodesic triangle with a right angle in hyperbolic space is given as follows:

Corollary 12 *Given that ΔABC is a geodesic triangle with three internal angles, $\alpha,\beta\leq\frac{\pi}{2}$ and $\gamma=\frac{\pi}{2}$ at the vertices A,B,C, then the hyperbolic form of Pythagoras' theorem is given by the following formula:*

$$\cosh\left(\sqrt{-\kappa}c\right)=\cosh\left(\sqrt{-\kappa}a\right)\cosh\left(\sqrt{-\kappa}b\right)$$

In addition, the following relations hold:

$$\begin{aligned}\tanh\left(\sqrt{-\kappa}b\right) &= \sinh\left(\sqrt{-\kappa}a\right)\tan\beta\\ \sinh\left(\sqrt{-\kappa}b\right) &= \sinh\left(\sqrt{-\kappa}c\right)\sin\beta\\ \tanh\left(\sqrt{-\kappa}a\right) &= \tanh\left(\sqrt{-\kappa}c\right)\cos\beta.\end{aligned}$$

An inscribed circle O in a geodesic triangle is a circle that touches each side of the triangle at exactly one point. The following theorem shows how to construct an inscribed circle in a geodesic triangle.

Theorem 56 *Given that ΔABC is a geodesic triangle with three internal angles, α,β,γ at the vertices A,B,C, respectively, then the three angle bisectors of ΔABC meet at a single point ζ in ΔABC. In addition, the radius R of the inscribed circle of ΔABC is given by*

$$\tanh^2 R=\frac{\cos^2\alpha+\cos^2\beta+\cos^2\gamma+2\cos\alpha\cos\beta\cos\gamma-1}{2\left(1+\cos\alpha\right)\left(1+\cos\beta\right)\left(1+\cos\gamma\right)}.$$

Lemma 16 *The area of a hyperbolic circle O whose radius has length R is given by the formula*

$$Area\,(O) = \frac{4\pi}{-\kappa}\sinh^2\left(\sqrt{-\kappa}\frac{R}{2}\right)$$

where $\kappa < 0$ is a sectional curvature.

Proof. It follows from equation ??? that in normal coordinate, the area of circle O whose radius has length R is given by

$$\begin{aligned} & \int_0^{2\pi}\int_0^R \frac{1}{\sqrt{-\kappa}}\sinh\left(\sqrt{-\kappa}r\right)drd\theta \\ = & \frac{2\pi}{-\kappa}\left(\cosh\left(\sqrt{-\kappa}R\right)-1\right) \\ = & \frac{4\pi}{-\kappa}\sinh^2\left(\sqrt{-\kappa}\frac{R}{2}\right). \end{aligned}$$

■

14.2 Slimness computation

The slimness of the geodesic triangle $\triangle ABC$ is defined as

$$\delta_S\left(\triangle ABC\right) := \max\left\{\delta_{[AB]}, \delta_{[BC]}, \delta_{[CA]}\right\}$$

where

$$\begin{aligned} \delta_{[AB]} &= \sup_{Z\in[AB]} d\left(Z, [BC]\cup[CA]\right) \\ \delta_{[BC]} &= \sup_{X\in[BC]} d\left(X, [CA]\cup[AB]\right) \\ \delta_{[CA]} &= \sup_{Y\in[CA]} d\left(Y, [AB]\cup[BC]\right). \end{aligned}$$

Given that $X \in [BC]$ and $Y \in [CA]$, then for each fixed $Z \in [AB]$, the distances $d\left(Z, X\right)$ and $d\left(Z, Y\right)$ are given as follows:

$$\begin{aligned} d\left(Z,X\right) &= \frac{1}{\sqrt{-\kappa}}\sinh^{-1}\left(\frac{\sin\beta\sinh\left(\sqrt{-\kappa}\left(c-z\right)\right)}{\sin\theta_X}\right) \\ d\left(Z,Y\right) &= \frac{1}{\sqrt{-\kappa}}\sinh^{-1}\left(\frac{\sin\alpha\sinh\left(\sqrt{-\kappa}z\right)}{\sin\theta_Y}\right). \end{aligned}$$

where $\theta_X = \angle BXZ$ and $\theta_Y = \angle AYZ$. Therefore $d\left(Z, [BC]\right) = \inf_{X\in[BC]} d\left(Z, X\right)$ and $d\left(Z, [CA]\right) = \inf_{Y\in[CA]} d\left(Z, Y\right)$ occured at the points X, Y where $\theta_X = \theta_Y = \frac{\pi}{2}$. In addition $[ZX]$ is the unique geodesic through Z which is orthogonal to $[BC]$ and $[ZY]$ is the unique geodesic through Z which is orthogonal to $[CA]$.

By the intermediate value theorem, there exists a point $\tilde{Z} \in [AB]$ such that

$$d\left(\tilde{Z}, [BC]\right) = d\left(\tilde{Z}, [CA]\right).$$

If $Z \in \left[\tilde{Z}A\right]$, then $d\left(Z, [BC] \cup [CA]\right) = d\left(Z, [CA]\right) \leq d\left(\tilde{Z}, [CA]\right)$ and if $Z \in \left[\tilde{Z}B\right]$, then $d\left(Z, [BC] \cup [CA]\right) = d\left(Z, [BC]\right) \leq d\left(\tilde{Z}, [BC]\right)$. Therefore,

$$\begin{aligned}
\delta_{[AB]} &= \sup_{Z \in [AB]} d\left(Z, [BC] \cup [CA]\right) \\
&= \sup_{Z \in [AB]} \inf\left\{d\left(Z, [BC]\right), d\left(Z, [CA]\right)\right\} \\
&= d\left(\tilde{Z}, [BC]\right) = d\left(\tilde{Z}, [CA]\right).
\end{aligned}$$

Given that $\tilde{Z} = d\left(\tilde{Z}, A\right)$, then with $d\left(\tilde{Z}, [BC]\right) = d\left(\tilde{Z}, [CA]\right)$, the following results can be computed.

$$\begin{aligned}
(\sin\alpha)\sinh\sqrt{-\kappa}\tilde{z} &= (\sin\beta)\sinh\sqrt{-\kappa}\,(c - \tilde{z}) \\
\frac{\sin\alpha}{\sin\beta}\sinh\left(\sqrt{-\kappa}\tilde{z}\right) &= \sinh\left(\sqrt{-\kappa}c\right)\cosh\left(\sqrt{-\kappa}\tilde{z}\right) - \cosh\left(\sqrt{-\kappa}c\right)\sinh\left(\sqrt{-\kappa}\tilde{z}\right) \\
\coth\left(\sqrt{-\kappa}\tilde{z}\right) &= \frac{1}{\sinh\left(\sqrt{-\kappa}c\right)}\left(\cosh\left(\sqrt{-\kappa}c\right) + \frac{\sin\alpha}{\sin\beta}\right) \\
1 + \left(\frac{1}{\sinh\left(\sqrt{-\kappa}\tilde{z}\right)}\right)^2 &= \frac{1}{\cosh^2\left(\sqrt{-\kappa}c\right) - 1}\left(\cosh\left(\sqrt{-\kappa}c\right) + \frac{\sin\alpha}{\sin\beta}\right)^2
\end{aligned}$$

The Cosine Rule II in $\triangle ABC$ yields

$$\begin{aligned}
1 + \left(\frac{1}{\sinh\left(\sqrt{-\kappa}\tilde{z}\right)}\right)^2 &= \frac{(\sin\alpha\sin\beta)^2}{(\cos\alpha\cos\beta + \cos\gamma)^2 - (\sin\alpha\sin\beta)^2}\left(\frac{\cos\alpha\cos\beta + \cos\gamma}{\sin\alpha\sin\beta} + \frac{\sin\alpha}{\sin\beta}\right)^2 \\
&= \frac{\left(\cos\gamma + \cos\alpha\cos\beta + 1 - \cos^2\alpha\right)^2}{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)} \\
\sinh^2\left(\sqrt{-\kappa}\tilde{z}\right) &= \frac{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)}{\left(2 + 2\cos\gamma - (\cos\alpha - \cos\beta)^2\right)\left(\sin^2\alpha\right)} \\
\left((\sin\alpha)\sinh\sqrt{-\kappa}\tilde{z}\right)^2 &= \frac{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)}{\left(2 + 2\cos\gamma - (\cos\alpha - \cos\beta)^2\right)}
\end{aligned}$$

Therefore,

$$\delta_{[AB]} = \frac{1}{\sqrt{-\kappa}}\sinh^{-1}\sqrt{\frac{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)}{\left(2 + 2\cos\gamma - (\cos\alpha - \cos\beta)^2\right)}}.$$

Similarly,

$$\delta_{[BC]} = \frac{1}{\sqrt{-\kappa}} \sinh^{-1} \sqrt{\frac{(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1)}{\left(2 + 2\cos\alpha - (\cos\beta - \cos\gamma)^2\right)}}$$

and

$$\delta_{[CA]} = \frac{1}{\sqrt{-\kappa}} \sinh^{-1} \sqrt{\frac{(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1)}{\left(2 + 2\cos\beta - (\cos\gamma - \cos\alpha)^2\right)}}$$

14.3 Insize computation

The insize of the geodesic triangle $\triangle ABC$ is defined as

$$\delta_I(\triangle ABC) := \sup\{d(i_A, i_B), d(i_B, i_C), d(i_C, i_A)\}$$

where $i_A \in [BC], i_B \in [CA], i_C \in [AB]$ are such that

$$\begin{aligned} d(i_A, B) &= d(i_C, B) = (A \cdot C)_B = \frac{c+a-b}{2} \\ d(i_B, C) &= d(i_A, C) = (B \cdot A)_C = \frac{a+b-c}{2} \\ d(i_C, A) &= d(i_B, A) = (C \cdot B)_A = \frac{b+c-a}{2}. \end{aligned}$$

The Cosine Law in $\triangle i_C B i_A$ yields

$$\begin{aligned} \cosh\left(\sqrt{-\kappa}d(i_C, i_A)\right) &= \cosh\left(\sqrt{-\kappa}d(i_A, B)\right)\cosh\left(\sqrt{-\kappa}d(i_C, B)\right) \\ &\quad - \sinh\left(\sqrt{-\kappa}d(i_A, B)\right)\sinh\left(\sqrt{-\kappa}d(i_C, B)\right)\cos\beta \\ &= \cosh^2\left(\sqrt{-\kappa}d(i_A, B)\right) - \left(\cosh^2\left(\sqrt{-\kappa}d(i_A, B)\right) - 1\right)\cos\beta \\ &= (1-\cos\beta)\cosh^2\left(\sqrt{-\kappa}d(i_A, B)\right) + \cos\beta. \end{aligned}$$

Since $2d(i_A, B) = c + a - b$, then

$$\cosh\left(2\sqrt{-\kappa}d(i_A, B)\right) = \cosh\left(\sqrt{-\kappa}(c+a-b)\right).$$

The left-hand side yields

$$\begin{aligned} \cosh\left(2\sqrt{-\kappa}d(i_A, B)\right) &= \cosh^2\left(\sqrt{-\kappa}d(i_A, B)\right) + \sinh^2\left(\sqrt{-\kappa}d(i_A, B)\right) \\ &= 2\cosh^2\left(\sqrt{-\kappa}d(i_A, B)\right) - 1. \end{aligned}$$

The right-hand side yields

$$\begin{aligned} \cosh\left(\sqrt{-\kappa}(c+a-b)\right) &= \cosh\left(\sqrt{-\kappa}a\right)\cosh\left(\sqrt{-\kappa}b\right)\cosh\left(\sqrt{-\kappa}c\right) \\ &\quad - \cosh\left(\sqrt{-\kappa}a\right)\sinh\left(\sqrt{-\kappa}b\right)\sinh\left(\sqrt{-\kappa}c\right) \\ &\quad + \cosh\left(\sqrt{-\kappa}b\right)\sinh\left(\sqrt{-\kappa}c\right)\sinh\left(\sqrt{-\kappa}a\right) \\ &\quad - \cosh\left(\sqrt{-\kappa}c\right)\sinh\left(\sqrt{-\kappa}a\right)\sinh\left(\sqrt{-\kappa}b\right). \end{aligned}$$

From the Cosine Rule I, the following expressions can be derived.

$$\begin{aligned}
\sinh\left(\sqrt{-\kappa}b\right)\sinh\left(\sqrt{-\kappa}c\right) &= \frac{\cosh\left(\sqrt{-\kappa}b\right)\cosh\left(\sqrt{-\kappa}c\right)-\cosh\left(\sqrt{-\kappa}a\right)}{\cos\alpha}\\
\sinh\left(\sqrt{-\kappa}c\right)\sinh\left(\sqrt{-\kappa}a\right) &= \frac{\cosh\left(\sqrt{-\kappa}c\right)\cosh\left(\sqrt{-\kappa}a\right)-\cosh\left(\sqrt{-\kappa}b\right)}{\cos\beta}\\
\sinh\left(\sqrt{-\kappa}a\right)\sinh\left(\sqrt{-\kappa}b\right) &= \frac{\cosh\left(\sqrt{-\kappa}a\right)\cosh\left(\sqrt{-\kappa}b\right)-\cosh\left(\sqrt{-\kappa}c\right)}{\cos\gamma}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\cosh\left(\sqrt{-\kappa}\,(c+a-b)\right)\\
=\;&\frac{\cosh\left(\sqrt{-\kappa}a\right)\cosh\left(\sqrt{-\kappa}b\right)\cosh\left(\sqrt{-\kappa}c\right)}{\cos\alpha\cos\beta\cos\gamma}\\
&\cdot\left(\cos\alpha\cos\gamma-\cos\alpha\cos\beta-\cos\beta\cos\gamma+\cos\alpha\cos\beta\cos\gamma\right)\\
&+\frac{1}{\cos\alpha\cos\beta\cos\gamma}\\
&\cdot\left(\cosh^2\left(\sqrt{-\kappa}a\right)\cos\beta\cos\gamma-\cosh^2\left(\sqrt{-\kappa}b\right)\cos\alpha\cos\gamma+\cosh^2\left(\sqrt{-\kappa}c\right)\cos\alpha\cos\beta\right).
\end{aligned}$$

The Cosine Rule II in $\triangle ABC$ yields

$$\begin{aligned}
\cosh\left(\sqrt{-\kappa}a\right) &= \frac{\cos\beta\cos\gamma+\cos\alpha}{\sin\beta\sin\gamma}\\
\cosh\left(\sqrt{-\kappa}b\right) &= \frac{\cos\gamma\cos\alpha+\cos\beta}{\sin\gamma\sin\alpha}\\
\cosh\left(\sqrt{-\kappa}c\right) &= \frac{\cos\alpha\cos\beta+\cos\gamma}{\sin\alpha\sin\beta}.
\end{aligned}$$

Hence

$$\begin{aligned}
&\cosh\left(\sqrt{-\kappa}\,(c+a-b)\right)\\
=\;&\left(\frac{1}{\cos\alpha\cos\beta\cos\gamma}\right)\left(\frac{1}{\sin\alpha\sin\beta\sin\gamma}\right)^2\left(\cos\alpha\cos\gamma-\cos\alpha\cos\beta-\cos\beta\cos\gamma+\cos\alpha\cos\beta\cos\right.\\
&\cdot\left(\cos\beta\cos\gamma+\cos\alpha\right)\left(\cos\gamma\cos\alpha+\cos\beta\right)\left(\cos\alpha\cos\beta+\cos\gamma\right)\\
&+\left(\frac{1}{\cos\alpha\cos\beta\cos\gamma}\right)\left(\frac{1}{\sin\alpha\sin\beta\sin\gamma}\right)^2\left(\cos\beta\cos\gamma+\cos\alpha\right)^2\left(1-\cos^2\alpha\right)\cos\beta\cos\gamma\\
&-\left(\frac{1}{\cos\alpha\cos\beta\cos\gamma}\right)\left(\frac{1}{\sin\alpha\sin\beta\sin\gamma}\right)^2\left(\cos\gamma\cos\alpha+\cos\beta\right)^2\left(1-\cos^2\beta\right)\cos\alpha\cos\gamma\\
&+\left(\frac{1}{\cos\alpha\cos\beta\cos\gamma}\right)\left(\frac{1}{\sin\alpha\sin\beta\sin\gamma}\right)^2\left(\cos\alpha\cos\beta+\cos\gamma\right)^2\left(1-\cos^2\gamma\right)\cos\alpha\cos\beta
\end{aligned}$$

$$\begin{aligned}
&2\cosh^2\left(\sqrt{-\kappa}d\left(i_A,B\right)\right)-1\\
=\;&\cosh\left(\sqrt{-\kappa}\left(c+a-b\right)\right)\\
=\;&\left(\frac{1}{\sin\alpha\sin\beta\sin\gamma}\right)^2\left(1-\cos\alpha\right)\left(1+\cos\beta\right)\left(1-\cos\gamma\right)\\
&\cdot\left(\cos\alpha-\cos\beta+\cos\gamma-\cos\alpha\cos\beta+\cos\alpha\cos\gamma-\cos\beta\cos\gamma+\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma\right)
\end{aligned}$$

$$\cosh^2\left(\sqrt{-\kappa}d\left(i_A,B\right)\right)=\frac{1}{2}\left(\frac{\cos\alpha-\cos\beta+\cos\gamma+1}{\sin\alpha\sin\beta\sin\gamma}\right)^2\left(1-\cos\alpha\right)\left(1+\cos\beta\right)\left(1-\cos\gamma\right)$$

Finally,

$$\begin{aligned}
\cosh\left(\sqrt{-\kappa}d\left(i_C,i_A\right)\right)&=\left(1-\cos\beta\right)\cosh^2\left(\sqrt{-\kappa}d\left(i_A,B\right)\right)+\cos\beta\\
&=\frac{1}{2}\frac{\left(\cos\alpha-\cos\beta+\cos\gamma+1\right)^2}{\left(1+\cos\alpha\right)\left(1+\cos\gamma\right)}+\cos\beta\\
&=\frac{\left(2\cos\alpha+2\cos\gamma+2\cos\alpha\cos\gamma+2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma+1\right)}{2\left(1+\cos\alpha\right)\left(1+\cos\gamma\right)}
\end{aligned}$$

$$d\left(i_C,i_A\right)=\frac{1}{\sqrt{-\kappa}}\cosh^{-1}\left(\frac{\left(2\cos\alpha+2\cos\gamma+2\cos\alpha\cos\gamma+2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma+1\right)}{2\left(1+\cos\alpha\right)\left(1+\cos\gamma\right)}\right)$$

Similarly,

$$\begin{aligned}
d\left(i_A,i_B\right)&=\frac{1}{\sqrt{-\kappa}}\cosh^{-1}\left(\frac{\left(2\cos\alpha+2\cos\beta+2\cos\alpha\cos\beta+2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma+1\right)}{2\left(1+\cos\alpha\right)\left(1+\cos\beta\right)}\right.\\
d\left(i_B,i_C\right)&=\frac{1}{\sqrt{-\kappa}}\cosh^{-1}\left(\frac{\left(2\cos\beta+2\cos\gamma+2\cos\beta\cos\gamma+2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma+1\right)}{2\left(1+\cos\beta\right)\left(1+\cos\gamma\right)}\right.
\end{aligned}$$

14.4 Thinness computation

The thinness of the geodesic triangle $\triangle ABC$ is defined as

$$\delta_T\left(\triangle ABC\right)=\sup\left\{\delta_A,\delta_B,\delta_C\right\}$$

where

$$\begin{aligned}
\delta_A&=\sup\left\{d\left(v,w\right):v\in\left[i_BA\right],w\in\left[i_CA\right],\text{and }d\left(v,A\right)=d\left(w,A\right)\right\}\\
\delta_B&=\sup\left\{d\left(w,u\right):w\in\left[i_CB\right],u\in\left[i_AB\right],\text{and }d\left(w,B\right)=d\left(u,B\right)\right\}\\
\delta_C&=\sup\left\{d\left(u,v\right):u\in\left[i_AB\right],v\in\left[i_CB\right],\text{and }d\left(u,C\right)=d\left(v,C\right)\right\}.
\end{aligned}$$

Given that $w\in\left[i_CB\right],u\in\left[i_AB\right]$,and $d\left(w,B\right)=d\left(u,B\right)$, then the Cosine law in ΔwBu yields

$$\begin{aligned}
\cosh\left(\sqrt{-\kappa}d\left(w,u\right)\right)&=\cosh\left(\sqrt{-\kappa}d\left(w,B\right)\right)\cosh\left(\sqrt{-\kappa}d\left(u,B\right)\right)\\
&\quad-\sinh\left(\sqrt{-\kappa}d\left(w,B\right)\right)\sinh\left(\sqrt{-\kappa}d\left(u,B\right)\right)\cos\beta\\
&=\left(1-\cos\beta\right)\cosh^2\left(\sqrt{-\kappa}d\left(u,B\right)\right)+\cos\beta\\
&\leq\left(1-\cos\beta\right)\cosh^2\left(\sqrt{-\kappa}d\left(i_A,B\right)\right)+\cos\beta
\end{aligned}$$

Therefore,

$$\delta_B = d\left(i_C, i_A\right)$$

Similarly,

$$\begin{aligned}\delta_A &= d\left(i_B, i_C\right)\\ \delta_C &= d\left(i_A, i_B\right)\end{aligned}$$

Hence

$$\delta_T\left(\triangle ABC\right) = \delta_I\left(\triangle ABC\right)$$

14.5 fatness and billiard dynamics

By definition, the billiard dynamics in a bounded domain Ω with piecewise C^1 boundary is a geodesic flow in Ω and at the boundary $\partial\Omega$ the trajectory is reflected with an angle equals to the incidence angle (see [53, Sec. 9.2]). In the case where $\partial\Omega$ is a geodesic triangle, the period 3 orbit of the billiard dynamics, that is, a periodic motion reflecting exactly once on each side of $\partial\Omega$, is the minimum perimeter inscribed triangle and as such provides the fatness of the triangle. As a warm up exercise, we begin with the Euclidean case.

Let $\triangle ABC$ be a geodesic (rectilinear) triangle in $\mathbb{E}^2$. We first consider the case where the triangle $\triangle ABC$ has no obtuse angles, that is, $\alpha, \beta, \gamma < \frac{\pi}{2}$. Finding the minimum perimeter triangle inscribed to $\triangle ABC$ is the celebrated *Fagnano problem* [28], which has the following solution: From A, draw the altitude AX, that is, the line segment such that $X \in [BC]$ and $[AX] \perp [BC]$. Likewise, draw the altitude $[BY]$ and $[CZ]$. As is well known, the three altitudes intersect at a single point, referred to as *orthocenter* H. It turns out that $\triangle XYZ$, refered to as *orthic triangle* [28, Sec. 1.6, p. 17] and shown in Figure 14.1, is the minimum perimeter inscribed triangle. The traditional Fermat principle of geometrical optics [98] is enough to prove that that $[XY] \cup [YZ] \cup [ZX]$ is the unique period 3 orbit of the billiard dynamics, in the sense that $\angle YXC = \angle ZXB$ with the same fact at the points Y, Z. We reassert the preceding more precisely in the following theorem:

Theorem 57 *For a Euclidean triangle without obtuse angle, the optimal solution (X, Y, Z) to (13.1) is a critical point of the mapping $(X, Y, Z) \mapsto d(X, Y) + d(Y, Z) + d(Z, X)$. Furthermore, this critical point is a periodic orbit of period 3 of the billard dynamics on the triangular table. Finally, this periodic orbit is given by the above construction and is unique.*

Proof. Since $d(X, Y) + d(Y, Z) + d(Z, X)$ should be minimized, it is evident that the solution is a critical point. It is easily seen that

$$\frac{\partial d(Z, X) + d(X, Y)}{\partial d(B, X)} = \cos\angle ZXB - \cos\angle YXC \qquad (14.1)$$

Therefore, at the critical point, $\angle ZXB = \angle YXC$. And a similar property holds at the points Y, Z. In other words, $[XY] \cup [YZ] \cup [ZX]$ is a periodic solution of

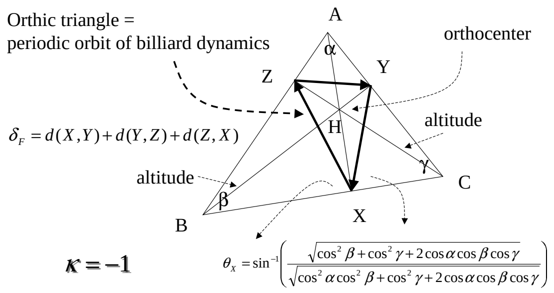

Orthic triangle =
periodic orbit of billiard dynamics
A
orthocenter
α
Z
Y
$\delta_F = d(X,Y) + d(Y,Z) + d(Z,X)$
H
altitude
altitude
β
γ
C
B
X
$\kappa = -1$
$\theta_X = \sin^{-1}\left(\frac{\sqrt{\cos^2\beta + \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}}{\sqrt{\cos^2\alpha\cos^2\beta + \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}}\right)$

Figure 14.1: The orthic triangle.

period 3 to the billiard dynamics. Next, we show that the points of reflection are given as the feet of the perpendicular lines from the vertices of the triangle to their respective opposite edges. Define the incidence angle and the reflection angle at X, respectively, as

$$\begin{aligned} \theta_x^i &= \angle ZXA \\ \theta_x^r &= \angle AXY \end{aligned}$$

Consider the quadrilateral $BZHX$. Since $\angle BZH = \frac{\pi}{2}$ and $\angle HXB = \frac{\pi}{2}$, it follows that this quadrilateral has a circumscribed circle of which $[BH]$ is the diameter. Since both angles $\angle ZHB$ and $\angle ZXB$ are supported by the same arc, they are equals (to $\frac{1}{2}$ the angle of the arc BZ expressed in radians). Hence $\theta_x^i = \frac{\pi}{2} - \angle ZHB$. The same argument on the quadrilateral $CYHX$ yields $\theta_x^r = \frac{\pi}{2} - \angle YHC$. Since, obviously, $\angle ZHB = \angle YHC$, it follows that $\theta_x^i = \theta_x^r$. To show that the critical solution is unique, assume that we have found three points, $X \in [BC]$, $Y \in [ac]$, and $Z \in [AB]$ such that at every such point the incidence angle equals the reflection angle. Let θ_x, θ_y, and θ_z be the respective angles. Construct the triangle XYZ and let H be the common intersection of the angle bisectors of the triangle XYZ. Clearly, $[HX] \perp [BC]$, but we do not as yet know whether $[AH]$ and $[HX]$ are aligned. To prove the latter, observe that $[HZ] \perp [AB]$, $[HY] \perp [AC]$. Using the argument of the circle circumscribed to the selected quadrilateral, we find that

$$\begin{aligned} \angle AHZ &= \frac{\pi}{2} - \theta_y \\ \angle ZHB &= \frac{\pi}{2} - \theta_x \\ \angle BHX &= \frac{\pi}{2} - \theta_z \end{aligned}$$

Furthermore, the sum of the internal angles of the triangle $[XYZ]$ amounts to π and this yields

$$\theta_x + \theta_y + \theta_z = \frac{\pi}{2}$$

Combining the above two, it follows that

$$\angle AHZ + \angle ZHB + \angle BHX = \frac{3\pi}{2} - \frac{\pi}{2} = \pi$$

and the line segments AH and HX are aligned. Likewise, the triples of points, Z, H, C and B, H, Y are colinear. It follows that the critical solution, if any, is given by the ortho-center construction. Since the latter is unique, so is the critical solution. ■

We can be somewhat more specific as to what the incidence and reflection angles are.

Corollary 13

$$\begin{aligned} \theta_x &= \frac{\pi}{2} - \alpha \\ \theta_y &= \frac{\pi}{2} - \beta \\ \theta_z &= \frac{\pi}{2} - \gamma \end{aligned}$$

Proof. An elementary geometric argument on rectangular triangles yields

$$\angle ZHB = \frac{\pi}{2} - \angle ZBH = \frac{\pi}{2} - (\frac{\pi}{2} - \alpha) = \alpha$$

Therefore, $\theta_x = \frac{\pi}{2} - \angle ZXB = \frac{\pi}{2} - \angle ZHB = \frac{\pi}{2} - \alpha$, with a similar relation at X, Y. ■

The first and second claims of the above theorem constitute the well-known *Fermat Principle* of optics, saying that a light ray from X to Z reflecting at $Y \in [BC]$ minimizes $\int_X^Z nds$ where n is the refraction index.

Consider now a triangle with an obtuse angle, say $\alpha \geq \frac{\pi}{2}$. As before, draw $[AX]$, with $X \in [BC]$ and such that $[AX] \perp [BC]$.

Theorem 58 *For an Euclidean triangle with an obtuse angle at A, the optimal solution (X, Y, Z) to (13.1) is $X, Y = Z = A$ and is unique.*

Proof. We first relax the constraints $Y \in [CA], Z \in [BA]$ and allow Y, Z to be on the *rays* $C/A, B/A$ emanating from C, B, respectively, and containing A (see [28, Section 12.2, page 179]). Under those conditions, the mapping $(X, Y, Z) \mapsto d(X, Y) + d(Y, Z) + d(Z, X)$ has a unique critical point achieved for $[AX] \perp [BC], [CY] \perp [BY], [BZ] \perp [CZ]$, which could not be the solution to (13.1). Hence the solution to (13.1) is to be sought at a nondifferentiable or boundary point. Under the relaxed constraints, the first order variation is still given by (14.1) and hence the mapping is differentiable, unless the angles are not defined. Assume $X \in (AB)$. The only situation where the angles are not defined is for $Z = Y = A$. Minimizing the performance index $d(X,A)+d(A,A)+d(A,X)$ requires X to be the foot of the altitude drawn from A. Next, for Y, Z slightly departing from A, keeping the first order variation relative to X null requires the incidence angle to be equal to the reflection angle at X. Let θ be this angle, considered positive when $Y \in [AC]$ and negative otherwise. Let us show that this point is a nondifferentiable minimum. Clearly, as $\theta \uparrow 0$, in other words, as $Y \to A, Z \to A$ on the ray $C/A, B/A$ with $A \in [YC], [BZ]$, respectively, $d(X,Y) + d(Y,Z) + d(Z,X)$ is decreasing. Now, take $\theta \downarrow 0$ and let us show that the right derivative is positive. As usual, let $z = d(Z,A)$ for $Z \in [AB]$ and let $z = -d(Z,A)$ otherwise; define $\bar{y} = d(A,Y)$, $h = d(A,X)$, $\alpha_1 = \angle BAX$, and $\alpha_2 = \angle XAC$. By the cosine rule,

$$d(Y,Z) + d(Z,X) + d(X,Y) = \\ \sqrt{z^2 + \bar{y}^2 - 2z\bar{y}\cos\alpha} + \sqrt{z^2 + h^2 - 2zh\cos\alpha_1} + \sqrt{\bar{y}^2 + h^2 - 2\bar{y}h\cos\alpha_2}$$

It follows that

$$\frac{d}{d\theta}\left(d(Y,Z) + d(Z,X) + d(X,Y)\right) = \\ \frac{zz' + \bar{y}\bar{y}' - z'\bar{y}\cos\alpha - z\bar{y}'\cos\alpha}{\sqrt{z^2 + \bar{y}^2 - 2z\bar{y}\cos\alpha}} + \frac{zz' - z'h\cos\alpha_1}{\sqrt{z^2 + h^2 - 2zh\cos\alpha_1}} + \frac{\bar{y}\bar{y}' - \bar{y}'h\cos\alpha_2}{\sqrt{\bar{y}^2 + h^2 - 2\bar{y}h\cos\alpha_2}}$$

where z' is a short for $\frac{dz}{d\theta}$ and $\bar{y}'$ is defined similarly. Clearly, as $\theta \downarrow 0$,

$$\frac{d}{d\theta}\left(d(Z,X) + d(X,Y)\right) = -z'\cos\alpha_1 - \bar{y}'\cos\alpha_2$$

Next, after some manipulation using the sine rule,

$$z = \frac{h\sin\theta}{\sin(\pi - \theta - \alpha_1)}, \quad \bar{y} = \frac{h\sin\theta}{\sin(\pi - \theta - \alpha_2)}$$

we find that

$$\frac{d}{d\theta}d(Y,Z) = z'\sqrt{1 + \frac{\sin^2\alpha_1}{\sin^2\alpha_2} - 2\frac{\sin\alpha_1}{\sin\alpha_2}\cos\alpha}$$

Hence

$$\frac{1}{z'}\frac{d}{d\theta}\left(d(Y,Z) + d(Z,X) + d(X,Y)\right) = \\ \sqrt{1 + \frac{\sin^2\alpha_1}{\sin^2\alpha_2} - 2\frac{\sin\alpha_1}{\sin\alpha_2}\cos\alpha} - \cos\alpha_1 - \frac{\sin\alpha_1}{\sin\alpha_2}\cos\alpha_2$$

Since $z' > 0$ for $\theta > 0$, it suffices to show that the right hand side of the above is positive, which in turn is equivalent to

$$1 + \frac{\sin^2\alpha_1}{\sin^2\alpha_2} - 2\frac{\sin\alpha_1\cos\alpha}{\sin\alpha_2} - \cos^2\alpha_1 - \frac{\sin^2\alpha_1}{\sin^2\alpha_2}\cos^2\alpha_2 - 2\frac{\cos\alpha_1\cos\alpha_2\sin\alpha_1}{\sin\alpha_2} > 0$$

Some elementary manipulation show that the right hand side is

$$-4\sin\alpha_1\sin\alpha_2\cos(\alpha_1+\alpha_2) = -4\sin\alpha_1\sin\alpha_2\cos\alpha > 0,$$

since the angle at A is obtuse. Since $d(X,Y)+d(Y,Z)+d(Z,X)$ is decreasing for $\theta < 0$ and strictly increasing for $\theta \geq 0$, $\theta = 0$, that is, $(X, Y = A, Z = A)$, is a nondifferentiable point, the only nondifferentiable point for $X \in (AB)$, and hence the optimum solution. It is also easily seen that the optimal solution could not have $X = A$ nor could it have $X = B$. Hence the optimum solution is (X, A, A).

■

The solution to the Euclidean Fagnano problem offered above was synthetic-analytic in nature and involved, in its synthetic part, a minimum amount of construction. Its drawback is that, while as we shall soon see Theorem 57 holds in hyperbolic geometry, the above proof does not lend itself to hyperbolic extension. Indeed, first, the existence of an orthocenter cannot be taken for granted in hyperbolic geometry, and, second and more importantly, the key feature of the proof-cycle that breaks down in hyperbolic geometry is the angle property $\angle ZHB = \angle ZXB$ in a quadrilateral $BZHX$ with two opposed right angles, $\angle HZB = \angle HXB = \frac{\pi}{2}$.

A construction that lends itself more easily to hyperbolic extension is that of Fejér. For the case of an Euclidean triangle without obtuse angles, first fix $X \in [BC]$ and find $Y \in [AC]$ and $Z \in [AB]$ such that the perimeter of XYZ is minimum, and then find X such that the perimeter of the previously constructed triangle is minimized. Draw $[AX]$ and reflect $[AX]$ across $[AB]$ to get $[AX']$. Likewise, reflect $[AX]$ across $[AC]$ to obtain $[AX'']$. Clearly, the perimeter of XYZ is $d(x',Z)+d(Z,Y)+d(Y,X'')$. Clearly, this length is minimized iff X',Z,Y,X'' are colinear, which clearly implies that $\angle AZY = \angle BZX$ and $\angle AYZ = \angle XYC$, that is, the Fermat principle. Now, it remains to find $X \in [BC]$ such that $d(X',X'')$ is minimized. Observe that $[X'X'']$ is the edge of the isoscles triangle $AX'X''$ opposed to the vertex A common to the two equal length edges. The length of the equal edges of this isoscles triangle is clearly $d(A,X)$. Furthermore, $\angle X'AX''$ is easily seen not to depend on X. Therefore, $d(X',X'')$ is minimized if $d(A,X') = d(A,X'') = d(A,X)$ is minimized. Clearly, this happens when $[AX] \perp [BC]$. Repeating the same argument for Z and Y yields the result.

This argument can be extended to hyperbolic geometry by substituting the hyperbolic geometry concept of inversion (which is conformal and hence preserves the angles [84, Th. 9]) for the Euclidean concept of reflection. That $d(X',X'')$ in the isoscles triangle $AX'X''$ is minimized whenever $d(A,X') = d(A,X'')$ is minimized requires, however, either a Jacobi field argument or an analytical argument based on hyperbolic trigonometry. The fact that $d(A,X)$ is minimized whenever $[AX] \perp [BC]$ again can easily be extended to the hyperbolic case. Observe that this approach does not need existence of the orthocenter. In fact, existence of the orthocenter is a corollary of this construction, in the sense that the altitudes of ABC are the angle bissectors of XYZ and since the latter are known to intersect, so do the former. Therefore, Theorem 57 remains true in hyperbolic geometry.

As in Euclidean geometry, the key synthetic geometry construction is to draw the altitudes. We outline a very explicit implementation of the construction in the Poincaré upper half-space conformal model $\mathbb{H}^2$ of the hyperbolic plane. Recall that, in that space, the metric is $ds^2 = \frac{dx^2+dy^2}{y^2}$ and the geodesics are half-circles with their centers on the x-axis. Let the edge $[BC]$ of the hyperbolic triangle lie on the half-circle c_{BC} with center O on the x-axis and radius r. Assume the vertex A lies outside the circle (the case where A lies inside the circle is very much the same). The foot X of the altitude AX is $c_{BC} \cap c$, where c is the circle with its center on the x-axis, passing through A, and orthogonal to c_{BC}. The problem is to construct the circle c. Recall that, if two circles are mutually orthogonal, the inversion relative to the center of one of them is an involution on the other [84]; for example, $i_O : c \to c$, where i_0 is defined as follows: For any $W \in c$, $W' = i_0(W)$ is defined to be the point on the ray OW such that $||OW||.||OW'|| = r^2$, where $||.||$ denotes the Euclidean norm. This means that to find another point besides A on the circle c it suffices to find $i_O(A)$. To find the latter, we draw the tangent AT to the cicle c_{BC} where T is the contact point of the tangent, and the projection of T on the ray OA yields a second point $i_O(A) = A'$. Once we know that c passes through A, A' and has its center on the x-axis, its construction is trivial. The same procedure is trivially repeated for the other altitudes.

Here, we provide a unified differential calculus solution to the hyperbolic Fagnano problem, by explicitly working out the hyperbolic trigonometry formula. The major advantage of this differential calculus approach, as opposed to the synthetic geometry methods, is that it provides an explicit characterization of the hyperbolic orthic triangle, and hence of the period three orbit, in terms of the $\triangle ABC$ data. In particular, Formula (14.2) is probably the most important result.

Before jumping to the main theorem, we need a lemma:

Lemma 17 *Let $\triangle ABC$ be a geodesic triangle in a manifold with constant sectional curvature $\kappa \leq 0$. Let $Z \in [AB], X \in [BC]$. Then*

$$\frac{\partial d(Z,X)}{\partial d(B,X)} = \cos \angle ZXB$$

Consequently, Equation (14.1) remains valid in Riemannian geometry.

Proof. The result is obvious in the Euclidean geometry case $\kappa = 0$; hence, it suffices to prove it for $\kappa < 0$. Let $x := d(B,X)$ and consider a point $X' \in [BC]$ such that $d(B,X') = x + dx$. Draw the geodesic $[ZX']$. Draw the hyperbolic circle with its center at Z and radius $d(Z,X)$. This circle intersects $[ZX']$ at a point P. Recall that, in the situation of two nearby geodesics $[ZX]$ and $[ZX']$ emanating from a common point Z, the Jacobi field is orthogonal to the nominal geodesic. Hence, treating $[ZX']$ as the nominal geodesic, it follows that

$[XP] \perp [ZX']$. Consider the right angle triangle $[XPX']$. By basic hyperbolic trigonometry,

$$\cos \angle PX'X = \frac{\tanh d(P,X')}{\tanh d(X,X')}$$

At the limit $X' \to X$, we get $\tanh d(P,X') = d(P,X')$ and $\tanh d(X,X') = d(X,X')$. Hence

$$\cos \angle ZXB = \cos \angle ZX'B = \cos \angle PX'X = \frac{d(P,X')}{d(X,X')} = \frac{\partial d(Z,X)}{\partial d(B,X)}$$

and the result follows. ■

Remark: The above lemma can also be rederived from the cosine rule, but at the expense of longer–and less conceptual–manipulation. Let $x = d(B,X)$, $z = d(A,Z)$, and $\beta = \angle ZBX$. Consider as in the lemma the triangle $[BZX]$. From the cosine rule

$$\cosh d(X,Z) = \cosh x \cosh(c-z) - \sinh x \sinh(c-z) \cos \beta$$

it is easily found that

$$\begin{aligned}
\frac{\partial d(X,Z)}{\partial x} &= \frac{\cosh(c-z)\sinh(x) - \sinh(c-z)\cos(\beta)\cosh(x)}{\sinh(d(z,x))} \\
&= \frac{\cosh(c-z)\sinh^2(x) - \sinh(c-z)\sinh(x)\cos(\beta)\cosh(x)}{\sinh(d(z,x))\sinh(x)} \\
&= \frac{\cosh(c-z)\cosh^2(x) - \sinh(c-z)\sinh(x)\cos(\beta)\cosh(x) - \cosh(c-z)}{\sinh(d(z,x))\sinh(x)} \\
&= \frac{(\cosh(c-z)\cosh(x) - \sinh(c-z)\sinh(x)\cos(\beta))\cosh(x) - \cosh(c-z)}{\sinh(d(z,x))\sinh(x)} \\
&= \frac{\cosh(d(z,x))\cosh(x) - \cosh(c-z)}{\sinh(d(z,x))\sinh(x)} \\
&= \cos(\angle ZXB)
\end{aligned}$$

Hence we have obtained an alternate proof of the lemma.

