

UNIVERSITY OF LJUBLJANA
FACULTY OF COMPUTER AND INFORMATION SCIENCE

Svetlana Nikić

**Topological Analysis of Data on 911
Calls in the Boston Area**

MASTERS THESIS

THE 2ND CYCLE MASTERS STUDY PROGRAMME
COMPUTER AND INFORMATION SCIENCE

SUPERVISOR: prof. dr. Neža Mramor Kosta

Ljubljana, 2015

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Svetlana Nikić

**Topološka analiza podatkov o klicih
na številko 911 na območju Bostona**

MAGISTRSKO DELO

MAGISTRSKI PROGRAM DRUGE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

MENTOR: prof. dr. Neža Mramor Kosta

Ljubljana, 2015

COPYRIGHT. The results of this Masters Thesis are the intellectual property of the author and the Faculty of Computer and Information Science, University of Ljubljana. For the publication or exploitation of the Masters Thesis results, a written consent of the author, the Faculty of Computer and Information Science, and the supervisor is necessary.

©2015 SVETLANA NIKIĆ

DECLARATION OF MASTERS THESIS AUTHORSHIP

I, the undersigned Svetlana Nikić am the author of the Master Thesis entitled:

Topological Analysis of Data on 911 Calls in the Boston Area

With my signature, I declare that:

- the submitted Thesis is my own unaided work under the supervision of prof. dr. Neža Mramor Kosta
- all electronic forms of the Masters Thesis, title (Slovenian, English), abstract (Slovenian, English) and keywords (Slovenian, English) are identical to the printed form of the Masters Thesis,
- I agree with the publication of the electronic form of the Masters Thesis in the collection "Dela FRI".

In Ljubljana, 23. June 2015

Author's signature:

ACKNOWLEDGMENTS

First and foremost, I would like to thank Prof. Neža Mramor Kosta, who devoted a lot of time and effort to the making of this thesis. Her guidance, patience and great ideas throughout the process were invaluable.

I would like to thank my committee members, Assoc. Prof. Janez Demšar, Assoc. Prof. Polona Oblak and Assoc. Prof. Tomaž Curk, for their unbelievable flexibility, support and suggestions.

A special thanks goes to Prof. Christopher Winship and Assist. Prof. Dan OBrien from the Department of Sociology at Harvard, without whom this thesis would not have been possible.

I would like to extend my sincere gratitude to my family. Their support through the years has made me who I am and made my road to this point much easier and more enjoyable.

I am also grateful to my friends, who were always ready to lend a helping hand and who were with me through all my ups and downs.

Svetlana Nikić, 2015

Contents

Povzetek	i
Abstract	v
1 Introduction	1
1.1 Related Work	1
1.2 Methodology	2
1.3 Chapter Overview	4
2 The Methods of Computational Topology	5
2.1 Homeomorphism	5
2.2 Simplicial Complex	7
2.3 Homotopy	11
2.4 Homology	12
3 Data Reconstruction	17
3.1 The Čech Complex	17
3.2 The Vietoris-Rips complex	20
4 Discrete Morse Theory	23
4.1 The Algorithm	26
5 Persistent Homology	33

CONTENTS

6	Implementation and results	37
6.1	Results Based on Critical Simplices	39
6.2	Results Based on Persistence Diagrams	46
7	Conclusions	59

Povzetek

Čedalje večje količine podatkov, ki se dnevno zbirajo, predstavljajo velik izziv za raziskovalce, ki iščejo metode za odkrivanje koristnih informacij, skritih v njih. Med novejšimi pristopi k analizi podatkov so topološke metode (TDA) [1], ki so se izkazale za zelo učinkovite, na primer pri analizi območja pokritosti senzorskih omrežij [2], v bioinformatiki [3] ali analizi trdnosti poroznih materialov [4].

V magistrskem delu smo s pomočjo diskretne Morsove teorije [5] in vztrajne homologije [6, 7] analizirali podatke s povsem nove domene, to je na podatkih o klicih na številko 911 v mestu Boston v ZDA. Pobuda za TDA na takšnih podatkih je prišla s strani sociološkega oddelka na Harvardu, kjer so jih do sedaj poskušali analizirati že z raznimi mehanskimi in matematičnimi modeli [8, 9], med drugim tudi z implementacijo mehanskega razmišljanja z uporabo usmerjenih acikličnih grafov (DAG) [10].

V poglavju 2 naloga govori o osnovnih matematičnih objektih, ki so temelj kasneje uporabljene teorije. V tem delu spoznamo simplekse, ki predstavljajo osnovno enoto za predstavitev podatkov in so lahko različnih dimenzij. Med seboj se povezujejo v simplicialne komplekse. To poglavje govori tudi o ekvivalencah, ki se pojavljajo v računski topologiji, s pomočjo katerih bomo lahko enačili pridobljene objekte z začetnimi.

V poglavju 3 so opisani načini za izgradnjo simplicialnih kompleksov iz podatkov. Dva izmed njih zgradita Čechov in Vietoris-Ripsov kompleks. Za gradnjo obeh je potrebno najprej uvesti razdaljo med podatki in neko določeno razdaljo r , za katero velja, da se bodo podatki, ki imajo medsebojno

razdaljo manjšo od r , povezovali. Vietoris-Ripsov kompleks je poenostavljena oblika Čechovega kompleksa, ki se zato večkrat uporablja pri implementaciji, saj je izgradnja hitrejša. V nalogi smo uporabili Vietoris-Ripsov kompleks na geografskih podatkih, da bi dobili reprezentativna področja po mestu, saj posamezni podatki ne nosijo dovolj informacij o svoji okolici. Tako dobljena področja smo predstavili s točkami, da smo lahko ponovno zgradili Vietoris-Ripsov kompleks, ki smo ga uporabljali v nadaljevanju.

Na tašnem kompleksu smo uporabili diskretno Morsovo teorijo, ki je opisana v poglavju 4. Že pri konstrukciji kompleksa vsakemu simpleksu priredimo vrednost Morsove funkcije, ki mora biti injektivna. Za Morsovo funkcijo velja, da bodo imeli simpleksi višjih dimenzij tudi višjo vrednost funkcije z možnimi izjemami. Za vsak simpleks pogledamo, če ima kateri izmed njegovih podsimpleksov, ki so za eno dimenzijo manjši, večjo vrednost in s tem predstavljajo izjemo. V tem primeru ta dva simpleksa združimo v par in dodamo puščico, usmerjeno od simpleksa višje dimenzije, k tistemu z manjšo. Simplekse, ki niso v paru, imenujemo kritični simpleksi in predstavljajo ekstreme ali ločnice med območji z različnimi lastnostmi. Da bi imeli takih simpleksov čim manj in s tem dobili bolj jasne rezultate, je implementiran algoritem, ki še dodatno združi kritične simplekse v pare, kjer je to mogoče.

Drug način analiziranja podatkov je vztrajnostna homologija, o kateri govori poglavje 5. Ta se uporablja, da vidimo kako kompleks nastaja glede na nek parameter. Recimo, da je ta parameter število klicev z nekega območja in gledamo njegovo padajočo vrednost. V tem primeru bodo najprej nastali simpleksi z največjim številom klicev. Postopoma se bodo pojavljali novi simpleksi, ki se pojavijo kot del komponente, ki že obstaja, ali pa ustvarijo novo. Vztrajnostni diagrami sledijo spremembam teh komponent. Ko se dve komponenti združita, mlaja umre. Čas, v katerem je ta komponenta obstajala, ponazarja vztrajnost komponente. Iz vztrajnostnega diagrama lahko vidimo, ali ima neko območje veliko komponent, ki so vztrajala dolgo časa, ali pa ima samo eno žarišče, h katerem so se pridružili ostali simpleksi.

Teoretičnem delu sledi poglavje 6, ki govori o sami implementaciji. Tu

najdemo komplekse s kritičnimi simpleksi in analizo diagramov vztrajnosti za različna območja Bostona. Kompleksi in diagrami so bili ustvarjeni z različnimi parametri, da bi pridobili čim širšo sliko.