Theorem 59 *Let $\triangle ABC$ be a geodesic triangle in a hyperbolic surface of constant negative curvature $\kappa = -1$. Let the altitudes $[AX], [BY], [CZ]$ be such that $x \in (BC)$, $y \in (AC)$ and $z \in (AB)$. Then the solution to (13.1) is a critical point of the mapping $(X,Y,Z) \mapsto d(X,Y) + d(Y,Z) + d(Z,X)$. Furthermore, this critical point is given by the feet X, Y, Z of the altitudes. In other words, $[XY] \cup [YZ] \cup [ZX]$ is the periodic orbit. Furthermore, this solution satisfies the second order variation test. Finally, this solution is unique.*

The proof is broken down in the following five subsections.

14.5.1 First order conditions

Given that ΔABC is a geodesic triangle in a constant negative curvature hyperbolic space M, then this triangle is uniquely specified up to isometry by the three internal angles, α, β, γ at the vertices A, B, C, respectively, provided that $\alpha + \beta + \gamma < \pi$. Given that a, b, c are the lengths of the sides opposite to the angles α, β, γ, respectively and X, Y, Z are arbitrary points in $[BC], [CA], [AB]$ respectively, then X can be defined as a mapping which maps $x \in [0, a]$ to the point $X(x) \in [BC]$, such that $d(X(x), B) = x$ with similar definitions for Y and Z. Given that

$$F(x, y, z) = d(X(x), Y(y)) + d(Y(y), Z(y)) + d(Z(z), X(x)),$$

it follows that

$$\delta_F(\Delta ABC) = \inf \left\{ F(x, y, z) : \begin{array}{c} 0 \le x \le a \\ 0 \le y \le b \\ 0 \le z \le c \end{array} \right\}$$

Clearly, the fatness is to be computed via the hyperbolic Cosine Rule I:

$$\begin{aligned}
\cosh\left(\sqrt{-\kappa}d(x, y)\right) &= \cosh\left(\sqrt{-\kappa}(a - x)\right)\cosh\left(\sqrt{-\kappa}(y)\right) - \sinh\left(\sqrt{-\kappa}(a - x)\right)\sinh\left(\sqrt{-\kappa}(y)\right)\cos(\gamma) \\
\cosh\left(\sqrt{-\kappa}d(y, z)\right) &= \cosh\left(\sqrt{-\kappa}(b - y)\right)\cosh\left(\sqrt{-\kappa}(z)\right) - \sinh\left(\sqrt{-\kappa}(b - y)\right)\sinh\left(\sqrt{-\kappa}(z)\right)\cos(\alpha) \\
\cosh\left(\sqrt{-\kappa}d(z, x)\right) &= \cosh\left(\sqrt{-\kappa}(c - z)\right)\cosh\left(\sqrt{-\kappa}(x)\right) - \sinh\left(\sqrt{-\kappa}(c - z)\right)\sinh\left(\sqrt{-\kappa}(x)\right)\cos(\beta)
\end{aligned}$$

where $d(x, y)$ is a short for $d(X(x), Y(y))$ with a similar convention for $d(y, z), d(z, x)$.
Taking partial derivatives for the hyperbolic Cosine Rule I yields the following result:

$$\begin{aligned}
\frac{\partial}{\partial x}\cosh(d(x, y)) &= \sinh(d(x, y))\frac{\partial}{\partial x}d(x, y) \\
&= \frac{\partial}{\partial x}(\cosh(a - x)\cosh(y) - \sinh(a - x)\sinh(y)\cos(\gamma)) \\
&= -\sinh(a - x)\cosh y + \cosh(a - x)\sinh y \cos\gamma
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial x}\cosh(d(z, x)) &= \sinh(d(z, x))\frac{\partial}{\partial x}d(z, x) \\
&= \frac{\partial}{\partial x}(\cosh(c - z)\cosh(x) - \sinh(c - z)\sinh(x)\cos(\beta)) \\
&= \cosh(c - z)\sinh x - \sinh(c - z)\cosh x \cos\beta
\end{aligned}$$

Manipulating the above and proceeding the same way for the other partial derivatives, the explicit expressions for the first order partial derivatives in terms of the triangle data are as follows:

$$\frac{\partial}{\partial x}d(x, y) = \frac{\sinh(-a + x)\cosh y + \cosh(-a + x)\sinh y \cos\gamma}{\sinh(d(x, y))}$$

$$\frac{\partial}{\partial y}d\left(x,y\right)=\frac{\cosh\left(-a+x\right)\sinh y+\sinh\left(-a+x\right)\cosh y\cos\gamma}{\sinh\left(d\left(x,y\right)\right)}$$

$$\frac{\partial}{\partial z}d\left(x,y\right)=0$$

$$\frac{\partial}{\partial y}d\left(y,z\right)=\frac{\sinh\left(-b+y\right)\cosh z+\cosh\left(-b+y\right)\sinh z\cos\alpha}{\sinh\left(d\left(y,z\right)\right)}$$

$$\frac{\partial}{\partial z}d\left(y,z\right)=\frac{\cosh\left(-b+y\right)\sinh z+\sinh\left(-b+y\right)\cosh z\cos\alpha}{\sinh\left(d\left(y,z\right)\right)}$$

$$\frac{\partial}{\partial x}d\left(y,z\right)=0$$

$$\frac{\partial}{\partial z}d\left(z,x\right)=\frac{\sinh\left(-c+z\right)\cosh x+\sinh\left(-c+z\right)\sinh x\cos\beta}{\sinh\left(d\left(z,x\right)\right)}$$

$$\frac{\partial}{\partial x}d\left(z,x\right)=\frac{\cosh\left(c-z\right)\sinh x-\sinh\left(c-z\right)\cosh x\cos\beta}{\sinh\left(d\left(z,x\right)\right)}$$

$$\frac{\partial}{\partial y}d\left(z,x\right)=0$$

The first order variation of $F\left(x,y,z\right)$ relative to x yields the following result:

$$\begin{aligned}0&=\frac{\partial}{\partial x}F\left(x,y,z\right)=\frac{\partial}{\partial x}\left(d\left(x,y\right)+d\left(y,z\right)+d\left(z,x\right)\right)\\&=\frac{\partial}{\partial x}\left(d\left(x,y\right)+d\left(z,x\right)\right)\\\frac{\partial}{\partial x}d\left(x,y\right)&=-\frac{\partial}{\partial x}d\left(z,x\right)\end{aligned}$$

The Cosine Rule I in ΔYXC yields

$$\begin{aligned}\cos\left(\angle YXC\right)&=\frac{\cosh\left(d\left(x,y\right)\right)\cosh\left(a-x\right)-\cosh\left(y\right)}{\sinh\left(d\left(x,y\right)\right)\sinh\left(a-x\right)}\\&=\frac{\left(\cosh\left(a-x\right)\cosh\left(y\right)-\sinh\left(a-x\right)\sinh\left(y\right)\cos\left(\gamma\right)\right)\cosh\left(a-x\right)-\cosh\left(y\right)}{\sinh\left(d\left(x,y\right)\right)\sinh\left(a-x\right)}\\&=\frac{\cosh^{2}\left(a-x\right)\cosh\left(y\right)-\sinh\left(a-x\right)\sinh\left(y\right)\cos\left(\gamma\right)\cosh\left(a-x\right)-\cosh\left(y\right)}{\sinh\left(d\left(x,y\right)\right)\sinh\left(a-x\right)}\\&=\frac{\sinh^{2}\left(a-x\right)\cosh\left(y\right)-\sinh\left(a-x\right)\sinh\left(y\right)\cos\left(\gamma\right)\cosh\left(a-x\right)}{\sinh\left(d\left(x,y\right)\right)\sinh\left(a-x\right)}\\&=\frac{\sinh\left(a-x\right)\cosh\left(y\right)-\sinh\left(y\right)\cos\left(\gamma\right)\cosh\left(a-x\right)}{\sinh\left(d\left(x,y\right)\right)}\\&=-\frac{\partial}{\partial x}d\left(x,y\right)\end{aligned}$$

The same Cosine Rule I in ΔZXB yields

$$\begin{aligned}\cos\left(\angle ZXB\right) &= \frac{\cosh\left(d\left(z,x\right)\right)\cosh\left(x\right)-\cosh\left(c-z\right)}{\sinh\left(d\left(z,x\right)\right)\sinh\left(x\right)}\\ &= \frac{\left(\cosh\left(c-z\right)\cosh\left(x\right)-\sinh\left(c-z\right)\sinh\left(x\right)\cos\left(\beta\right)\right)\cosh\left(x\right)-\cosh\left(c-z\right)}{\sinh\left(d\left(z,x\right)\right)\sinh\left(x\right)}\\ &= \frac{\cosh\left(c-z\right)\cosh^2\left(x\right)-\sinh\left(c-z\right)\sinh\left(x\right)\cos\left(\beta\right)\cosh\left(x\right)-\cosh\left(c-z\right)}{\sinh\left(d\left(z,x\right)\right)\sinh\left(x\right)}\\ &= \frac{\cosh\left(c-z\right)\sinh^2\left(x\right)-\sinh\left(c-z\right)\sinh\left(x\right)\cos\left(\beta\right)\cosh\left(x\right)}{\sinh\left(d\left(z,x\right)\right)\sinh\left(x\right)}\\ &= \frac{\cosh\left(c-z\right)\sinh\left(x\right)-\sinh\left(c-z\right)\cos\left(\beta\right)\cosh\left(x\right)}{\sinh\left(d\left(z,x\right)\right)}\\ &= \frac{\partial}{\partial x}d\left(z,x\right)\end{aligned}$$

The preceding three results imply that

$$\cos\left(\angle YXC\right)=\cos\left(\angle ZXB\right)=:\cos\left(\theta_x\right)$$

This is the *hyperbolic Fermat principle*, saying that a light ray emanating from Y, reflecting at $X\in[BC]$, to reach Z would have its reflection angle equal to its incidence angle. Next, cancelling the first order variation relative to y yields

$$\cos\left(\angle ZYA\right)=\cos\left(\angle XYC\right)=:\cos\left(\theta_y\right)$$

Finally, cancelling the first order variation relative to z yields

$$\cos\left(\angle XZB\right)=\cos\left(\angle YZA\right)=:\cos\left(\theta_z\right)$$

For the optimization problem to be a differentiable one, it is hence necessary that there exists an inscribed geodesic triangle ΔXYZ such that the reflection angles of its edges on ΔABC equal the corresponding incidence angles. In Euclidean geometry, this is equivalent to saying that ΔABC has acute angles only. The argument in hyperbolic geometry is, however, more complicated.

14.5.2 Hyperbolic orthocenter construction

It is easily seen that, for a hyperbolic geodesic triangle ΔABC, there exists a point $X\in[BC]$ such that $[AX]\perp[BC]$ if the angles $\angle ABX$ and $\angle ACX$ are acute. Therefore, if the triangle ΔABC has no obtuse angles, there are points $X\in[BC]$, $Y\in[AC]$, $Z\in[AB]$ such that $[AX]\perp[BC]$, $[BY]\perp[AC]$, $[CZ]\perp[AB]$, respectively. Even though it is not known at present whether $[AX]\cap[BY]\cap[CZ]\neq\emptyset$, this construction yields an inscribed triangle ΔXYZ, which has the property that its reflection angles on the edges of ΔABC equal the corresponding incidence angles.

Before proceeding to the proof of the equality between the incidence and reflection angles, a very explicit implementation of the construction in the

Poincaré upper half-space conformal model $\mathbb{H}^2$ of the hyperbolic plane can be derived as follows: Recall that, in that space, the metric is $ds^2 = \frac{dx^2+dy^2}{y^2}$ and the geodesics are half-circles with their centers on the x-axis. Let the edge $[BC]$ of the hyperbolic triangle lie on the half-circle c_{BC} with center O on the x-axis and radius r. Assume the vertex A lies outside the circle (the case where A lies inside the circle is very much the same). The foot X of the altitude AX is $c_{BC} \cap c$, where c is the circle with its center on the x-axis, passing through A, and orthogonal to c_{BC}. The problem is to construct the circle c. Recall that, if two circles are mutually orthogonal, the inversion relative to the center of one of them is an involution on the other [84]; for example, $i_O : c \to c$, where i_0 is defined as follows: For any $W \in c$, $W' = i_0(W)$ is defined to be the point on the ray OW such that $||OW||.||OW'|| = r^2$, where $||.||$ denotes the Euclidean norm. This means that to find another point besides A on the circle c it suffices to find $i_O(A)$. To find the latter, draw the tangent AT to the cicle c_{BC} where T is the contact point of the tangent, and the projection of T on the ray OA yields a second point $i_O(A) = A'$ on the circle c. Once knowing that c passes through A, A' and has its center on the x-axis, its construction is trivial. The same procedure is trivially repeated for the other altitudes.

Lemma 18 *Given that ΔABC is a geodesic triangle with three internal angles, $\alpha, \beta \leq \frac{\pi}{2}$ and $\gamma = \frac{\pi}{2}$ at the vertices A, B, C, then the hyperbolic form of Pythagoras' theorem is given by the following formula:*

$$\cosh\left(\sqrt{-\kappa}c\right) = \cosh\left(\sqrt{-\kappa}a\right)\cosh\left(\sqrt{-\kappa}b\right)$$

In addition, the following relations hold.

$$\begin{aligned}\tanh\left(\sqrt{-\kappa}b\right) &= \sinh\left(\sqrt{-\kappa}a\right)\tan\beta\\ \sinh\left(\sqrt{-\kappa}b\right) &= \sinh\left(\sqrt{-\kappa}c\right)\sin\beta\\ \tanh\left(\sqrt{-\kappa}a\right) &= \tanh\left(\sqrt{-\kappa}c\right)\cos\beta.\end{aligned}$$

Hyperbolic trigonometry of the right-angled subtriangles of ΔABC yields

$$\begin{aligned}\tanh x &= \tanh c\cos\beta\\ \tanh y &= \tanh a\cos\gamma\\ \tanh z &= \tanh b\cos\alpha\end{aligned}$$

$$\begin{aligned}\tanh(a-x) &= \tanh b\cos\gamma\\ \tanh(b-y) &= \tanh c\cos\alpha\\ \tanh(c-z) &= \tanh a\cos\beta\end{aligned}$$

From the Cosine Rule 1 applied to the triangles ΔZBX and ΔYCX, $d(z,x)$ and $d(x,y)$ can be expressed as follows:

$$\begin{aligned}\cosh d(z,x) &= (\cosh(c-z)\cosh x - \sinh(c-z)\sinh x\cos\beta)\\ &= (\cosh(c-z)\cosh x)(1-\tanh(c-z)\tanh x\cos\beta)\\ &= (\cosh(c-z)\cosh x)\left(1-\tanh c\tanh a\cos^3\beta\right)\end{aligned}$$

$$\begin{aligned}\cosh d\,(x,y) &= (\cosh\,(a-x)\cosh y - \sinh\,(a-x)\sinh y\cos\gamma)\\ &= (\cosh\,(a-x)\cosh y)\,(1-\tanh\,(a-x)\tanh y\cos\gamma)\\ &= (\cosh\,(a-x)\cosh y)\left(1-\tanh a\tanh b\cos^3\gamma\right)\end{aligned}$$

Given that θ_x^l denotes $\angle ZXB$ and θ_x^r denotes $\angle YXC$, then the Sine Rule in triangle ΔZBX and ΔYCX yield the following results:

$$\begin{aligned}\sin^2\theta_x^l &= \left(\sin^2\beta\right)\frac{\sinh^2(c-z)}{\sinh^2 d\,(z,x)} = \left(\sin^2\beta\right)\frac{\sinh^2(c-z)}{\cosh^2 d\,(z,x)-1}\\ &= \left(\sin^2\beta\right)\frac{\sinh^2(c-z)}{(\cosh\,(c-z)\cosh x)^2\,(1-\tanh a\tanh c\cos^3\beta)^2-1}\\ \sin^2\theta_x^r &= \left(\sin^2\gamma\right)\frac{\sinh^2(y)}{\sinh^2 d\,(x,y)} = \left(\sin^2\gamma\right)\frac{\sinh^2(y)}{\cosh^2 d\,(x,y)-1}\\ &= \left(\sin^2\gamma\right)\frac{\sinh^2(y)}{(\cosh\,(a-x)\cosh y)^2\,(1-\tanh a\tanh b\cos^3\gamma)^2-1}\end{aligned}$$

Now, using the above expressions yield the following computation:

$$\begin{aligned}\sinh^2(c-z) &= \frac{\tanh^2(c-z)}{1-\tanh^2(c-z)} = \frac{(\tanh a\cos\beta)^2}{1-(\tanh a\cos\beta)^2}\\ \cosh^2(c-z) &= \frac{1}{1-\tanh^2(c-z)} = \frac{1}{1-(\tanh a\cos\beta)^2}\end{aligned}$$

$$\begin{aligned}\sinh^2 y &= \frac{\tanh^2 y}{1-\tanh^2 y} = \frac{(\tanh a\cos\gamma)^2}{1-(\tanh a\cos\gamma)^2}\\ \cosh^2 y &= \frac{1}{1-\tanh^2 y} = \frac{1}{1-(\tanh a\cos\gamma)^2}\end{aligned}$$

$$\begin{aligned}\cosh^2 x &= \frac{1}{1-\tanh^2 x} = \frac{1}{1-(\tanh c\cos\beta)^2}\\ \cosh^2(a-x) &= \frac{1}{1-\tanh^2(a-x)} = \frac{1}{1-(\tanh b\cos\gamma)^2}\end{aligned}$$

$$\begin{aligned}\tanh^2 a &= \frac{\cosh^2 a-1}{\cosh^2 a}\\ &= \frac{(\cos\beta\cos\gamma+\cos\alpha)^2-\left(1-\cos^2\beta\right)\left(1-\cos^2\gamma\right)}{(\cos\beta\cos\gamma+\cos\alpha)^2}\\ &= \frac{\left(2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma-1\right)}{(\cos\beta\cos\gamma+\cos\alpha)^2}\end{aligned}$$

$$\tanh^2 b = \frac{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)}{\left(\cos\gamma\cos\alpha + \cos\beta\right)^2}$$

$$\tanh^2 c = \frac{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)}{\left(\cos\alpha\cos\beta + \cos\gamma\right)^2}$$

$$\begin{aligned}
\sin^2\theta_x^l &= \left(\sin^2\beta\right)\frac{\frac{(\tanh a\cos\beta)^2}{1-(\tanh a\cos\beta)^2}}{\left(\frac{1}{1-(\tanh a\cos\beta)^2}\,\frac{1}{1-(\tanh c\cos\beta)^2}\right)(1-\tanh a\tanh c\cos^3\beta)^2 - 1} \\
&= \frac{\left(\sin^2\beta\right)(\tanh a)^2\left(1-(\tanh c\cos\beta)^2\right)}{\big(\tanh^2 a - 2\tanh a\tanh c\cos\beta + \tanh^2 c - \tanh^2 a\tanh^2 c\cos^2\beta + \tanh^2 a\tanh^2 c\cos^4\beta} \\
\sin^2\theta_x^r &= \left(\sin^2\gamma\right)\frac{\frac{(\tanh a\cos\gamma)^2}{1-(\tanh a\cos\gamma)^2}}{\left(\frac{1}{1-(\tanh b\cos\gamma)^2}\,\frac{1}{1-(\tanh a\cos\gamma)^2}\right)(1-\tanh a\tanh b\cos^3\gamma)^2 - 1} \\
&= \frac{\left(\sin^2\gamma\right)(\tanh a)^2\left(1-(\tanh b\cos\gamma)^2\right)}{\big(\tanh^2 a - 2\tanh a\tanh b\cos\gamma + \tanh^2 b - \tanh^2 a\tanh^2 b\cos^2\gamma + \tanh^2 a\tanh^2 b\cos^4\gamma}
\end{aligned}$$

Finally substituting $\tanh a, \tanh b, \tanh c$ for their values given as:

$$\begin{aligned}
\tanh^2 a &= \frac{\cosh^2 a - 1}{\cosh^2 a} \\
&= \frac{(\cos\beta\cos\gamma + \cos\alpha)^2 - \left(1-\cos^2\beta\right)\left(1-\cos^2\gamma\right)}{(\cos\beta\cos\gamma + \cos\alpha)^2} \\
&= \frac{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)}{(\cos\beta\cos\gamma + \cos\alpha)^2}
\end{aligned}$$

$$\tanh^2 b = \frac{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)}{\left(\cos\gamma\cos\alpha + \cos\beta\right)^2}$$

$$\tanh^2 c = \frac{\left(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\right)}{\left(\cos\alpha\cos\beta + \cos\gamma\right)^2}$$

$$\sin^2\theta_x^l = \sin^2\theta_x^r = \frac{\cos^2\beta + \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}{\cos^2\alpha + \cos^2\beta + \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}$$

yields,

$$\begin{aligned}
\theta_x^l &= \theta_x^r && (14.2) \\
&= \arcsin\left(\frac{\sqrt{\cos^2\beta + \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}}{\sqrt{\cos^2\alpha + \cos^2\beta + \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}}\right) && (14.3)
\end{aligned}$$

This proves that the reflection angle at X equals the incidence angle at the same point. The same fact is easily proved for the points Y, Z. Therefore, the hyperbolic orthocenter construction yields an inscribed triangle with its incidence angles equal the corresponding reflection angles.

It follows from the preceding this that the altitudes $[AX], [BY], [CZ]$ of the triangle ΔABC are the angle bisectors of the triangle ΔXYZ and hence intersect in a single point.

Corollary 14 *The altitudes of a constant curvature hyperbolic geodesic triangle intersect at a single point, called hyperbolic orthocenter.*

14.5.3 Second order variation

The second order partial derivatives at the critical point can be computed as follows:

$$
\begin{aligned}
&\frac{\partial^2}{\partial x^2} d\left(x,y\right)\\
&= \frac{\partial}{\partial x}\left(\frac{\sinh\left(-a+x\right)\cosh y+\cosh\left(-a+x\right)\sinh y\sin\gamma}{\sinh\left(d\left(x,y\right)\right)}\right)\\
&= \frac{\left(\cosh\left(-a+x\right)\cosh y+\sinh\left(-a+x\right)\sinh y\sin\gamma\right)}{\sinh\left(d\left(x,y\right)\right)}\\
&\quad -\frac{\left(\sinh\left(-a+x\right)\cosh y+\cosh\left(-a+x\right)\sinh y\sin\gamma\right)}{\left(\sinh\left(d\left(x,y\right)\right)\right)^2}\cosh\left(d\left(x,y\right)\right)\frac{\partial}{\partial x}d\left(x,y\right)\\
&= \frac{1}{\tanh\left(d\left(x,y\right)\right)}-\frac{\left(\sinh\left(-a+x\right)\cosh y+\cosh\left(-a+x\right)\sinh y\cos\gamma\right)^2}{\left(\sinh\left(d\left(x,y\right)\right)\right)^2}\frac{\cosh\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\\
&= \frac{\sin^2\left(\theta_x\right)}{\tanh\left(d\left(x,y\right)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial y^2} d\left(x,y\right)\\
&= \frac{\partial}{\partial y}\left(\frac{\cosh\left(-a+x\right)\sinh y+\sinh\left(-a+x\right)\cosh y\cos\gamma}{\sinh\left(d\left(x,y\right)\right)}\right)\\
&= \frac{\left(\cosh\left(-a+x\right)\cosh y+\sinh\left(-a+x\right)\sinh y\cos\gamma\right)}{\sinh\left(d\left(x,y\right)\right)}\\
&\quad -\frac{\left(\cosh\left(-a+x\right)\sinh y+\sinh\left(-a+x\right)\cosh y\cos\gamma\right)}{\left(\sinh\left(d\left(x,y\right)\right)\right)^2}\cosh\left(d\left(x,y\right)\right)\frac{\partial}{\partial y}d\left(x,y\right)\\
&= \frac{1}{\tanh\left(d\left(x,y\right)\right)}-\frac{\left(\cosh\left(-a+x\right)\sinh y+\sinh\left(-a+x\right)\cosh y\cos\gamma\right)^2}{\left(\sinh\left(d\left(x,y\right)\right)\right)^2}\frac{\cosh\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\\
&= \frac{\sin^2\left(\theta_y\right)}{\tanh\left(d\left(x,y\right)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial x\partial y}d\left(x,y\right)\\
&= \frac{\partial}{\partial y}\left(\frac{\sinh\left(-a+x\right)\cosh y+\cosh\left(-a+x\right)\sinh y\cos\gamma}{\sinh\left(d\left(x,y\right)\right)}\right)\\
&= \frac{\left(\sinh\left(-a+x\right)\sinh y+\cosh\left(-a+x\right)\cosh y\cos\gamma\right)}{\sinh\left(d\left(x,y\right)\right)}\\
&\quad -\frac{\left(\sinh\left(-a+x\right)\cosh y+\cosh\left(-a+x\right)\sinh y\cos\gamma\right)}{\left(\sinh\left(d\left(x,y\right)\right)\right)^2}\cosh\left(d\left(x,y\right)\right)\frac{\partial}{\partial y}d\left(x,y\right)\\
&= \frac{\left(-\sinh\left(a-x\right)\sinh y+\cosh\left(a-x\right)\cosh y\cos\gamma\right)}{\sinh\left(d\left(x,y\right)\right)}+\cos\left(\theta_x\right)\frac{\cosh\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\cos\left(\theta_y\right)\\
&= \frac{\left(-\sinh\left(a-x\right)\sinh y+\left(\cosh\left(d\left(x,y\right)\right)+\sinh\left(a-x\right)\sinh\left(y\right)\cos\left(\gamma\right)\right)\cos\gamma\right)}{\sinh\left(d\left(x,y\right)\right)}\\
&\quad +\cos\left(\theta_x\right)\frac{\cosh\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\cos\left(\theta_y\right)\\
&= \frac{-\sinh\left(a-x\right)\sinh y+\cosh\left(d\left(x,y\right)\right)\cos\gamma+\sinh\left(a-x\right)\sinh\left(y\right)\cos^2\left(\gamma\right)}{\sinh\left(d\left(x,y\right)\right)}\\
&\quad +\cos\left(\theta_x\right)\frac{\cosh\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\cos\left(\theta_y\right)\\
&= \frac{\cosh\left(d\left(x,y\right)\right)\cos\gamma-\sinh\left(a-x\right)\sinh\left(y\right)\sin^2\left(\gamma\right)}{\sinh\left(d\left(x,y\right)\right)}+\cos\left(\theta_x\right)\frac{\cosh\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\cos\left(\theta_y\right)\\
&= \frac{\cosh\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\left(\cos\left(\theta_x\right)\cos\left(\theta_y\right)+\cos\gamma\right)-\frac{\sinh\left(a-x\right)\sinh\left(y\right)\sin^2\left(\gamma\right)}{\sinh\left(d\left(x,y\right)\right)}\\
&= \frac{\cosh^2\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\sin\left(\theta_x\right)\sin\left(\theta_y\right)-\sinh\left(y\right)\sin\left(\gamma\right)\sin\left(\theta_y\right)\\
&= \frac{\cosh^2\left(d\left(x,y\right)\right)}{\sinh\left(d\left(x,y\right)\right)}\sin\left(\theta_x\right)\sin\left(\theta_y\right)-\sinh\left(d\left(x,y\right)\right)\sin\left(\theta_x\right)\sin\left(\theta_y\right)\\
&= \frac{\sin\left(\theta_x\right)\sin\left(\theta_y\right)}{\sinh\left(d\left(x,y\right)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial y^2} d\left(y,z\right)\\
&\quad = \quad \frac{\partial}{\partial y}\left(\frac{\sinh\left(-b+y\right)\cosh z+\cosh\left(-b+y\right)\sinh z\cos\alpha}{\sinh\left(d\left(y,z\right)\right)}\right)\\
&\quad = \quad \frac{\left(\cosh\left(-b+y\right)\cosh z+\sinh\left(-b+y\right)\sinh z\cos\alpha\right)}{\sinh\left(d\left(y,z\right)\right)}\\
&\qquad\quad -\frac{\left(\sinh\left(-b+y\right)\cosh z+\cosh\left(-b+y\right)\sinh z\cos\alpha\right)}{\left(\sinh\left(d\left(y,z\right)\right)\right)^2}\cosh\left(d\left(y,z\right)\right)\frac{\partial}{\partial y}d\left(y,z\right)\\
&\quad = \quad \frac{1}{\tanh\left(d\left(y,z\right)\right)}-\frac{\left(\sinh\left(-b+y\right)\cosh z+\cosh\left(-b+y\right)\sinh z\cos\alpha\right)^2}{\left(\sinh\left(d\left(y,z\right)\right)\right)^2}\frac{\cosh\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\\
&\quad = \quad \frac{\sin^2\left(\theta_y\right)}{\tanh\left(d\left(y,z\right)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial z^2} d\left(y,z\right)\\
&\quad = \quad \frac{\partial}{\partial z}\left(\frac{\cosh\left(-b+y\right)\sinh z+\sinh\left(-b+y\right)\cosh z\cos\alpha}{\sinh\left(d\left(y,z\right)\right)}\right)\\
&\quad = \quad \frac{\left(\cosh\left(-b+y\right)\cosh z+\sinh\left(-b+y\right)\sinh z\cos\alpha\right)}{\sinh\left(d\left(y,z\right)\right)}\\
&\qquad\quad -\frac{\left(\cosh\left(-b+y\right)\sinh z+\sinh\left(-b+y\right)\cosh z\cos\alpha\right)}{\sinh^2\left(d\left(y,z\right)\right)}\cosh\left(d\left(y,z\right)\right)\frac{\partial}{\partial z}d\left(y,z\right)\\
&\quad = \quad \frac{1}{\tanh\left(d\left(y,z\right)\right)}-\frac{\left(\cosh\left(-b+y\right)\sinh z+\sinh\left(-b+y\right)\cosh z\cos\alpha\right)^2}{\sinh^2\left(d\left(y,z\right)\right)}\frac{\cosh\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\\
&\quad = \quad \frac{\sin^2\left(\theta_z\right)}{\tanh\left(d\left(y,z\right)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial y \partial z} d\left(y,z\right) \\
&= \frac{\partial}{\partial z}\left(\frac{\sinh\left(-b+y\right)\cosh z + \cosh\left(-b+y\right)\sinh z \cos\alpha}{\sinh\left(d\left(y,z\right)\right)}\right) \\
&= \frac{\left(\sinh\left(-b+y\right)\sinh z + \cosh\left(-b+y\right)\cosh z \cos\alpha\right)}{\sinh\left(d\left(y,z\right)\right)} \\
&\quad -\frac{\left(\sinh\left(-b+y\right)\cosh z + \cosh\left(-b+y\right)\sinh z \cos\alpha\right)}{\left(\sinh\left(d\left(y,z\right)\right)\right)^2}\cosh\left(d\left(y,z\right)\right)\frac{\partial}{\partial z} d\left(y,z\right) \\
&= \frac{\left(-\sinh\left(b-y\right)\sinh z + \cosh\left(b-y\right)\cosh z \cos\alpha\right)}{\sinh\left(d\left(y,z\right)\right)} + \cos\left(\theta_y\right)\frac{\cosh\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\cos\left(\theta_z\right) \\
&= \frac{\left(-\sinh\left(b-y\right)\sinh z + \left(\cosh\left(d\left(y,z\right)\right) + \sinh\left(b-y\right)\sinh\left(z\right)\cos\left(\alpha\right)\right)\cos\alpha\right)}{\sinh\left(d\left(y,z\right)\right)} \\
&\quad + \cos\left(\theta_y\right)\frac{\cosh\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\cos\left(\theta_z\right) \\
&= \frac{-\sinh\left(b-y\right)\sinh z + \cosh\left(d\left(y,z\right)\right)\cos\alpha + \sinh\left(b-y\right)\sinh\left(z\right)\cos^2\left(\alpha\right)}{\sinh\left(d\left(y,z\right)\right)} \\
&\quad + \cos\left(\theta_y\right)\frac{\cosh\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\cos\left(\theta_z\right) \\
&= \frac{\cosh\left(d\left(y,z\right)\right)\cos\alpha - \sinh\left(b-y\right)\sinh\left(z\right)\sin^2\left(\alpha\right)}{\sinh\left(d\left(y,z\right)\right)} + \cos\left(\theta_y\right)\frac{\cosh\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\cos\left(\theta_z\right) \\
&= \frac{\cosh\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\left(\cos\left(\theta_y\right)\cos\left(\theta_z\right) + \cos\alpha\right) - \frac{\sinh\left(b-y\right)\sinh\left(z\right)\sin^2\left(\alpha\right)}{\sinh\left(d\left(y,z\right)\right)} \\
&= \frac{\cosh^2\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\sin\left(\theta_y\right)\sin\left(\theta_z\right) - \sinh\left(z\right)\sin\left(\alpha\right)\sin\left(\theta_z\right) \\
&= \frac{\cosh^2\left(d\left(y,z\right)\right)}{\sinh\left(d\left(y,z\right)\right)}\sin\left(\theta_y\right)\sin\left(\theta_z\right) - \sinh\left(d\left(y,z\right)\right)\sin\left(\theta_y\right)\sin\left(\theta_z\right) \\
&= \frac{\sin\left(\theta_y\right)\sin\left(\theta_z\right)}{\sinh\left(d\left(y,z\right)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial z^2} d\left(z,x\right)\\
&\quad = \quad \frac{\partial}{\partial z}\left(\frac{\sinh\left(-c+z\right)\cosh x+\cosh\left(-c+z\right)\sinh x\cos\beta}{\sinh\left(d\left(z,x\right)\right)}\right)\\
&\quad = \quad \frac{\left(\cosh\left(-c+z\right)\cosh x+\sinh\left(-c+z\right)\sinh x\cos\beta\right)}{\sinh\left(d\left(z,x\right)\right)}\\
&\qquad\quad -\frac{\sinh\left(-c+z\right)\cosh x+\cosh\left(-c+z\right)\sinh x\cos\beta}{\left(\sinh\left(d\left(z,x\right)\right)\right)^2}\cosh\left(d\left(z,x\right)\right)\frac{\partial}{\partial z}d\left(z,x\right)\\
&\quad = \quad \frac{1}{\tanh\left(d\left(z,x\right)\right)}-\frac{\left(\sinh\left(-c+z\right)\cosh x+\cosh\left(-c+z\right)\sinh x\cos\beta\right)^2}{\left(\sinh\left(d\left(z,x\right)\right)\right)^2}\frac{\cosh\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\\
&\quad = \quad \frac{\sin^2\left(\theta_z\right)}{\tanh\left(d\left(z,x\right)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial x^2} d\left(z,x\right)\\
&\quad = \quad \frac{\partial}{\partial x}\left(\frac{\sinh\left(-c+z\right)\cosh x+\sinh\left(-c+z\right)\sinh x\cos\beta}{\sinh\left(d\left(z,x\right)\right)}\right)\\
&\quad = \quad \frac{\sinh\left(-c+z\right)\sinh x+\sinh\left(-c+z\right)\cosh x\cos\beta}{\sinh\left(d\left(z,x\right)\right)}\\
&\qquad\quad -\frac{\left(\sinh\left(-c+z\right)\cosh x+\sinh\left(-c+z\right)\sinh x\cos\beta\right)}{\left(\sinh\left(d\left(z,x\right)\right)\right)^2}\cosh\left(d\left(z,x\right)\right)\frac{\partial}{\partial x}d\left(z,x\right)\\
&\quad = \quad \frac{1}{\tanh\left(d\left(z,x\right)\right)}-\frac{\left(\sinh\left(-c+z\right)\cosh x+\sinh\left(-c+z\right)\sinh x\cos\beta\right)^2}{\left(\sinh\left(d\left(z,x\right)\right)\right)^2}\frac{\cosh\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\\
&\quad = \quad \frac{\sin^2\left(\theta_x\right)}{\tanh\left(d\left(z,x\right)\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial z\partial x}d\left(z,x\right)\\
&= \frac{\partial}{\partial x}\left(\frac{\sinh\left(-c+z\right)\cosh x+\cosh\left(-c+z\right)\sinh x\cos\beta}{\sinh\left(d\left(z,x\right)\right)}\right)\\
&= \frac{\left(\sinh\left(-c+z\right)\sinh x+\cosh\left(-c+z\right)\cosh x\cos\beta\right)}{\sinh\left(d\left(z,x\right)\right)}\\
&\quad -\frac{\left(\sinh\left(-c+z\right)\cosh x+\cosh\left(-c+z\right)\sinh x\cos\beta\right)}{\left(\sinh\left(d\left(z,x\right)\right)\right)^2}\cosh\left(d\left(z,x\right)\right)\frac{\partial}{\partial x}d\left(z,x\right)\\
&= \frac{\left(-\sinh\left(c-z\right)\sinh x+\cosh\left(c-z\right)\cosh x\cos\beta\right)}{\sinh\left(d\left(z,x\right)\right)}+\cos\left(\theta_z\right)\frac{\cosh\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\cos\left(\theta_x\right)\\
&= \frac{\left(-\sinh\left(c-z\right)\sinh x+\left(\cosh\left(d\left(z,x\right)\right)+\sinh\left(c-z\right)\sinh\left(x\right)\cos\left(\beta\right)\right)\cos\beta\right)}{\sinh\left(d\left(z,x\right)\right)}\\
&\quad +\cos\left(\theta_z\right)\frac{\cosh\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\cos\left(\theta_x\right)\\
&= \frac{-\sinh\left(c-z\right)\sinh x+\cosh\left(d\left(z,x\right)\right)\cos\beta+\sinh\left(c-z\right)\sinh\left(x\right)\cos^2\left(\beta\right)}{\sinh\left(d\left(z,x\right)\right)}\\
&\quad +\cos\left(\theta_z\right)\frac{\cosh\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\cos\left(\theta_x\right)\\
&= \frac{\cosh\left(d\left(z,x\right)\right)\cos\beta-\sinh\left(c-z\right)\sinh\left(x\right)\sin^2\left(\beta\right)}{\sinh\left(d\left(z,x\right)\right)}+\cos\left(\theta_z\right)\frac{\cosh\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\cos\left(\theta_x\right)\\
&= \frac{\cosh\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\left(\cos\left(\theta_z\right)\cos\left(\theta_x\right)+\cos\beta\right)-\frac{\sinh\left(c-z\right)\sinh\left(x\right)\sin^2\left(\beta\right)}{\sinh\left(d\left(z,x\right)\right)}\\
&= \frac{\cosh^2\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\sin\left(\theta_z\right)\sin\left(\theta_x\right)-\sinh\left(x\right)\sin\left(\beta\right)\sin\left(\theta_x\right)\\
&= \frac{\cosh^2\left(d\left(z,x\right)\right)}{\sinh\left(d\left(z,x\right)\right)}\sin\left(\theta_z\right)\sin\left(\theta_x\right)-\sinh\left(d\left(z,x\right)\right)\sin\left(\theta_z\right)\sin\left(\theta_x\right)\\
&= \frac{\sin\left(\theta_z\right)\sin\left(\theta_x\right)}{\sinh\left(d\left(z,x\right)\right)}
\end{aligned}
$$

In summary, at the critical point, the second order partial derivatives are as follows:

$$
\begin{aligned}
\frac{\partial^2}{\partial x^2}d\left(x,y\right) &= \frac{\sin^2\left(\theta_x\right)}{\tanh\left(d\left(x,y\right)\right)}\\
\frac{\partial^2}{\partial y^2}d\left(x,y\right) &= \frac{\sin^2\left(\theta_y\right)}{\tanh\left(d\left(x,y\right)\right)}\\
\frac{\partial^2}{\partial x\partial y}d\left(x,y\right) &= \frac{\sin\left(\theta_x\right)\sin\left(\theta_y\right)}{\sinh\left(d\left(x,y\right)\right)}
\end{aligned}
$$

$$\begin{aligned}
\frac{\partial^2}{\partial y^2} d(y,z) &= \frac{\sin^2(\theta_y)}{\tanh(d(y,z))} \\
\frac{\partial^2}{\partial z^2} d(y,z) &= \frac{\sin^2(\theta_z)}{\tanh(d(y,z))} \\
\frac{\partial^2}{\partial y \partial z} d(y,z) &= \frac{\sin(\theta_y)\sin(\theta_z)}{\sinh(d(y,z))}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial z^2} d(z,x) &= \frac{\sin^2(\theta_z)}{\tanh(d(z,x))} \\
\frac{\partial^2}{\partial x^2} d(z,x) &= \frac{\sin^2(\theta_x)}{\tanh(d(z,x))} \\
\frac{\partial^2}{\partial z \partial x} d(z,x) &= \frac{\sin(\theta_z)\sin(\theta_x)}{\sinh(d(z,x))}
\end{aligned}$$

With those second order derivatives, the Hessian matrix H can be easily seen to be equal to

$$H = A_{xy} + A_{yz} + A_{zx}$$

$$A_{xy} = \begin{bmatrix} \frac{\sin^2(\theta_x)}{\tanh(d(x,y))} & \frac{\sin(\theta_x)\sin(\theta_y)}{\sinh(d(x,y))} & 0 \\ \frac{\sin(\theta_x)\sin(\theta_y)}{\sinh(d(x,y))} & \frac{\sin^2(\theta_y)}{\tanh(d(x,y))} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$A_{yz} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{\sin^2(\theta_y)}{\tanh(d(y,z))} & \frac{\sin(\theta_y)\sin(\theta_z)}{\sinh(d(y,z))} \\ 0 & \frac{\sin(\theta_y)\sin(\theta_z)}{\sinh(d(y,z))} & \frac{\sin^2(\theta_z)}{\tanh(d(y,z))} \end{bmatrix}$$

$$A_{zx} = \begin{bmatrix} \frac{\sin^2(\theta_x)}{\tanh(d(z,x))} & 0 & \frac{\sin(\theta_z)\sin(\theta_x)}{\sinh(d(z,x))} \\ 0 & 0 & 0 \\ \frac{\sin(\theta_z)\sin(\theta_x)}{\sinh(d(z,x))} & 0 & \frac{\sin^2(\theta_z)}{\tanh(d(z,x))} \end{bmatrix}$$

To prove that H is positive definte, let $\tilde{A}_{xy}, \tilde{A}_{yz}, \tilde{A}_{zx}$ be defined as follows:

$$\begin{aligned}
\tilde{A}_{xy} &= \begin{bmatrix} \frac{\sin^2(\theta_x)}{\tanh(d(x,y))} & \frac{\sin(\theta_x)\sin(\theta_y)}{\sinh(d(x,y))} \\ \frac{\sin(\theta_x)\sin(\theta_y)}{\sinh(d(x,y))} & \frac{\sin^2(\theta_y)}{\tanh(d(x,y))} \end{bmatrix} \\
\tilde{A}_{yz} &= \begin{bmatrix} \frac{\sin^2(\theta_y)}{\tanh(d(y,z))} & \frac{\sin(\theta_y)\sin(\theta_z)}{\sinh(d(y,z))} \\ \frac{\sin(\theta_y)\sin(\theta_z)}{\sinh(d(y,z))} & \frac{\sin^2(\theta_z)}{\tanh(d(y,z))} \end{bmatrix} \\
\tilde{A}_{zx} &= \begin{bmatrix} \frac{\sin^2(\theta_x)}{\tanh(d(z,x))} & \frac{\sin(\theta_z)\sin(\theta_x)}{\sinh(d(z,x))} \\ \frac{\sin(\theta_z)\sin(\theta_x)}{\sinh(d(z,x))} & \frac{\sin^2(\theta_z)}{\tanh(d(z,x))} \end{bmatrix}
\end{aligned}$$

To prove that $\tilde{A}_{xy}, \tilde{A}_{yz}, \tilde{A}_{zx} > 0$, it is required that θ_x, θ_y, θ_z cannot be equal to zero. Indeed, $\theta_x = 0$ would imply, by uniqueness of the geodesics in hyperbolic space, that $Z = B$ and $Y = C$. The latter would in turn imply that $\alpha = \beta = \frac{\pi}{2}$ and furthermore $\alpha+\beta+\gamma \geq \pi$, a violation of the hyperbolic condition. It follows that $\frac{\sin^2(\theta_x)}{\tanh(d(x,y))} > 0$, because $\theta_x > 0$, and $d(x, y) < \infty$ by minimality. Hence the diagonal entries of $\tilde{A}_{xy}, \tilde{A}_{yz}, \tilde{A}_{zx}$ are positive. Next, observe the following:

$$\begin{aligned}
\det\left(\tilde{A}_{xy}\right) &= \frac{\sin^2(\theta_x)}{\tanh(d(x,y))}\frac{\sin^2(\theta_y)}{\tanh(d(x,y))} - \left(\frac{\sin(\theta_x)\sin(\theta_y)}{\sinh(d(x,y))}\right)^2 \\
&= \left(\frac{\sin(\theta_x)\sin(\theta_y)}{\sinh(d(x,y))}\right)^2 \left(\cosh^2(d(x,y)) - 1\right) \\
&= \left(\sin(\theta_x)\sin(\theta_y)\right)^2 > 0
\end{aligned}$$

Thus $\tilde{A}_{xy}, \tilde{A}_{yz}, \tilde{A}_{zx}$ are positive definite. Hence $H = A_{xy} + A_{yz} + A_{zx} \geq 0$. To show that it is positive definite, let

$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

be a vector such that

$$\begin{aligned}
0 &= w^T H w \\
0 &= w^T \left(A_{xy} + A_{yz} + A_{zx}\right) w \\
0 &= w^T A_{xy} w + w^T A_{yz} w + w^T A_{zx} w
\end{aligned}$$

The above implies that

$$\begin{aligned}
0 &= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \tilde{A}_{xy} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \\
0 &= \begin{bmatrix} w_2 & w_3 \end{bmatrix} \tilde{A}_{yz} \begin{bmatrix} w_2 \\ w_3 \end{bmatrix} \\
0 &= \begin{bmatrix} w_1 & w_3 \end{bmatrix} \tilde{A}_{zx} \begin{bmatrix} w_1 \\ w_3 \end{bmatrix}
\end{aligned}$$

Since $\tilde{A}_{xy}, \tilde{A}_{yz}, \tilde{A}_{zx} > 0$, it follows that

$$w = 0$$

Therefore, the Hessian matrix

$$H = \begin{bmatrix}
\frac{\sin^2(\theta_x)}{\tanh(d(x,y))} + \frac{\sin^2(\theta_x)}{\tanh(d(z,x))} & \frac{\sin(\theta_x)\sin(\theta_y)}{\sinh(d(x,y))} & \frac{\sin(\theta_z)\sin(\theta_x)}{\sinh(d(z,x))} \\
\frac{\sin(\theta_x)\sin(\theta_y)}{\sinh(d(x,y))} & \frac{\sin^2(\theta_y)}{\tanh(d(x,y))} + \frac{\sin^2(\theta_y)}{\tanh(d(y,z))} & \frac{\sin(\theta_y)\sin(\theta_z)}{\sinh(d(y,z))} \\
\frac{\sin(\theta_z)\sin(\theta_x)}{\sinh(d(z,x))} & \frac{\sin(\theta_y)\sin(\theta_z)}{\sinh(d(y,z))} & \frac{\sin^2(\theta_z)}{\tanh(d(y,z))} + \frac{\sin^2(\theta_z)}{\tanh(d(z,x))}
\end{bmatrix}$$

is positive definite and the second order variation test passes.

The proof is complete.

14.5.4 Uniqueness

In this section, the uniqueness of the inscribed triangle that satisfies the first order variation is shown. That is, the only inscribed triangle that has its incidence angles equal to the corresponding reflection angles at X, Y, Z is the orthic triangle.

Given that ΔXYZ is an inscribed triangle which incidence angles equal to the corresponding reflection angles at X, Y, Z, and denoted these angles by $\theta_x, \theta_y, \theta_z$ respectively, then the Cosine Rule II for ΔXCY yields

$$\begin{aligned}\cosh d\left(X, C\right) &= \frac{\cos\theta_x \cos\gamma + \cos\theta_y}{\sin\theta_x \sin\gamma} \\ \cosh d\left(Y, C\right) &= \frac{\cos\theta_y \cos\gamma + \cos\theta_x}{\sin\theta_y \sin\gamma}.\end{aligned}$$

$$\begin{aligned}\sinh d\left(X, C\right) &= \sqrt{\cosh^2 d\left(X, C\right) - 1} \\ &= \frac{\sqrt{(2\cos\gamma\cos\theta_x\cos\theta_y + \cos^2\gamma + \cos^2\theta_x + \cos^2\theta_y - 1)}}{\sin\theta_x \sin\gamma} \\ \sinh d\left(Y, C\right) &= \frac{\sqrt{(2\cos\gamma\cos\theta_x\cos\theta_y + \cos^2\gamma + \cos^2\theta_x + \cos^2\theta_y - 1)}}{\sin\theta_y \sin\gamma}\end{aligned}$$

Recall that the Cosine Rule II for ΔABC yields the following formula:

$$\begin{aligned}\cosh\left(a\right) &= \frac{\cos\left(\beta\right)\cos\left(\gamma\right) + \cos\left(\alpha\right)}{\sin\left(\beta\right)\sin\left(\gamma\right)} \\ \cosh\left(b\right) &= \frac{\cos\left(\gamma\right)\cos\left(\alpha\right) + \cos\left(\beta\right)}{\sin\left(\gamma\right)\sin\left(\alpha\right)} \\ \cosh\left(c\right) &= \frac{\cos\left(\alpha\right)\cos\left(\beta\right) + \cos\left(\gamma\right)}{\sin\left(\alpha\right)\sin\left(\beta\right)}\end{aligned}$$

$$\begin{aligned}\sinh\left(a\right) &= \frac{\sqrt{2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1}}{\sin\left(\beta\right)\sin\left(\gamma\right)} \\ \sinh\left(b\right) &= \frac{\sqrt{2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1}}{\sin\left(\gamma\right)\sin\left(\alpha\right)} \\ \sinh\left(c\right) &= \frac{\sqrt{2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1}}{\sin\left(\alpha\right)\sin\left(\beta\right)}\end{aligned}$$

Then

$$\begin{aligned}
\cosh\left(d\left(X,B\right)\right) &= \cosh\left(a-d\left(X,C\right)\right)=\cosh\left(a\right)\cosh\left(d\left(X,C\right)\right)-\sinh\left(a\right)\sinh\left(d\left(X,C\right)\right)\\
&= \frac{\cos\left(\beta\right)\cos\left(\gamma\right)+\cos\left(\alpha\right)}{\sin\left(\beta\right)\sin\left(\gamma\right)}\frac{\cos\left(\gamma\right)\cos\left(\theta_x\right)+\cos\left(\theta_y\right)}{\sin\left(\gamma\right)\sin\left(\theta_x\right)}\\
&\quad -\frac{\sqrt{2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma-1}}{\sin\left(\beta\right)\sin\left(\gamma\right)}\\
&\quad \cdot\frac{\sqrt{2\cos\gamma\cos\theta_x\cos\theta_y+\cos^2\gamma+\cos^2\theta_x+\cos^2\theta_y-1}}{\sin\left(\gamma\right)\sin\left(\theta_x\right)}
\end{aligned}$$

$$\begin{aligned}
\cosh\left(d\left(Y,A\right)\right) &= \cosh\left(b-d\left(Y,C\right)\right)=\cosh\left(b\right)\cosh\left(d\left(Y,C\right)\right)-\sinh\left(b\right)\sinh\left(d\left(Y,C\right)\right)\\
&= \frac{\cos\left(\gamma\right)\cos\left(\alpha\right)+\cos\left(\beta\right)}{\sin\left(\gamma\right)\sin\left(\alpha\right)}\frac{\cos\left(\gamma\right)\cos\left(\theta_y\right)+\cos\left(\theta_x\right)}{\sin\left(\gamma\right)\sin\left(\theta_y\right)}\\
&\quad -\frac{\sqrt{2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma-1}}{\sin\left(\gamma\right)\sin\left(\alpha\right)}\\
&\quad \cdot\frac{\sqrt{2\cos\gamma\cos\theta_x\cos\theta_y+\cos^2\gamma+\cos^2\theta_x+\cos^2\theta_y-1}}{\sin\left(\gamma\right)\sin\left(\theta_y\right)}
\end{aligned}$$

The Cosine Law II for ΔZBX and ΔZAY yield

$$\begin{aligned}
\cos\left(\theta_z\right) &= \cosh\left(d\left(X,B\right)\right)\sin\left(\beta\right)\sin\left(\theta_x\right)-\cos\left(\beta\right)\cos\left(\theta_x\right)\\
\cos\left(\theta_z\right) &= \cosh\left(d\left(Y,A\right)\right)\sin\left(\alpha\right)\sin\left(\theta_y\right)-\cos\left(\alpha\right)\cos\left(\theta_y\right).
\end{aligned}$$

Substituting $\cosh\left(d\left(X,B\right)\right)$ and $\cosh\left(d\left(Y,A\right)\right)$ by their values given above yields

$$\begin{aligned}
\cos\left(\theta_z\right) &= \frac{\left(\left(\cos\left(\beta\right)\cos\left(\gamma\right)+\cos\left(\alpha\right)\right)\left(\cos\left(\gamma\right)\cos\left(\theta_x\right)+\cos\left(\theta_y\right)\right)\right)-\left(1-\cos^2\gamma\right)\cos\left(\beta\right)\cos\left(\theta_x\right)}{\left(\sin\left(\gamma\right)\right)^2}\\
&\quad -\frac{1}{\left(\sin\left(\gamma\right)\right)^2}\sqrt{2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma-1}\\
&\quad \cdot\sqrt{2\cos\gamma\cos\theta_x\cos\theta_y+\cos^2\gamma+\cos^2\theta_x+\cos^2\theta_y-1}
\end{aligned}$$

$$\begin{aligned}
\cos\left(\theta_z\right) &= \frac{\left(\cos\left(\gamma\right)\cos\left(\alpha\right)+\cos\left(\beta\right)\right)\left(\cos\left(\gamma\right)\cos\left(\theta_y\right)+\cos\left(\theta_x\right)\right)-\left(1-\cos^2\gamma\right)\cos\left(\alpha\right)\cos\left(\theta_y\right)}{\left(\sin\left(\gamma\right)\right)^2}\\
&\quad -\frac{1}{\left(\sin\left(\gamma\right)\right)^2}\sqrt{2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma-1}\\
&\quad \cdot\sqrt{2\cos\gamma\cos\theta_x\cos\theta_y+\cos^2\gamma+\cos^2\theta_x+\cos^2\theta_y-1}
\end{aligned}$$

Equating the two expressions for θ_z yields

$$\begin{aligned}
&\left(\cos\alpha\cos\theta_y-\cos\beta\cos\theta_x+\cos\alpha\cos\gamma\cos\theta_x+\cos\beta\cos\gamma\cos\theta_y+2\cos\beta\cos^2\gamma\cos\theta_x\right)\\
=\;&\left(\cos\beta\cos\theta_x-\cos\alpha\cos\theta_y+\cos\alpha\cos\gamma\cos\theta_x+\cos\beta\cos\gamma\cos\theta_y+2\cos\alpha\cos^2\gamma\cos\theta_y\right)
\end{aligned}$$

That is

$$\begin{aligned}(2\cos\beta)\left(1-\cos^2\gamma\right)\cos\theta_x &= (2\cos\alpha)\left(1-\cos^2\gamma\right)\cos\theta_y\\ 2\left(1-\cos^2\gamma\right)(\cos\beta\cos\theta_x-\cos\alpha\cos\theta_y) &= 0\end{aligned}$$

Therefore, either $\cos\gamma = 1$ or

$$\begin{aligned}0 &= \cos\beta\cos\theta_x-\cos\alpha\cos\theta_y\\ \frac{\cos\theta_x}{\cos\alpha} &= \frac{\cos\theta_y}{\cos\beta}\end{aligned}$$

By symmetry (in case $\alpha,\beta,\gamma \neq 0$),

$$\frac{\cos\theta_x}{\cos\alpha}=\frac{\cos\theta_y}{\cos\beta}=\frac{\cos\theta_z}{\cos\gamma}$$

This is the first relation, which might be called Cosine Rule for inscribed triangle.

Next,

$$\begin{aligned}d(Z,A) &= c-d(Z,B)\\ \cosh(d(Z,A)) &= \cosh(c-d(Z,B))\\ &= \cosh(c)\cosh(d(Z,B))-\sinh(c)\sinh(d(Z,B))\end{aligned}$$

The Cosine Law II for ΔZAY and ΔZBX yields

$$\begin{aligned}\cosh(d(Z,A)) &= \frac{\cos\alpha\cos\theta_z+\cos\theta_y}{\sin\alpha\sin\theta_z}\\ \sinh(d(Z,A)) &= \frac{\sqrt{2\cos\alpha\cos\theta_y\cos\theta_z+\cos^2\alpha+\cos^2\theta_y+\cos^2\theta_z-1}}{\sin(\alpha)\sin(\theta_z)}\\ \\ \cosh(d(Z,B)) &= \frac{\cos\beta\cos\theta_z+\cos\theta_x}{\sin\beta\sin\theta_z}\\ \sinh(d(Z,B)) &= \frac{\sqrt{2\cos\beta\cos\theta_x\cos\theta_z+\cos^2\beta+\cos^2\theta_x+\cos^2\theta_z-1}}{\sin(\beta)\sin(\theta_z)}\end{aligned}$$

Therefore, equating the two expressions for $\cosh(d(Z,A))$ yields

$$\begin{aligned}&(\cosh(c))\,\frac{\cos\beta\cos\theta_z+\cos\theta_x}{\sin\beta\sin\theta_z}-\frac{\cos\alpha\cos\theta_z+\cos\theta_y}{\sin\alpha\sin\theta_z}\\ =\ &(\sinh(c))\,\frac{\sqrt{2\cos\beta\cos\theta_x\cos\theta_z+\cos^2\beta+\cos^2\theta_x+\cos^2\theta_z-1}}{\sin(\beta)\sin(\theta_z)}\end{aligned}$$

Substituting the expression for $\cosh(c)$ and $\sinh(c)$ into the previous yields the following expression:

$$\begin{aligned}&(\cos\alpha\cos\beta+\cos\gamma)(\cos\beta\cos\theta_z+\cos\theta_x)-\left(1-\cos^2\beta\right)(\cos\alpha\cos\theta_z+\cos\theta_y)\\ =\ &\sqrt{2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma-1}\\ &\cdot\sqrt{2\cos\beta\cos\theta_x\cos\theta_z+\cos^2\beta+\cos^2\theta_x+\cos^2\theta_z-1}\end{aligned}$$

Then

$$\begin{aligned} & \big(\cos\gamma\cos\theta_x - \cos\alpha\cos\theta_z - \cos\theta_y + \cos\alpha\cos\beta\cos\theta_x + \cos\beta\cos\gamma\cos\theta_z + \cos^2\beta\cos\theta_y + 2\cos \\ = & \big(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\big)\big(2\cos\beta\cos\theta_x\cos\theta_z + \cos^2\beta + \cos^2\theta_x + \cos^2 \end{aligned}$$

From the Cosine Rule for the inscribed triangle, $\cos\theta_x$ and $\cos\theta_y$ can be expressed as

$$\begin{aligned} \cos\theta_x &= \frac{\cos\theta_z}{\cos\gamma}\cos\alpha \\ \cos\theta_y &= \frac{\cos\theta_z}{\cos\gamma}\cos\beta. \end{aligned}$$

Then substituting the expression for $\cos\theta_x$ and $\cos\theta_y$ yields

$$\begin{aligned} & \cos^2\beta\cos^2\theta_z\big(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma - 1\big) \\ = & \big(2\cos\alpha\cos\beta\cos\gamma\cos^2\theta_z - \cos^2\gamma + \cos^2\beta\cos^2\gamma + \cos^2\alpha\cos^2\theta_z + \cos^2\gamma\cos^2\theta_z\big) \end{aligned}$$

$$\begin{aligned} & \big(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma\big)\big(1 - \cos^2\beta\big)\cos^2\theta_z \\ = & \big(1 - \cos^2\beta\big)\cos^2\gamma \end{aligned}$$

Finally,

$$\begin{aligned} \cos^2\theta_z &= \frac{\cos^2\gamma}{(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma)} \\ \sin^2\theta_z &= \frac{2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta}{(2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma)} \end{aligned}$$

Clearly, θ_z, and by the same taken $\cos\theta_x$, $\cos\theta_y$, are uniquely defined once the incidence angles are set equal to the reflection angles. Observe that the expression for θ_z is consistent with the one obtained from the orthic triangle.

14.5.5 Fatness formula

Given that ΔXYZ is the inscribed triangle that satisfied the first order variation, then the internal angle of ΔXYZ at X, Y, Z are $\pi - 2\theta_x, \pi - 2\theta_y, \pi - 2\theta_z$, respectively, where

$$\begin{aligned} \cos^2\theta_x &= \frac{\cos^2\alpha}{2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma} \\ \cos^2\theta_y &= \frac{\cos^2\beta}{2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma} \\ \cos^2\theta_z &= \frac{\cos^2\gamma}{2\cos\alpha\cos\beta\cos\gamma + \cos^2\alpha + \cos^2\beta + \cos^2\gamma}. \end{aligned}$$

Then

$$\begin{aligned}\cos\left(\pi-2\theta_x\right) &= 1-2\cos^2\theta_x\\ &= \frac{2\cos\alpha\cos\beta\cos\gamma+\cos^2\beta+\cos^2\gamma-\cos^2\alpha}{\left(2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma\right)}\end{aligned}$$

$$\begin{aligned}\cos\left(\pi-2\theta_y\right) &= 1-2\cos^2\theta_y\\ &= \frac{2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\gamma-\cos^2\beta}{\left(2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma\right)}\end{aligned}$$

$$\begin{aligned}\cos\left(\pi-2\theta_z\right) &= 1-2\cos^2\theta_z\\ &= \frac{2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta-\cos^2\gamma}{\left(2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma\right)}.\end{aligned}$$

The Cosine Rule II for ΔXYZ yield the following result:

$$\begin{aligned}\cosh d\left(X,Y\right) &= \frac{\cos\left(\pi-2\theta_x\right)\cos\left(\pi-2\theta_y\right)+\cos\left(\pi-2\theta_z\right)}{\sin\left(\pi-2\theta_x\right)\sin\left(\pi-2\theta_y\right)}\\ &= \frac{\left(1-2\cos^2\theta_x\right)\left(1-2\cos^2\theta_y\right)+1-2\cos^2\theta_z}{\left(2\sin\theta_x\cos\theta_x\right)\left(2\sin\theta_y\cos\theta_y\right)}\\ &= \frac{\left(2\cos^2\theta_x\cos^2\theta_y-\cos^2\theta_x-\cos^2\theta_y-\cos^2\theta_z+1\right)}{2\left(\sin\theta_x\cos\theta_x\right)\left(\sin\theta_y\cos\theta_y\right)}\\ \sinh d\left(X,Y\right) &= \sqrt{\cosh^2 d\left(X,Y\right)-1}\\ &= \frac{\sqrt{\left(\cos^2\theta_x+\cos^2\theta_y+\cos^2\theta_z-1\right)^2-\left(2\cos\theta_x\cos\theta_y\cos\theta_z\right)^2}}{2\left(\sin\theta_x\cos\theta_x\right)\left(\sin\theta_y\cos\theta_y\right)}\end{aligned}$$

$$\begin{aligned}\cosh d\left(Y,Z\right) &= \frac{\left(2\cos^2\theta_y\cos^2\theta_z-\cos^2\theta_x-\cos^2\theta_y-\cos^2\theta_z+1\right)}{2\left(\sin\theta_y\cos\theta_y\right)\left(\sin\theta_z\cos\theta_z\right)}\\ \sinh d\left(Y,Z\right) &= \frac{\sqrt{\left(\cos^2\theta_x+\cos^2\theta_y+\cos^2\theta_z-1\right)^2-\left(2\cos\theta_x\cos\theta_y\cos\theta_z\right)^2}}{2\left(\sin\theta_y\cos\theta_y\right)\left(\sin\theta_z\cos\theta_z\right)}\end{aligned}$$

$$\begin{aligned}\cosh d\left(Z,X\right) &= \frac{\left(2\cos^2\theta_z\cos^2\theta_x-\cos^2\theta_x-\cos^2\theta_y-\cos^2\theta_z+1\right)}{2\left(\sin\theta_z\cos\theta_z\right)\left(\sin\theta_x\cos\theta_x\right)}\\ \sinh d\left(Z,X\right) &= \frac{\sqrt{\left(\cos^2\theta_x+\cos^2\theta_y+\cos^2\theta_z-1\right)^2-\left(2\cos\theta_x\cos\theta_y\cos\theta_z\right)^2}}{2\left(\sin\theta_z\cos\theta_z\right)\left(\sin\theta_x\cos\theta_x\right)}\end{aligned}$$

Therefore,

$$\begin{aligned}\sinh\left(d\left(X,Y\right)+d\left(Y,Z\right)+d\left(Z,X\right)\right) &= \cosh d\left(X,Y\right)\cosh d\left(Y,Z\right)\sinh d\left(Z,X\right) \\ &\quad +\cosh d\left(X,Y\right)\cosh d\left(Z,X\right)\sinh d\left(Y,Z\right) \\ &\quad +\cosh d\left(Y,Z\right)\cosh d\left(Z,X\right)\sinh d\left(X,Y\right) \\ &\quad +\sinh d\left(X,Y\right)\sinh d\left(Y,Z\right)\sinh d\left(Z,X\right) \\ &= \frac{\left(1-\cos^2\theta_x-\cos^2\theta_y-\cos^2\theta_z\right)}{2\left(\cos^2\theta_x\right)\left(\cos^2\theta_y\right)\left(\cos^2\theta_z\right)} \\ &\quad \cdot\sqrt{\left(\cos^2\theta_x+\cos^2\theta_y+\cos^2\theta_z-1\right)^2-\left(2\cos\theta_x\cos\theta_y\cos\theta_z\right.}\end{aligned}$$

Finally, substituting the expression for the $\cos\theta_x, \cos\theta_y, \cos\theta_z$ yields the expression for the fatness as follows:

$$\begin{aligned}\sinh\left(d\left(X,Y\right)+d\left(Y,Z\right)+d\left(Z,X\right)\right) &= 2\sqrt{\left(2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma\right)} \\ &\quad \cdot\sqrt{\left(2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma-1\right)}\end{aligned}$$

$$\delta_F\left(\Delta ABC\right)=\frac{1}{\sqrt{-\kappa}}\sinh^{-1}\left(\begin{array}{c}2\sqrt{\left(2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma\right)} \\ \cdot\sqrt{\left(2\cos\alpha\cos\beta\cos\gamma+\cos^2\alpha+\cos^2\beta+\cos^2\gamma-1\right)}\end{array}\right)$$

14.6 bibliographical and historical notes

Fagnano's original solution involved differential calculus and, two centuries thereafter, several *purely synthetic geometry* solutions were developed. In particular, Fejér's solution [28, Sec. 1.8] and Schwarz's solution utilize the concept of reflection across edges of the triangle, a concept that extends to hyperbolic geometry as the *Schwarz reflection.*

Via a Möbius transformation, which does not affect the hyperbolic geometry, let us place the vertex A at the center of the unit disk model, in which case the geodesics $[AB]$ and $[AC]$ become radial lines. If we assume that the angles of the triangle $\triangle ABC$ satisfy [69, p. 312]

$$\alpha=\frac{\pi}{m},\beta=\frac{\pi}{n},\gamma=\frac{\pi}{p};\quad \frac{1}{m}+\frac{1}{n}+\frac{1}{p}<1$$

then by repeatedly applying Schwarz inversions relative to the sides of the triangles, one obtains a *Dirichlet tessellation* of the unit disk [69, Fig. 41], [60, Fig. 5]. An even number of such transformations preserve the orientation and hence form a *Fuchsian group* Γ, a discrete subgroup of $PSL(\mathbb{R}^2)$. (The Schwarzian conformal mapping $S : ABC \to$ Upper Half Plane is automorphic relative to Γ [69, Sec. VI.5].) Let $I_{[BC]} : ABC \to A'BC$ be the inversion [24, Sec. 139] relative to the circle $[BC]$. Under those circumstances, the "light ray" XY, instead of reflecting on BC, can be considered as crossing the edge BC of the tessellation, becoming YZ', and the reflected ray in ABC is recovered as $I^{-1}_{[BC]}(YZ')$ [79]. Therefore, to compute XYZ, one follows the geodesic ray XY

as it goes through the tessellation until it hits an edge which, mapped back to BC, yields the reflection law at a fixed point.

Amazingly, it is not known whether an arbitrary polygonal billiard table has a periodic orbit. However, in the case of a triangular table, the above construction, sometimes referred to as *Fagnano periodic orbit*, provides an affirmative answer.

Chapter 15

random versus hyperbolic graphs

In this chapter, the connections between random graphs and hyperbolic metric spaces are discussed. As a warm up exercise, we first utilize the fatness as measure of hyperbolicity. Unfortunately, while it is very intuitive, this measure is hard to compute, so that it does not lead us that far in our numerical exploration. For that reason, we then switch to the Gromov product δ as a more computationally convenient measure, allowing us to do the metric hyperbolic analysis on much larger graphs. Since simulation can only produce finite graphs, the hyperbolic properties of such graphs would really manifest themselves when the δ is much smaller than the diameter and for that reason we have adopted $\frac{\delta}{\text{diam}}$ as the relevant measure.

The purpose of this simulation study is to compare the hyperbolic property of several random graph generators defined in the previous section. Each random graph generator generates a finite graph where each edge has unit distance without direction. The δ_G of the Gromov product, which is computed from the 4-point condition, can serve as the measure of hyperbolicity. However, the numerical simulation can only create finite graphs all of which have finite diameter and hence have finite δ_G. Therefore, every graph can be considered as a hyperbolic metric space. Observe that the hyperbolic property would manifest itself only if the δ_G is significantly smaller than the diameter of graph. It follows that the mathematical expectation of the normalized delta $E\left(\frac{\delta_G}{diam}\right)$, where *diam* is the diameter of the graph, is the key hyperbolic measure for random graph generator. The simulation methodology is set up as follows:

15.1 setting up the various graph generators

The random graphs as modeled by Erdős and Rényi considered here are the $G(n,m)$ models. Observe that the connectivity of $G(n,m)$ depends on the number of edges; hence to avoid disconnected graphs in simulation, the $G(n,m)$

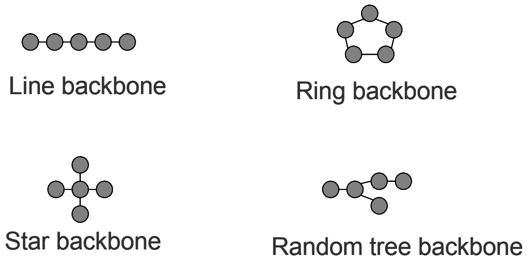


Figure 15.1: The backbone graphs in the simulation.

model is generated based on the following procedure: First, the backbone graph which is the connected graph of order n and of size m_0 is generated; then the rest of the $m - m_0$ edges are randomly selected from the $\binom{n}{2} - m_0$ available edges. In fact, the backbone graph can be a deterministic or a random graph of order n. Given that the vertex set of the backbone graph is $\{1, 2, \ldots, n\}$, then the 4 backbone graphs in the simulation are defined as follows (see Figure 15.1):

1. Line backbone: $G_{Line}(n)$ is a graph in which $E_{G_{Line}(n)} = \{(i, i+1) : i = 1, \ldots, n-1\}$;

2. Ring backbone: $G_{Ring}(n)$ is a graph in which $E_{G_{Ring}(n)} = \{(i, i+1) \bmod n : i = 1, \ldots, n\}$;

3. Star backbone: $G_{Star}(n)$ is a graph in which $E_{G_{Star}(n)} = \{(1, i) : i = 2, \ldots, n\}$;

4. Random tree backbone: $G_{Rand}(n)$ is a graph G_n obtained from the evolution $\{G_t\}_{t=1}^{\infty}$ where G_1 is a single vertex 1 and G_t is recursively obtained from G_{t-1} by adding a new vertex t to G_{t-1} and a new edge from a vertex randomly selected from $\{1, \ldots, t-1\}$ to vertex t.

Clearly, $G_{Line}(n)$, $G_{Star}(n)$, and $G_{Rand}(n)$ are trees of size $n-1$ and $G_{Ring}(n)$ is of size n. Hence the $G(n, m)$ models with line, star or random tree backbone have $m-n+1$ edges randomly selected from $\binom{n}{2}-n+1$ possible edges and $G(n, m)$ models with ring backbone have $m-n$ edges randomly selected from $\binom{n}{2} - n$ possible edges.

The small worlds graphs as modeled by Watts-Strogatz considered in the simulation are the β-models where the β parameter varies as $0, 0.1, \ldots, 1$. Observe that as β varies from 0 to 1, the random graph varies from regular graph to approximately purely random graph. The resulting graphs are of order n and of size $\left\lfloor \bar{k}\frac{n}{2} \right\rfloor$. In the simulation, the effects of the parameter β on the hyperbolicity as well as the average degree $\bar{k}$ are observed.

The scale free graphs in the simulation are generated by the Barabási-Albert approach in which the starting graphs are line, ring, star, and random tree backbones with n_0 vertices. The graphs are continuously evolving from the previous graphs by the growth and preferential attachment until the resulting graphs are of order n. To study the effect of preferential attachment on hyperbolicity, graphs generated from growth and uniform attachment are also considered in the test bed. In contrast to the preferential attachment, in the uniform attachment, the probability $\Pi(i)$ that a vertex i is connected to a new vertex is equal among all vertices $1, \ldots, i-1$. The degree distribution $P(k)$ at time t for a random graph with growth and uniform attachment can be computed by the following formula:

$$P(k) = \frac{n_0 + t - 1}{(n_0 + t)\, l} \exp\left(1 - \frac{k}{l}\right).$$

This formula has been derived from the continuum approach by Barabási-Albert [10]. As $t \to \infty$,

$$P(k) = \frac{1}{l} \exp\left(1 - \frac{k}{l}\right),$$

hence the uniform attachment provides an exponentially decaying degree distribution which depends on the parameter l of the graphs.

In the scale free and growth with uniform attachment random graph generators, the resulting graphs are of order n and of size $n_0 - 1 + (n - n_0)\, m$ for line, star, and random tree backbones and $n_0 + (n - n_0)\, m$ for ring backbone.

In each random graph generator, each model generates a different topological graph structure and has different parameters. To understand the effect of random graph generator on hyperbolicity, the parameters for each generator are determined so that the resulting graph for each generator is of the same order and approximately the same size. In the simulation, the average degree $\bar{k}$ in small world generator is approximately twice the number of edges m for each additional vertex. Then the random graphs from different generators all have approximately the same size. Hence in this simulation the parameter $\bar{k}$ is set as $\bar{k} = 2, \ldots, 2n_0$ in the small world generator, and the parameter m is set as $m = 1, \ldots, n_0$ in the scale free and growth with uniform attachment random graph generators. The total number of edges in the Erdős and Rényi random graph generators is set to be the same as the size of the random graphs generated by the other generators so that the comparison among all random graph generators can be made.

In this simulation, the total number of vertices n is set to 50 and 100 and the parameter n_0 are equal to $\frac{n}{5}$. In addition, the number of simulations is equal to 100. The simulations reveal the following conclusions:

15.2 fatness analysis

We begin with the most intuitive δ_F measure of hyperbolicity. While the fatness is easy to interpret, it has the drawback that it is cumbersome to compute and

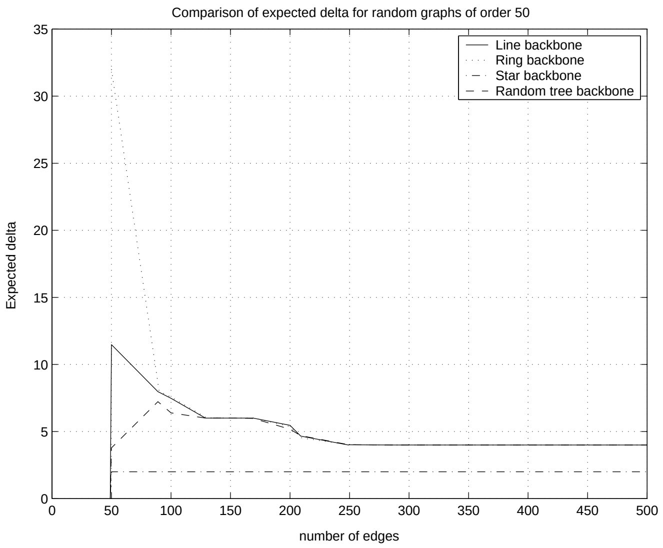


Figure 15.2: $\max \delta_F$ versus the number of links for random graphs.

as such imposes a practical limit of 50 on the order of the graphs that are manageable with the algorithm developed in Section 13.6.1.

15.2.1 Erdős and Rényi Random graphs

15.2.2 Watts-Strogatz Small world graphs

15.2.3 Barabási-Albert scale free graphs

15.2.4 Growth with uniform attachment graphs

15.2.5 Comparison among all graph generators

15.3 Gromov product analysis

The δ_G measure of hyperbolicity is much easier to compute than the δ_F, and as such it displaces the upper limit on the order on the graphs that can be managed to about 100. The problem is that δ_G is by far less trivial to interpret than δ_F.

15.3.1 Erdős and Rényi Random graphs

Although the random graph generators in this simulation construct random graphs on top of the backbone graphs whereas the random graphs generated by Erdős and Rényi are purely random graphs, there are no significant deviation

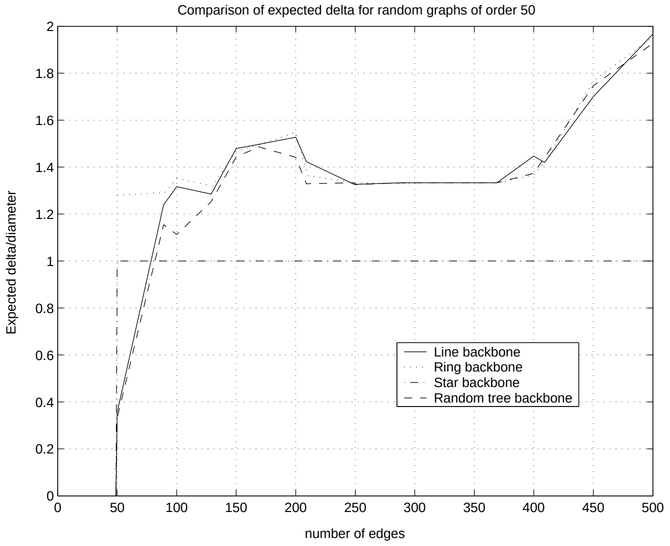


Figure 15.3: $\max \frac{\delta_F}{\text{diam}}$ versus the number of links for random graphs.

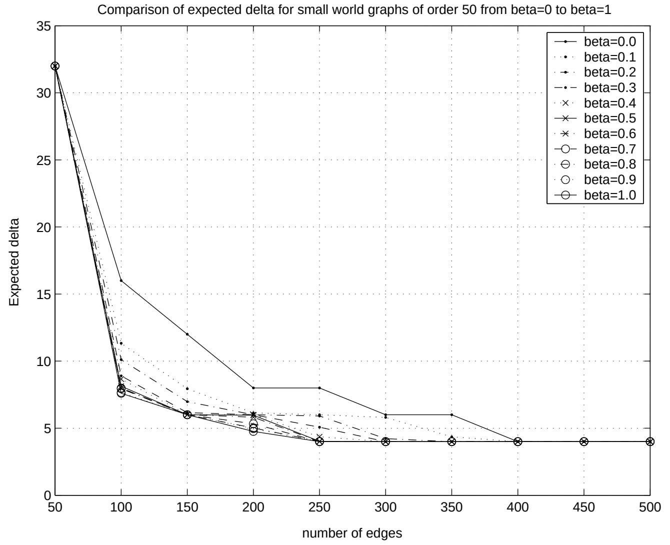


Figure 15.4: $\max \delta_F$ versus the number of links for small world graphs.