Zadnje poglavje vsebuje zaključke in predloge za nadaljnje delo.

Ključne besede

topološka analiza podatkov, klici v sili, diskretna Morsova teorija, vztrajna homologija, Vietoris-Ripsov kompleks

Abstract

Among the newer approaches to data analysis are topological methods (TDA) [1], which proved to be effective in analyzing data [2, 3, 4]. In this thesis we analyze data on 911 calls that include a large number of calls. Firstly, we prepare data by grouping calls together using the Vietoris-Rips complex [11, 12]. We do this because it enables us to also analyze smaller areas and connect them. We analyze this complex in two ways: by using Morse theory [5] and persistent homology [7]. Morse theory is used to acquire critical simplices from the complex. They give us new information about the data. Using persistent homology, we produce persistent diagrams that illustrate how homology of a complex changes depending on a parameter. The initiative to use the TDA on such data came from the Department of Sociology at Harvard, where they had already tried to analyze this data by using various mechanical and mathematical models [8, 9, 10].

Keywords

topological data analysis, emergency calls, discrete Morse theory, persistent homology, the Vietoris-Rips complex

Chapter 1

Introduction

An increasing amount of data that is collected daily represents a great challenge for researchers who are looking for methods for finding useful information hidden in them. Among the newer approaches to data analysis are topological methods (TDA) [1], which proved to be very effective, for example, in analyzing the coverage area of sensor networks [2], in bioinformatics [3] and image analysis [4].

In this thesis discrete Morse theory [5] and persistent homology [7] are used to analyze data from a new domain, namely 911 calls in Boston, MA. The initiative to use the TDA on such data came from the Department of Sociology at Harvard, where they had already tried to analyze this data by using various mechanical and mathematical models [8, 9], including the causal implementation of mechanical thinking with the use of DAG [10].

1.1 Related Work

A thorough review of the application of topological methods to data is given in the books *Topology for Computing* [13] and *Computational Topology: An Introduction* [14] and in the paper *Topology and Data* [1].

In topological data analysis, data is represented as a set of points (vectors) in Euclidian space of some dimension and is connected into simplicial com-

plexes or other topological objects using object reconstruction algorithms. In this thesis, our reconstruction models will be the Vietoris-Rips and the Čech simplicial complexes [11, 12, 15].

These complexes are then analyzed using topological methods, for example homology and cohomology [2], persistent homology [6, 7], and discrete Morse theory [5].

An additional advantage of topological methods is that they enable analysis of data in different resolutions by constructing complexes on individual data points or on groups of data points and by varying the parameters of the reconstructions.

Similar approaches for calculating Morse field and critical simplices have been used in *Generating Discrete Morse Functions from Point Data* [16] and were optimized in the paper *Theory and Algorithms for Constructing Discrete Morse Complexes from Grayscale Digital Images* [4]. The authors of this paper used 2- and 3-Dimensional images, which are most naturally represented with a cubic complex, while for the scattered data, such as 911 calls, this complex is unsuitable. The method itself, however, remains similar.

Analysis of the complex can also be done by using persistent homology as discussed in the paper *Morse Theory for Filtrations and Efficient Computation of Persistent Homology* [7]. The theoretical paper *Reducing Complexes in Multidimensional Persistent Homology Theory* [17] talks about a similar problem but mainly focuses on multi-dimensional filtration, which no longer guarantees that the resulting Morse complex is ideal.

1.2 Methodology

In our case, the data points are vectors obtained from 911 calls in the Boston area. They contain a number of parameters, for example the geographic location of the call, the time of the call, the type (vandalism, shootings, domestic violence) etc.

In our analysis, we consider different combinations of parameters. Using

geographic coordinates as data points, we build the Vietoris-Rips complex based on the geographic distances between points. Firstly, we connect points in simplices. A subset of points forms a simplex if the distance between each pair is lower than some threshold. Changing this threshold changes the resolution at which we look at the data.

Firstly, the threshold was set so that the resulting simplicial complex has several connected components representing the areas with connected calls. We represent each component as a new point with parameters: the approximate coordinates of the area and the total number of calls from that area.

For a closer look at the individual areas we build a Vietoris-Rips complex from the points representing areas. This can be done in different ways by using different distances.

Once the simplicial complex is constructed, we use two approaches for further analysis, discrete Morse theory and persistence.

In order to apply discrete Morse theory, we add function values to the vertices and extend these to a discrete Morse function on the complex, which associates a value to each simplex. A simplex of dimension n has a higher value than its faces of dimension $n-1$ with at most one exception per simplex. If such an exceptional face exists, we pair them and add an arrow pointing from the higher value to the lower. Simplices that do not have a pair are called critical and tell us the most about the data. In some cases critical simplices arise from the noise in the data or unimportant details. In order to lower the number of critical simplices and thus give us clearer results, an algorithm is implemented for cancelling pairs of critical simplices and thus reducing the complexity.

The second method for analyzing data used in this thesis is based on persistent homology. We want to know how the complex is built simplex by simplex based on some parameter. If this parameter is time, the simplices with the earliest calls will be born first. Gradually, other simplices will show up, either as a part of already existing components or as a new one. When two components merge, the younger one dies.

The interval between the birth and the death of a component illustrates the component's importance. We illustrate these changes with persistence diagrams in order to obtain insight into the structure of calls in the individual areas.

The data used in this thesis, including a table of 1.26 million calls, was provided by Professor Christopher Winship's team from the Department of Sociology at Harvard.

1.3 Chapter Overview

In chapter 2, basic topology that is needed for further work is described. This includes homeomorphism, simplices, complexes, homotopy and homology. The next chapter talks about data reconstruction - in other words, it addresses the question of how we can get an object from point data. Morse theory and an algorithm for constructing Morse field are described in chapter 4. Chapter 5 discusses how persistent components are and how this can be viewed on diagrams. How this theory was implemented and the results are described in chapter 6. Possible future work and conclusions can be found in chapter 7.

Chapter 2

The Methods of Computational Topology

This chapter includes definitions from and is based on [13]. Topology is a field of mathematics that studies shapes and topological spaces and assumes two objects are the same if one can be turned into another only with continuous deformations, such as folding and stretching but not tearing and gluing. Topology is interested in values, such as how many parts the object is constructed from, how many holes it has, but it does not care about specific measurements.

2.1 Homeomorphism

Definition 2.1 *Sets A and B are homeomorphic, $A \cong B$, if there exists a homeomorphism, that is, a continuous bijective map $f : A \rightarrow B$ such that $f^{-1} : B \rightarrow A$ is also continuous.*

Example: The circle C , given with the equation $x^2 + y^2 = 1$ and a square K , given with the equation $\max\{|x|, |y|\} = 1$ are homeomorphic.

Homeomorphism $f : C \rightarrow K$ and its inverse $f^{-1} : K \rightarrow C$ are

$$f(x, y) = \frac{1}{\sqrt{x^2 + y^2}}(x, y),$$



Figure 2.1: Objects that are not homeomorphic to each other.



Figure 2.2: Two homeomorphic objects.

$$f^{-1}(x, y) = \frac{1}{\max\{|x|, |y|\}}(x, y).$$

Another example is shown in Figure 2.1 shows the letter A written in different fonts. Even though they all represent the same letter, the corresponding subsets of \mathbb{R}^2 are not homeomorphic. Two objects are not homeomorphic if they have a different number of holes. This means that the third and the fourth letter are not homeomorphic to any of the other letters since they have a unique number of holes. The first two numbers both have one hole and they both have two tails, but the second A also has two tails on each of its tails that cannot be pushed inside with a continuous map. The last two As do not have any holes, but the last one has more tails and as such they are not homeomorphic.

Figure 2.2, however, shows two homeomorphic objects, even though the two objects do not look very similar at the first sight. They both have one hole and two tails that originate from two different points of the circle.