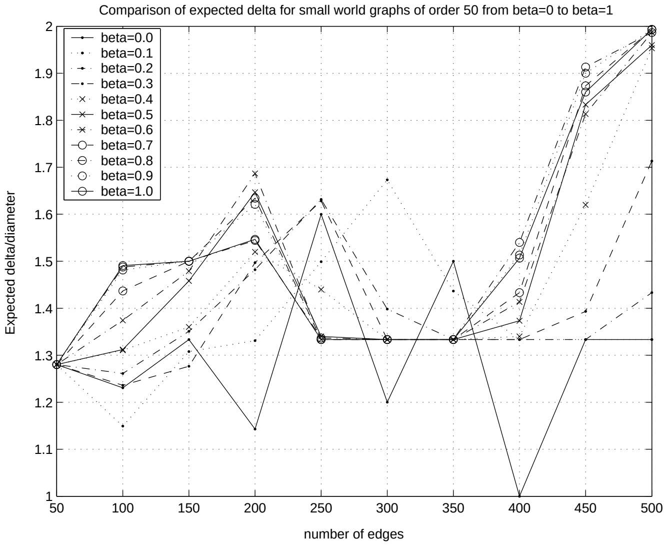


Figure 15.5: $\max \frac{\delta_F}{\text{diam}}$ versus the number of links for small world graphs.

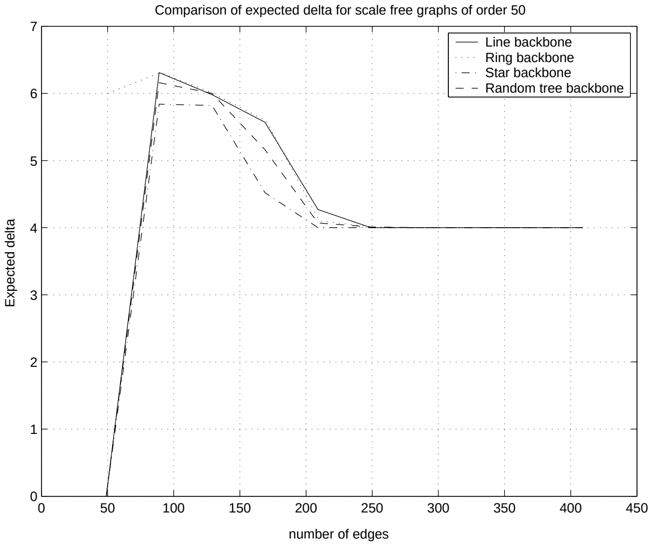


Figure 15.6: $\max \delta_F$ versus the number of links for scale free graphs.

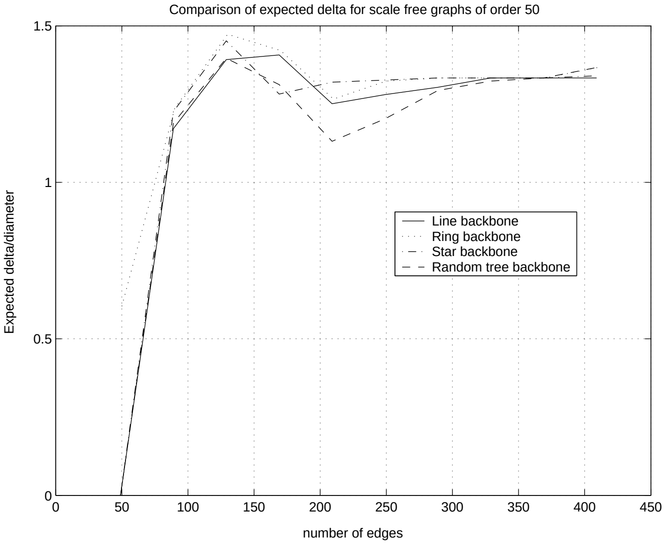


Figure 15.7: $\max \frac{\delta_F}{\text{diam}}$ versus the number of links for scale free graphs.

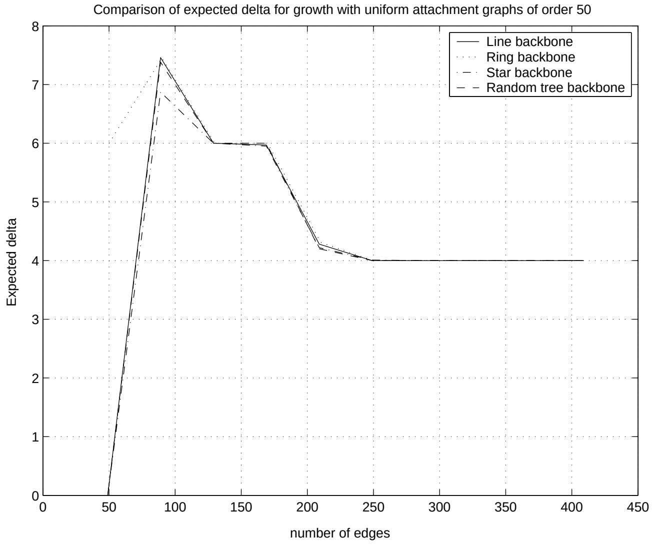


Figure 15.8: $\max \delta_F$ versus the number of links for growth with uniform attachment graphs.

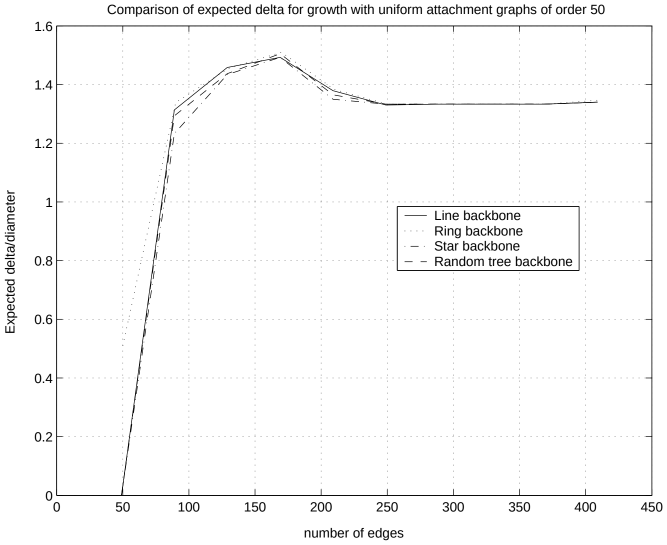


Figure 15.9: $\max \frac{\delta_F}{\text{diam}}$ versus the number of links for growth with uniform attachment graphs.

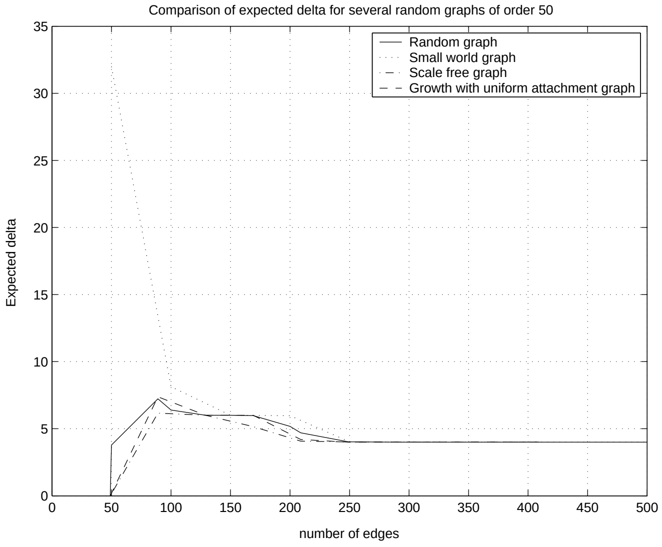


Figure 15.10: Comparison of $\max \delta_F$ versus the number of links for various graphs.

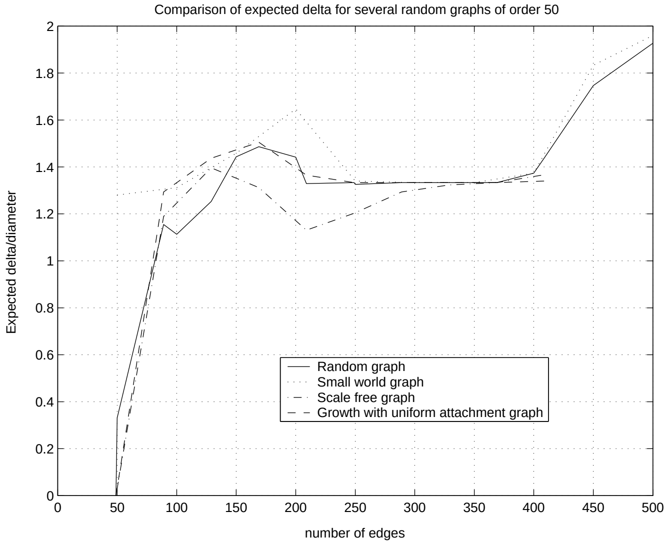


Figure 15.11: Comparison of max $\frac{\delta_F}{\text{diam}}$ versus the number of links for various graphs.

in the degree distribution and density functions between these two methods, except in the star backbone as shown in Figure 15.12 through Figure 15.13 where the parameters for these generators are $n = 3000$, and $m = 147,550$ for random graphs with Ring backbone, $m = 147,549$ for random graphs without backbone and random graphs with Line, Star, and Random tree backbones. The deviation of the star backbone from the other cases follows from the fact that the star backbone has a center vertex that connects to all other vertices. This provides the existence of a vertex with high degree which contradicts the Poisson degree distribution. In the other cases, the degree distribution for the random graphs generated by this method are roughly the same as in the Erdős and Rényi model which theoretically has an exponential decay.

The $E\left(\delta_G\right)$ and $E\left(\frac{\delta_G}{diam}\right)$ for random graphs with several backbones are shown in Figures 15.16, 15.17 for $n = 50$ and in Figures 15.18, 15.19 for $n = 100$.

In the random graphs generated with different backbones, as m increases, the expected delta and the expected normalized delta are approximately the same among these backbones, except for the star backbone. As the number of edges increases, the expected delta decreases to 1. However the diameter of the random graphs decreases. This yields an increase of the expected normalized delta. After increasing the number of edges from the Line, Ring, and Random tree backbone graphs, the minimum expected normalized delta occurs at a number of edges roughly equal to 7 times the number of vertices. This suggests a network parameter that yields a good hyperbolic graph.

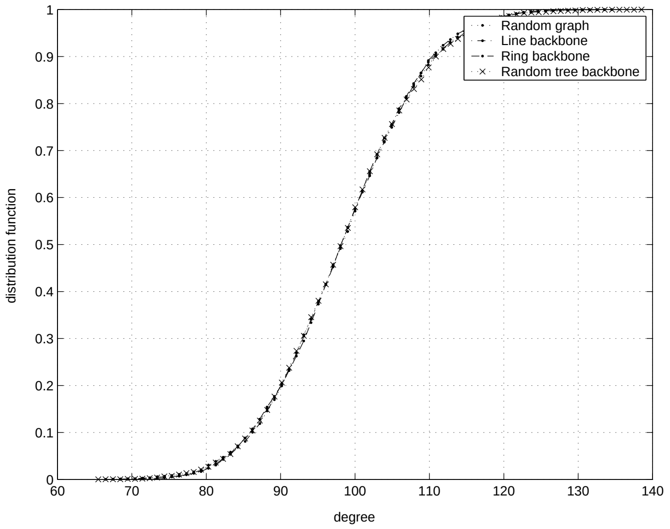


Figure 15.12: The degree probability distribution function that results from the numerical simulation of random graphs with different backbones.

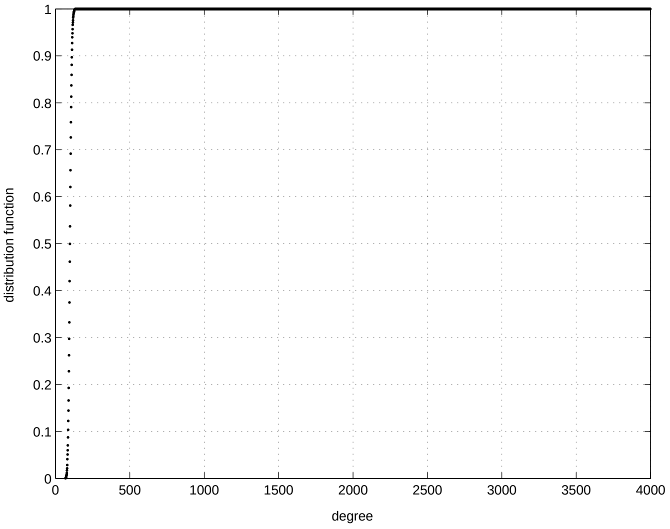


Figure 15.13: The degree probability distribution function that results from the numerical simulation of random graphs with Star backbone.

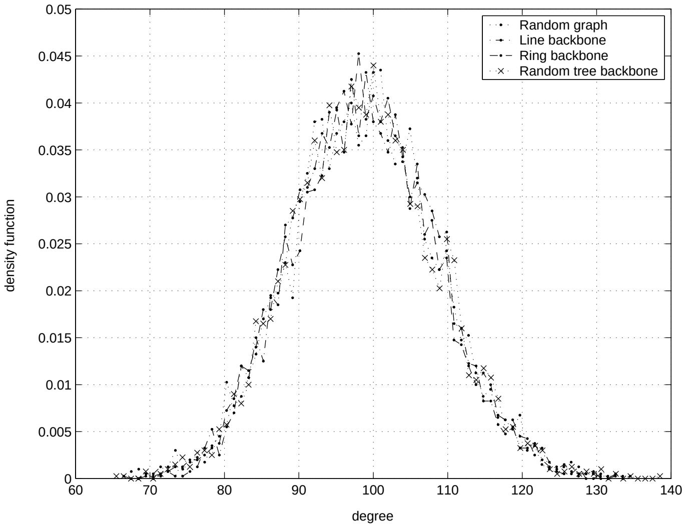


Figure 15.14: The degree probability density function that results from the numerical simulation of random graphs with different backbones.

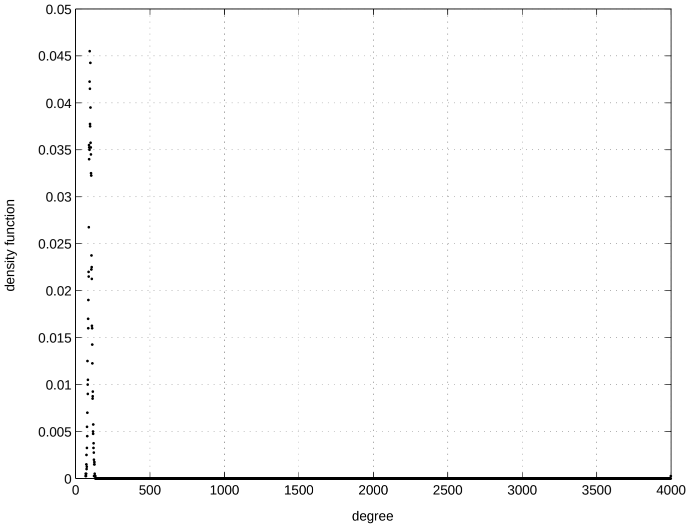


Figure 15.15: The degree probability density function that results from the numerical simulation of random graphs with Star backbone.

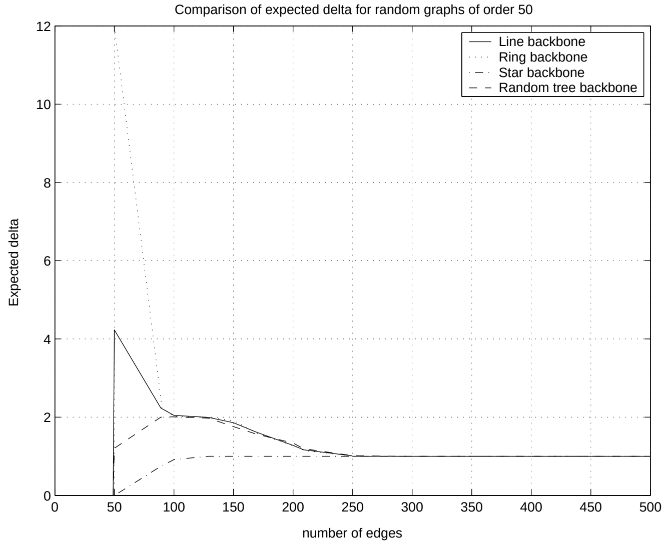


Figure 15.16: Comparison of $E\left(\delta_G\right)$ for random graphs of order 50.

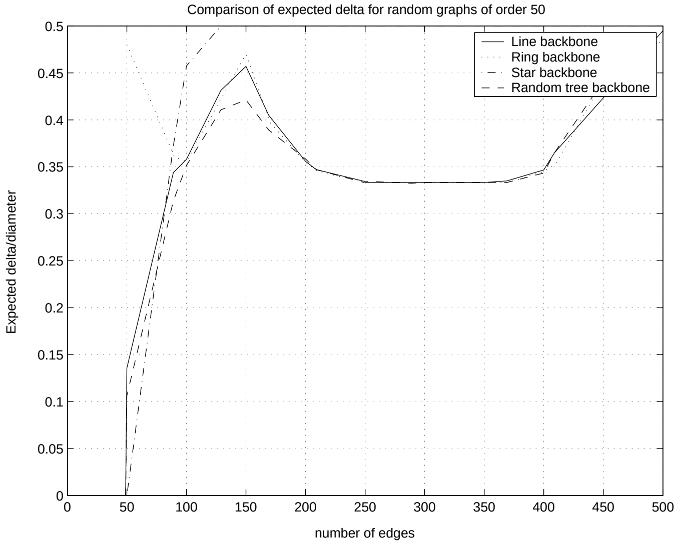


Figure 15.17: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for random graphs of order 50.

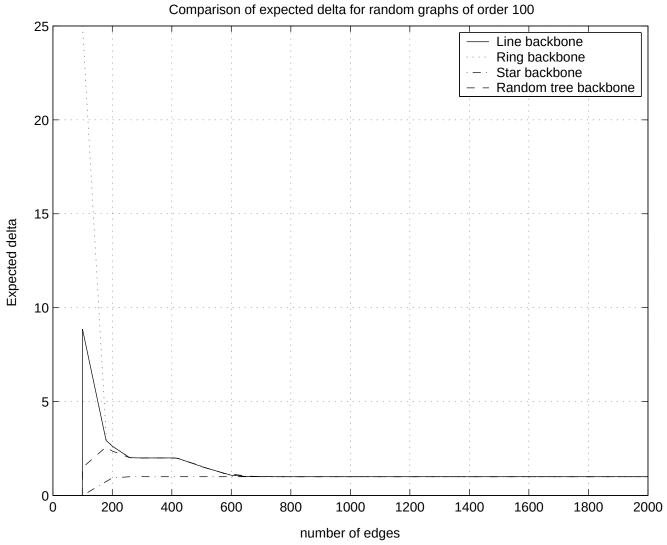


Figure 15.18: Comparison of $E\left(\delta_G\right)$ for random graphs of order 100.

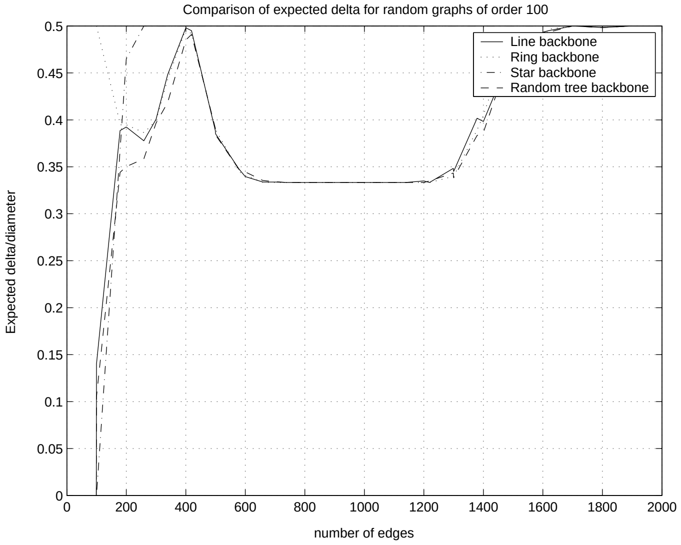


Figure 15.19: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for random graphs of order 100.

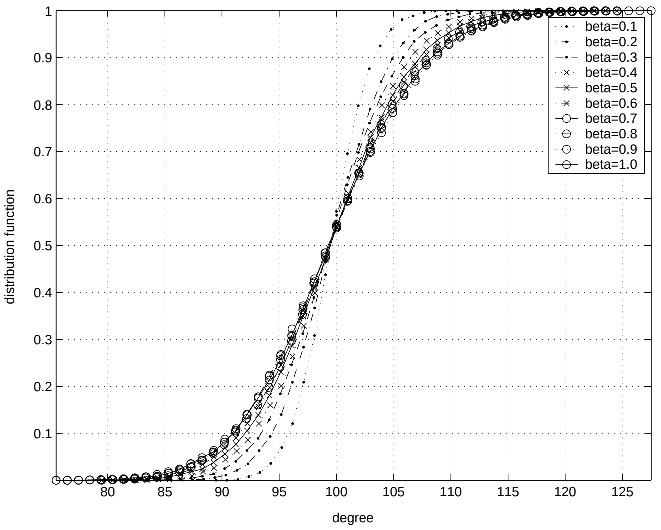


Figure 15.20: The degree probability distribution function that results from the numerical simulation of small world graphs with different parameter β.

15.3.2 Watts-Strogatz Small world graphs

The degree distribution and density functions for the small world graphs with the β parameter varying from 0 to 1 are shown in Figures 15.20 and 15.21, respectively where the parameters for these generators are $n = 3000$ and $m = 150,000$. As the parameter β varies from 0 to 1, the larger the tail of the density functions.

The $E\left(\delta_G\right)$ and $E\left(\frac{\delta_G}{diam}\right)$ for the small world graphs with the β parameter varying from 0 to 1 are shown in Figures 15.22, 15.23 for $n = 50$ and in Figures 15.24, 15.25 for $n = 100$.

The simulation shows that as β is increasing from 0 to 1, the expected delta is decreasing. However, the diameter of a graph is not monotonically varying with the parameter β. Given that $\beta > 0$ and fixed, then after increasing the number of edges, the expected normalized delta reaches a minimum of about 0.33. The number of edges at the minimum expected normalized delta depends upon the parameter β. The larger the parameter β, the smaller the number of edges at the minimum expected normalized delta. After continuously increasing the number of edges, the delta of a graph will reach its minimum. Finally, after continuously increasing the size of a graph, the diameter of a graph is decreasing. This result yields an increasing expected normalized delta. This is a behavior similar to that of the previous random graph generator.

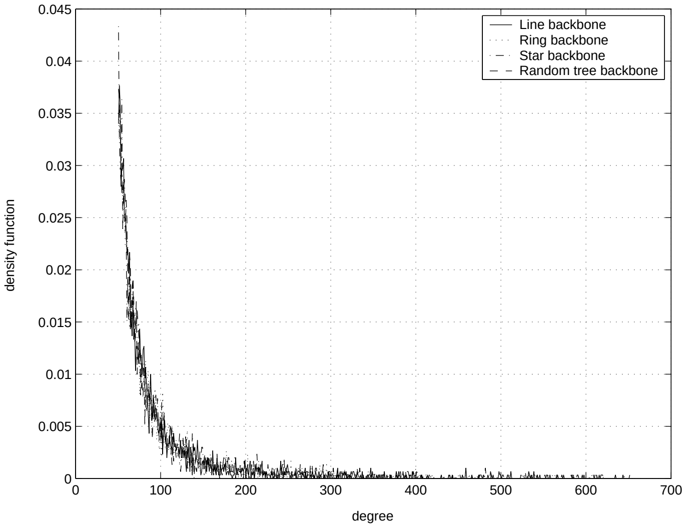


Figure 15.21: The degree probability density function that results from the numerical simulation of small world graphs with different parameter β.

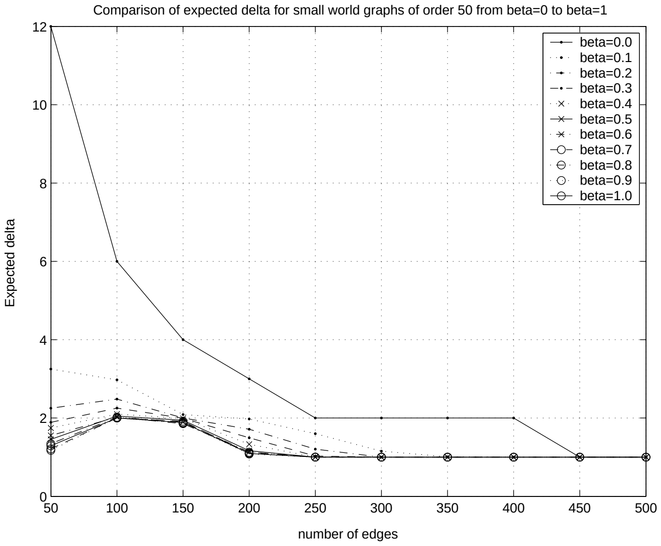


Figure 15.22: Comparison of $E(\delta_G)$ for small world graphs of order 50.

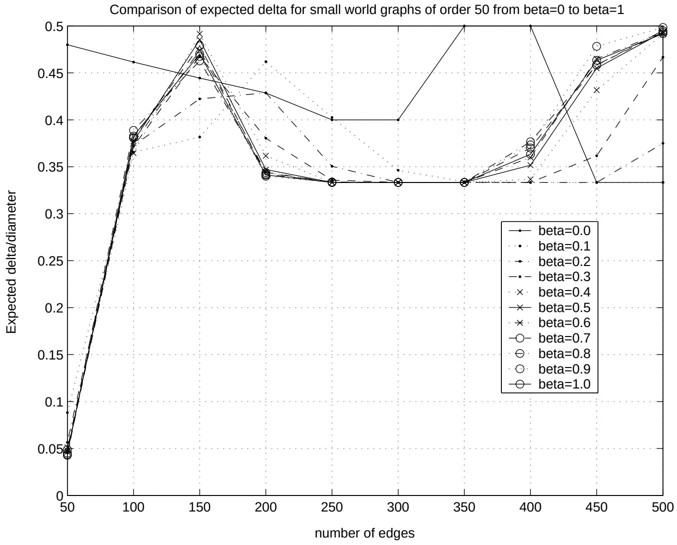


Figure 15.23: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for random graphs of order 50.

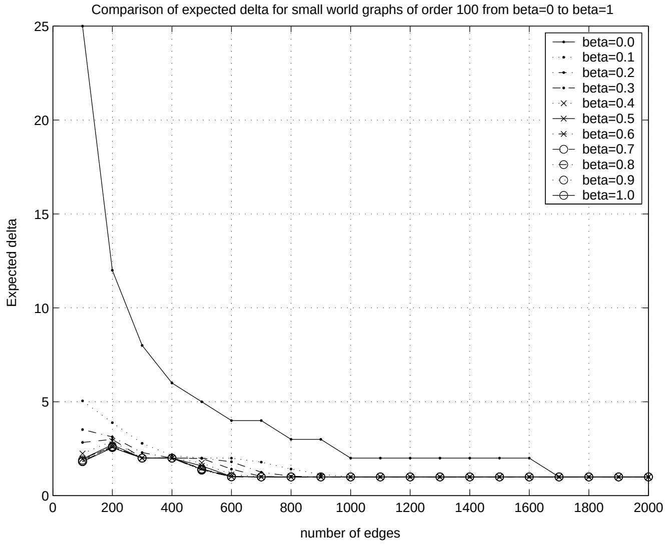


Figure 15.24: Comparison of $E\left(\delta_G\right)$ for small world graphs of order 100.

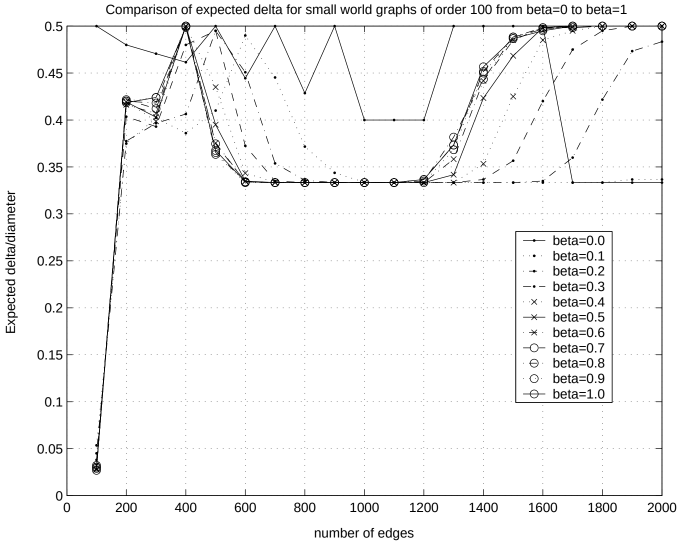


Figure 15.25: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for random graphs of order 100.

15.3.3 Barabási-Albert scale free graphs

The degree distribution and density functions for the Barabási-Albert scale free graphs with different backbones are shown in Figures 15.26 and 15.27, respectively, where the parameters for these generators are $n = 3000$, and $m = 147,550$ for random graphs with Ring backbone, $m = 147,549$ for random graphs with Line, Star, and Random tree backbones.

The $E(\delta_G)$ and $E\left(\frac{\delta_G}{diam}\right)$ for the scale free graphs with different backbones are shown in Figures 15.28, 15.29 for $n = 50$ and in Figures 15.30, 15.31 for $n = 100$.

The simulation shows that although the star backbone has the smallest expected delta compared with the other backbones, the random tree backbone has the smallest expected normalized delta in the middle range of the sizes of the graphs. The other backbones yield slightly different expected normalized delta's.

15.3.4 Growth with uniform attachment graphs

The degree distribution and density functions for the growth with uniform attachment graphs with different backbones are shown in Figures 15.32 and 15.33, respectively, where the parameters for these generators are $n = 3000$, and $m = 147,550$ for random graphs with Ring backbone, $m = 147,549$ for random graphs with Line, Star, and Random tree backbones.

The $E(\delta_G)$ and $E\left(\frac{\delta_G}{diam}\right)$ for growth with uniform attachment graphs with

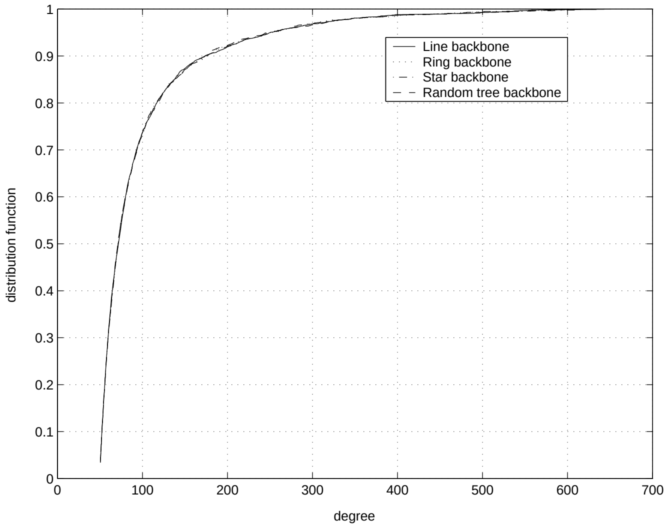


Figure 15.26: The degree probability distribution function that results from the numerical simulation of scale free graphs with different backbones.

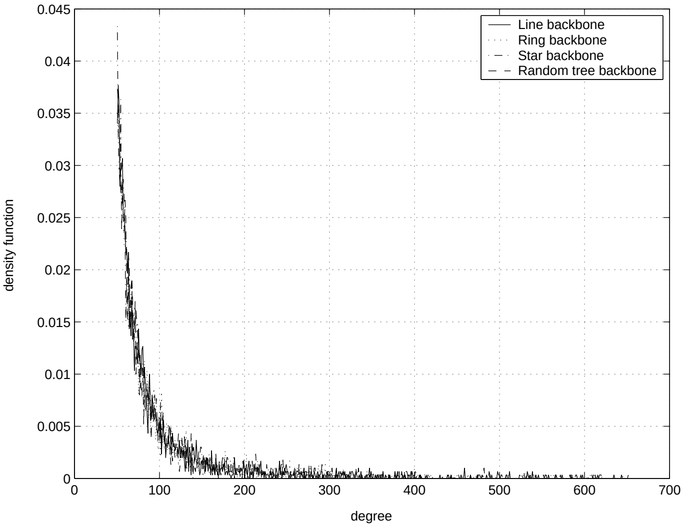


Figure 15.27: The degree probability density function that results from the numerical simulation of scale free graphs with different backbones.

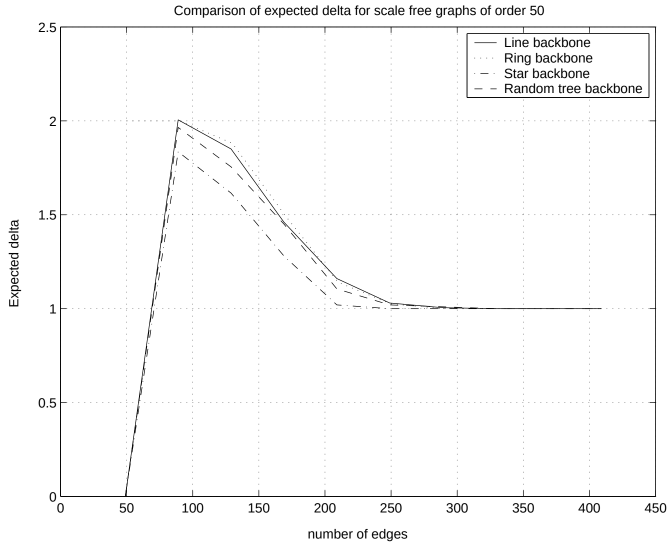


Figure 15.28: Comparison of $E\left(\delta_G\right)$ for scale free graphs of order 50.

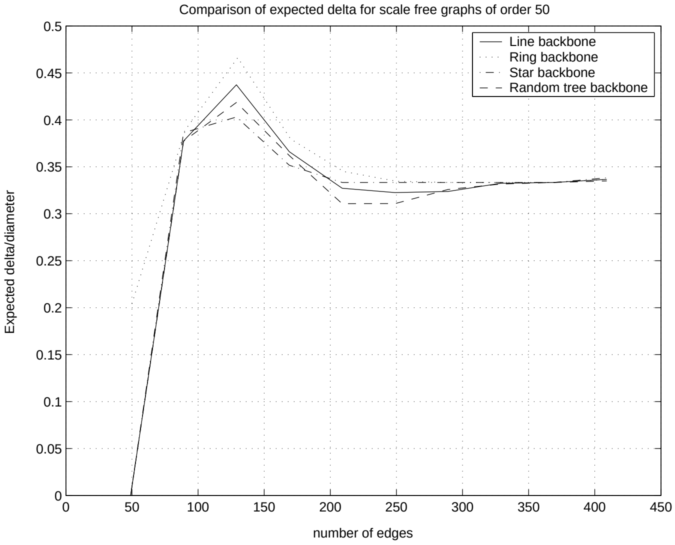


Figure 15.29: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for scale free graphs of order 50.

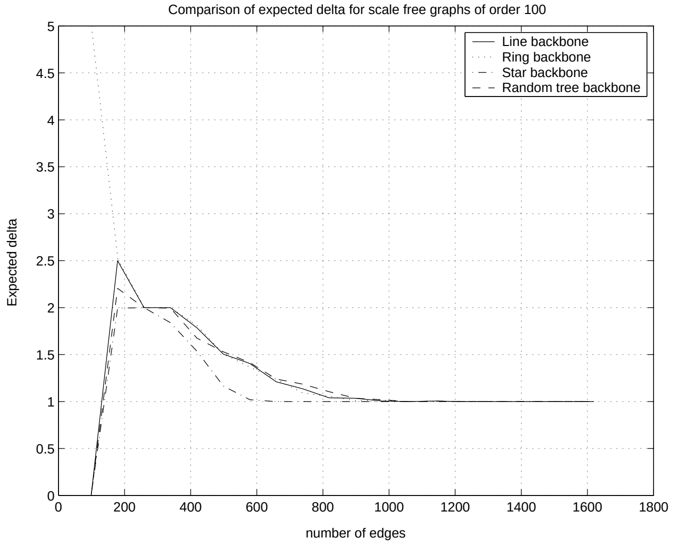


Figure 15.30: Comparison of $E\left(\delta_G\right)$ for scale free graphs of order 100.

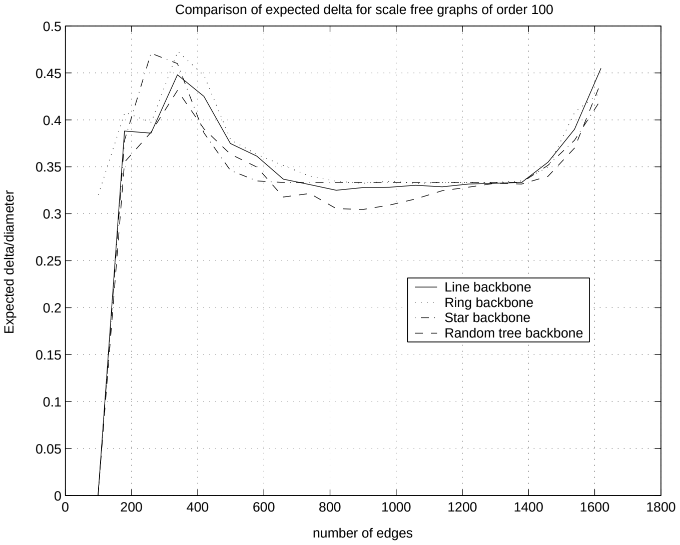


Figure 15.31: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for scale free graphs of order 100.

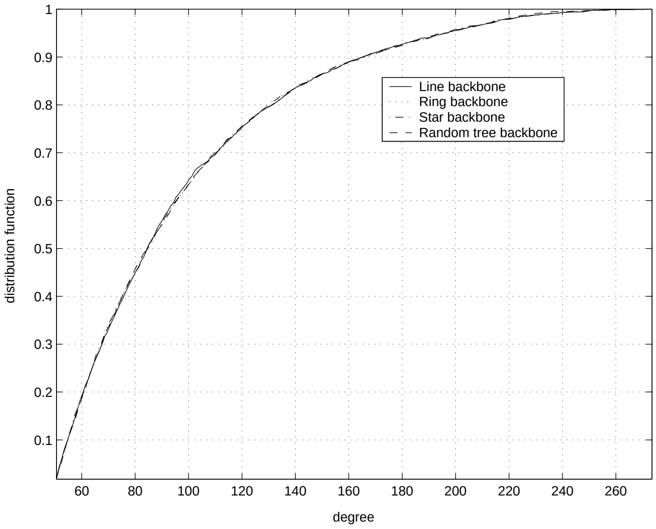


Figure 15.32: The degree probability distribution function that results from the numerical simulation of growth with uniform attachment graphs with different backbones.

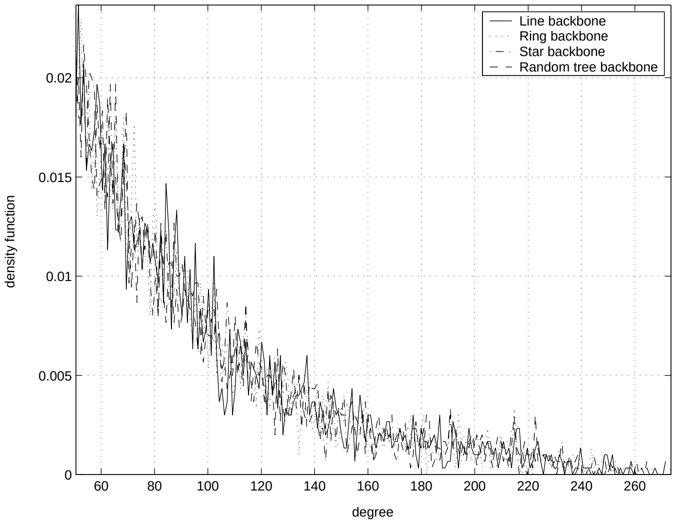


Figure 15.33: The degree probability density function that results from the numerical simulation of growth with uniform attachment graphs with different backbones.