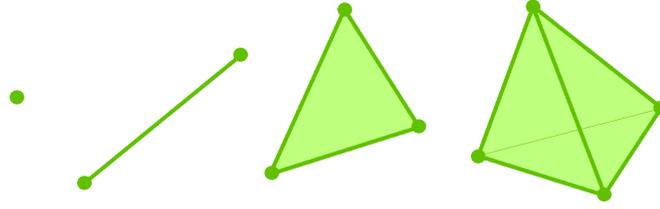


Figure 2.3: A vertex, an edge, a triangle and a tetrahedron.

2.2 Simplicial Complex

Definition 2.2 Let $S = \{p_0, p_1, \dots, p_k\} \subset \mathbb{R}^d$. A linear combination is $x = \sum_{i=0}^k \lambda_i p_i$ for some $\lambda_i \in \mathbb{R}$. An affine combination is linear combination with $\sum_{i=0}^k \lambda_i = 1$. A convex combination is an affine combination with $\lambda_i \geq 0$, for all i . The set of all convex combinations is the convex hull.

Definition 2.3 A set S is linearly (affinely) independent if no point in S is a linear (affine) combination of the other points in S .

Definition 2.4 A k -simplex is a convex hull of $k + 1$ affinely independent points $S = \{v_0, v_1, \dots, v_k\}$. The points in S are the vertices of the simplex.

A k -simplex σ can also be denoted as σ^k . σ is a k -dimensional subspace of \mathbb{R}^d , $\dim \sigma = k$.

Simplices are basic units used to represent data. The simplices of lowest dimensions have special names: a 0-simplex is also called a *vertex*, a 1-simplex is called an *edge*, a 2-simplex is a *triangle* and a 3-simplex a *tetrahedron*.

Definition 2.5 Let σ be a k -simplex defined by $S = \{v_0, v_1, \dots, v_k\}$. A simplex τ defined by $T \subseteq S$ is a face of σ and has σ as a coface. The relationship is denoted with $\sigma \geq \tau$ and $\tau \leq \sigma$. Note that $\sigma \geq \sigma$ and $\sigma \leq \sigma$.

τ is called a proper face if $\tau \neq \sigma$ and we write $\tau < \sigma$.

A k -simplex σ is a convex hull of a set of $k + 1$ points. Let us call this set S . There are 2^{k+1} possible subsets of S including S itself and an empty

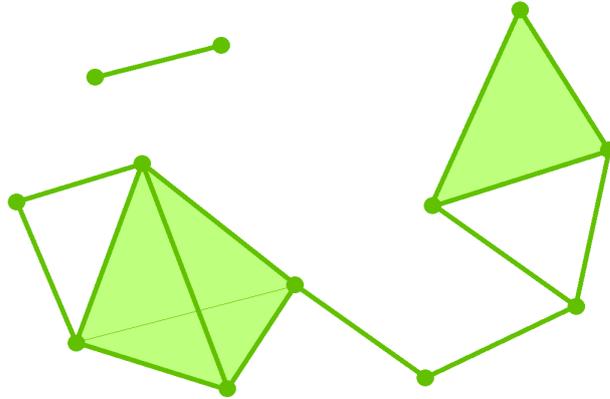


Figure 2.4: An example of a simplicial complex.

set. This means that every k -simplex σ has $2^{k+1} - 2$ proper faces and a face σ itself.

Definition 2.6 *The boundary $\partial\sigma$ of a simplex σ is the union of all proper faces of σ . The interior of σ is the simplex without its boundary, that is, $\text{int}(\sigma) = \sigma - \partial\sigma$.*

To represent multiple data we connect simplices in a bigger structure called a simplicial complex.

Definition 2.7 *A geometric simplicial complex K is a family of simplices that satisfies two conditions:*

- *If $\sigma \in K$ and $\tau \leq \sigma \Rightarrow \tau \in K$.*
- *If $\sigma_1, \sigma_2 \in K$ and $\sigma_1 \cap \sigma_2 \neq \emptyset \Rightarrow \sigma_1 \cap \sigma_2 \in K$, that is, the intersection is a face of both simplices.*

Definition 2.8 *The dimension of a simplicial complex K is the highest dimension of any simplex in K , $\dim K = \max\{\dim\sigma \mid \sigma \in K\}$.*

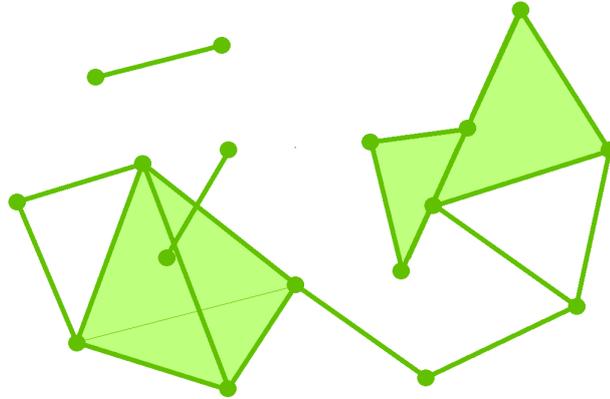


Figure 2.5: Because of the added simplices to the simplicial complex from Figure 2.4 this is not a complex any more. They are both added in a way that violates the second part of the complex definition, that is, the intersections are not a part of the complex.

Definition 2.9 *The n -skeleton of a simplicial complex K , $K(n)$ is the family of all simplices of K of dimension $\leq n$. The special case $K(0)$, that is, the set of all vertices of K , is also denoted as $V(K)$.*

Geometric simplicial complexes depend on the geometric position of the points. In order to represent our data, we do not need and do not want to be restricted by geometric representation. That is why we use abstract simplicial complexes, which are purely combinatorial description of the geometric notion of a simplicial complex.

Definition 2.10 *An abstract simplicial complex is a set K , together with a collection S of subsets of K called (abstract) simplices such that:*

- *For all $v \in K$, $\{v\} \in S$. We call the sets $\{v\}$ the vertices of K .*
- *If $\tau \subseteq \sigma \in S$, then $\tau \in S$.*

A subset $L \subseteq K$ which is a simplicial (geometric or abstract) complex is a subcomplex of K .

When it is clear from the context what S is, we refer to K as a complex. We say σ is a k -simplex of dimension k if $|\sigma| = k + 1$. If $\tau \subseteq \sigma$, τ is a face of σ and σ is a coface of τ .

Even though we usually work with abstract simplicial complexes, sometimes the transformation between geometric and abstract complexes is needed.

Transforming a geometric simplicial complex K into an abstract simplicial complex is easy. We only keep the sets of vertices of simplices and forget the geometric characteristics.

The transformation is also possible the other way around. If the dimension of the ambient space is high enough, we can always construct a geometric simplicial complex from an abstract simplicial complex K and call it the *geometric realization* of K .

Theorem 2.11 *For every abstract simplicial complex K of dimension d there exists a geometric realization in \mathbb{R}^{2d+1} .*

Proof. Let us say that we have a function f that is an injection of vertices from K to \mathbb{R}^{2d+1} . We can assume points lie in general position, that is, any $2d + 2$ points are affinely independent.

In order to demonstrate that we get a geometric simplicial complex, we have to show that the intersection of two simplices is either empty or a face of both. So let α_1 and α_2 be two simplices in A . The size of their union $\alpha_1 \cup \alpha_2$ is $\text{card}(\alpha_1) + \text{card}(\alpha_2) - \text{card}(\alpha_1 \cap \alpha_2) \leq 2d + 2$, which means points in $f(\alpha_1 \cup \alpha_2)$ are affinely independent and every convex combination of points can only be written in one way. Let us denote simplices in the geometric complex $\sigma_1 = \text{conv}(f(\alpha_1))$ and $\sigma_2 = \text{conv}(f(\alpha_2))$ and let x be a convex combination in their intersection. x is unique, so if it belongs to both of them, it has to belong to $\text{conv}(f(\alpha_1 \cap \alpha_2))$. The intersection is then either empty or the simplex $\text{conv}(f(\alpha_1 \cap \alpha_2))$, that is, the face of both simplices. This means we have a geometric simplicial complex. \square

2.3 Homotopy

In topology, another equivalence between two spaces is used. It is called homotopy equivalence. This equivalence is weaker than topological (homeomorphic) equivalence.