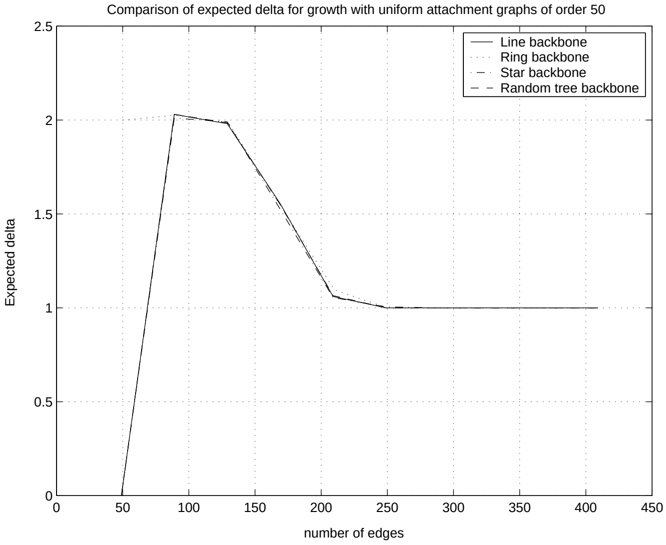


Figure 15.34: Comparison of $E\left(\delta_G\right)$ for growth with uniform attachment graphs of order 50.

different backbones are shown in Figures 15.34, 15.35 for $n = 50$ and in Figures 15.36, 15.37 for $n = 100$.

In contrast to the scale free graph, the random graph generated by growth with uniform attachment does not depend upon its backbone topology. The four backbones seem to provide nearly indistinguishable results for the expected delta and expected normalized delta as the size of the graph is increasing.

15.3.5 Comparison among all graph generators

The comparisons of the $E\left(\delta_G\right)$ and $E\left(\frac{\delta_G}{diam}\right)$ among random graph generators are shown in Figures 15.38, 15.39 for $n = 50$ and in Figures 15.40, 15.41 for $n = 100$. Here, the backbone graphs in random graphs, scale free graphs, and growth with uniform attachment generators are the random trees. In addition, the β parameter for small world graph is 0.5.

The simulation suggests the following conclusions:

1. As the size of the graph is increasing, the expected delta is decreasing. In contrast, the expected normalized delta is first decreasing as the expected delta is decreasing, then approximately constant after its reaches its minimum; finally increasing as the diameter is decreasing. This shows that the hyperbolic property occurs in the middle range of the sizes of the graphs. In the beginning, a graph is a tree (except for small world graph) where the δ_G vanishes. As the size of graph is increasing, the expected delta abruptly increases to a certain value and then continuously decreases. In

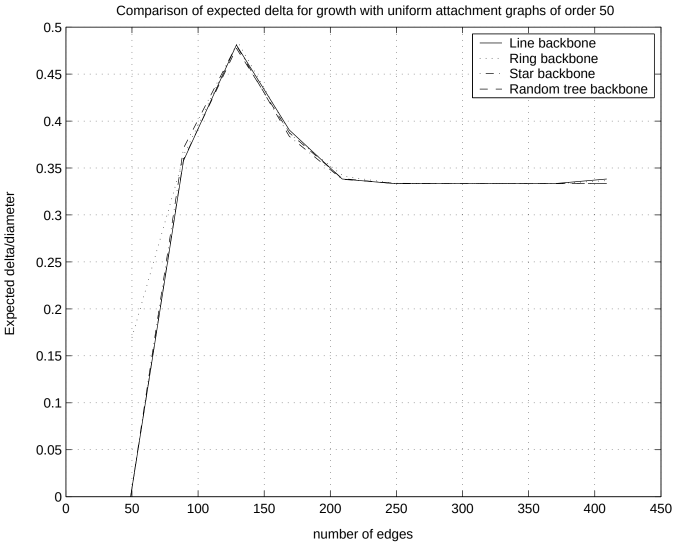


Figure 15.35: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for growth with uniform attachment graphs of order 50.

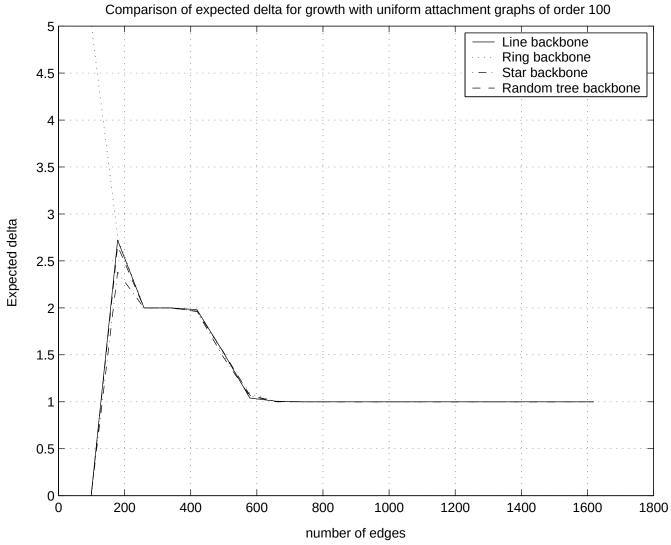


Figure 15.36: Comparison of $E\left(\delta_G\right)$ for growth with uniform attachment graphs of order 100.

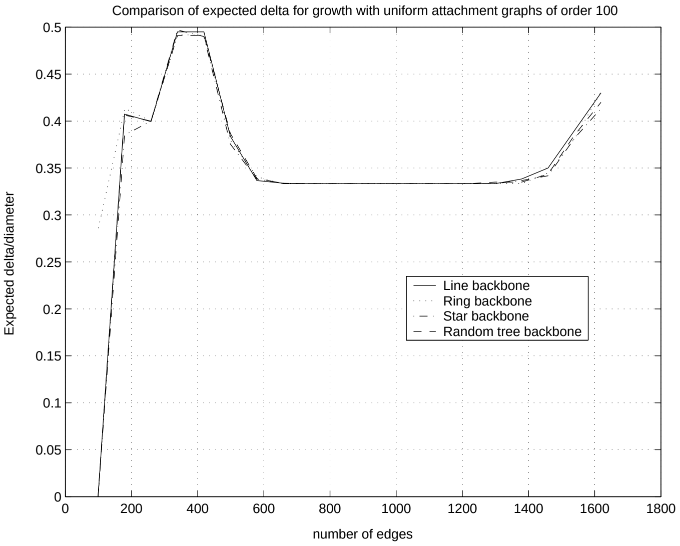


Figure 15.37: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for growth with uniform attachment graphs of order 100.

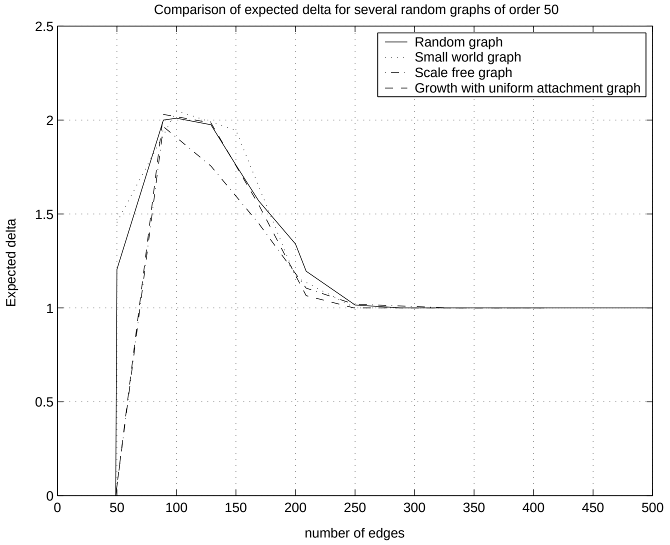


Figure 15.38: Comparison of $E\left(\delta_G\right)$ for all random graph generators of order 50.

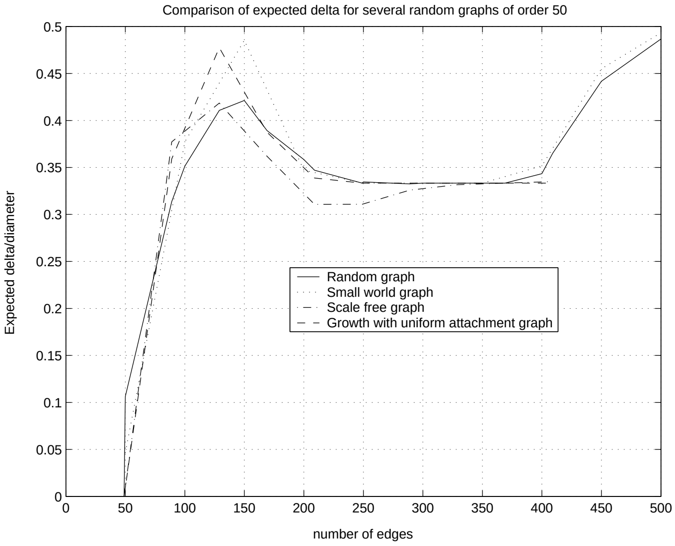


Figure 15.39: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for all random graph generators of order 50.

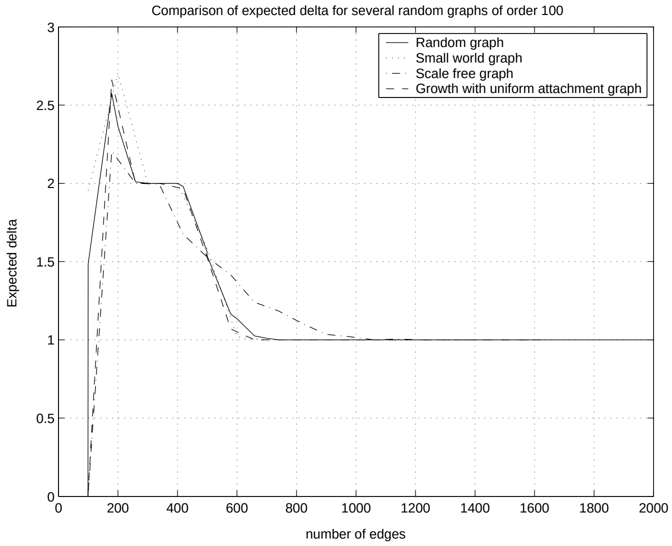


Figure 15.40: Comparison of $E\left(\delta_G\right)$ for all random graph generators of order 100.

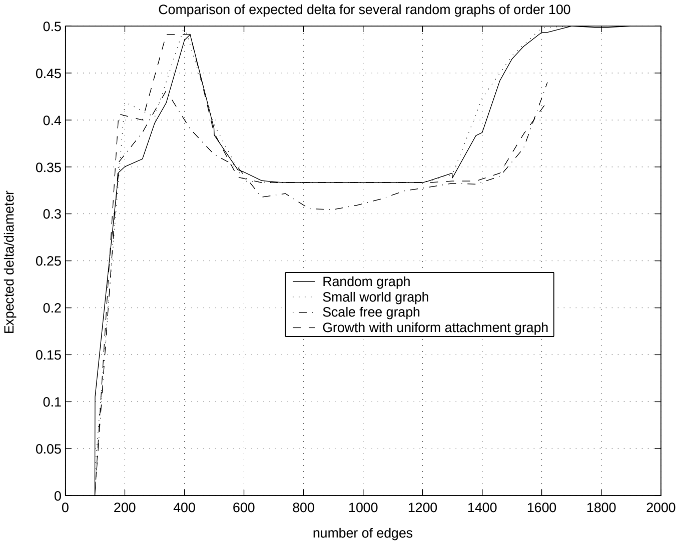


Figure 15.41: Comparison of $E\left(\frac{\delta_G}{diam}\right)$ for all random graph generators of order 100.

the middle range of the sizes of the graphs, the expected delta reaches a minimum and no longer decreases. This explains the flat curve in the middle range. As the size of the graph is continuously increasing, the graph becomes more of a complete graph and the diameter is decreasing until it is comparable to the δ_G. This yields an increase of the expected normalized delta. Therefore the graph is not large enough to observe the hyperbolic property.

2. Although the expected delta in the scale free generator is not less than that of the other random graph generators, *the expected normalized delta for the scale free generator is the minimum among all generators.* This suggests that the scale free graphs are more hyperbolic (in the sense of δ_G) than the other random graphs.

3. The expected normalized delta's in the scale free and in the growth with uniform attachment cases have longer middle ranges than the random graphs and small world graphs. This suggests that the graphs generated from the growth process seem to provide longer range of hyperbolic properties than the graphs without growth process.

4. All random graph generators have the hyperbolic property occurring around the middle range of the sizes of the graphs. This corresponds to a graph which intuitively has the probability p of an edge between two different vertices around $0.15 - 0.25$. This follows from the fact that p is roughly

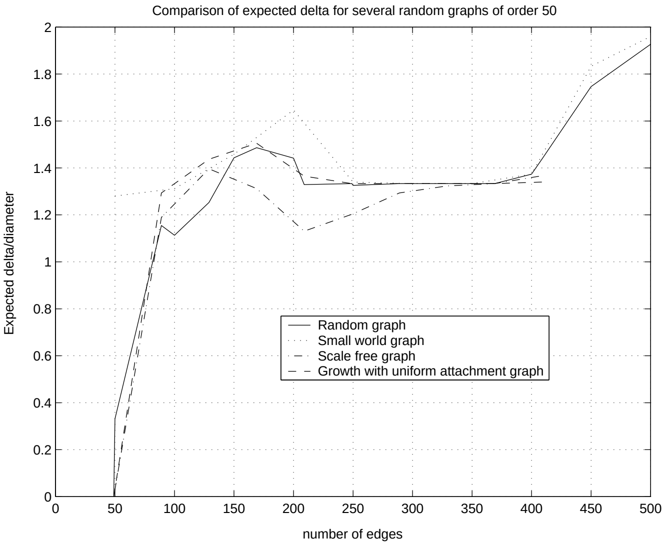


Figure 15.42: The mathematical expectation of $\delta_F/diam$ versus the total number of edges for all 4 graph generators. Observe that the scale free graph is the most hyperbolic.

the fraction of the size of the graph to the total possible number of edges.

15.4 comparison between δ_F and δ_G analyses

Part IV

Flows on large-scale structures

Chapter 16

Robustness against Curvature Uncertainty

16.1 Introduction

A fundamental feature of a communication network is that every link is assigned a nominal link cost [85, p. 402], which, depending on the Type of Service (TOS), reflects communication delay, bandwidth, etc. This link cost defines the metric d, which in turn defines the curvature. During the operation of the network, such highly uncertain factors as recently observed outages, congestion recently observed over the link, delay, packet error rate, etc. are used to readjust the weights in a feedback scheme, the stability of which is still problematic.

If the network has been modeled as a continuous geometric structure as outlined in Chap. **??**, this link cost uncertainty leads to the generic problem of understanding whether a geodesic field, or a geodesic lamination [19], is robust against curvature uncertainty. From a control perspective, a first way to approach this curvature uncertainty is to view it as a simple parameter variation problem, although it is more accurately modeled as an uncertain feedback.

In traditional Riemannian geometry, the issue of the sensitivity of the geodesics *to the end points* is well known and is approached using the Jacobi field concept; however, such a thing as sensitivity to curvature does not appear to have been formalized. We are adding a new dimension to coarse geometry in the sense that the "coarseness" of the space here stems neither from the noncommutativity of the algebra used to represent it, nor from its definition up to a coarsening operator [76, p. 14], but from its imprecisely defined curvature.

Another generic problem that results from these considerations is the understanding of the curvature dependency of control problems over a Riemannian manifold. For example, in the Linear Quadratic Dynamically Varying tracking problem over a Riemannian manifold, as explained in Chapter **??**, we have managed to strip the solution from most of its g_{ij} dependency by measuring the quadratic cost as $\langle x_0, X_{\theta_0} x_0 \rangle$ relative to the bilinear pairing $\langle \cdot, \cdot \rangle$ induced by the metric g_{ij} on the tangent space. In fact, in local coordinates, the Christoffel symbols *completely drop* in both the partial differential Riccati equation (**??**) and the partial differential operator (**??**), meaning that the curvature is involved only in the long range behavior of the control problem. Furthermore, nowhere in the (long range) stability analysis is there any g_{ij} involved, except in the conditions of stabilizability and detectability. We already proved that these conditions are robust against variation of the dynamical map f small in the Whitney topology [38]. From this, it appears that the problem of robustness of stabilizability and detectability against uncertainty in g_{ij} is manageable.

16.2 sensitivity to Christoffel's symbols

If we view the uncertainty as a mere parameter variation acting on the Christoffel symbols as $\Gamma^i_{jk}+\epsilon\Delta^i_{jk}$ and if we model the resulting variation of the geodesic in contravariant coordinates as $\gamma^i+\epsilon\beta^i$, the variation of the geodesic is given, in a local coordinate patch, by

$$\frac{d^2\beta^i}{ds^2}+\left(\left(\Gamma^i_{jk}+\Gamma^i_{kj}\right)\frac{d\gamma^k}{ds}\right)\frac{d\beta^j}{ds}=\frac{d\gamma^j}{ds}\frac{d\gamma^k}{ds}\Delta^i_{jk} \tag{16.1}$$

The feedback modeling of Δ is put aside for the time being, but whatever the outcome, the error field β is given by a linear differential equation with its coefficients varying according to the nominal dynamics γ, and as such the above in a Linear Dynamically Varying (LDV) system in the sense of Chapter **??**. Therefore, the strong stability/stabilizability results developed for these systems are ready to be applied [16, 17, 18].

16.3 sensitivity of geodesics in negative curvature space

This section deals with the archetypical fact that geodesics are well behaved in negative curvature spaces. By well behaved, we mean that the variation of the geodesics remains bounded by a quantity depending on the variation of the curvature. The argument to justify the latter proceeds as follows: The geodesics for a perturbed metric are quasi-geodesics for the nominal metric, and as such the geodesics for the perturbed metric are within a bounded distance of the corresponding geodesics for the nominal metric [23, p. 290]. This of course holds under the assumption that the perturbation of the negatively curved metric remains negatively curved. In case of an idealized infinite network, the Gromov-hyperbolic property as defined in Sec. 13.1 is invariant under quasi-isometry [21, Th. III.1.9], so that, in case the metric variation is a quasi-isometry, the perturbed metric will stay negatively curved.

The corollary of the above general principle is that the "fluttering" problem (see [85, p. 405]) in a networks that happens to be hyperbolic is not due to some extreme sensitivity of the geodesic to the metric, but some feedback instability.

Chapter 17

Worm propagation

17.1 Introduction

A worm can be defined to be a malicious piece of code that self replicates as it self propagates through a network by exploiting software vulnerabilities. The distinction between "worm" and "virus" has its roots in biological epidemiology, where the terminology of "virus" means an organism that consists of a "malicious" DNA or RNA encapsulated in a protein shell and that, as such, is unable to replicate on its own, but will replicate once it has invaded a host cell [34, pp. 20-21]. Likewise, a computer virus needs a program on which it attaches, whereas a worm just propagates on its own.

Here, we look at worm propagation as it relates to the topology of the underlying graph that serves as propagation medium. By the definition of this graph, its nodes are either infected, contagious agents or noninfected, susceptible recipients and its links represent some kind of contacts between nodes that could transmit the pathogenic agent. There are many worms, propagating in different ways, and hence there are many propagation graphs. For such worms as Code-Red choosing their targets by uniform random scanning [66] of the 32 bit address space, the propagation graph would be random in the sense of Chapter 2, if the scanning were truly random. However, in most cases, the "random" scanning is implemented using a pseudorandom number generator and as such the propagation graph is a traveling salesman path visiting all nodes. In the case of the faulty pseudo-random number generator of the Slammer/Sapphire worm [65], the nature of the propagation graph is less clear. For such worms as Code-Red II and Nimda, which scan preferentially the local subnet [66], the propagation graph is more towards the physical graph and hence deterministic. As such, in one time tick, the worm could either travel through several sites before reaching its target, or it could just attack its immediate neighbor. For an e-mail worm [70], the graph is the logical e-mail graph in the sense of Chapter A. Some simulation of Code-Red v2 have used the Autonomous System (AS) graph where the worm was jumping at random from one AS to another [66].

There are two accepted propagation models that somehow refer to the propagation mode: the Epidemiological model and the Analytical Active Worm Propagation (AAWP) model [25]. The former is a traditional, generic model deeply rooted in public health epidemiology, whereas the former is specifically devised for computer worms propagating by random scanning.

In this chapter, we study a different aspect of the propagation, in the sense that it is more relevant to the topology oriented propagation of e-mail worms and peer-to-peer worms. In this case, the underlying graph structure, e.g., the mail logical graph in case of e-mail worm, plays a predominant role. This study is more relevant to such worms as Code Red II and Nimda that preferentially attack neighbors than to worms that randomly choose their targets. The basic epidemiological feature that the propagation depends on the fraction of uninfected machines remains in this study.

The most specific aspects that are investigated here are the specific features of propagation on a hyperbolic graph. This of course creates the problem of generating a great many hyperbolic graphs, so that the general pattern of the propagation of the infection could be understood. The latter problem is approached using Cayley graphs of combinatorial group theory as prototype of hyperbolic graphs.

17.2 epidemiological model

Qualitatively, the epidemiological model is a *homogeneous* mode of propagation where *any* contagious agent could transmit the pathogen to *any other* susceptible recipient with *uniform* probability p. This is also sometimes referred to as *homogeneous mixing* model. Since any infected node in a vulnerable population n could potentially infect any uninfected node with probability p, it can be said that the propagation occurs on a $G^{p,n}$ random graph, or the complete graph K_n where every link has a probability p of carrying the infection. Quantitatively, the premise of the epidemiological model is that the infection rate is proportional to the product of the number of infected nodes and the fraction of uninfected nodes. That the rate depends on the number of infected agents is obvious. The dependency on the number of uninfected subjects relates to the fact that the infection rate depends on the number of subjects susceptible of being infected. Specifically, let $\beta(t)$ be the number of infected subjects in a "vulnerable" population of n (the rationale for the notation $\beta(t)$ will become clearer later). Define the fraction of uninfected subjects as

$$1 - \frac{\beta(t)}{n}$$

The epidemiological model is sometimes referred to as the famous "logistic" equation,

$$\dot{\beta}(t) = r\beta(t)\left(1 - \frac{\beta(t)}{n}\right) - d\beta(t)$$

where r is the contact rate, that is, the average number of times an infected agent will infect a susceptible subject per unit time, and d is the death rate, that is, the rate at which infected agents die and are no longer considered "infected," nor are they contagious. For example, in case of Code-Red, which does uniform scanning on the 32 bits IP address space, $r = s\frac{n}{2^{32}}$, where s is the scanning rate. To study this infection, one could plot the number of infected nodes $\beta(t)$ versus t, but we prefer to plot the *rate of infection per agent* v, that is, the average number of subjects that will soon become infected divided by the number of contagious agents, that is, assuming $d = 0$,

$$v = \frac{\dot{\beta}}{\beta} = r\left(1 - \frac{\beta}{n}\right)$$

Clearly, for a discrete-time evolutionary model, the normalized infection rate would be measure as

$$v = \frac{\beta_{k+1} - \beta_k}{\beta_k}$$

17.3 Analytical Active Worm Propagation

The AAWP model is more accurate than the epidemiological model, in the sense that it incorporates the feature that most worms do a network scanning to choose their next target. The AAWP model is better suited to totally random scanning, although it can be modified to model the preferential scanning limited to the local subnet. In addition to the fundamental feature of the epidemiological model, the AAWP model incorporates some elementary model of the patching that users are putting in place when they are warned that a worm is crawling throughout the network. The model is based on the following lemma:

Lemma 19 *Let n be the total number of vulnerable machines and let b_k be the number of infected machines at time k. Set $b_{k+1} = b_k + \Delta_k$. Then if $E(\Delta_k|j)$ denotes the expectation of Δ_k under j scans,*

$$E(\Delta_k|j) = (n - b_k)\left(1 - \left(1 - \frac{1}{2^{32}}\right)^j\right)$$

Proof. The proof [25, Theorem 1] is by induction on j. The formula is obviously satisfied for $j = 0$. Furthermore,

$$\begin{aligned}
&E(\Delta_k|j+1) \\
&= (E(\Delta_k|j)+1)\frac{n - b_k - E(\Delta_k|j)}{2^{32}} + E(\Delta_k|j)\left(1 - \frac{n - b_k - E(\Delta_k|j)}{2^{32}}\right) \\
&= \frac{n - b_k - E(\Delta_k|j)}{2^{32}} + E(\Delta_k|j) \\
&= \frac{n - b_k}{2^{32}} + (1 - \frac{1}{2^{32}})(n - b_k)(1 - (1 - \frac{1}{2^{32}})^j) \\
&= \frac{n - b_k}{2^{32}}(1 - (1 - \frac{1}{2^{32}})^{j+1})
\end{aligned}$$

■

Since there are sb_k scans during one time tick, the model takes the form

$$b_{k+1} = b_k + (n - b_k)\left(1 - \left(1 - \frac{1}{2^{32}}\right)^{sb_k}\right)$$

The above model can be refined by introducing a death rate d and a patch rate u,

$$b_{k+1} = (1 - d - u)b_k + \left((1-u)^k n - b_k\right)\left(1 - \left(1 - \frac{1}{2^{32}}\right)^{sb_k}\right)$$

The graph over which the propagation occurs is again the $G^{p,n}$ graph in the random scanning case.

17.4 comparison between two models

The AAWP model is closer to the epidemiological model than one might first believe, at least for $sb_k << 2^{32}$. Indeed, under this condition, we have

$$\left(1 - \frac{1}{2^{32}}\right)^{sb_k} \approx 1 - \frac{sb_k}{2^{32}}$$

so that the AAWP model becomes

$$b_{k+1} = (1 - d - u)b_k + \left((1-u)^k n - b_k\right)\frac{sb_k}{2^{32}}$$

Under the no patch ($u = 0$) condition, and if T denotes the time between two clock ticks, the above becomes,

$$\frac{b_{k+1} - b_k}{T} = \frac{sn}{T2^{32}} b_k \left(1 - \frac{b_k}{n}\right) - \frac{d}{T} b_k$$

The above is clearly equivalent to the epidemiological model.

Chapter 18

Groups, Cayley graphs, and Cayley Complexes

Let $\langle g_1, ..., g_r | R_1, ..., R_m \rangle$ be a presentation of a group Γ by generators and relators. The Cayley graph of this presentation is the graph rooted at the identity element 1, with branches corresponding to right multiplication by g_i, g_j^{-1}. The relators R_i destroy the tree structure of the graph and create loops. The distance between two words $d(w_1, w_2)$ is the minimum number of generators g_i, g_j^{-1} needed to construct $w_1^{-1}w_2$. It is easily seen that this distance is symmetric and that the Cayley graph is a geodesic space. Of course, a communication network graph is far from a Cayley graph, the chief difference being that communication network graphs are more heterogeneous; nevertheless, Cayley graphs because they are so well understood provide an ideal testbed of new theories. For example, the rate of propagation of a worm on a graph appears to be slowed down by loops; this phenomenon is most easily analyzed on Cayley graphs of "small cancellation" groups (see [59, Chap. V]), because the number of loops and their sizes are easily controlled by the relators and their word length.

18.1 basic definitions

A group $(\Gamma, \times)$ is a set Γ endowed with an internal law $\times$ satisfying the conditions

- Associativity: $(a \times b) \times c = a \times (b \times c)$.
- Neutral Element: $a \times 1 = 1 \times a = a$.
- Inverse Element: $a \times a^{-1} = a^{-1} \times a = 1$.

The internal composition law is written as a multiplication to indicate that it need not be commutative. If, however, the internal composition law satisfies the extra condition

- Commutativity: $a \times b = b \times a$,

then the group is said to be Abelian. From here on we will drop the × symbol and write the "multiplicative" law as a mere juxtaposition, viz., $a \times b = ab$.

The order of a group Γ is the cardinality of Γ as a set. This cardinality is written $|G|$. The order of an element $\gamma \in \Gamma$ is the order of the subgroup it generates. Clearly the subgroup generated by γ is $\{1, \gamma, \gamma^2, ...\}$. If m is the least integer such that $\gamma^m = 1$, the element γ is said to be of order m.

$H \subseteq \Gamma$ is said to be a subgroup if it is closed under the multiplicative law. $H\gamma = \{h\gamma : h \in H\}$ is called the right coset of H in Γ. A similar definition holds for the left coset. Two cosets Ha, Hb with $a \neq b$ are either equal or disjoint. It is easily seen that the group Γ can be partitioned as $\Gamma = Ha \cup Hb \cup Hc \cup ...$ for some selected elements $a, b, c, ... \in \Gamma$. It is easily seen that to each partitioning in right cosets there is a corresponding partitioning in left cosets, and vice versa. The number of (right or left) cosets in a partitioning is called the index of the subgroup H in Γ and is written $[\Gamma : H]$. This notation is motivated by the fact that, if Γ is finite, then so is H and the order of H divides the order of Γ and the quotient $\frac{|\Gamma|}{|H|}$ is the index.

If Γ is not Abelian, it is easily seen that the set of right cosets does not form a group, which we would call the quotient group Γ/H. This motivates the definition of a normal subgroup $N \subseteq \Gamma$ to be a subgroup such that $\gamma^{-1}n\gamma \in N$, $\forall \gamma \in \Gamma$, $\forall n \in N$. In other words, given $n\gamma$, $n \in N$, there always exists an $n' \in N$ such that $n\gamma = \gamma n'$. It follows that for a normal subgroup, the right and left cosets are the same, viz., $N\gamma = \gamma N$. Given two cosets $N\gamma_1 = \gamma_1 N$, $N\gamma_2 = \gamma_2 N$, the multiplication of any two representatives $n_1\gamma_1 = \gamma_1 n_1'$, $n_2\gamma_2 = \gamma_2 n_2'$ yields a representative $n_1\gamma_1\gamma_2 n_2'$ of the coset $N\gamma_1\gamma_2 = \gamma_1\gamma_2 N$. Hence the cosets unambiguously form a group, called quotient group and written Γ/N.

Conversely, there arises the question as to whether a group can be reconstructed from a normal subgroup and its quotient by the normal subgroup. More specifically, if we are given two groups N and Q, can we find a group Γ such that N is normal in Γ and $Q = \Gamma/N$? In other words, can we find a group Γ that fits within the short exact sequence

$$0 \rightarrow N \rightarrow \Gamma \rightarrow Q \rightarrow 0$$

The answer is yes and is given by the direct product of N and Q. More generally, the group extension problem addresses the problem of exhausting all groups Γ fitting within the above short exact sequence. More involved however is the problem of reconstructing the cohomology of Γ from the cohomology of N and Γ/N. The answer is given by a successive approximation procedure, called the Lyndon spectral sequence.

A set of generators for a group is a subset of group elements $a_1, ..., a_n$ such that every element of the group can be reconstructed as the product of selected generators and their inverses in some selected order. Such a product is called a word. The group so generated by $a_1, ..., a_n$ is called the free group on n generators and is written $< a_1, ..., a_n >$. It is easily seen that all free groups on n generators are isomorphic, so that any such group is sometimes rewritten as

F_n. In most of the interesting cases, however, the group is restricted by some constraints among its generators. These constraints are called relators and are usually written as $R_i(a_1, ..., a_n) = 1$, where R_i is a monomial in the a_i's and their inverses. The typical example is an Abelian group in which the generators are related by $a_i a_j a_i^{-1} a_j^{-1} = 1$. The group defined by the set of generators $A = \{a_1, ..., a_n\}$ and relators $R = \{R_1, ..., R_p\}$ is written $\Gamma =< A, R >$, the presentation of the group in terms of generators and relators.

Even when a group is defined by only one relation, it should be observed that the cyclic conjugates of the relation word are also relations. For example, if $\Gamma =< a, b, c|abc >$, the relation $abc = 1$ yields $bc = a^{-1}$ and hence $bca = 1$. Observe that (b, c, a) is a cyclic permutation of (a, b, c); hence the terminology that bca is the cyclic congugate of abc. Observe that, in this example, there is still one more cyclic congugate of the relation, $cab = 1$. If a group is presented as $< A|R >$, define F be the free group on the generators A and let N be the normal closure of R (and its cyclic congugates). The normal closure is the smallest normal group containing the relations (and their cyclic congugates). It follows that the group Γ can also be described as the quotient F/N.

18.2 examples of groups

Groups are abundant in mathematics and in physics. A group is most easily defined by a textual description of its elements and its composition law. However, in most cases, the textual description in argot language can be translated in the formal language of generators and relators.

18.2.1 polynomial matrices groups

Recall that a square polynomial matrix $P(s)$ is said to be unimodular whenever $\det(P(s)) = 1$, $\forall s$. It is well known, and easily verified, that the group of unimodular matrices is freely generated by the identity matrix I and all all matrices $G_{ij}^{\pm}$, $i, j = 1, ..., n$, where $G_{ij}^{\pm}$ is the matrix composed of one's on the diagonal, $\pm s$ in the (i, j)-position, and zeros everywhere else.

18.2.2 The modular group

Recall that a Möbius transformation is a bilinear mapping of the complex plane into itself

$$\begin{array}{rcl} \mathbb{C} & \to & \mathbb{C} \\ z & \mapsto & \dfrac{az+b}{cz+d} \end{array}$$

and such that $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \neq 0$. If, in addition, $a, b, c, d \in \mathbb{Z}$ and $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = 1$, then the transformation is said to be modular. The set of such transformations form the modular group, also referred to as the projective special linear group $PSL(2,\mathbb{Z})$ of transformations of $\mathbb{C}^2$ with integer coefficients. If we define the generators,

$$f(z) = -\frac{1}{z}, \quad g(z) = \frac{1}{-z+1}$$

then a presentation of the modular group is

$$PSL(2,\mathbb{Z}) = \left\langle f, g \,\middle|\, f^2, g^3 \right\rangle$$

(See [60, Sec. 1.4].)

18.2.3 combinatorial groups

The permutation group Σ_S of the set S is the set of bijections $\pi : S \to S$ under composition of transformations. The symmetric group Σ_n is the group of all permutation of $\{1, 2, ..., n\}$. To define $\Sigma_{\{a,b,c\}}$ in terms of generators and relators, define the generator p to be the cyclic permutation $a \mapsto b \mapsto c$ and q to be the cyclic permutation $a \mapsto c$. Then (see [60, Sec. 1.1])

$$\Sigma_{\{a,b,c\}} =< p, q : p^3, q^2, pq = qp^2 >$$

Another presentation is in terms of the generator r defined to be the cyclic permutation $a \mapsto b$ and the generator s defined as the cyclic permutation $a \mapsto c$. Indeed, it can be shown that (see [60, Sec. 1.1])

$$\Sigma_{\{a,b,c\}} =< r, s : r^2, s^2, (rs)^3 >$$

A transposition t_{ij} is an elementary permutation of $\{1, 2, ..., n\}$ fixing all elements except i, j, that is, $t_{ij}(1, 2, ..., i, ..., j, ..., n) = (1, 2, ..., j, ..., i, ..., n)$. The alternating group $A_n \subset \Sigma_n$ is the group of transformations $\{1, 2, ..., n\} \to \{1, 2, ...n\}$ consisting of an even number of transpositions. Equivalently, it is the subgroup of Σ_n such that $\Pi_{1 \le i<j \le n}(x_i - x_j) = \Pi_{1 \le i<j \le n}(x_{\pi(i)} - x_{\pi(x_j)})$.

The dihedral group D_n is the group of symmetries of a regular n-gon. A representation of D_{2n} is the group of matrices

$$\begin{pmatrix} \pm 1 & k \\ 0 & 1 \end{pmatrix}$$

with their elements in $\mathbb{Z} \bmod k$ and under matrix multiplication. If we define as generators

$$a = \begin{pmatrix} +1 & 1 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

then the presentation is

$$D_{2n} = \langle a, b | a^n, b^2, ba = a^{-1}b \rangle$$

18.2.4 Braid group

It is probably out of knot theory [1, 4, 56] that presentation of groups in terms of generators and relators comes out most naturally. Consider n strings hanging vertically from their top end points $t_1, t_2, ..., t_n$ arranged from left to right on a horizontal line. The strings go in a "box" in which they are crossed in such a way that along every string the height of a point is a decreasing function of the string length measured from the top end point, and the strings get out of the box in the sequence $b_1, ...b_n$ of bottom end points arranged from left to right on a horizontal line parallel to the top one. Such a crossing of the strings as they go from the top line to the bottom line is called a braid. Formally, a braid is the track of an isotopy $\{t_1, ..., t_n\} \to \{b_1, ..., b_n\}$ (see [45, p. 111]). No distinction is made between two braid that can be deformed into each other without crossing. Formally, no distinction is made between two braids that can be transformed into each other by an isotopy with fixed top end points and bottom end points (see [56, p. 9]). Clearly, such (isotopy classes of) braids can be composed to yield the braid group B_n on n strings. For this group, define generators $\sigma_1, ..., \sigma_{n-1}$, where σ_i, $i = 1, ..., n-1$ is the braid in which strings i and $i+1$ cross once, with string i on top of string $i+1$, while all other strings do not cross. Observe that σ_i is *not* the transposition $t_{i,i+1}$ because the latter contains no information as to whether string i goes on top or underneath string $i+1$. Then it can be shown (see [56, p. 10]) that a presentation of the braid group B_n is given by

$$B_n = \left\langle \sigma_1, ..., \sigma_{n-1} \,\middle|\, \begin{array}{l} \sigma_i\sigma_j = \sigma_j\sigma_i, \text{ for } |i-j| \geq 2 \\ \sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1} \end{array} \right\rangle$$

The connection with links and knots is established via the concept of closed braids. Recall that a m component link is a smooth embedding of m disjoint circles into $\mathbb{R}^3$. A knot is a smooth embedding $S^1 \to \mathbb{R}^3$. Clearly, a knot is a 1-component link. A closed braid is obtained by connecting the top and bottom end points of a braid box by the identity braid. The celebrated Alexander theorem asserts that every link or knot has a representation as a closed braid (see [1, Sec. 5.4] for an elementary proof).

18.2.5 the fundamental group

Intuitively a surface S might have closed paths that cannot be shrunken to a point. Furthermore, nontrivial closed paths can be combined to produce more complicated closed paths; for example, combining a closed path around the great circle of a torus with several closed paths around the small circle yields a "coil of a ferrite." The fundamental group of a surface or manifold is the group of all nontrivial closed paths, up to continuous deformations, endowed with the composition of paths operation.

Formally, given a topological space X along with a reference point x_0, called base-point, a path on X hinged at x_0 is a continuous function

$$\begin{array}{rcl} f:[0,1] & \rightarrow & X \\ t & \mapsto & f(t) \end{array}$$

such that

$$f(0)=f(1)=x_0$$

Two paths f,g on x_0 are said to be base point preserving homotopic if there exists a function

$$\begin{array}{rcl} F:[0,1]\times[0,1] & \rightarrow & X \\ (s,t) & \mapsto & F_s(t) \end{array}$$

such that

$$F_0(t)=f(t), F_1(t)=g(t)$$

$$F_s(0)=F_s(1)=x_0$$

We shall not make a distinction between two paths that are homotopic. Given two paths f,g, their product or composition is the path $f*g$ defined as

$$\begin{array}{rcl} f*g:[0,1] & \rightarrow & X \\ \left[0,\frac{1}{2}\right]\ni t & \mapsto & f(2t) \\ \left[\frac{1}{2},1\right]\ni t & \mapsto & g(2t-1) \end{array}$$

The composition path so defined is required to pass through the base point for $t=\frac{1}{2}$, but the paths base point preserving homotopic to $f*g$ are not. The homotopy class of $f*g$ is written $\{f*g\}$. It is easily seen that the latter can be obtained from any representative of the homotopy classes of f,g so that that $\{f*g\}=\{f\}*\{g\}$. The identity path is the constant mapping into x_0. The inverse path is simply the path traversed in reverse direction. The fundamental group $\pi_1(X,x_0)$ of the space X relative to the base point x_0 is the group of all homotopy classes of closed paths endowed with the composition operation $*$. In general, this group is not Abelian.