Definition 2.12 *Two functions $f, g : A \rightarrow B$ are homotopic $f \simeq g$ if there exists a family of functions $\{h_t : A \rightarrow B; t \in [0, 1]\}$, such that:*

- $h_0 = f$
- $h_1 = g$
- $H : A \times [0, 1] \rightarrow B, H(x, t) = f_t(x)$ is continuous.

Function H is called homotopy.

Definition 2.13 *We say that sets $A, B \subset \mathbb{R}^d$ are homotopy equivalent or have the same homotopy type, $A \simeq B$, if there exist functions $f : A \rightarrow B$ and $g : B \rightarrow A$, such that $f \circ g \simeq id_B$ and $g \circ f \simeq id_A$.*

A special case of a homotopy equivalence is a deformation retract.

Definition 2.14 *A subset $B \subseteq A$ is a retract of A if there exists a retraction, that is, a function $r : A \rightarrow B$, such that $r(x) = x$, for all $x \in B$. If r and the identity on A are homotopic, $r \simeq id_A$, then B is called a deformation retract and r a deformation retraction.*

If B is a point and a deformation retract of A , then we say A is contractible.

In Figure 2.6, we again have letters in different fonts but these are all homotopy equivalent. Here we basically just count the number of holes.

Two objects that are homeomorphic are also homotopy equivalent.

If $h : A \rightarrow B$ is a homeomorphism, then $h \circ h^{-1}$ and $h^{-1} \circ h$ are the identities of B and A , respectively, so the homotopy equivalence of A and B is an immediate consequence of the reflexivity of that relation.



Figure 2.6: Homotopy equivalent objects.

2.4 Homology

Homology tells us how points are connected and what kind of holes they create. For every dimension we know how many holes we have. Conceptually 1-dimensional holes are created by empty circles, 2-dimensional holes by empty spheres and so on.

Definition 2.15 Let K be a simplicial complex of dimension n and n_p be the number of simplices of dimension p . A p -chain is a formal sum

$$c = \sum_{i=1}^{n_p} a_i \sigma_i,$$

where a_i are coefficients and σ_i are simplices of dimension p .

Coefficients can be elements of an arbitrary field or a ring, but for our needs we will use modulo 2 coefficients, which can only have values 0 or 1.

The sum of two p -chains is calculated in the same way as vector sum,

$$\sum_{i=1}^{n_p} a_i \sigma_i + \sum_{i=1}^{n_p} b_i \sigma_i = \sum_{i=1}^{n_p} (a_i + b_i) \sigma_i.$$

The set of all p -chains $C_p(K)$ with addition forms the (Abelian) group of p -chains.

Definition 2.16 Let K be a simplicial complex and $\sigma = [v_0, \dots, v_p]$ a simplex with vertices v_0, \dots, v_p . The boundary of σ is the $(p-1)$ -chain

$$\partial_p \sigma = \sum_{i=0}^p [v_0, \dots, \hat{v}_i, \dots, v_p],$$

where \hat{v}_i means v_i is omitted.

The boundary map is extended to p -chains as a homomorphism that maps a p -chain to $(p - 1)$ -chain, $\partial_p : C_p \rightarrow C_{p-1}$, $\partial(\sum_{i=0}^n \alpha_i \tau_i) = \sum_{i=0}^n \alpha_i \partial \tau_i$.

In addition $C_p = 0$ for $p > \dim K$ and $p < 0$.

Theorem 2.17 *For every p , the composition $\partial_p \partial_{p+1}$ is the trivial homomorphism: $p + 1$ -chain c and every integer p it stands that $\partial_p \partial_{p+1} c = 0$.*

Proof. A p -chain is a full sum of simplices, so since ∂_p and ∂_{p+1} are homomorphisms it is enough to show that $\partial_p \partial_{p+1} \sigma = 0$ for any $(p + 1)$ -simplex σ .

Firstly, we calculate the boundary ∂_{p+1} on σ . It gives us all the faces of that simplex of dimension p . We then calculate the boundary ∂_p on the faces. For every p -face we get all the $(p - 1)$ -faces of that face. Since every $(p - 1)$ -face lies in exactly two p -faces, they nullify each other when adding them up (we only use modulo 2 coefficients). This means $\partial_p \partial_{p+1} \sigma = 0$. \square

Definition 2.18 *The chain complex is the sequence of chain groups connected by boundary homomorphisms,*

$$\dots \xrightarrow{\partial_{p+2}} C_{p+1}(K) \xrightarrow{\partial_{p+1}} C_p(K) \xrightarrow{\partial_p} C_{p-1}(K) \xrightarrow{\partial_{p-1}} \dots \xrightarrow{\partial_1} C_0(K) \rightarrow 0.$$

Now we introduce two types of chains needed to define homology groups.

Definition 2.19 *A p -chain c is a p -cycle if it has empty boundary, $\partial c = 0$.*

The set of p -cycles $Z_p(K) = \ker \partial_p$ is a subset of the group $C_p(K)$.

Definition 2.20 *A p -chain c is a p -boundary if it is a boundary of some $p + 1$ -chain, $c = \partial c'$, $c' \in C_{p+1}(K)$.*

The set of p -boundaries $B_p(K) = \text{im} \partial_{p+1}$ is a subset of the group $C_p(K)$.

∂ is a homomorphism, so it commutes with addition. This means p -cycles form a group of p -cycles which are subgroup of p -chains. We denote them as $Z_p(K)$. $Z_p(K)$ is the kernel of ∂_p .

Similarly p -boundaries form a group $B_p(K)$, which is also a subgroup. $B_p(K)$ is the image of ∂_{p+1} .

The consequence of the theorem 3.2 is that $B_p \subseteq Z_p$ or that every p -boundary is also a p -cycle.

Now we can talk about partitioning cycle groups into equivalence classes consisting of cycles that differ from each other by a boundary. In order to do that, we first introduce equivalence classes.

Definition 2.21 *Let A be a commutative group and $B \subseteq A$ a subgroup. The quotient group A/B consists of equivalence classes of elements of A with respect to the relation $a \sim b$ if and only if $a - b \in B$ and with addition defined by $[a] + [b] = [a + b]$.*

So if we move back to homology, we partition cycles into equivalence classes according to boundaries.

Definition 2.22 *The p -th homology group is the quotient group of the p -th cycle group over the p -th boundary group,*

$$H_p = Z_p/B_p.$$

Cycles a and b that belong to the same homology class are said to be homologous.

An element of a homology group is an equivalence class of cycles and can be represented by an arbitrary cycle from the class. We obtain each element by adding a boundary to a given cycle, $c + B_p$, $c \in Z_p$.

Definition 2.23 *The homology of a simplicial complex K is the sequence of its homology groups*

$$H(K) = (H_0(K), H_1(K), \dots, H_p(K), \dots)$$

Groups $Z_p(K)$, $B_p(K)$ and $H_p(K)$ are vector spaces over \mathbb{Z}_2 . The dimension of the vector space is also called the *rank* of the group and it represents the number of independent generators of the group.

Definition 2.24 *Let K be a simplicial complex. The p -th Betti number of K is the rank of the p -th homology group of K ,*

$$\beta_p(K) = \text{rang}H_p(K).$$

Example: In dimension 0, a chain is a sum of vertices and the boundary homomorphism is 0, so all chains are cycles. Two vertices that can be connected by a chain of edges in K represent the boundary of the chain, so they are homologous. H_0 thus corresponds to the connected components of K and $\beta_0(K)$ is the number of connected components.

Chapter 3

Data Reconstruction

In the analysis, the data is given as a discrete set of vectors, where each vector represents a single emergency call. We represent these vectors as points in Euclidian space \mathbb{R}^d , where d is the number of parameters that we consider. We need to connect the points in order to get a shape that the data represents and then use topological methods to get useful information about the data. In this chapter, we describe two reconstruction methods, which connect the given points in two simplicial complexes: the Čech and the Vietoris-Rips complexes. The Čech complexes were based on the theory of Čech homology, introduced by Eduard Čech [15]. A few years before simplified version, that is, Vietoris-Rips complex made the first appearance in [18] and was later discussed in [19].

3.1 The Čech Complex

The Čech complex is a special case of a general topological construction called the nerve construction. It associates a simplicial complex with a family of subsets in \mathbb{R} .

Definition 3.1 *Given a finite family of sets $\mathcal{A} = \{A_1, \dots, A_n\}$, the nerve $\mathcal{N}(\mathcal{A})$ is an abstract simplicial complex with the elements of \mathcal{A} as vertices.*

The simplices are the subsets of \mathcal{A} with non-empty common intersections.

$$\mathcal{N}(\mathcal{A}) = \{\mathcal{F} \subseteq \mathcal{A}; \bigcap \mathcal{F} \neq \emptyset\}$$

The geometric realization of the nerve of a family of sets is under certain conditions homotopy equivalent to the union of the sets. This gives us some control over the correctness of the reconstruction. There are several versions of the required conditions. For us, the relevant version is the following [20].

Theorem 3.2 (Nerve theorem) *Let \mathcal{A} be a finite family of closed subsets in \mathbb{R}^d such that all intersections $A_{i_1} \cap \dots \cap A_{i_k}$, $A_{i_j} \in \mathcal{A}$ of subfamilies of \mathcal{A} are contractible. Then*

$$\mathcal{N}(\mathcal{A}) \simeq \bigcup_{A \in \mathcal{A}} A.$$

Proof includes other topological definitions and theorems not covered in this thesis. They can be found in [20].

The Čech complex on a finite set of points $S \subseteq \mathbb{R}^d$ is the nerve of the family of balls with radius $r > 0$ and centers in the points of S .

Definition 3.3 *Let S be a set of points and radius $r > 0$. \mathcal{F} is a family of balls with the center in points from S and radius r , $\mathcal{F} = \{B_r(x) : x \in S\}$.*

The Čech complex of S is a nerve of \mathcal{F} :

$$\check{C}_r(S) = \mathcal{N}(\mathcal{F}),$$

$$\check{C}_r(S) = \{A \subseteq S; \bigcap_{x \in A} B_x(r) \neq \emptyset\}.$$

The Čech complex is an abstract simplicial complex that usually does not have a geometric realization in \mathbb{R}^d but of higher dimension.

If $r_1 < r_2$, then $\check{C}_{r_1}(S) \subseteq \check{C}_{r_2}(S)$. This means that with increasing r we, get a family of simplicial complexes with increasing dimension, but according to the nerve theorem, their geometric realizations are homotopy equivalent to the union of the balls.

3.1.1 Constructing the Čech Complex

In order to construct the Čech complex, we need to find out which points form simplices, that is, which balls with these points in the center have common intersection.

To say that balls have non-empty intersection is the same as saying that points (centers of the balls) all lie inside some ball with the same radius and the main idea of a fast algorithm to construct the Čech complex [21] relies on this fact.

The algorithm checks if a subset of points A forms a simplex in the Čech complex by computing the smallest closed ball that contains A . If this ball, we call it miniball, has radius equal or smaller than r , then A forms a simplex.

The algorithm 1 is recursive and takes two arguments, τ and ν . τ is a list of points that lie in the interior of miniball and ν is a list of points that lie on its boundary.

Algorithm 1 MiniBall(τ, ν)

```

1: if  $\tau = \emptyset$  then
2:   compute the miniball  $B$  of  $\nu$  directly
3: else
4:   choose a random point  $u \in \tau$ 
5:    $B \leftarrow \text{MiniBall}(\tau - \{u\}, \nu)$ 
6:   if  $u \notin B$  then
7:      $B \leftarrow \text{MiniBall}(\tau - \{u\}, \nu \cup \{u\})$ 
8:   end if
9: end if
10: return  $B$ 

```

At the beginning, we do not know which points lie on the boundary. That is why the algorithm with an empty set for ν (boundary) and the whole subset for τ is called.

Let us show that the expected time complexity is constant per point. ν has at most $d + 1$ points, where d is the dimension, and assuming that d is a constant, computing miniball from points in ν takes constant time. The

time complexity depends on the number of points n in τ , which is arbitrary, and on the number of points in ν that are still indefinite. Let j be that maximum possible number, that is, $j = d + 1 - |\nu|$. Let us say that the current iteration needs $t(n, j)$ time to calculate the miniball. The probability that the call $u \notin B$ returns true is $\frac{j}{n}$, which gives us the equation $t(n, j) \leq 1 + 1 + t(n - 1, j) + 1 + \frac{j}{n} \cdot t(n - 1, j - 1)$. Iterating over j we see $j = 0$ gives us linear time in n and so on to maximum j , which is $\leq d + 1$. If dimension d is a constant, algorithm takes expected linear time in n .

3.2 The Vietoris-Rips complex

The Vietoris-Rips complex represents a more efficient construction of a simplicial complex on a set S .

To simplify computing, the Vietoris-Rips complex is often used instead of the Čech complex. We only check distances between two points and if they are close enough we add the edge containing them to the complex. If three edges form a boundary of a triangle, the triangle is also added to the complex and so on to higher dimensions.

Definition 3.4 The Vietoris-Rips complex *is defined as*

$$VR_r(S) = \{\sigma \subseteq S; \text{diam}(\sigma) \leq 2r\}.$$

The drawbacks of this construction are that the dimension of the complex increases faster and that the Nerve theorem is not valid any more. This means that we have no control over the correctness. But we will show that the Vietoris-Rips complex is always contained in the Čech complex with a bigger radius.

Lemma 3.5 *Let S be a finite set of points and r radius, $r \geq 0$. Then $\check{C}_r(S) \subseteq VR_r(S) \subseteq \check{C}_{\sqrt{2}r}(S)$.*

Proof. The first part, $\check{C}_r(S) \subseteq VR_r(S)$, is obvious. For the same radius the Čech and Vietoris-Rips complexes contain the same edges. The Vietoris-Rips complex contains all the simplices that can be built from these edges, automatically including all the possible simplices in the Čech complex.

Let us now prove the second part, $VR_r(S) \subseteq \check{C}_{\sqrt{2}r}(S)$. We need to show that every simplex in the Vietoris-Rips complex fits into a ball with radius $\sqrt{2}r$. All the edges in the Vietoris-Rips complex have the length at most $2r$. The ball for a simplex of dimension d needs to be the biggest if the simplex is regular, that is, all the edges are of the same length, and all the edges are of length $2r$. For a regular simplex, the smallest ball is the circumscribed ball and its radius is $\sqrt{\frac{d}{2(d+1)}}a$, where a is the length of the edge. This equation is always smaller but limits to $\sqrt{\frac{1}{2}}a = \frac{\sqrt{2}}{2}2r = \sqrt{2}r$ as dimension goes to infinity. Every simplex in the Vietoris-Rips complex is indeed contained in a ball with radius $\sqrt{2}r$ and thus part of the Čech complex with the same radius. \square

3.2.1 The Algorithm for Constructing the Vietoris-Rips Complex

The following algorithm was proposed in [12].

The construction of the Vietoris-Rips complex is divided into two parts - making a neighborhood graph from points (generating simplices of dimension 1) and expanding this to the whole complex (generating other dimension simplices from that graph).

Constructing a neighborhood graph is a well-known problem, solved with different approaches, such as brute force, scanning, kd-trees etc.

In order to expand it to the whole complex, we will use an inductive algorithm, which glues higher dimensional simplices to lower ones.

Method 2 calculates neighbors of the point with lower values and is used in the main method 3.

The proof that this algorithm returns the Vietoris-Rips complex can be

Algorithm 2 LowerNeighbors(K, h, u)

```

1: lowerNeighbors  $\leftarrow \emptyset$ 
2: for each vertex  $u$  in  $K_0$  do
3:   if  $[u v]$  is an edge in  $K$  and  $h(v) < h(u)$  then
4:     lowerNeighbors  $\leftarrow$  lowerNeighbors  $\cup v$ 
5:   end if
6: end for
7: return lowerNeighbors

```

Algorithm 3 Expanding(K, h)

```

1: for  $d \leftarrow 1, \dots, \dim K$  do
2:   for each  $i$ -simplex  $\tau \in K$  do
3:      $N \leftarrow \bigcap_{u \in \tau} \text{LowerNeighbors}(K, h, u)$ 
4:     for each  $v \in N$  do
5:       add  $\tau * v$  to  $K$ 
6:     end for
7:   end for
8: end for

```

found in [12].