The fundamental group in a sense depends only weakly on the base point. Formally, a change of base point from x_0 to x_1 changes the fundamental group by no more than a group isomorphism

$$\pi(X,x_0)\rightarrow\pi(X,x_1)$$

It is customary to find generators for the fundamental group. For example, for a “8” with the cross point chosen as base point, the two generators could be the top and bottom loops of the “8,” the top loop with clockwise direction and the bottom loop with counterclockwise direction. The composition of the two generators would be the whole figure “8.” For a torus, the two generators would be a closed loop around the small circle and a closed loop around the large circle. For a surface of genus 2, things are a little more complicated. Clearly, if we weld two tori, there are 4 generators, two generators a_1, a_2 around the two small circles and two generators b_1, b_2 around the two large circles. The problem is that these generators are not independent. In fact, it takes a little bit of geometric intuition to observe that $a_1 b_1 a_1^{-1} b_1^{-1}$ is homotopic to $b_2 a_2 b_2^{-1} a_2^{-1}$, so that we have a relation $a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} = 1$. In other words,

$$\pi(S_2, s_0) =< a_1, b_1, a_2, b_2 | a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} >$$

In general, for a surface of genus g,

$$\pi(S_g, s_0) =< a_1, b_1, a_2, b_2, ..., a_g b_g | a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} ... a_g b_2 a_g^{-1} b_g^{-1} >$$

In particular, for the usual torus,

$$\pi_1(S_1, s_0) =< a_1, b_1, a_2, b_2 | a_1 b_1 a_1^{-1} b_1^{-1} >$$

that is, the fundamental group is Abelian.

18.3 The Burnside problem

A finitely presented group may or may not be infinite and it is in general not easy to determine whether a finitely presented group is finite. In particular, if Γ is a finitely presented group such that every element γ is nilpotent, that is, $\gamma^{m(\gamma)} = 1$, where $m(\gamma)$ might depend on γ, it is true that the group is finite? This is the celebrated Burnside problem. The answer is that the group is not in general finite.

The solution to the Burnside problem is very complicated and in fact was only recently solved by Zelmanov, who was awarded the Fields Award (the "Nobel Prize" in Mathematics) for the solution to the problem.

A somewhat restricted version of the Burnside problem is the case of a finitely presented group such that there exists an m such that $\gamma^m = 1$, $\forall \gamma$. Observe that, here, m does not depend on the word γ. Again, the answer is that the group is not finite in general.

More formally, define

$$B(e, n) = F_n / N$$

where N is the normal subgroup of F_n generated by all eth powers of elements of F_n. $B(e, n)$ is called Burnside group on n generators and exponent e. For $e = 2$, we have $a^2 = 1$ and hence $a = a^{-1}$; furthermore $(ab)(ab) = aba^{-1}b^{-1} = 1$ so that the group is Abelian. Therefore, the group is finite.

We can illustrate the Burnside problem for the case where $m = 2$ by computing the number of words of a certain lenth goes and showing that it goes to zero as the length increases. Consider the group

$$\Gamma =< a, b, c, ... | \gamma^2 = 1 >$$

on n generators, where γ is an arbitrary word, and let us prove that it is finite. Let $N(k)$ be the number of words of minimal length k. We find a recursion on $N(k)$. Consider a word $w = a_1 ... a_k$ of length k and let us add a generator ? so as to create a word $w?$ of length $k+1$. Clearly, ? could not be the rightmost generator a_k in w for otherwise we would get a word of length $k-1$ rather than $k+1$. Therefore,

$$N(k+1) \leq N(k)(n-1)$$

Next, we have to make sure that $a_{k-2}a_{k-1} \neq a_k?$ for otherwise we would get a word of length $k-3$. Therefore,

$$N(k+1) \leq N(k)(n-1) - N(2)N(k+1-4)$$

Next, we want to make sure that $a_{k-4}a_{k-3}a_{k-2} \neq a_{k-1}a_k?$ for otherwise we get a word of lenth $k-5$. Hence

$$N(k+1) \leq N(k)(n-1) - N(2)N(k+1-4) - N(3)N(k+1-6)$$

The recursion should now be obvious. We keep on removing unacceptable ?'s until we have exhausted all unacceptable solutions and this yields

$$N(k+1) = N(k)(n-1) - \sum_{\substack{x=2 \\ 2x \leq k+1}}^{\infty} N(x)N(k+1-2x)$$

It is clearly seen that for whatever number of generators, $N(k)$ eventually goes to zero.

For example, if $n = 3$, we get

$$\begin{aligned}
N(1) &= 3 \\
N(2) &= 6 \\
N(3) &= 6 \times 2 = 12 \\
N(4) &= 12 \times 2 - N(2) = 18 \\
N(5) &= 18 \times 2 - N(2)N(1) = 18 \\
N(6) &= 18 \times 2 - N(2)N(2) - N(3) \times 1 = -12
\end{aligned}$$

Therefore, it appears that with 3 generators the maximum word length is 5. As such the group is finite.

18.4 Commutator calculus

NonAbelian groups are, from the point of view of computation and complexity, extremely difficult to handle. The concept of solvability of a given group comes as some kind of a reassurance that despite the fact that the group is nonAbelian, it is computationally tractable. Define the commutator of two elements a, b to be $[a, b] = aba^{-1}b^{-1}$. The derived group Γ' is defined to be the subgroup generated by all commutators of Γ. Equivalently, $\Gamma' = [\Gamma, \Gamma]$. It is easily seen that Γ' is a normal subgroup and that Γ/Γ' is Abelian. A group Γ is said to be solvable if the derived series

$$\Gamma \supseteq \Gamma' \supseteq \Gamma'' \supseteq ...$$

eventually terminates with an Abelian group. Equivalently, there exists a sequence of normal subgroups

$$\Gamma \supseteq N_1 \supseteq N_2 \supseteq ...$$

such that N_i/N_{i+1} is Abelian.

Γ is said to be nilpotent if the lower central series

$$\Gamma \supseteq [\Gamma, \Gamma] \supseteq [\Gamma, [\Gamma, \Gamma]] \supseteq ...$$

eventually terminates with 1. A nilpotent group is solvable.

Solvability of groups is crucially related to such traditional problems as solution of equations by radicals and compass and ruler constructions. This is the so-called Galois theory. In particular the Galois theory states that an equation is solvable by radicals if the symmetry group of its roots is solvable. It can be shown that $S_n, n \geq 5$ is not solvable. Therefore, we obtain the celebrated result that the quintic equation is not solvable by radicals. Another celebrated traditional problem–the fact that an angle cannot be trisected–has to do with the fact that the Galois group of a related field extension is not solvable.

18.5 Cayley graphs

The Cayley graph of a finitely presented groups is the graph, the nodes of which are all (reduced) words of the groups and two words w_1, w_2 are linked iff there exists a generator a such that either $w_1 a = w_2$ or $w_1 a^{-1} = w_2$. Each link is bidirectional in the sense that going along the link in one direction corresponds to multiplication by, say, a while going in the reverse direction would be multiplication by a^{-1}. Therefore, the Cayley graph is defined to be undirected. The Cayley graph is rooted at the identity element 1.

The Cayley graph of a free group is a tree. The Cayley graph of an Abelian group on n generators is the n-dimensional cubical lattice.

The Cayley graph is made a metric space as follows: Each link is assigned a weight of 1. The distance between two words, w_1, w_2 is defined as

$$d(w_1, w_2) = \min\{m : w_1 a_{i_1} a_{i_2} ... a_{i_m} = w_2, a_i \text{ or } a_i^{-1} \in A\}$$

In other words, $d(w_1, w_2)$ is the minimum number of hops between w_1 and w_2. It is easily seen that this is a distance, i.e., that it satisfies the triangle inequality. The length of a word w is defined to be its distance from 1,

$$\ell(w) = d(1, w)$$

in other words, it is the minimum number of generators (or their inverses) that is needed to construct the word.

18.6 growth functions

The growth function is defined as

$$\beta(k) = \#\{\gamma : d(1, \gamma) \leq k\}$$

The growth series is defined as

$$B(z) = \sum_{k=0}^{\infty} \beta(k) z^k$$

The spherical growth function is defined as

$$\sigma(k) = \beta(k) - \beta(k-1) = \#\{\gamma : d(1, \gamma) = k\}$$

The spherical growth series is defined as

$$\Sigma(z) = \sum_{k=0}^{\infty} \sigma(k) z^k$$

It is easily seen that

$$\beta(k) = \sum_{\ell=0}^{k} \sigma(\ell)$$

from which it follows that

$$B(z) = \Sigma(z)(1 + z + z^2 + ...) = \frac{\Sigma(z)}{1 - z}$$

Clearly $\beta(k)$ is of exponential growth iff the radius of convergence of the $B(z)$ series is < 1. A similar statement would hold for the spherical growth. In fact, because of the relation between the growth series, it follows that $\beta(k)$ is exponential iff $\sigma(k)$ is exponential.

Clearly, if we take two groups elements γ_1 and γ_2 and multiply them, some cancellations might occur. It follows that the length is sublogarithmic, that is,

$$\ell(\gamma_1 \gamma_2) \leq \ell(\gamma_1) + \ell(\gamma_2)$$

Now, consider the following string:

$$\begin{array}{rcl} \{\gamma : \ell(\gamma) \leq k_1 + k_2\} & = & \{\gamma_1 \gamma_2 : \ell(\gamma_1) \leq k_1, \ell(\gamma_2) \leq k_2\} \\ & \subseteq & \{\gamma_1 : \ell(\gamma_1) \leq k_1\} \times \{\gamma_2 : \ell(\gamma_2) \leq k_2\} \end{array}$$

The subset inclusion stems from the fact that there are many ways of breaking a word γ as $\gamma_1 \gamma_2$. Taking the resulting subset inclusion and going to the cardinality we find that the growth function is subexponential, that is,

$$\beta(k_1 + k_2) \leq \beta(k_1) \beta(k_2)$$

Theorem 60 *The (spherical) growth series of a finitely presented Gromov hyperbolic group is rational.*

Theorem 61 *A group Γ has polynomial growth if and only if if contains a nilpotent group of finite index.*

Chapter 19

worm propagation on Cayley graphs

19.1 growth function propagation model

Assume a worm is born at the identity element of a Cayley graph at time $k = 0$ and that from every node it will infect in one time tick all nodes within a unit distance of an infected node. As such, the worm propagates from the identity element through the whole graph. At time k, there are $\beta(k)$ infected nodes, where $\beta(k)$ is the growth function introduced in Section 18.6. The problem is that this kind of infection on a Cayley graph is not of the *homogeneous mixing* type, because among the $\beta(k)$ infected nodes, only $\sigma(k)$ of them, where $\sigma(k)$ is the spherical growth function of Section 18.6, are "active," or "contagious," and only those nodes could infect those at a distance $(k+1)$ of 1. The lack of homogeneity is two-fold: first, the contagious agents are inhomogeneous in the infected population and, second, the contagious agents can only target a population of recipient inhomogeneous in the uninfected population. In this inhomogeneous case, given that $\beta(k+1)-\beta(k)$ subject will be infected by $\sigma(k)$ agents, the infection rate, normalized per unit contagious agent, is

$$v = \frac{\beta(k+1) - \beta(k)}{\sigma(k)} = \frac{\sigma(k+1)}{\sigma(k)}$$

The quantity $\frac{\sigma(k+1)}{\sigma(k)}$ can be referred to as *number of uninfected neighbors.* Our objective is to achieve an analytical understanding of the normalized speed of propagation, given the analytical understanding of the growth function. Specifically, we would like to plot $\frac{\sigma(k+1)}{\sigma(k)}$ as a function of k, that is, as a function of the time. Another possibility is to plot $\frac{\sigma(k+1)}{\sigma(k)}$ as a function of $\beta(k)$, that is, as a function of the total number of infected nodes.

19.2 Elementary examples of worm propagation

19.2.1 free group

For a free group on n generators, the unit element has n outflowing links, each producing a distinctive new element of length 1. Therefore, $\sigma(1) = 2n$. However, at any nontrivial node γ, that is, a node at a distance $k \geq 1$ from the identity, there are only $2n-1$ links that yield an element of length $k+1$; indeed, the link that multiplies by the inverse of the rightmost generator in γ yields a word of length $k-1$, not $k+1$. It follows that $\sigma(k) = 2n(2n-1)^{k-1}$, $k \geq 1$. The spherical growth series is

$$\begin{aligned}\Sigma(z) &= 1 + 2nz + 2n(2n-1)z^2 + 2n(2n-1)^2z^3 + ... \\ &= 1 + 2nz\left(1 + (2n-1)z + (2n-1)^2z^2 + ...\right) \\ &= 1 + 2nz\frac{1}{1-(2n-1)z} \\ &= \frac{1+z}{1-(2n-1)z}\end{aligned}$$

and the growth series is

$$B(z) = \frac{1+z}{(1-z)(1-(2n-1)z)}$$

It follows that the worm propagation speed $\frac{\sigma(k)}{\sigma(k-1)}$ is constant, equal to $(2n-1)$ for $k \geq 1$.

19.2.2 Abelian group

We now consider the growth problem in an Abelian group $\Gamma =< a_1, ..., a_n | a_i a_j a_i^{-1} a_j^{-1} >$ on n generators. First, observe that if Γ is Abelian, $[\Gamma, \Gamma] = 1$, so that an Abelian group is nilpotent. Therefore, Γ contains a nilpotent subgroup (Γ) of finite index (1), so that by Theorem 61, the growth should be polynomial.

In the case of an Abelian group, any word is of the form $\gamma = a_1^{x_1} ... a_n^{x_n}$, $x_i \in \mathbb{Z}$ and can be viewed as a point with coordinates $(x_1, ..., x_n)$ in $\mathbb{R}^n$. It follows that $\beta(k)$ and $\sigma(k)$ are the cardinalities of the sets

$$\begin{aligned}\{\gamma \in \Gamma : \ell(\gamma) \leq k\} &= \{(x_1, ..., x_n) : \sum_{i=1}^n |x_i| \leq k\} \\ \{\gamma \in \Gamma : \ell(\gamma) = k\} &= \{(x_1, ..., x_n) : \sum_{i=1}^n |x_i| = k\}\end{aligned}$$

Clearly the points of the first (the second) set are within the rhombus (the boundary of the rhombus) with vertices

$$(\pm k, 0, ..., 0), (0, \pm k, ..., 0), ..., (0, 0, ..., \pm k) \tag{19.1}$$

Let $\beta_n(k)$ ($\sigma_n(k)$) be the number of word points in the rhombus (on the boundary of the rhombus) of size k in dimension n. The recursion for the σ's is easily seen to be

$$\sigma_n(k+1) = \sigma_n(k) + \sigma_{n-1}(k+1) + \sigma_{n-1}(k)$$

The intuition behind the above is the following: The $(k+1)$ boundary shell of the rhombus, which has $\sigma_n(k+1)$ elements, can be obtained from the k boundary, which has $\sigma_n(k)$ elements, by slicing the k boundary into two pieces, $x_{\geq}0$ and$x_1 < 0$, and moving the pieces of k boundaries one step along the x-axis so as to fit the pieces with the (k+1) boundary at apexes $(\pm(k+1), 0, ..., 0)$. To do this symmetrically, we need another version of the $x_0 = 0$ basis of the k-shell, which has $\sigma_{n-1}(k)$ elements. After doing so, we are missing the $x_0 = 0$ basis of the $(k+1)$ shell, which has $\sigma_{n-1}(k+1)$ elements. This yields the above formula.

Now, multiplying the above by z^k and taking the summation over all k's yields,

$$\frac{1}{z}\Sigma_n(z) = \Sigma_n(z) + \frac{1}{z}\Sigma_{n-1}(z) + \Sigma_{n-1}(z)$$

that is,

$$\frac{\Sigma_n(z)}{\Sigma_{n-1}(z)} = \frac{1+z}{1-z}$$

and finally

$$\Sigma_n(z) = \left(\frac{1+z}{1-z}\right)^n$$

from which $\sigma_n(k)$ and $\beta_n(k)$ can easily be found.

We put in the following table some of the $\sigma_n(k)$'s. The column index is the length k, starting at $k = 0$, while the row index is number n of generators.

	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$
$\sigma_n(0)$	1	1	1	1	1	1
$\sigma_n(1)$	4	6	8	10	12	14
$\sigma_n(2)$	8	18	32	50	72	98
$\sigma_n(3)$	12	38	88	170	292	462
$\sigma_n(4)$	16	66	192	450	912	1666
$\sigma_n(5)$	20	102	360	1002	2364	4942
$\sigma_n(6)$	24	146	608	1970	5336	12642
$\sigma_n(7)$	28	198	952	3530	10836	28814
$\sigma_n(8)$	32	258	1408	5890	20256	59906
$\sigma_n(9)$	36	326	1992	9290	35436	115598

	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$
$\beta_n(0)$	1	1	1	1	1	1
$\beta_n(1)$	5	7	9	11	13	15
$\beta_n(2)$	13	25	41	61	85	113
$\beta_n(3)$	25	63	129	231	377	575
$\beta_n(4)$	41	129	321	681	1289	2241
$\beta_n(5)$	61	231	681	1683	3653	7183
$\beta_n(6)$	85	377	1289	3653	8989	19825
$\beta_n(7)$	113	575	2241	7183	19825	48639
$\beta_n(8)$	145	833	3649	13073	40081	108545
$\beta_n(9)$	181	1159	5641	22363	75517	224143

A plot of $\frac{\sigma_4(k)}{\sigma_4(k-1)}$ versus k is shown in figure 19.1.

Plots of $\frac{\sigma_4(k)}{\sigma_4(k-1)}$ versus $\beta(k)$ are also available in Figure 19.2 and Figure 19.3. Figure 19.2 shows the large scale behavior, while Figure 19.3 rather focuses on the details of the area where the number of infected nodes is not too large.

It is claimed that, $\forall n$, $\lim_{k\to\infty} \frac{\sigma_n(k)}{\sigma_n(k-1)} = 1$. To show this, we proceed to evaluate the spherical growth asymptotically for large k's. Clearly, for large k's $\beta(k)$ is the volume of the rhombus R_k with vertices given by 19.1. Clearly,

$$\mathrm{vol}(R_k) = \mathrm{vol}(R_1)k^n$$

so that

$$\beta(k) = \mathrm{vol}(R_1)k^n$$

Clearly, the growth is polynomial in k. As far as worm propagation is concerned, we get

$$\frac{\sigma(k)}{\sigma(k-1)} = \frac{k^n - (k-1)^n}{(k-1)^n - (k-2)^n} \overset{k\to\infty}{\longrightarrow} 1$$

as claimed.

Clearly, the propagation in the case of an Abelian group is drastically reduced compared with the case of a free group.

19.3 cancellation groups

19.3.1 one versus two cancellation

So far we have been looking at two extremes cases–free groups (no relations) and Abelian groups ($\begin{pmatrix} n \\ 2 \end{pmatrix}$) relations. We now look at those intermediate cases characterized by one single relation. Even in the case of one single relation, there are plenty of groups structures, depending mainly on the amount of cancellation involved between the relation and its cyclic conjugates. Here we look at small cancellation groups [59]; more specifically, one and two cancellation groups.

The case studies involve groups with $n = 2$ generators and a variety of small cancellation relations as shown in Figure 19.4. The overall shape of the

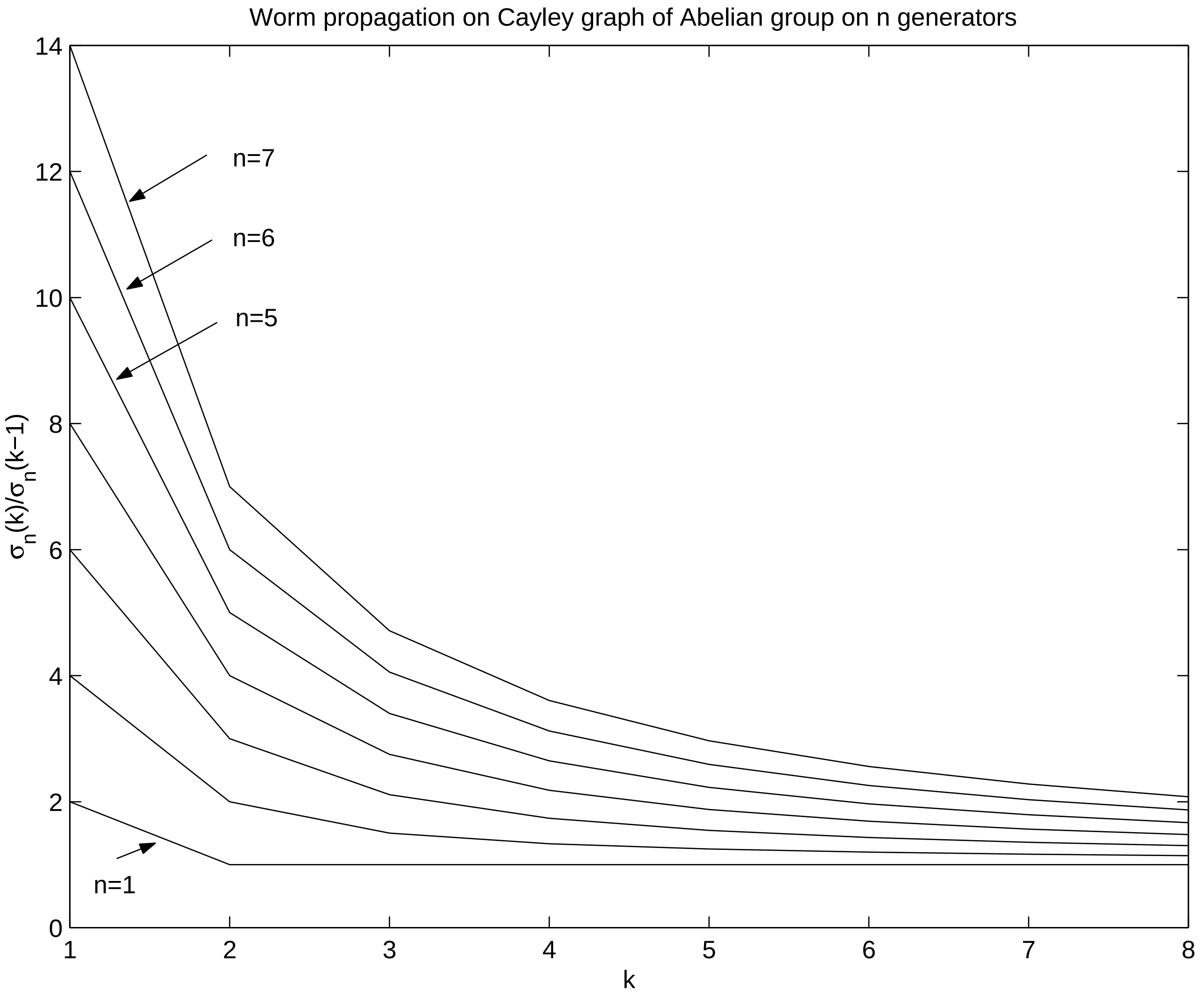


Figure 19.1: The number of uninfected nodes versus the time (or distance) k.

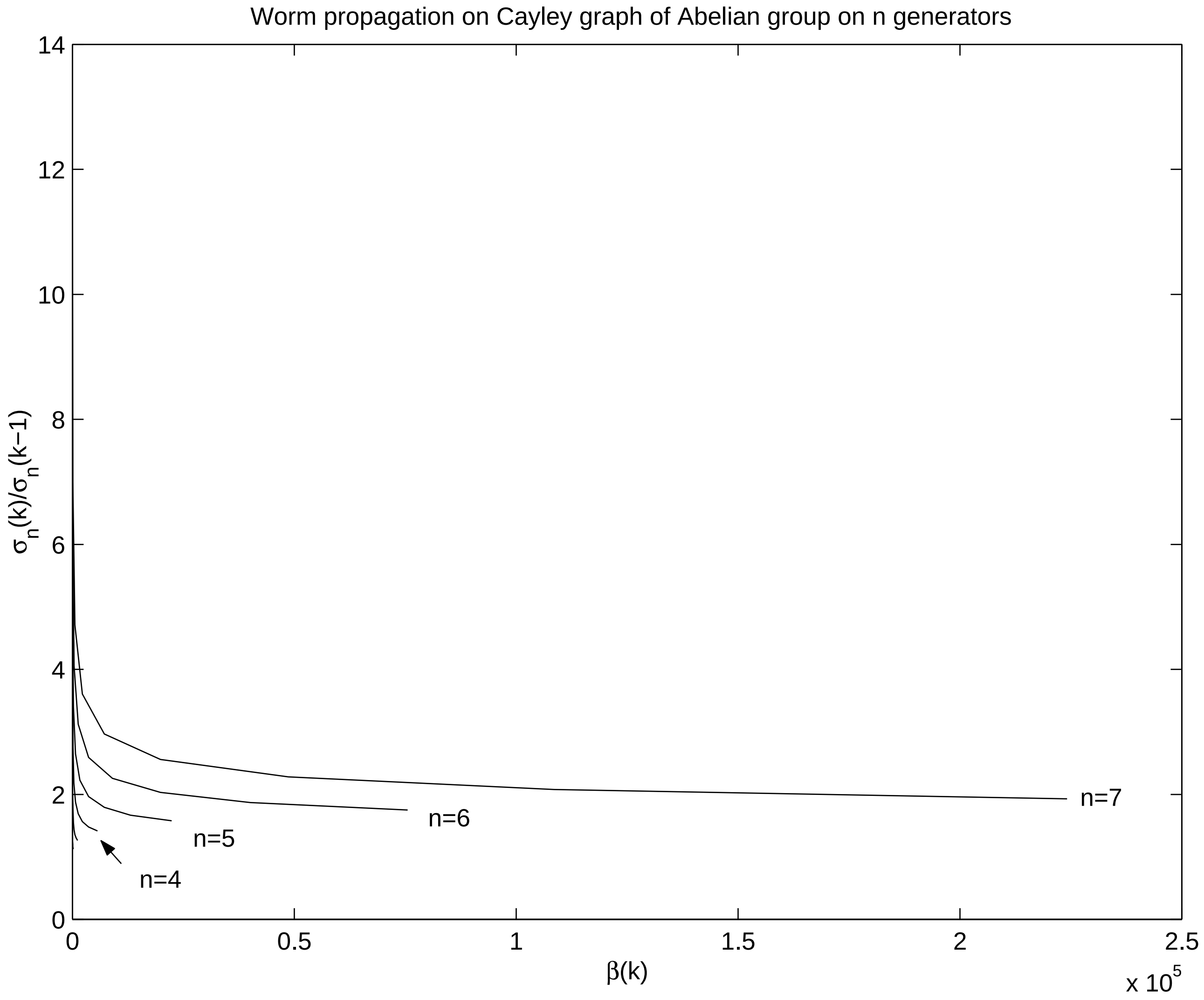


Figure 19.2: Large scale behavior of the number of uninfected nodes versus the total number of infected nodes.

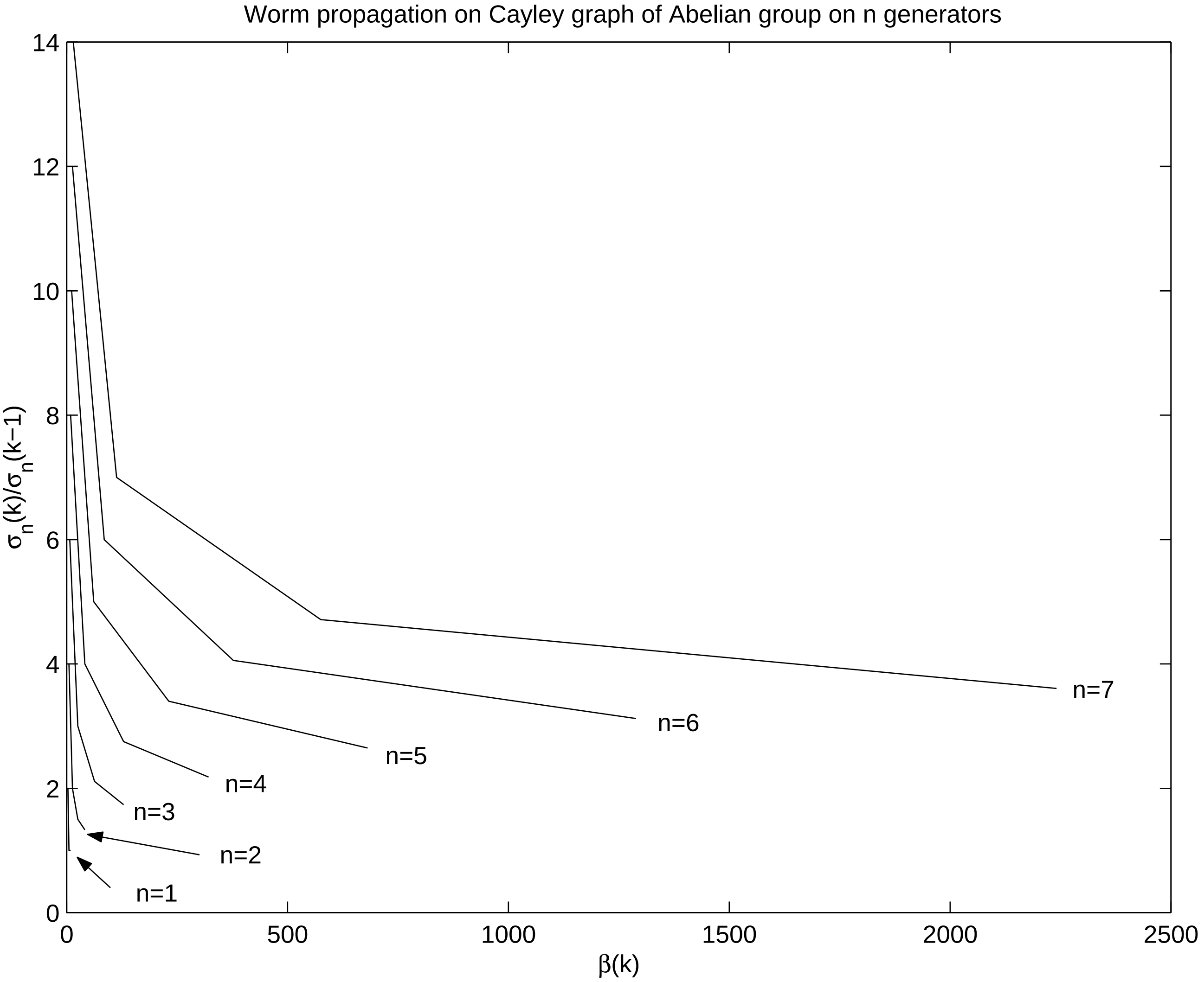


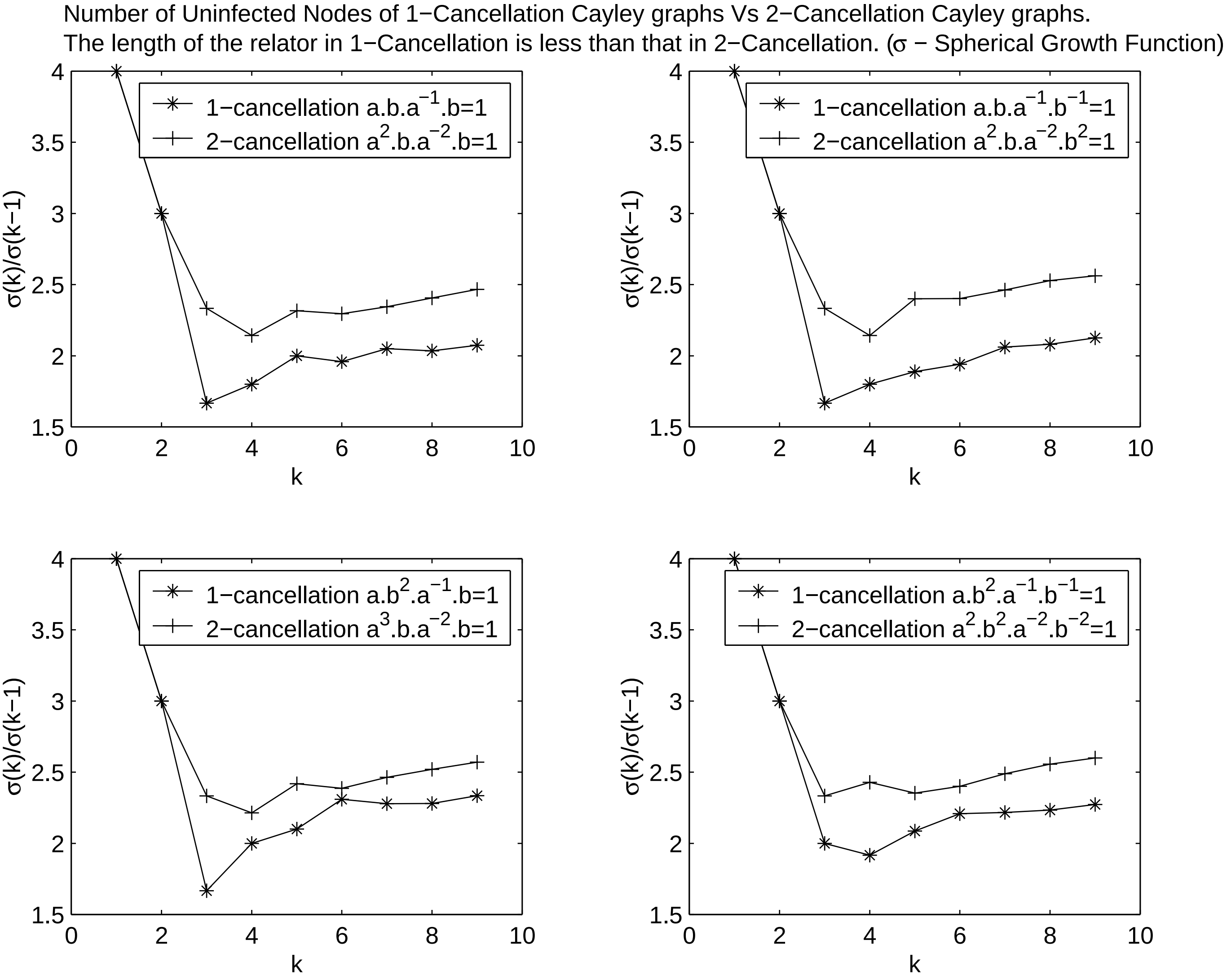
Figure 19.3: Details of the behavior of the number of uninfected nodes versus the total number of infected nodes near the origin.

curve is still the one established empirically by simulation, that is, first a fast propagation of the worm, followed by a time of real slow propagation, and then a possibly small flare-up in the propagation and eventually a leveling off.

To be slightly more specific, for the very first steps of the propagation, that is, for $k < [\ell(R)/2]$ for $ell(R)$ even and $k \leq [\ell(R)/2]$ for $\ell(r)$ odd, the relator R has not yet manifested itself, so that the propagation $\sigma(k)/\sigma(k-1)$ is exactly the same as that of the free group F_n on the generators. Thereafter, asymptotically as k gets larger, as shown in Figures 19.4,19.8,19.9, the speed of propagation appears to be dictated more by the length of the relator than by the cancellation. In case of equal length relators, however, Figures 19.6,19.7 indicate that the propagation is slightly faster for the 1-cancellation case.

19.3.2 two versus three cancellation

The situation is pretty much the same as in the previous case. Early on, as long as the relator has not yet kicked in, the propagation follows the free propagation rule of Section 19.2.1. Asymptotically as the propagation speed levels off, Figures 19.10,19.11 indicate that the propagation depends essentially on the length of the relator. On the other hand, Figures 19.12 indicates that with the same relator length, the 3-cancellation case propagates faster that the 2-cancellation case.



Number of Uninfected Nodes of 1−Cancellation Cayley graphs Vs 2−Cancellation Cayley graphs.
The length of the relator in 1−Cancellation is less than that in 2−Cancellation. (σ − Spherical Growth Function)

Figure 19.4: Number of uninfected nodes versus time of one versus two cancellation groups. OneVsTwo1.

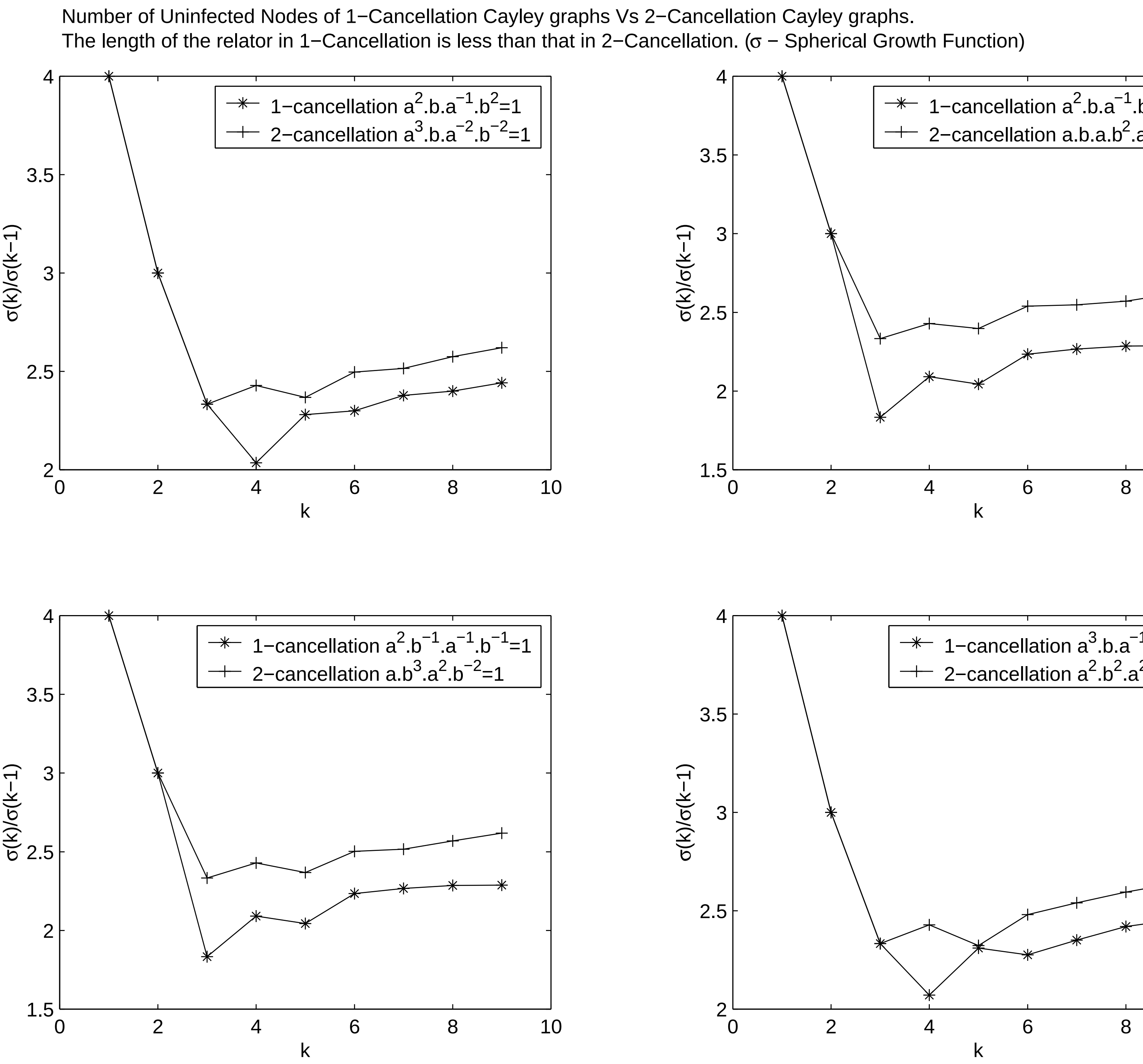


Figure 19.5: Number of uninfected nodes versus time of one versus two cancellation groups. OneVsTwo2.

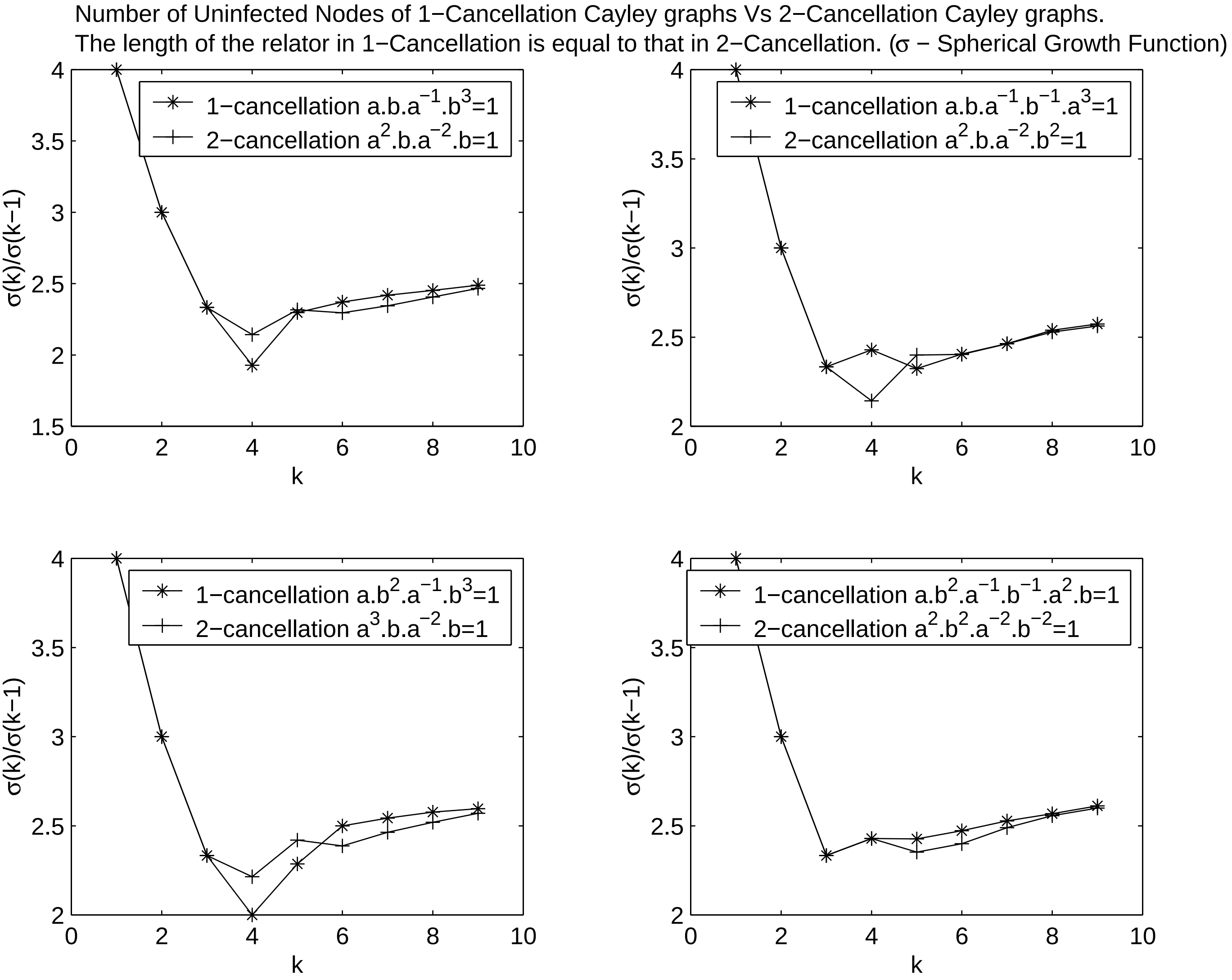


Figure 19.6: Number of uninfected nodes versus time of one versus two cancellation groups. OneVsTwo3.

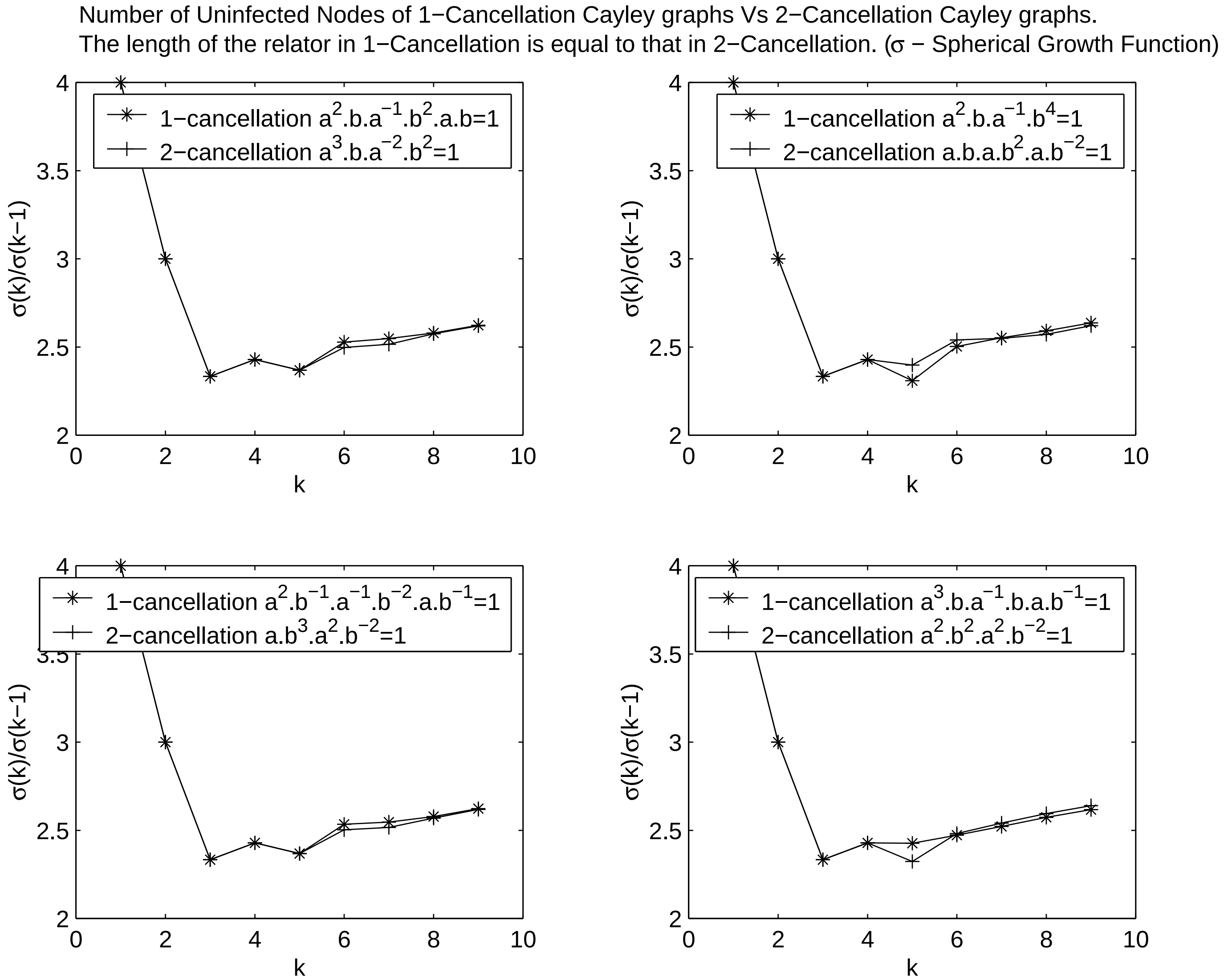
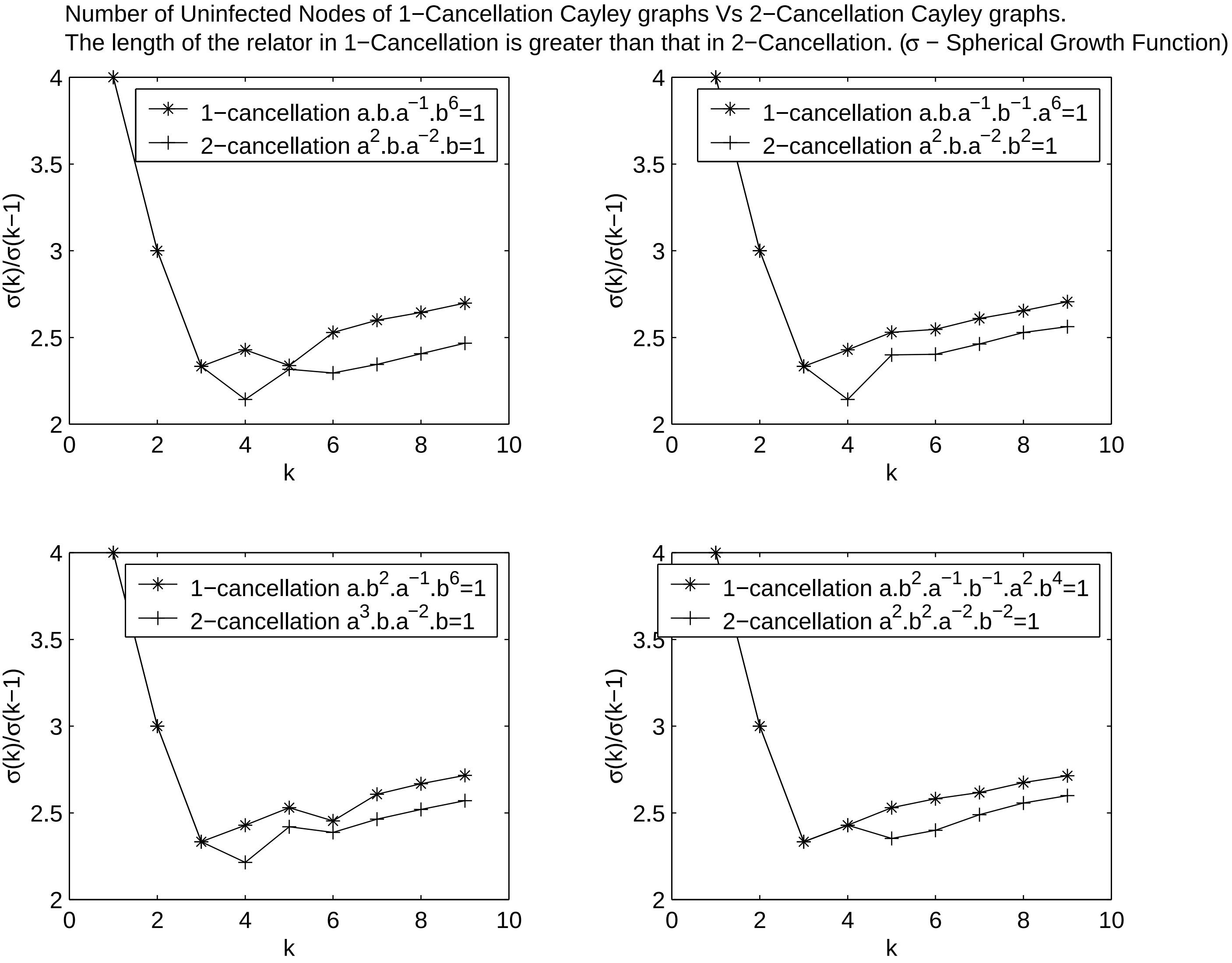


Figure 19.7: Number of uninfected nodes versus time of one versus two cancellation groups. OneVsTwo4.



Number of Uninfected Nodes of 1−Cancellation Cayley graphs Vs 2−Cancellation Cayley graphs.
The length of the relator in 1−Cancellation is greater than that in 2−Cancellation. (σ − Spherical Growth Function)

Figure 19.8: Number of uninfected nodes versus time of one versus two cancellation groups. OneVsTwo5.

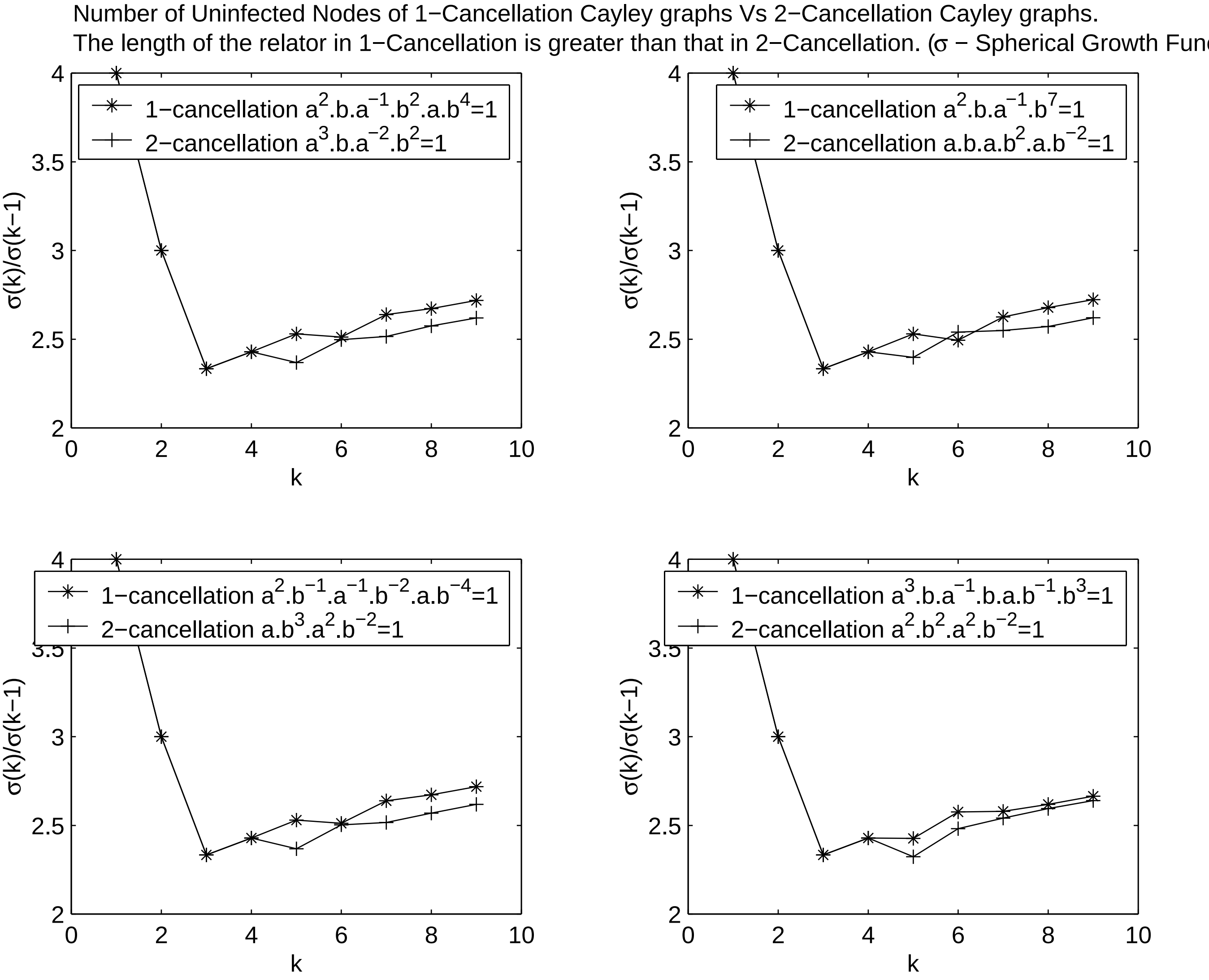


Figure 19.9: Number of uninfected nodes versus time of one versus two cancellation groups. OneVsTwo6.

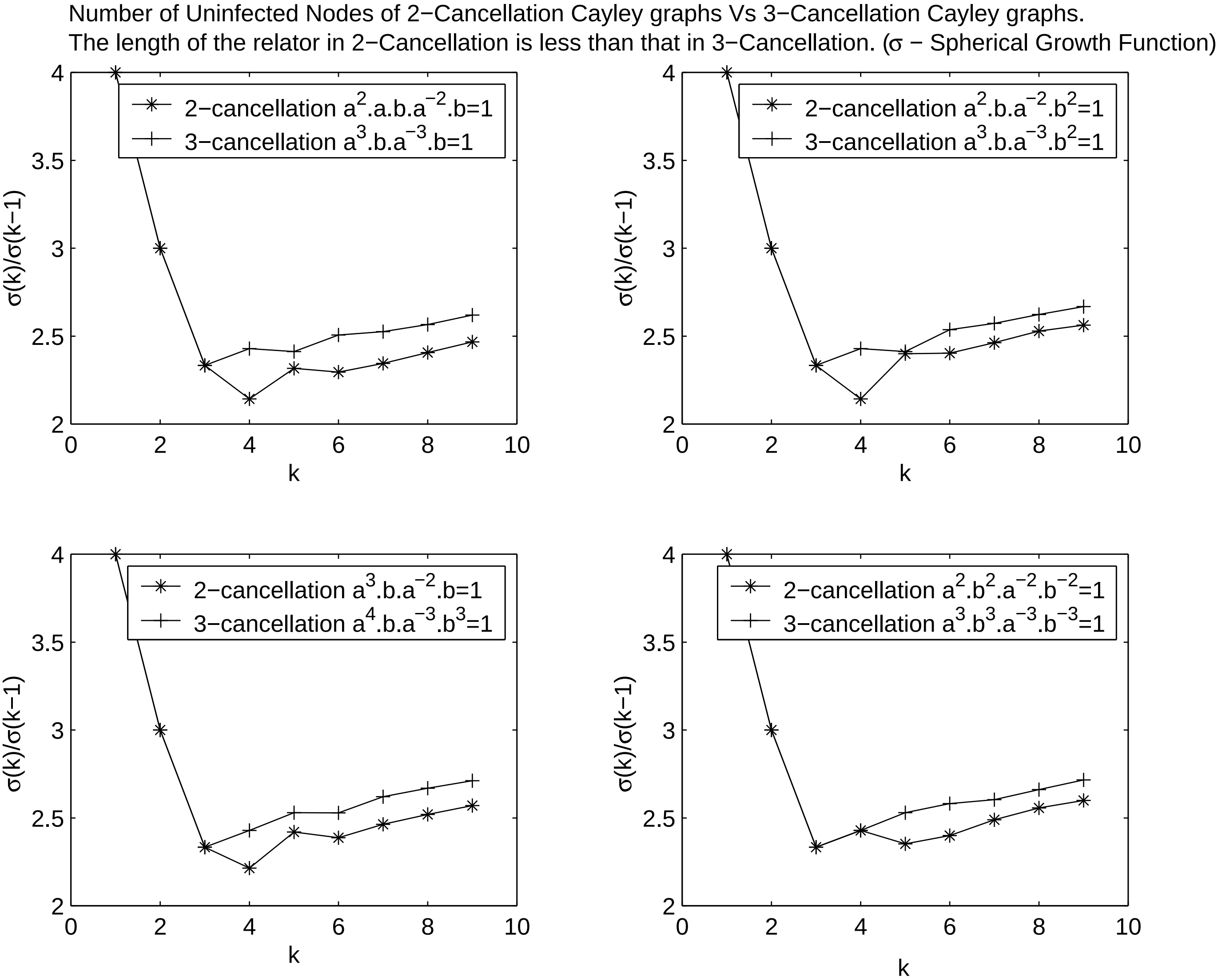


Figure 19.10: Number of uninfected nodes versus time of two versus three cancellation groups. TwoVsThree1.

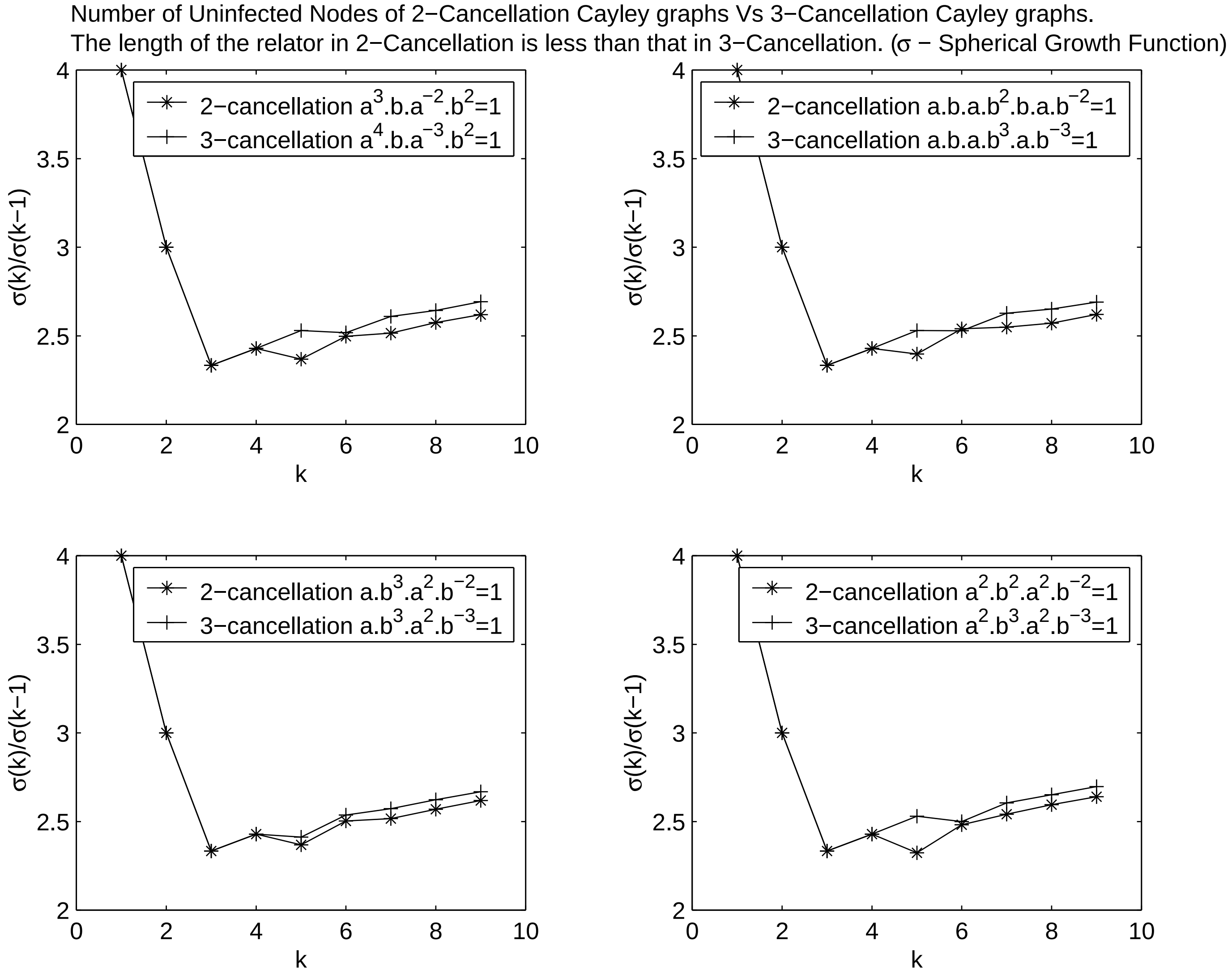


Figure 19.11: Number of uninfected nodes versus time of two versus three cancellation groups. TwoVsThree2.

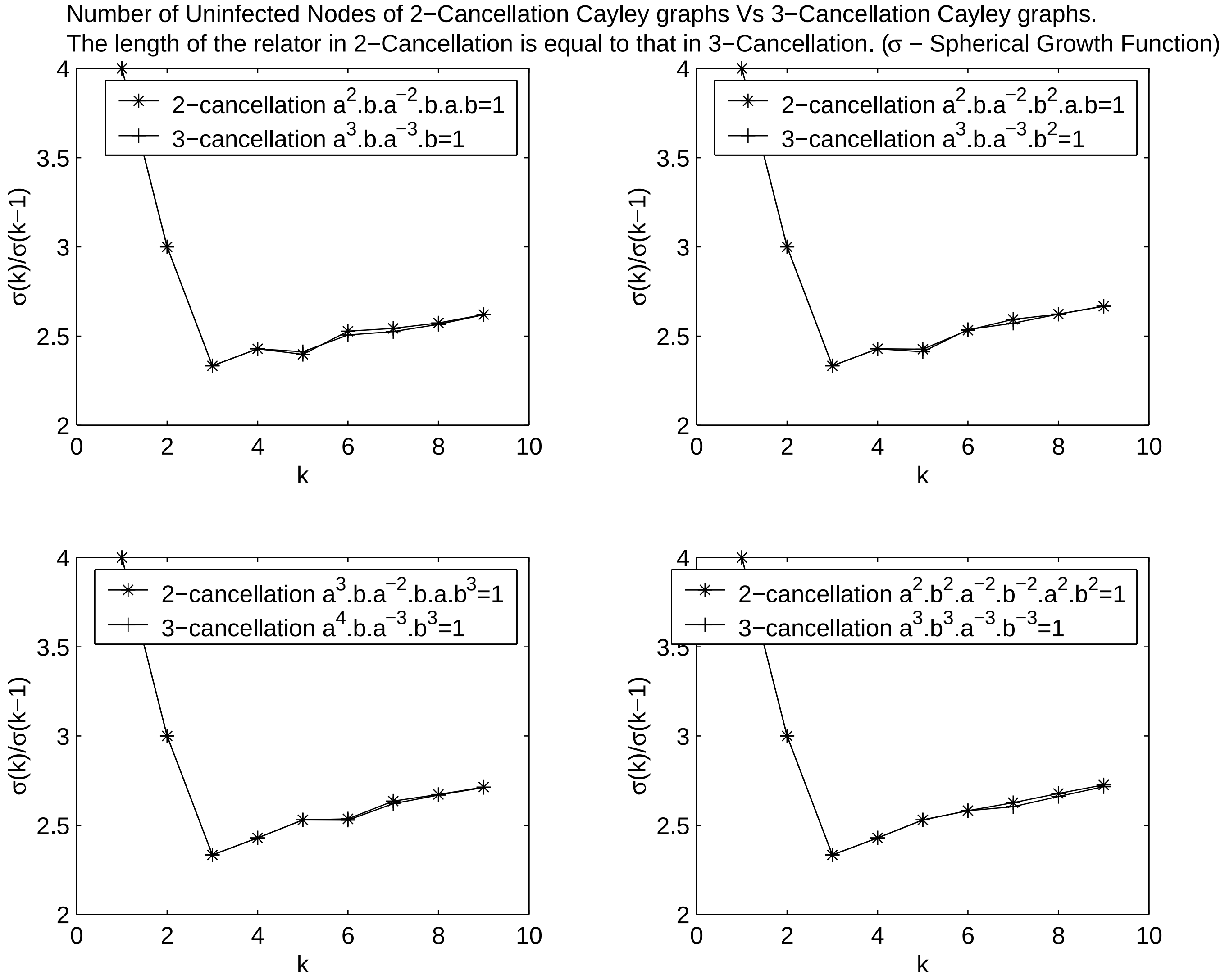


Figure 19.12: Number of uninfected nodes versus time of two versus three cancellation groups. TwoVsThree3.

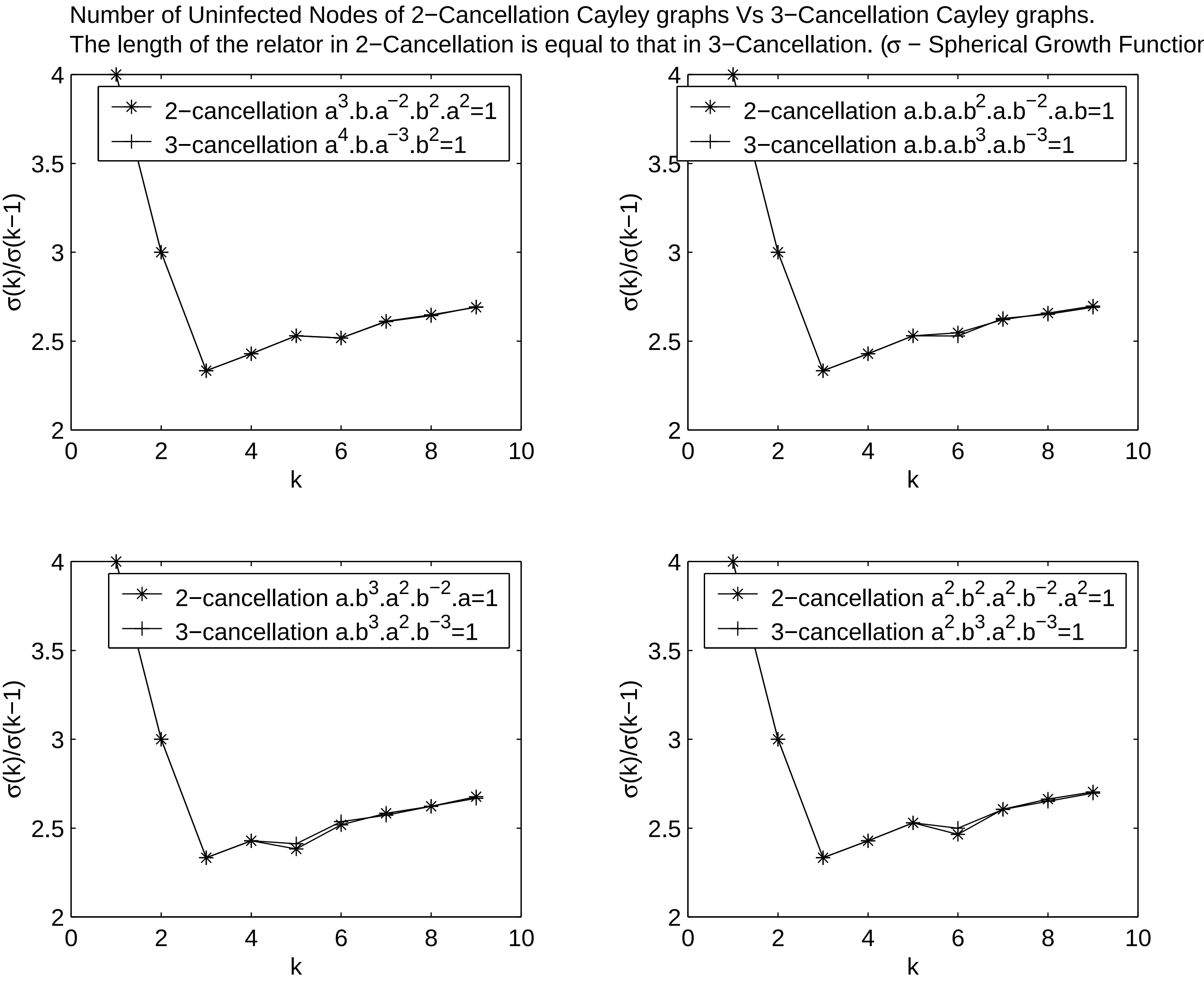


Figure 19.13: Number of uninfected nodes versus time of two versus three cancellation groups. TwoVsThree4.

Chapter 20

Worm defense

Such a mailer as Microsoft Outlook might be viewed as a feedback from the POP3 to the SMTP servers, as shown in Fig. 20.1. This feedback is highly uncertain, as it is meant to model how a user reacts to an incoming e-mail message (e.g., Is the user likely to respond?) Clearly, a mail worm would pass straight through the mailer as it would send a new infected message to a destination it has found in the address book. A simple mail worm defense strategy would consist in shutting down the mailer (opening up the loop) whenever it receives more than twice the same e-mail message from the same "suspicious" sender within a small time interval. The issue is whether this would slow down the propagation of the worm; more precisely, whether this would change *qualitatively* the propagation speed. As said, propagation depends strongly on graph topology, in particular on the curvature, as defined by the δ-hyperbolic property. The key issue is then to compare the curvature properties of the network, along with the mailers, with and without the defense strategy implemented. Assume that the network is infinite and δ-hyperbolic. Assume that the number of mailers is finite. It turns out that the two graphs (with and without defense strategy) are quasi-isometric and hence if one is δ-hyperbolic the other is δ'-hyperbolic [21, Th. III.H.1.9], with an identifiable relation between δ and δ'. Consequently, regardless as to whether or not this simple defense strategy is implemented, the propagation would remain, qualitatively, the same, i.e., it would remain the propagation on a δ-hyperbolic graph while, quantitatively, the propagation would go from that on a δ-hyperbolic graph to that on a δ'-hyperbolic graph. *This is an illustration on the fundamental limitation on worm propagation speed that a simple defense strategy could possibly accomplish.*

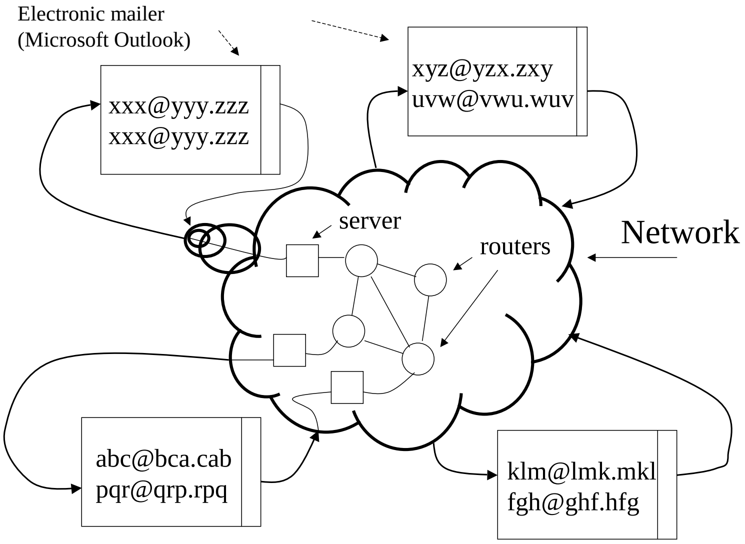


Figure 20.1: Mailers viewed as feedback on the backbone network.

Appendix B

Coarse Flow Control by C^*-Algebra Dynamics

Next, the development makes an algebraic detour with some concept of noncommutative geometry in which a space is given a coarse representation by a noncommutative C^*-algebra. The connection with the Gromov hyperbolic concept resides in the fact that only for those space can the preceding statement be made precise.

The traditional time-window averaging of the number of packets flowing along a link–say, the bottleneck link of the dumbbell topology–is already a coarsening in disguise of the dynamics of the traffic relative to the time parameter. Here, we are rather referring to the coarsening relative to the space parameter. In other words, our objective is to formalize the concept of a coarse geodesic traffic flow in the sense that it evolutes over an imprecisely defined space. In Section A.4, we discussed how to coarsen the geometrical object that supports the flow using C^*-algebra techniques. The problem now is to coarsen the flow, in a way consistent with the coarsening of the space. Since the space was coarsened using C^*-algebra techniques, it appears most natural to use the same techniques to coarsen the flow. Such a possibility is offered by the theory of C^*-dynamical systems, in which a dynamical system is viewed as a shift on an algebra.

B.1 Topological Dynamics

Here, we review traditional topological dynamics and, as a warm up exercise, show how it historically led to C^*-algebra dynamics [35, p. 133].

A topological dynamical system is a triple (G, α, X), where G is a locally compact Hausdorff topological group, X is a locally compact Hausdorff space, and α is an action of G on X, that is, a map $\alpha : G \times X \to X$, $(g, x) \mapsto \alpha_g(x) = gx$. In a continuous-time evolution, the dynamical shift g_t is defined $\forall t \in \mathbb{R}$ and satisfies the composition law $g_{t_2} \circ g_{t_1} = g_{t_1+t_2}$, so that the relevant group of shifts $G = (\{g_t : t \in \mathbb{R}\}, \circ)$ is clearly isomorphic to $G = (\mathbb{R}, +)$. The action will be more specifically written $\alpha_g(x) = gx$. Likewise, for a discrete time-evolution, the group of shifts $(\{g_k : k \in \mathbb{Z}\}, \circ)$ is isomorphic to $(\mathbb{Z}, +)$.

A C^*-dynamical system [35, Sec. 3.8,8.1] is a triple (G, α, A) where G a topological group, A is a C^*-algebra, and α an action of G on A, that is, a homomorphism $\alpha : G \to \mathrm{Aut}(A)$ such that $g \to \alpha_g(a)$ is continuous for all $a \in A$.

B.1.1 Commutative Case

Given a topological dynamical system (G, α, X), we associate with it a commutative C^*-algebra dynamics by setting $A = C^0(X)$ and by defining the action of G on $C^0(X)$ to be $\alpha_g(f)(x) = f(g^{-1}x)$.

Clearly, topological dynamics is a spinoff of Halmos' shift operator $f(x) \mapsto f(g_t x)$ of ergodic theory, except for a time reversal (see [78],[88],[73] for connections between traditional ergodic theory and the modern algebraic point of view.)

Probably the best way to describe the situation of classical versus topological dynamics is by using the categorical language:

$$\begin{tikzcd}
\cdots & X \arrow[l] \arrow[d] & X \arrow[l, "g_{t_1}^{-1}=g_{-t_1}"'] \arrow[d] & X \arrow[l, "g_{t_2}^{-1}=g_{-t_2}"'] \arrow[d] & \cdots \arrow[l] \\
\cdots \arrow[r] & C^0(X) \arrow[r, "g^*_{-t_1}=\alpha_{t_1}"] & C^0(X) \arrow[r, "g^*_{-t_2}=\alpha_{t_2}"] & C^0(X) \arrow[r] & \cdots
\end{tikzcd} \tag{B.1}$$

The top row is the category of traditional dynamical systems running over X, the "downarrow" is the contravariant functor of "going to the space of continuous functions defined on X," leading to the bottom row depicting the commutative C^*-dynamical systems running over X.

Consider the usual time-invariant system

$$\frac{dx}{dt} = Ax$$

with state transition matrix Φ_t. To reformulate the above system, well known to control theoreticians, in the abstract language of topological dynamics, here, the locally compact Hausdorff space is taken as $X = \mathbb{R}^n$, the topological group of dynamical shifts is taken as $G = (\mathbb{R}, +)$, the group of *continuous* dynamical shifts, and the action is in fact given by the state transition, viz., $g_t = \Phi_t$. With this topological dynamical system, we associate a commutative C^*-algebra dynamical system $(C(X), \alpha, G)$, where the action is defined by

$$\alpha_t(x) = f(g_t^{-1}x) = f(\Phi_t^{-1}x), \quad f \in C(X)$$

It is convenient to define this action via a partial differential equation. To this end, define

$$\varphi(x, t) = f(\Phi_t^{-1}x)$$

Then we get

$$\begin{aligned}\frac{\partial\varphi}{\partial t} &= -\left(\begin{array}{ccc}\frac{\partial f}{\partial x_1} & \dots & \frac{\partial f}{\partial x_n}\end{array}\right)\Phi_t^{-1}Ax \\ &= -\left(\begin{array}{ccc}\frac{\partial \varphi}{\partial x_1} & \dots & \frac{\partial \varphi}{\partial x_n}\end{array}\right)Ax\end{aligned}$$

Therefore, the PDE formulation of the commutative C^*-algebra dynamics is

$$\frac{\partial\varphi}{\partial t} = -\left(\begin{array}{ccc}\frac{\partial \varphi}{\partial x_1} & \dots & \frac{\partial \varphi}{\partial x_n}\end{array}\right)Ax$$

Next, consider the usual linear time-invariant (LTI) control system,

$$\dot{x}(t) = Ax(t) + Bu(t)$$

This a particular case of a Linear Dynamically Varying (LDV) control as defined in Chapter **??**. We develop the commutative C^*-algebra version of LTI control. In this case, the action of G on X, g_t, is the state transition matrix Φ_t defined by $x(t) = \Phi_t x(0)$ and f can be interpreted as the "probability density" of the state. As in the free case, we develop a PDE formulation of the action

$$\alpha_{t,u}(f)(x) = f\left(\Phi_t^{-1}x - \int_0^t \Phi_{-\tau}Bu(\tau)d\tau\right)$$

To this end, define

$$\varphi(x,t) = \alpha_{t,u}(f)(x)$$

We have

$$\begin{aligned}\frac{\partial\varphi}{\partial t} &= \left(\begin{array}{ccc}\frac{\partial f}{\partial x_1} & \dots & \frac{\partial f}{\partial x_n}\end{array}\right)\left(-\Phi_t^{-1}Ax - \Phi_t^{-1}Bu\right) \\ &= \left(\begin{array}{ccc}\frac{\partial f}{\partial x_1} & \dots & \frac{\partial f}{\partial x_n}\end{array}\right)\Phi_t^{-1}\left(-Ax - Bu\right) \\ &= \left(\begin{array}{ccc}\frac{\partial \varphi}{\partial x_1} & \dots & \frac{\partial \varphi}{\partial x_n}\end{array}\right)\left(-Ax - Bu\right)\end{aligned}$$

Hence the commutative C^*-algebra system associated with the LTI system is

$$\frac{\partial\varphi}{\partial t} = -\left(\begin{array}{ccc}\frac{\partial \varphi}{\partial x_1} & \dots & \frac{\partial \varphi}{\partial x_n}\end{array}\right)(Ax + Bu) \tag{B.2}$$

Now, we can develop a commutative C^*-algebra version of LTI control. Whereas in LTI theory the system is stabilized around some point $x \in X$, in C^*-algebra dynamics the system is stabilized around some desired $\varphi_d \in C(X)$, which to simplify is taken to be an unperturbed motion, that is, it satisfies Equation (B.2) for $u = 0$ Let φ_a be the actual controlled coarse state. Let $\varphi = \varphi_a - \varphi_d$ be the "error," which is easily seen to satisfy the same equation as (B.2). The feedback would be $u = K_x\varphi$, where K_x is in general a partial differential operator. Should it have constant coefficients only, then $K\varphi$ might be called *coarse feedback*, because it only requires a coarse description of the state. An explicit dependency on x is a *pointwise component* of the feedback in the sense that it requires a *precise knowledge* of the state. Plugging the general feedback in (B.2) yields

$$\frac{\partial \varphi}{\partial t} = -\begin{pmatrix} \frac{\partial \varphi}{\partial x_1} & \dots & \frac{\partial \varphi}{\partial x_n} \end{pmatrix} (Ax + BK_x\varphi) \tag{B.3}$$

It is immediately seen that the chief difference between classical LTI control and the above is that, even though the control is linear and the dynamics in its traditional interpretation is linear, we end up with a *nonlinear* partial differential equation!