We used this algorithm to build different Vietoris-Rips complexes and analyzed them by using different topological approaches.

Chapter 4

Discrete Morse Theory

Discrete Morse theory was introduced by Robin Forman [5]. It provides an insight into the behavior of functions that are defined (or sampled) on the vertices of a simplicial complex, as well as into the shapes of the complex itself.

In the case of 911 calls, we can consider some of the parameters as data points and build a simplicial complex, while one chosen parameter can be considered as a function value. A way to analyze the behavior of the function on the whole complex is provided by discrete Morse theory.

We already have a complex in which every vertex represents a call or a group of calls and are connected in higher dimension simplices. Now we need to assign a value to every simplex. We do this by using a discrete Morse function.

Definition 4.1 *Let K be a simplicial complex and $F : K \rightarrow \mathbb{R}$ a function that assigns a value to every simplex. The function F is a discrete Morse function if for every simplex σ from K the following conditions hold:*

- *the number of faces $\tau \subset \sigma$, where $F(\tau) > F(\sigma)$ is at most 1,*
- *the number of cofaces $\nu \supset \sigma$, where $F(\nu) < F(\sigma)$ is at most 1.*

It does not take us long to realize that both numbers cannot be equal to 1 at the same time. Let us say there exist such a simplex σ . It has a

face τ with higher value $F(\tau) > F(\sigma)$ and also a coface ν with smaller value $F(\nu) < F(\sigma)$. Then a simplex σ' that is a different face of ν but also contains τ has to have a higher (or equal) value than τ and smaller (or equal) value than ν resulting in $F(\tau) \leq F(\sigma') \leq F(\nu) < F(\sigma) < F(\tau)$, which is not possible.

Except in at most one direction, a discrete Morse function increases with dimension.

An example of a discrete Morse function is given in Figure 4.1.

Definition 4.2 *Let $F : K \rightarrow \mathbb{R}$ be a discrete Morse function. A simplex is critical if the following conditions hold:*

- *number of faces $\tau \subset \sigma$, where $F(\tau) > F(\sigma) = 0$,*
- *number of cofaces $\nu \supset \sigma$, where $F(\nu) < F(\sigma) = 0$.*

Otherwise the simplex is called regular.

Critical simplices are the ones that describe the features of the data, as well as the shape of the complex.

Two regular simplices that form an exception to increasing with dimension are paired and we draw an arrow between them. The arrow points from higher to lower value. Figure 4.2 shows the arrows associated with the discrete Morse function from Figure 4.1.

Once we have the arrows, we can forget the values since all we need to know is how they relate to each other. Simplices that are neither the head nor the tail of an arrow are critical.

Now we introduce discrete vector fields, which can be considered a version of gradient vector fields of a smooth function.

Definition 4.3 *Let K be a simplicial complex. A discrete vector field V on K is a collection of pairs $(\tau^{(p)}, \sigma^{(p+1)})$, where τ and σ are simplices of K , τ is a face of σ and every simplex is in at most one pair of V .*

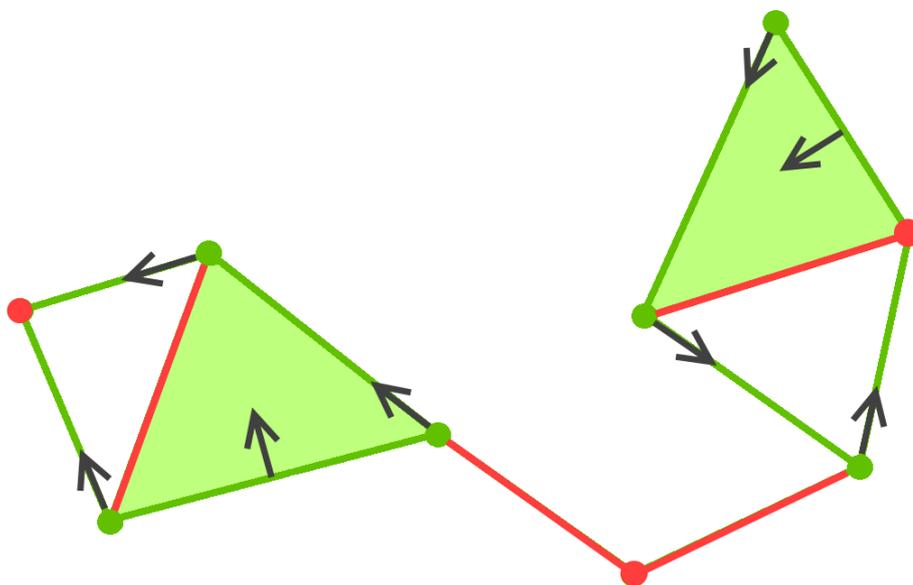


Figure 4.2: A complex with arrows instead of values.

for $i = 1, \dots, n$ and

$$F(\tau_0) \geq F(\sigma_0) > F(\tau_1) \geq F(\sigma_1) > \dots \geq F(\sigma_n) > F(\tau_{n+1}).$$

Theorem 4.6 *A discrete vector field V is a gradient vector field of some discrete Morse function F iff there are no nontrivial closed V -paths.*

4.1 The Algorithm

In [16] an algorithm is proposed that will be used to calculate the critical simplices. The great thing about this algorithm is that it only needs an injective function on the vertices and expands it to the Morse function on all simplices.

Before we go into the details of the algorithm, we need to define a few structures used in it.

Definition 4.7 *Let K be a simplicial complex and v a vertex of K . The star of v , $St(v)$, is the minimal subcomplex of K that contains all the simplices that contain v .*

Definition 4.8 *Let K be a simplicial complex, v a vertex of K and $St(v)$ its star. The link of v is the subcomplex of $St(v)$ that contains all the simplices that do not contain v .*

Let F be a Morse function on K . The lower link of v contains all the simplices of the link that have lower value of F on all its vertices.

In the algorithm, a notation for joining two simplices will be used. Let us say we have a simplicial complex K and two disjoint simplices $\sigma_1, \sigma_2 \in K$. $\sigma_1 * \sigma_2$ is either a simplex of K that has the union of vertices of σ_1 and σ_2 for its vertices or is undefined if this simplex does not exist in K .

The *Morse* algorithm takes two arguments - a simplicial complex K and an injective function h with vertices of K for its domain. It returns three sets A , B and C and a map $r : B \rightarrow A$. Sets A and B will contain the tails and the heads of arrows (the paired simplices) and r will define which simplices are paired. Of every pair the simplex with smaller dimension will be in A and the simplex with bigger dimension in B . In C there will be critical (unpaired) simplices.

The algorithm returns a relatively large set of critical simplices. In order to get fewer critical simplices, we try to pair those even further while maintaining the Morse guidelines. This is done by finding a gradient path between two simplices and reversing the arrows. Obviously, this only works if there is only one such path between them. Otherwise we would get a loop. This procedure is called cancelling critical simplices.

We call the cancelling algorithm for every dimension of the simplicial complex. We assume we have A, B, C and r calculated. Cancelling critical simplices can be done in theory for all unique paths in dimensions bigger than 1. In dimension 0 and 1 we have to be careful since it changes the order of function values. In order to avoid this problem we only cancel critical

Algorithm 4 Morse(K, h)

```

1:  $A, B, C \leftarrow \emptyset$ 
2: for each vertex  $v \in K$  do
3:    $L \leftarrow$  lower link of  $v$ 
4:   if  $L$  is empty then
5:     add  $v$  to  $C$ 
6:   else
7:     add  $v$  to  $A$ 
8:      $(A', B', C', r') \leftarrow$  Morse( $L, h'$ ), where  $h'$  is the restriction of  $h$ 
9:      $w \leftarrow$  the vertex in  $C'$  with the smallest function value
10:    add  $v * w$  to  $B$ 
11:    define  $r(v * w) = v$ 
12:    for each simplex  $\sigma \in C' - w$  do
13:      add  $v * \sigma$  to  $C$ 
14:    end for
15:    for each simplex  $\sigma \in B'$  do
16:      add  $v * \sigma$  to  $B$ 
17:      add  $v * r'(\sigma)$  to  $A$ 
18:      define  $r(v * \sigma) = v * r'(\sigma)$ 
19:    end for
20:  end if
21: end for

```

simplices that have function values differing by less than a threshold. In the algorithm 5 $maxh(\sigma)$ denotes the maximum function value of all of the vertices of σ .

Algorithm 5 PairCritical(K, h, p)

```

1: for  $d \leftarrow 1, \dots, dimK$  do
2:   for each critical  $d$ -simplex  $\sigma$  do
3:     find all gradient paths from  $\sigma$  to  $d-1$ -simplices  $\tau_i$ , where  $maxh(\tau) > maxh(\sigma) - p$ 

4:     for each  $\tau_i$  do
5:       if  $\tau_i$  does not equal any other  $\tau_j$  then
6:          $m_i = maxh(\tau_i)$ 
7:       end if
8:     end for
9:     if there exists any  $m_i$  then
10:      find  $M$ , such that  $m_M = min m_i$ 
11:      find the unique gradient path from  $\sigma$  to  $\tau_M$ :
            $\sigma = \sigma_1 \rightarrow \tau_1 \rightarrow \sigma_2 \rightarrow \dots \rightarrow \sigma_j \rightarrow \tau_j = \tau_M$ 
12:      delete  $\sigma$  and  $\tau_M$  from  $C$ 
13:      add  $\sigma$  to  $B$ 
14:      add  $\tau_M$  to  $A$ 
15:      for  $k = 1, \dots, j$  do
16:        redefine  $r(\sigma_k) = \tau_k$ 
17:      end for
18:    end if
19:  end for
20: end for

```

The next two theorems allow us to construct a discrete Morse function on a simplicial complex.

Theorem 4.9 *The gradient field produced by Morse algorithm has no directed loops.*

Proof. Let us say there is a directed loop. Since arrows never point to a higher value, it means every simplex has the same value $maxh$. The h function is injective, which means all the simplices in the loop share a vertex

v . The algorithm applies the same structure to the higher dimension, which means simplices in the lower link of v also form a loop, which is by induction impossible. \square

Theorem 4.10 *The PairCritical algorithm does not produce any directed loops.*

Proof. We have A, B, C and r produced by the Morse algorithm with no loops. The algorithm can only produce a loop when reversing arrows between two critical simplices, from σ to τ . They are joined by a unique gradient path. Let us say we produced a loop α . Since there was no loop before, a part of the loop γ coincides with a part of the original gradient path η . But that means we could replace the part of the gradient path η with $\alpha - \gamma$ and get another path from σ to τ , which is not possible. \square

The PairAlgorithm searches for paths that start in the higher and end in the lower dimensional simplices. The following lemma shows we can always start in a higher dimensional simplex and will not stop at the simplex of the same dimension.

The following lemma is used in PairCritical algorithm and helps us reduce the number of gradient paths between critical simplices.

Lemma 4.11 *There is no such simplex $\sigma \in C_i, i > 0$, such that all its codimension-one faces are in B_{i-1} .*

Proof. Let us say we have an i -simplex $\sigma \in C_i, i > 0$ all its $(i - 1)$ -faces τ_0, \dots, τ_i are in B_{i-1} . Then we have $(i - 2)$ -simplices v_0, \dots, v_i , such that $r(\tau_j) = v_j$ for $j = 0, \dots, i$. Every v_j belongs to exactly two $(i - 1)$ -faces of σ , which means we can construct a loop

$$v_0 \rightarrow \tau_0 \rightarrow v_{j_1} \rightarrow \dots \rightarrow \tau_{j_i} \rightarrow v_0,$$

which is not possible according to the previous theorems. \square

In the algorithm, we only used the function values on the vertices. The next theorem shows we can construct a discrete Morse function from the field we generate with the algorithm.

This algorithm calculates Morse field without extending h to Morse function. In the article [16], they proved a theorem that h can be extended to a Morse function F that produces the same Morse field as the Morse algorithm.

Chapter 5

Persistent Homology

One of the most fruitful methods of topological data analysis is persistent homology or, in short, persistence. Persistence is a method which allows analyzing data at different resolutions. The idea is to build the simplicial complex step by step by defining a filtration, and to study how homology classes are born and die during the process. Classes that persist longer correspond to stronger features in the data, while classes with short persistence correspond to details or even noise.

Let K_i be subcomplexes of K , such that

$$\emptyset = K_0 \subset K_1 \subset \cdots \subset K_n = K.$$

Inclusion of K_i into K_{i+1} produces a homomorphism of homology groups $H_p(K_i) \rightarrow H_p(K_{i+1})$ for all p .

For every p we then get a sequence of homomorphisms

$$0 = H_p(K_0) \rightarrow H_p(K_1) \rightarrow \cdots \rightarrow H_p(K_n) = H_p(K),$$

and when we go from K_i to K_{i+1} we gain new homology classes and we lose those that merged with other classes or became trivial. We collect the classes that are born at or before a given threshold and die after another threshold in groups.

Definition 5.1 *The p -th persistent homology group is defined as $H_p^{i,j} = f_*^{i,j}(H_p(K_i)) \subseteq H_p(K_j)$ or equivalent $H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i))$.*

The p -th persistent Betti number is the rank of corresponding group, $\beta_p^{i,j} = \text{rank} H_p^{i,j}$.

A homology class $\gamma \in H_p(K_i)$ is *born* in K_i if $\gamma \notin H_p^{i-1,j}$.

A homology class born in K_i *dies* in K_j , $j > i$, if it merges with an older class in K_j , $f_*^{i,j-1}(\gamma) \in H_p(K_{j-1})$, $f_*^{i,j}(\gamma) = f_*^{i',j}(\gamma') \in H_p(K_j)$.

If at K_j two classes merge, the older one remains.

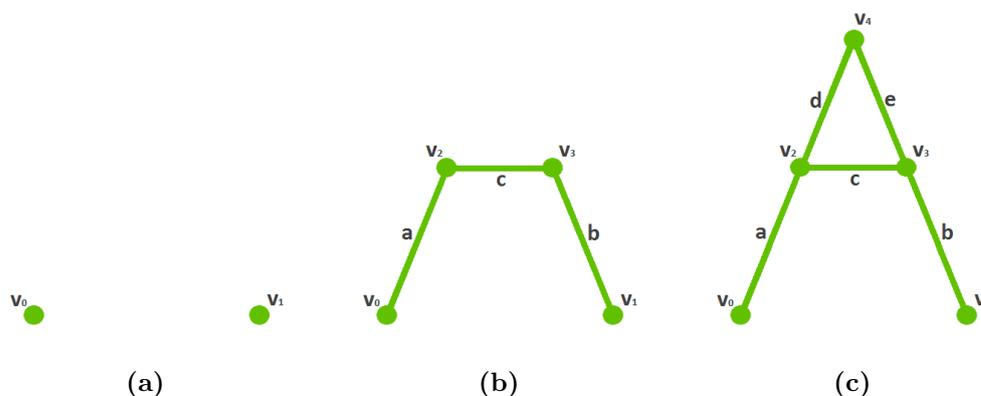


Figure 5.1: Building of letter A.

In the Figure 5.1, we can see how the letter A is built if each complex in the sequence includes all simplices with height h or lower. At the first stage (a), two 0-dimensional homology classes v_0 and v_1 are born. After that, no new classes are born until we reach the height of the 'bridge' of A (stage (b)). At this height, class v_1 dies because it merges with v_0 . The last stage (c) shows the complete letter A. In dimension 0, nothing changes, we still have only one class. But in dimension 1, the cycle $c + d + e$ represents a new 1-dimensional homology class. The Betti numbers of the letter A are $\beta_0(A) = 1$ and $\beta_1(A) = 1$ and the generators are represented by the homology classes v_0 and $c + d + e$, which never die.