Here is an illustration as how to proceed. Consider the simple 1-dimensional case subject to the following control:

$$u = \frac{1}{B}\left(-Ax + \varphi - k\frac{\varphi_{xx}}{\varphi_x}\right), \quad k > 0$$

The motivation for this control is that the closed-loop system (B.3) becomes the celebrated (viscous) Burgers equation (see [3, Chap. VI, Sec. 1], [87, p. 72], [86]), that is,

$$\frac{\partial \varphi}{\partial t} + \varphi\frac{\partial \varphi}{\partial x} = k\frac{\partial^2 \varphi}{\partial x^2}$$

Remarkably, this equation is integrable via the Cole-Hopf transformation:

$$\varphi = -2k\frac{\Psi_x}{\Psi}$$

Injecting this transformation in the equation for φ yields

$$\left(\frac{\Psi_t}{\Psi}\right)_x = k\left(\frac{\Psi_{xx}}{\Psi}\right)_x$$

Integrating yields

$$\Psi_t = k\Psi_{xx} + C(t)\Psi$$

where $C(t)$ is a function of the time, independent of x. Next, consider the transformation

$$\Psi = Te^{\int_0^t C(\tau)d\tau}$$

Injecting the above in the equation for Ψ, it follows that T is a solution of the heat equation,

$$\frac{\partial T}{\partial t} = k\frac{\partial^2 T}{\partial x^2}$$

It follows that the error dynamics is given by

$$\varphi = -2k\frac{T_x}{T}$$

Since the solution to the heat equation yields, asymptotically as $t \to \infty$, a constant temperature $T > 0$ distribution over a compact medium, it follows that $\varphi \to 0$ as $t \to \infty$, as claimed.

B.1.2 Noncommutative Case

The prototype of a noncommutative C^*-dynamical system builds on the algebra A of bounded integral operators $L^2(X,\mu) \to L^2(X,\mu)$ defined as

$$(Kf)(x) = \int_X k(x,y)f(y)d\mu(y)$$

where μ is a measure defined on X. The multiplicative law of the algebra A is the composition of integral operators. Specifically, the composition of the operators K, L is defined by

$$(KL)(x,y) = \int_X k(x,w)l(w,y)d\mu(w)$$

The $*$-involution is defined as

$$(K^*f)(x) = \int_X k^*(y,x)f(y)d\mu(y)$$

The norm on A is the usual operator induced norm, $||K|| := \sup \frac{||Kf||}{||f||}$, from which it is easily seen that the C^*-identity $||KK^*|| = ||K||^2$ is satisfied. The topological group of shifts is the one of continuous time dynamics. The action is given by

$$\alpha_t(K)(x,y) = K(g_{-t}x, g_t y)$$

where g_t is the action, the shift, of the underlying topological dynamics.

It is possible to define functorially noncommutative C^*-dynamical systems for the algebra of operators $L^2(X,\mu) \to L^2(X,\mu)$. Specifically, passing from the category of classical dynamical systems to the category of noncommutative C^*-dynamical systems is bifunctorial, contravariant in the first component and covariant in the second component. This bifunctoriality is obvious from the definition of the action.

Next, consider a topological dynamics defined by the LTI control system

$$\dot{x} = Ax + Bu$$

The noncommutative dynamics induced by the above can be formulated in terms of a PDE. To this effect, define

$$\kappa(x,y,t) = K(g_{-t}x, g_t y)$$

Then it is easily verified that the PDE dynamics is given by

$$\frac{\partial \kappa}{\partial t} = -(x'A' + u'B') \begin{pmatrix} \frac{\partial \kappa}{\partial x_1} \\ \vdots \\ \frac{\partial \kappa}{\partial x_n} \end{pmatrix} + \begin{pmatrix} \frac{\partial \kappa}{\partial y_1} & \cdots & \frac{\partial \kappa}{\partial y_n} \end{pmatrix} (Ay + Bu)$$

We leave it up to the reader the plug the control terms in the above and verify that stabilizability is achieved via a Burgers-like equation.

Clearly, both commutative and noncommutative C^*-algebra formulations of the LDV theory sketched in Chapter **??** are in sight.

B.2 C^*-Algebra Control

In the previous sections, we proceeded from a traditional dynamical system running over a traditional topological space and showed how to construct from this data a C^*-algebra dynamics that is consistent, up to a given extend, with the topological dynamics. Here, we assume that the the space over which the dynamics evolutes is defined, possibly coarsely, by a, possibly noncommutative, C^*-algebra and we attempt to understand what "dynamics" can be defined on this, possibly coarse, structure.

One way to proceed would be to derive a concrete representation of the algebra in terms of "continuous functions defined on some topological space" or some "integral operators defined on some Hilbert space," from which the dynamics could be defined as in the preceding sections. Obtaining concrete representations of a given C^*-algebra is a traditional problem referred to as "Gelfand representation."

B.2.1 Commutative case

The Gelfand representation of A provides some (at least local) coordinates $(x_1, ..., x_n)$ for the space X.

B.2.2 Noncommutative case

The simplest way to provide the "kernel dynamics" of Section B.1.2 with an interpretation that does not proceed from a topological dynamics is to invoke the second Gelfand-Naimark representation theorem.

Insofar as bounded linear operators on a Hilbert space can be represented by integral operators, the C^*-algebra dynamics over a space defined by a noncommutative C^*-algebra would take the same form as that of Section B.1.2.

B.3 Functorial Coarsening

Besides C^*-algebra techniques, there is another approach to coarsening dynamical systems, based on the fact that coarsening is functorial (see [76, p. 14]). By applying the coarsening (covariant) functor to the top row category of the diagram (B.1), one obtains a coarsening of the dynamical system. This new functor can in turn be made contravariant by going to the category whose objects are coarse functions from coarse spaces to (some coarsening of) $\mathbb{R}$. The connection between the two approaches and whether some natural equivalence can be established is open.

Part V

Appendices: wireline, wireless, quantum and power networks

Appendix A

Wireline Physical and Logical Graphs

A.1 physical graphs

The physical graph of a network consists of nodes that are devices transmitting, processing or receiving data and of links that are the (wired or wireless) communication media carrying the data from/to the nodes. The global physical graph can be thought of as being built by the piecemeal process of interconnecting local graphs in a hierarchical structure. This hierarchy spans all the way from Local Area Networks (LANs) to the Internet.

A Local Area Network (LAN) is a group of computers and associated devices that share a common communications line or wireless link within a small geographic area (an office, a home, or a building). The number of users served by a LAN could be anywhere from two to several thousands. Major LAN technologies include Ethernet, Token Ring, FDDI (Fiber Distributed Data Interface).

A Metropolitan Area Network (MAN) is a network that interconnects users with computer resources in a geographic area or region larger than that covered by a large LAN but smaller than the area covered by a Wide Area Network (WAN). Typically, this term refers to the interconnection of networks within a city into a single larger network, which may then also offer efficient connection to a wide area network. It also refers to the interconnection of several LANs by bridging them with backbone lines. The latter usage is also sometimes referred to as campus network.

A Wide Area Network (WAN) is a geographically dispersed telecommunications network, as opposed to a geographically confined network. A wide area network may be privately owned or rented, but the term usually connotes the inclusion of public (shared user) networks.

An internet is the interconnection of many networks (LANs and/or WANs). This interconnection process is referred to as internetworking. The constituting networks, as long as they keep their own identity, are referred to as subnetworks.

The largest internet is called Internet.

A.1.1 nodes

A node could be either an end system (workstation or server) or an intermediate devise (router, bridge, brouter).

A bridge is a device that connects one LAN to another LAN using the same protocol. A bridge usually connects a LAN to exactly one neighboring LAN. A bridge examines each message on a LAN, “passing” those known to be within the same LAN, and forwarding those known to be on the other interconnected LAN (or LANs). In bridging networks, computer or node addresses have no specific relationship to location. For this reason, messages are sent out to every address on the network and accepted only by the intended destination node. Bridges learn which addresses are on which network and develop a learning table so that subsequent messages can be forwarded to the right network.

A router is a device that connect a network to other networks that usually make a WAN. In packet-switched networks such as the Internet, a router is a device that determines the next network point to which a packet should be forwarded toward its destination. The router is connected to at least two networks and decides which way to send each information packet based on its current understanding of the state of the networks it is connected to.

A brouter is a router and a bridge combined together.

A.1.2 links

A link could be wired or wireless. Communication along the links can be either simplex, half-duplex or full duplex.

Simplex communication is permanent unidirectional communication. Some of the very first serial connections between computers were simplex connections. For example, mainframes sent data to a printer and never checked to see if the printer was available or if the document printed properly since this was left to a human operator. Simplex links are built so that the transmitter sends a signal and it is up to the receiving device to figure out what was sent and to correctly do what it was told. No traffic is possible in the other direction across the same connection; hence no acknowledgment or return traffic is possible over a simplex circuit. Satellite communication provides a more modern example of simplex communication. A radio signal is transmitted and it is up to the receiver to correctly determine what message has been sent, whether it arrived intact, and if not how to fix it. On a broader scale, simplex communication is the standard in broadcast media such as radio, television and public announcement systems, since these systems do not (as yet) talk back to the broadcasting stations.

A half duplex link can communicate in only one direction at a time. Two way communication is possible, but not simultaneously. Walkie-talkies and CB radios are early examples of this kind of communication, in the sense that a person transmitting cannot hear the other person if (s)he is transmitting at the same time. Communication between aircraft and Air Traffic Control (ATC) is also half duplex, in the sense that either the pilot or the air traffic controller must depress the “push to talk” button to transmit the carrier and as such blocks the frequency for him (her) to transmit. The reason why a pilot and an air traffic controller cannot transmit at the same time is that the local carrier oscillators in the aircraft and the ATC center are not synchronized, so that under simultaneous push to talk the two carriers destroy each other.

Full duplex (or bidirectional) communication is two-way communication achieved over a link that has the ability to communicate in both directions simultaneously. An easy way to create a full duplex circuit is to use two separate physical connections each running in half duplex mode or simplex mode. With most electrical, fiber optic, two-way radio and satellite links, full duplex is usually achieved with more than one physical connection. Two way satellite communication is achieved using two simplex connections. However, the real challenge is to achieve full duplex communication across a single physical link (e.g., a pair of electric wires). For example, in telephone networks, where the two way conversation is carried by a single pair of wires, the full duplex operation is achieved with a directional coupler, a device that carries the microphone signal of one user to the speaker of the other user, while blocking the transmission of the microphone signal to the speaker of the same user. One advantage of this approach is that the full duplex link can theoretically provide twice the bandwidth of the half duplex mode.

A.1.3 packet switched versus circuit switched network

The Public Telephone Network (PTN) is circuit switched, in the sense that, before the voice exchange starts, a physical path of wired and/or wireless channels is constructed by means of switches and assigned to the conversation. Such a network is also called connection oriented. Data network are, however, packet switched, in the sense that the data stream is broken down in small entities, called packets, with their destination address incorporated in a header, which are sent one by one through the network where they have to find their way in some autonomous fashion to their destination. Such network are called connectionless in the sense that two packets from the same message may take different routes.

The motivation for the choice of packet switched architecture for data network is that the PTN, while able to provide excellent Quality of Service (QoS), does not make efficient utilization of the infrastructure. Indeed, a telephone conversation always has idle moments, during which time the circuit is still assigned to the conversation and hence becomes underutilized. The idea in a packet switched network is to make maximum utilization of the infrastructure by allowing several end-to-end users to access the same resource, say a link, in some sort of multiplexing fashion.

A.1.4 protocol stack

The hierarchical structure of the Internet clearly calls for some kind of hierarchical structure of the communication protocol. It indeed appears necessary to be able to communicate either within the confines of the same LAN or from a LAN across the Internet, without major redesign. However, the layered structure of the protocol does much more than just duplicate the hierarchy of the Internet. Each layer indeed performs, *independently of the other layers*, some specific functions necessary for successful end to end communication. In a way, the design of one module should be independent of the other modules. The typical PCP/IP protocol involves five layers, independent in the sense that the top layers need not be concerned with the bottom layers. These five layers are, from top to bottom,

- the application layer
- the transport layer
- the Internet layer
- the network access layer
- the physical layer

Let us start from the bottom:

The physical layer deals with the communication medium and its related hardware.

The network access layer is concerned with the exchange of data between an end user (typically a workstation) and the network it is directly attached to (typically a LAN). The bridges are typically operating at that layer.

The Internet layer deals with the routing of the data through the various networks of the Internet. The routers are typically operating at that layer. The protocol at that layer is referred to as Internet Protocol (IP).

The transport layer deals with fast and reliable communications. The protocol basically attempts to send data at the rate the fastest possible consistent with the bandwidth of the links and the congestion downstream, which are not accurately known to the sender. In doing so, the transmission protocol sometimes overflows queues, causing data to be lost. Therefore, the same protocol has to check whether the packets that have been sent have reached their destination and, if not, to retransmit them. It also checks whether the packets arrive in the right order. This protocol is called Transmission Control Protocol (TCP) and is essentially an end-to-end protocol, in the sense that the end hosts exchange transmission information without knowledge of what has happened along the route. Connectionless protocols such as TCP require half or full-duplex.

Finally, the application layer deals with specific application; for example, file transfer protocol (FTP), hyper text transfer protocol (http), simple mail transfer protocol (smtp), post office protocol (pop3), etc. Application data must not only be sent to the correct end-user as specified by its IP address, but they must more precisely be sent to the correct process running on the end user platform; this latter is specified by the port number.

A.2 logical graphs

In a logical graph, the nodes need not be end users or intermediate devices; they could be any abstract element; the links need not be physical communication channels; however, for two abstract nodes to be logically connected, there has to be some physical link, or a path of physical links, ensuring that the abstract elements can be somehow hardware connected. Examples are abundant.

1. Clustering. To simplify a complex graph, it is convenient to partition it into highly connective subsets of nodes in such a way that the connectivity within a subset is much higher than the connectivity across different subsets. This cluster computation can be performed using a modified version of the Ford-Fulkerson algorithm [27, 36]. A "reduced model" of the graph is then obtained by declaring every such highly connective subset to be a logical node. More formally, the reduced model is an overlaying, in which each node of the overlaying is a representative of a highly connective cluster and two connective clusters, say a^0, a^1, are connected by a logical link if there are enough physical paths joining the representatives of the connective clusters.

2. Autonomous System (AS) graph. The Autonomous System (AS) graph is the result of a particular kind of clustering. An Autonomous System is a connected subset of routers and subnetworks, sharing a common routing protocol, and managed by a identifiable organization. As such, an Autonomous System is the clustering of a subset of physical nodes. Two Autonomous Systems AS1, AS2 are linked iff there exists a physical path from one node of AS1 to node of AS2 that does not pass through any other Autonomous System.

3. The http graph. The nodes of the http graph are the web sites and two web sites are joined by a logical link if one has a hyperlink to the other. If there is a need to specify that site A has a hyperlink to site B, then A and B are linked by a directed edge, $A \to B$ (see [33]).

4. The Mail Graph. The nodes are the client mailboxes and two mailboxes are joined by a logical link if one of them has the other in its address book.

A.3 overlaying

Given a network graph P (where P stands for Physical layer), an overlay A (where A stands for Application layer) is a graph defined by a subset of physical nodes together with logical links. A logical link a^0a^1 joining two nodes a^0, a^1 of the overlay is a path of physical links and nodes joining them in the physical layer. (See Figure A.1.) Since there might be many physical paths corresponding to a given logical link, the latter is rather defined to be the collection of all physical paths joining them. It should be observed that the overlay is not, in general, a sub-graph of P . Applications of the concept include the following:

1. The Survivable Overlay. The physical graph is in most cases not survivable in the sense that, under a link failure or a node compromised by an attack, the graph might become disconnected and hence some clients nodes cannot get service and some servers cannot provide services. To determine what part of the network is survivable, it is convenient to construct an overlay in which every logical link is supported by enough physical links so as to survive some failure. To be more specific, if every logical link is supported by at least k paths of physical nodes and links, then the overlay would remain unchanged under $k-1$ physical link failures. Such an overlay is said to be k-(edge)connected.

2. ATM Networks. In a certain sense, ATM networks borrow some of the overlaying concepts. Recall that ATM is connection-oriented in the sense that, before transmitting data from a source node to a destination node, a path of physical nodes and physical links from the source to the destination is set up. This path is called a virtual channel connection (VCC). Next, at a higher layer, a bundle of VCC's connecting the same end points is called a virtual path connection (VPC) (see [85, Sec. 4.2]). Clearly, the concept of virtual path connection is unmistakably the same as that of a logical link in the overlaying layer.

It should be clear that the concept of overlaying provides, in addition to the already complex graph of physical routers and links, an incredibly rich array of related graphs.

A.4 Topological Aspects of Overlaying

The Internet graph P can be viewed as a 1-D simplicial complex. From that point of view, the overlay is a 1-D simplicial complex as well. It should already be stressed that the overlay is not, in general, a sub-complex of the complex of the physical graph. As we have already argued, it is convenient to interpolate the graph P with a manifold $M(P)$ so that the information flow can be viewed as a vector field on the manifold. Besides the mere convenience and a way to get around the curse of dimensionality, this "manifold" point of view allows for the concept of vorticity of the flow. The latter might be an indication of an attack initiated from many different points and narrowing down on some critical part of the network. Somehow, the vorticity of the information flow is a global, topological anomaly of the flow, as opposed to the local, statistical anomaly that can be detected by, say, a wavelet analysis of the number of packets at a point of the network.

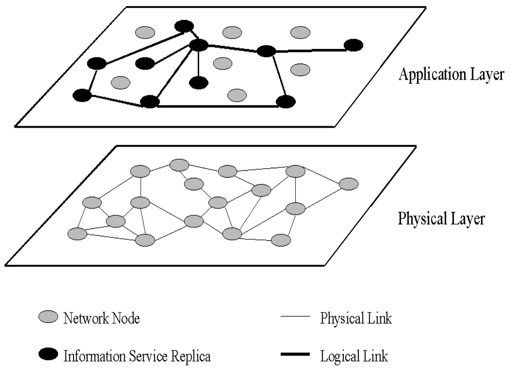


Figure A.1: Overlaying

It follows from the above discussion that there are two ways to go about the topology of an overlay–either the graph-theoretical or the geometrical approach. In the following sections, we develop both aspects. The section dealing with "Combinatorial Topological Aspects" is somehow the hinge between the graph-theoretic and the geometrical approaches.

A.4.1 Graph Theoretic Aspects of Overlaying

Given two nodes a^0, a^1 of the overlay graph connected by a logical link, the logical link is not supposed to see the difference between two acceptable physical edge paths joining a^0, a^1 . Two such physical edge paths determine a cycle, which, because of the indistinguishability of the two physical edge paths as seen from the logical network, should be set to zero. It appears therefore that the logical network graph is the physical network graph quotiented out by a subset of cycles.

To be more precise, let Z be the subset of cycles of the graph P . Since P is a 1-D simplicial complex, deformation of paths cannot happen, and therefore the subset of cycles, endowed with a group structure by the composition of cycles, is the same as the first or fundamental group of the graph $\pi_1(P)$. Let $Z(A_0)$ be the subset of cycles on the overlay nodes A_0, that is, the subset of physical cycles passing through two logical nodes. Clearly, $Z(A_0) \subseteq \pi_1(P)$. Therefore, it follows that the overlay graph A is defined as

$$A = P/Z(A_0)$$

The above in turn defines a natural quotient map

$$P \xrightarrow{\preceq} A$$

which can be read as "A is an overlay of P." Conceptually, given an Internet graph P, we are, in general, able to associate many overlays. We define, over the set of all possible overlays of P, a partial ordering or arrow. We write

$$A_1 \xrightarrow{\preceq} A_2$$

to denote that "A_1 is finer than A_2," or that "A_2 is an overlay of A_1." Clearly, this is a mathematical conceptualization of scalability. Now, we can define a category as follows: The objects are the overlays of P and the arrows are partial ordering relations.

A.4.2 Combinatorial Topological Aspects of Overlaying

The graphs P, A are 1-D simplicial complexes and as such their geometrical significance is limited. The key to endowing them with a richer geometric structure is to extend them to higher-dimensional simplicial complexes as follows: A collection of p nodes of P such that every pair of such nodes is physically linked will be declared a p-simplex. As such, with P, we associate an n-dimensional simplicial complex $K(P)$, where n is the maximum of all p's. The same procedure is repeated with the logical graph to yield a m-dimensional simplicial complex $K(A)$.

The simplicial complexes $K(P), K(A)$ share in common vertices only, so that they are intertwined in some subtle way. Observe, however, that links and hence simplexes of $K(A)$ are equivalent classes–actually simple homotopy classes–of paths in $K(P)$. It follows that $K(P), K(A)$ have more in common than what might be seen at a first look. At the extreme, one could envision situation where $K(P) \to K(A)$ is a simple homotopy equivalence , but that does not seems to be the case in general.

The concept of simple homotopy equivalence, along with elementary collapses and expansions, provides a mathematical conceptualization of adaptability. Clearly, one should be able to go from one overlay to another one by a series of very elementary operations creating a succession of overlays, two consecutive overlays differing by some highly localized features. By definition, a simple homotopy equivalence is the same idea–a succession of elementary collapses and expansions that allow for a succession of very simple transformations amounting to a transformation that is not that simple. As we have already seen, one elementary operation to go from one overlay to another is to cluster several nodes. From there on, it can be seen that clustering many physical nodes into one logical node can be viewed as an elementary collapse, whereas adding a new logical node can be viewed as an elementary expansion.

A.4.3 Geometrical Aspects of Overlaying

Recall a premise of our approach: Interpolate the graph P with a manifold $M(K(P))$ so that the information flow could be viewed as a tangent vector field on the manifold. The rationale is that, instead of being overwhelmed by the curse of dimensionality of the Internet graph, the key features of the network behavior–including failure and attacks–might be revealed by some PDE's on the manifold. The manifold $M(K(P))$ is not in general unique; it could be a stratified manifold; or it could simply not exist. In fact, assuming that $K(P)$ is a simple Poincaré complex, we define the set of structures $S(K(P))$ to be the set of simple homotopy equivalence classes of manifolds that can be associated with $K(P)$. Since the overlay A is also a graph, one could equally associate a manifold with the overlay, $M(K(A))$, and define the set of all equivalence classes of such manifolds as $S(K(A))$. From here on, we could define an attack tolerant overlay if none of the manifolds in $S(K(A))$ shows the vorticity indicative of a global anomaly on the manifold of the physical layer.

Closely related to the concept of attack tolerant overlay is the question as to what is the relation between the manifolds $M(K(P)), M(K(A))$? Could we have an inclusion relation $M(K(P)) \overset{\subseteq}{\to} M(K(A))$? There is certainly one case in which this happens, although in some extreme sense. If $K(A)$ is a collapse of $K(P)$, then $M(K(P)), M(K(A))$ are homotopically equivalent and hence the same up to homotopy equivalence.

At a conceptual level, the issue is the following: Consider the physical layer and a string of application layers related by partial ordering, as shown in the top line of the diagram below. The top line of the diagram shown below could be viewed as the category of all overlays of the "world graph" P together with partial ordering relation or "arrow." Somehow, we must find the image of this string under $S \circ K$. Since K and S are fairly well understood operations, the real challenge is to find the connecting morphism linking the bottom objects. Whether there is some category structure to support the bottom line and whether $S \circ K$ could be viewed as a (covariant or contravariant?) functor remain open.

$$\begin{tikzcd}
P \arrow[r, "\preceq"] \arrow[d, "S \circ K"] & A_1 \arrow[r, "\preceq"] \arrow[d, "S \circ K"] & A_2 \arrow[d, "S \circ K"] \\
S(K(P)) \arrow[r, leftrightarrow, "?"] & S(K(A_1)) \arrow[r, leftrightarrow, "?"] & S(K(A_2))
\end{tikzcd}$$

Appendix B

wireless

Appendix C

quantum

C.1 Spintronic networks

C.2 Entanglement networks

Appendix D

power

Bibliography

[1] Colin C. Adams. *The Knot Book–An Elementary Introduction to the Mathematical Theory of Knots.* W. H. Freeman and Company, New York, 1994.

[2] J. F. Adams. *Stable Homotopy and Generalised Homology.* Chicago Lectures in Mathematics. The University of Chicago Press, Chicago and London, 1974.

[3] V. I. Arnold and B. A. Khesin. *Topological Methods in Hydrodynamics.* Springer-Verlag, 1999.

[4] Michael Atiyah. *The Geometry and Physics of Knots.* Cambridge University Press, Cambridge, 1990.

[5] T. Aubin. Equations différentielles non linéaires et problème de Yamabe concernant la courbure scalaire. *J. Math. Pures Appl.*, 55(9)(3):269–296, 1976.

[6] R. Banirazi, E. Jonckheere, and B. Krishnamachari. Heat diffusion algorithm for resource allocation and routing in multihop wireless networks. In *Globecom, Session WN16: Routing and Multicasting*, pages 5915–5920, Anaheim, California, USA, December 3-7 2012. Available at `http://eudoxus2.usc.edu`.

[7] R. Banirazi, E. Jonckheere, and B. Krishnamachari. Dirichlet's principle on multiclass multihop wireless networks: Minimum cost routing subject to stability. In *MSWiM'14, Proceedings of the 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 31–40, Montréal, Canada, September 22-25 2014.

[8] R. Banirazi, E. Jonckheere, and B. Krishnamachari. Heat-Diffusion: Pareto optimal dynamic routing for time-varying wireless networks. In *IEEE INFOCOM—IEEE Conference on Computer Communications*, pages 325–333, Toronto, Canada, April 27- May 02 2014. Available at `http://eudoxus2.usc.edu`.

[9] R. Banirazi, E. Jonckheere, and B. Krishnamachari. Minimum delay in class of throughput-optimal control policies on wireless networks. In *American*

Control Conference (ACC), pages 2668–2675, Portland, OR, June 04-06 2014. Available at `http://eudoxus2.usc.edu`.

[10] Albert-Laszlo Barabasi, Reka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272:173–187, 1999. published by Elsevier.

[11] Alan F. Beardon. *The Geometry of Discrete Groups*, volume 91 of *Graduate Texts in Mathematics*. Springer, New York, Berlin, 1983.

[12] Marcel Berger. *Riemannian Geometry During the Second Half of the Twentieth Century*, volume 17 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2000.

[13] R. L. Bishop and S. I. Goldberg. *Tensor Analysis on Manifolds*. Dover, 1980.

[14] Bruce Blackadar. *K-Theory for Operator Algebras*, volume 5 of *Mathematical Sciences Research Institute Publications*. Cambridge University Press, Cambridge, England, 1998.

[15] L. M. Blumenthal. *Theory and Applications of Distance Geometry*. Oxford at the Clarendon Press, London, 1953.

[16] S. Bohacek and E. A. Jonckheere. Linear dynamically varying LQ control of nonlinear systems over compact sets. *IEEE Transaction on Automatic Control*, 46:840–852, June 2001.

[17] S. Bohacek and E. A. Jonckheere. Nonlinear tracking over compact sets with linear dynamically varying H^∞ control. *SIAM J. Control and Optimization*, 40(4):1042–1071, 2001.

[18] S. Bohacek and E. A. Jonckheere. Structural stability of linear dynamically varying (LDV) controllers. *Systems and Controls Letters*, 44:177–187, 2001.

[19] F. Bonahon. Geodesic laminations on surfaces. *Contemporary Mathematics*, 269:1–37, 2001.

[20] G. E. Bredon. *Topology and Geometry*. Springer-Verlag, New York, 1993.

[21] Martin R. Bridson and André Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319 of *A Series of Comprehensive Surveys in Mathematics*. Springer, New York, NY, 1999.

[22] M. Buchanan. *Nexus–Small Worlds and the Groundbreaking Science of Networks*. Norton, New York, London, 2002.

[23] D. Burago, Y. Burago, and S. Ivanov. *A Course in Metric Geometry*, volume 33 of *Graduate Study in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2001.

[24] C. Carathéodory. *Conformal Representation.* Dover, 1998.

[25] Z. Chen, L. Gao, and K. Kwiat. Modeling the spread of active worms. In *IEEE INFOCOM*, 2003.

[26] A. Connes. *Noncommutative Geometry.* Academic Press, London and San Diego, 1994.

[27] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms.* MIT Press, 1992.

[28] H. S. M. Coxeter. *Introduction to Geometry.* Wiley Classics Library. John Wiley & Sons, New York, 1989.

[29] J. Cuntz. K-theory for certain C^*-algebras. *Annals of Mathematics*, 113:181–197, 1981.

[30] E. Deza and M. Deza. Dictionary of distances. Technical report, ???, 2006.

[31] M. do Carmo. *Differential Curves and Surfaces.* Prentice-Hall, New Jersey, 1976.

[32] C. H. Dowker. Homology groups of relations. *Annals of Mathematics*, 56(1):84–95, 1952.

[33] J.-P. Eckmann and E. Moses. Curvature of co-links uncovers hidden thematic layers in the world wide web. *PNAS*, 99(9), April 2002.

[34] N. A. Edwards and K. A. Hassall. *Biochemistry and Physiology of the Cell–An Introductory Text.* McGraw Hill Book Company (UK) Limited, London, 1980. Second Edition.

[35] Peter A. Fillmore. *A User's Guide to Operator Algebras.* Canadian Mathematical Society Series of Monographs and Advanced Texts. Wiley, New York, 1996.

[36] L. R. Ford and D. R. Fulkerson. *Flows in Networks.* Princeton University Press, Princeton, 1963.

[37] N. Fowler, M. Laca, and I. Raeburn. The C^*-algebras of infinite graphs. *Proc. Amer. Math. Soc.*, pages 2319–2327, 2000.

[38] M. Golubitsky and V. Guillemin. *Stable Mappings and Their Singularities*, volume 14 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1973.

[39] J. M. Gracia-Bondia, J. C. Varilly, and H. Figueroa. *Elements of Noncommutative Geometry.* Birkhauser, Boston, 2001.

[40] M. Gromov. Hyperbolic groups. In S. M. Gersten, editor, *Essays in Group Theory*, volume 8 of *Mathematical Sciences Research Institute Publication*, pages 75–263. Springer-Verlag, New York, 1987.

[41] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*, volume 152 of *Progress in Mathematics.* Springer-Verlag, 2001.

[42] J. L. Gross and T. W. Tucker. *Topological Graph Theory.* Dover, Mineola, New York, 2001.

[43] F. Harary. *Graph Theory.* Addison-Wesley, Reading, MA, 1987.

[44] M. Henle. *Modern Geometries: Non-Euclidean, Projective, and Discrete Geometry.* College Div. Prentice Hall, 2001. 2nd edition.

[45] M. W. Hirsch. *Differential Topology*, volume 33 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1976.

[46] D. Husemoller. *Fibre Bundles (Third Edition).* Springer-Verlag, New York, 1994.

[47] K. K. Jensen and K. Thomsen. *Elements of KK-Theory.* Birkhauser, Boston, Basel, Berlin, 1991.

[48] E. A. Jonckheere. *Algebraic and Differential Topology of Robust Stability.* Oxford, New York, 1997.

[49] Edmond Jonckheere, Mingji Lou, Francis Bonahon, and Yuliy Baryshnikov. Euclidean versus hyperbolic congestion in idealized versus experimental networks. *Internet Mathematics*, 7(1):1–27, March 2011.

[50] J. Jost. *Nonpositive Curvature: Geometric and Analytic Aspects.* Lectures in Mathematics. Birkhauser, Basel-Boston-Berlin, 1997.

[51] J. Jost. *Riemannian Geometry and Geometric Analysis.* Universitext. Springer, Berlin, Heidelberg, New York, 1998. Second Edition.

[52] Michael Kapovich. *Hyperbolic Manifolds and Discrete groups*, volume 183 of *Progress in Mathematics.* Birkhauser, Boston, MA, 2001.

[53] A. Katok and B. Hasselblatt. *Introduction to the Modern Theory of Dynamical Systems.* Cambridge, 1997.

[54] S. Kobayashi and K. Nomizu. *Foundation of Differential Geometry*, volume 2. Wiley, 1996.

[55] B. Korte and J. Vygen. *Combinatorial Optimization.* Number 21 in Algorithms and Combinatorics. Springer-Verlag, Berlin, New York, 2000.

[56] W. B. Raymond Lickorish. *An Introduction to Knot Theory.* Number 175 in Graduate Text in Mathematics. Springer, New York, 1997.

[57] Feng Luo. On a problem of Fenchel. *Geometriae Dedicata*, 64:277–282, 1997.

[58] Feng Luo. Combinatorial Yamabe flow on surfaces. *Communications in Contemporary Mathematics*, 6(5):765–780, 2004.

[59] R. C. Lyndon and P. E. Schupp. *Combinatorial Group Theory.* Classics in Mathematics. Springer, Berlin, Heildelberg, New York, 2001. Reprint of the 1977 edition.

[60] W. Magnus, A. Karras, and D. Solitar. *Combinatorial Group Theory—Presentations of Groups in Terms of Generators and Relations.* Dover, New York, 1976.

[61] W. Massey. *Singular Homology Theory.* Springer-Verlag, New York, 1980.

[62] W. S. Massey. *A basic course in algebraic topology.* Number 127 in GTM. Springer, Berlin, New York, 1991.

[63] J. P. May. *A Concise Course in Algebraic Topology.* Chicago Lectures in Mathematics. The University of Chicago Press, Chicago and London, 1999.

[64] B. Mohar and C. Thomassen. *Graphs on Surfaces.* The Johns Hopkins University Press, Baltimore, London, 2001.

[65] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver. The spread of the sapphire/slammer worm. http://www.cs.berkeley.edu/ nweaver/sapphire/, 2003. Cooperative Association for Internet Data Analysis and University of California, San Diego.

[66] D. Moore, C. Shannon, G. M. Volker, and S. Savage. Internet quarantine: requirements for containing self-propagating code. In *IEEE INFOCOM*, 2003.

[67] J. Munkres. *Elements of Algebraic Topology.* Addison-Wesley, Reading, MA, 1984.

[68] O. Narayan and I. Saniee. Large-scale curvature of networks. *Physical Review E*, 84:066108–1–8, 2011.

[69] Z. Nehari. *Conformal Mapping.* Dover, New York, 1952.

[70] M. E. J. Newman, S. Forrest, and J. Balthrop. Email network and the spread of computer viruses. *Physical Review E*, 66:035101–1–4, 2002.

[71] Y. Ollivier. Ricci curvature on Markov chains on metric spaces. *J. Functional Analysis*, 256(3):810–864, 2009.

[72] I. R. Porteous. *Clifford Algebras and the Classical Groups*, volume 50 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, 1995.

[73] S. C. Power. Simplicity of c^*-algebras of minimal dynamical systems. *J. London Math. Soc.*, 18:534–538, 1978.

[74] Stephen C. Power. *Limit Algebras: An Introduction to Subalgebras of C^*-Algebras*, volume 278 of *Pitman Research Notes in Mathematics Series.* Longman Scientific & Technical, Essex, England, 1992.

[75] I. Rivin. Some observations on the simplex. arXiv:math.MG/0308239v1, 2003.

[76] John Roe. *Index Theory, Coarse Geometry, and Topology of Manifolds.* Number 90 in Conference Board of the Mathematical Sciences (CBMS); Regional Conference Series in Mathematics. American Mathematical Society, Providence, RI, 1996.

[77] J. Rosenberg. *Algebraic K-Theory and Its Applications.* Springer-Verlag, New York, 1994.

[78] K. Schmidt. *Algebraic Ideas in Ergodic Theory.* Number 76 in Conference Board of the Mathematical Sciences (CBMS); Regional Conference Series in Mathematics. American Mathematical Society, Providence, RI, 1989.

[79] T. Schmidt. Invariants of flat surfaces lattice Veech groups. In *Southwest Regional Workshop on New Directions in Dynamical Systems*, University of Southern California, November 2000. National Science Foundation. http://www-rcf.usc.edu/ dynamics/ab12.htm.

[80] R. Schoen. Conformal deformation of a Riemannian metric to constant scalar curvature. *J. Differential Geometry*, 20:479–495, 1984.

[81] L. Shalalfeh, P. Bogdan, and E. Jonckheere. Evidence of long range-dependence in power grid. In *Power and Energy Society General Meeting (PESGM)*, Boston, MA, July 2016. Available at `http://eudoxus2.usc.edu`.

[82] L. Shalalfeh, P. Bogdan, and E. Jonckheere. Modeling of PMU data using ARFIMA models. In *Clemson University Power System Conference*, Charleston, SC, September 2018. Paper Session T-M II: Phasor Measurement Units (PMUs).

[83] Andrzej Skowron and Jaroslaw Stepaniuk. Tolerance approximation spaces. *Fundamenta Informaticae*, 27(2,3):245–253, 1996.

[84] A. S. Smogorzhevsky. *Lobachevskian Geometry.* Little Mathematics Library. Mir Publishers, Moscow, 1982.

[85] W. Stallings. *High-Speed Networks: TCP/IP and ATM Design Principles.* Prentice Hall, Upper Saddle River, NJ, 1998.

[86] T. Taniuti and K. Nishihara. *Nonlinear waves*, volume 15 of *Monographs and Studies in Mathematics.* Pitman, Boston, London, Melbourne, 1977.

[87] M. Toda. *Nonlinear Waves and Solitons.* Mathematics and Its Applications. Kluwer Academic Publishers, Doordrecht,Boston, London, 1989.

[88] J. Tomiyama. *Invitation to C^*-Algebras and Topological Dynamics*, volume 3 of *Advanced Series in Dynamical Systems.* World Scientific, Singapore, New Jersey, Hong Kong, 1987.

[89] N. S. Trudinger. Remarks concerning the conformal deformation of Riemannian structures on compact manifolds. *Ann. Scuola Norm. Sup. Pisa*, 22(2):265–274, 1968.

[90] Chi Wang, E. Jonckheere, and R. Banirazi. Wireless network capacity versus Ollivier-Ricci curvature under Heat Diffusion (HD) protocol. In *American Control Conference (ACC)*, pages 3536–3541, Portland, OR, June 04-06 2014. Available at `http://eudoxus2.usc.edu`.

[91] Chi Wang, E. Jonckheere, and R. Banirazi. Interference constrained network performance control based on curvature control. In *2016 American Control Conference (ACC)*, pages 6036–6041, Boston, MA, July 6-8 2016.

[92] D. J. Watts and S. H. Strogatz. Collective dynamics of small world networks. *Nature*, 393:440–442, 1998.

[93] N. E. Wegge-Olsen. *K-Theory and C^*-Algebras.* Oxford Science Publications. Oxford University Press, New York, 1993.

[94] S. Weinberger. *The Topological Classification of Stratified Spaces.* Chicago Lectures in Mathematics, Chicago and London, 1994.

[95] Arthur T. White. *Graphs, Groups, and Surfaces*, volume 8 of *Mathematical Studies.* North-Holland, Amsterdam, New York, Oxford, 1984.

[96] Walter Willinger and Vern Paxson. Where mathematics meets the internet. *Notices of the American Mathematical Society*, pages 961–970, 1998.

[97] H. Yamabe. On a deformation of Riemannian structures on compact manifolds. *Osaka Math. J.*, 12:21–37, 1960.

[98] W. Yourgrau and S. Mandelstam. *Variational Principles in Dynamics and Quantum Theory.* Dover, New York, 1968.