Definition 5.2 Persistence is the difference $\text{pers}(\gamma) = j - i$, where γ is born at K_i and dies entering K_j .

If γ is born but never dies, its persistence is ∞ .

In order to visualize persistence, we use *persistence diagrams*. For a homology class with persistence $j - i$ we draw a point on a plane with coordinates (i, j) . If several classes have persistence (i, j) , we label the point in the diagram with its multiplicity which represents the number of such classes. If the stages K_i of the filtration correspond to function values, persistence can also be measured according to the difference of the values.

For the homology classes that were born at K_i but never die, we draw a ray upwards from (a_i, a_i) .

Lemma 5.3 For every $p \geq 0$ and every pair of indices $k \leq l$ in a filtration

$$\emptyset = K_0 \xrightarrow{f^0} K_1 \xrightarrow{f^1} \dots \xrightarrow{f^{n-1}} K_n = K$$

the p -th Betti number is

$$\beta_p^{k,l} = \sum_{i \leq k} \sum_{j > l} \mu_p^{i,j}.$$

Figure 5.2 shows a persistence diagram for the letter A built as in Figure 5.1, (a) for dimension 0 and (b) for dimension 1.

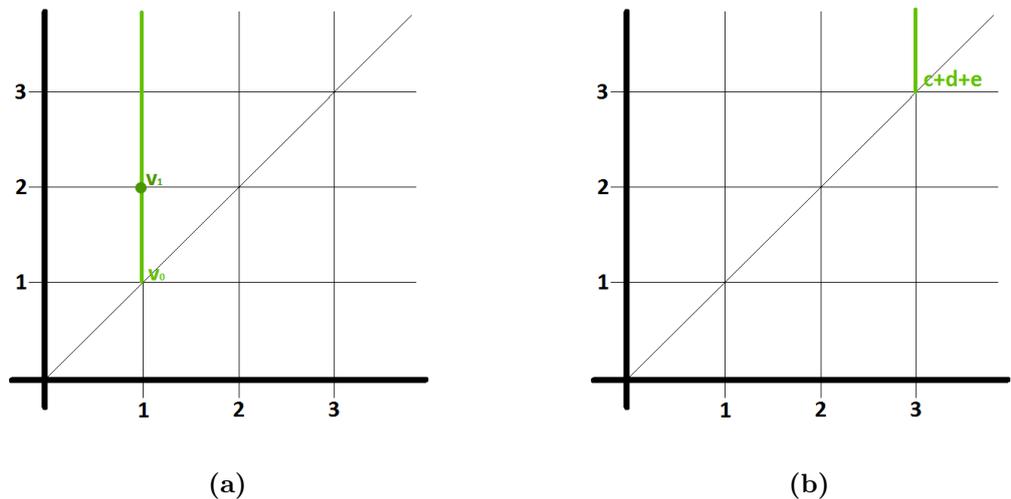


Figure 5.2: Building of letter A.

Chapter 6

Implementation and results

In this thesis we try to extract information from the data on 911 calls. We obtained the data from the Department of Sociology at Harvard University. There are approximately 1.2 million 911 calls from November 2010 to November 2012. Each call has many characteristics, these are some of them:

- id of the call,
- type - what is the reason behind the call, e.g. vandalism in progress, cardiac disorder, vehicle alarm triggered, etc.
- priority on a scale from 1 to 9, 1 being the highest priority
- dates and times - when did the call occur, when did the dispatch arrive on the scene, etc.
- address
- neighborhood
- coordinates

Our basic reconstruction model was the Vietoris-Rips complex. The distance between calls or groups of calls can be computed based on the geographic proximity or some other measure, for example, the difference in the

number of calls, the distance between the call vectors incorporating various parameters, etc. The Vietoris-Rips complex can be built on vertices that represent individual calls or groups of similar calls, grouped either by proximity or other similarity measures. This enables studying the distribution of calls at different resolutions and at the same time vastly reduces the number of vertices and the complexity of calculations.

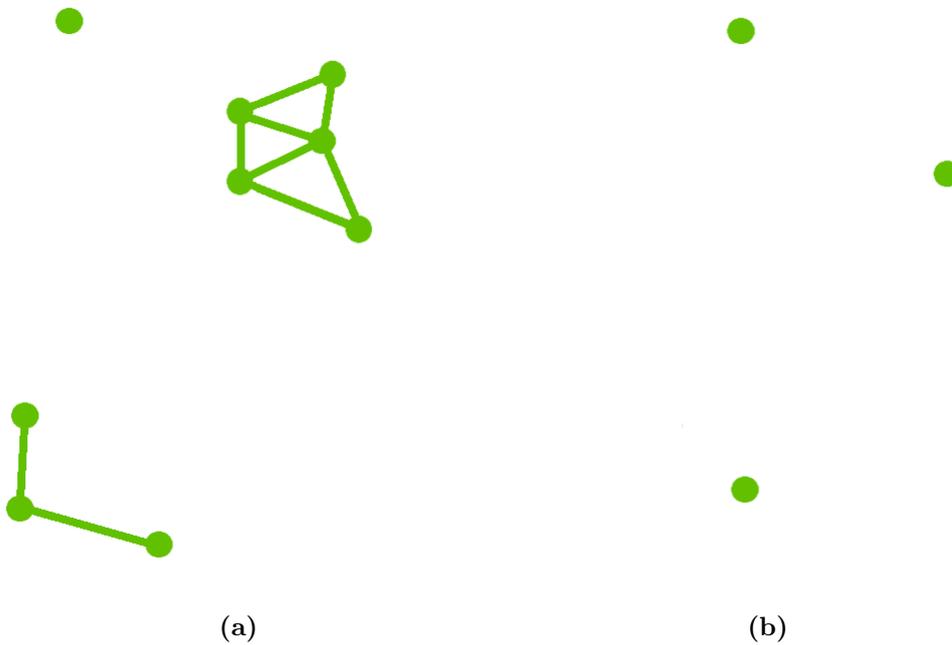


Figure 6.1: Each component of the Vietoris-Rips complex represented by a single point.

An algorithm for constructing the Vietoris-Rips complex based on [12] was implemented. In order to apply discrete Morse theory to the data, the algorithm was upgraded so that the function values were added simultaneously with the construction. Since the algorithm for expanding these values to a discrete vector field requires an injective function on the vertices. To achieve injectivity, noise was added in such a way that no numbers repeated themselves.

The algorithm for calculating Morse field was implemented as described in chapter 4.

Persistence diagrams were produced by running the algorithm with different parameter values.

Our algorithms are implemented in Java. The number of vertices changes, depending on the information we are looking for. Most often approximately 100.000 calls at a time are analyzed, which takes a few minutes to calculate on a standard laptop.

We performed multiple experiments, which can be arranged into two groups based on the methods used: analyzing the critical simplices produced by the Morse algorithm and analyzing persistence diagrams.

6.1 Results Based on Critical Simplices

Boston is the largest city and capital of Massachusetts in the United States of America. It has an estimated population of 650.000 people and covers an area of $124km^2$. During work hours, there are about 1.2 million people in the city. The median age of citizens is 30.8 years. About 45% of its population is Non-Hispanic White, 24% Black, 18% Hispanic or Latino, 9% Asian [22, 23].

The critical simplices of a complex describe the characteristic features of the complex. For example, the critical simplices with the highest values represent local maxima and point to areas with the most disturbances. On the other side, simplices with the lowest values point at seemingly most peaceful areas. Saddles are the critical simplices where in one direction values rise and in a different direction values fall. They might represent borders between more peaceful areas and areas with more disturbances.

The critical simplices of a complex can be analyzed in different ways. We can look at each critical simplex independently and look at its characteristics, compare the critical simplices, find the simplices representing the local maxima and minima, look at the whole structure and see where they arise etc.

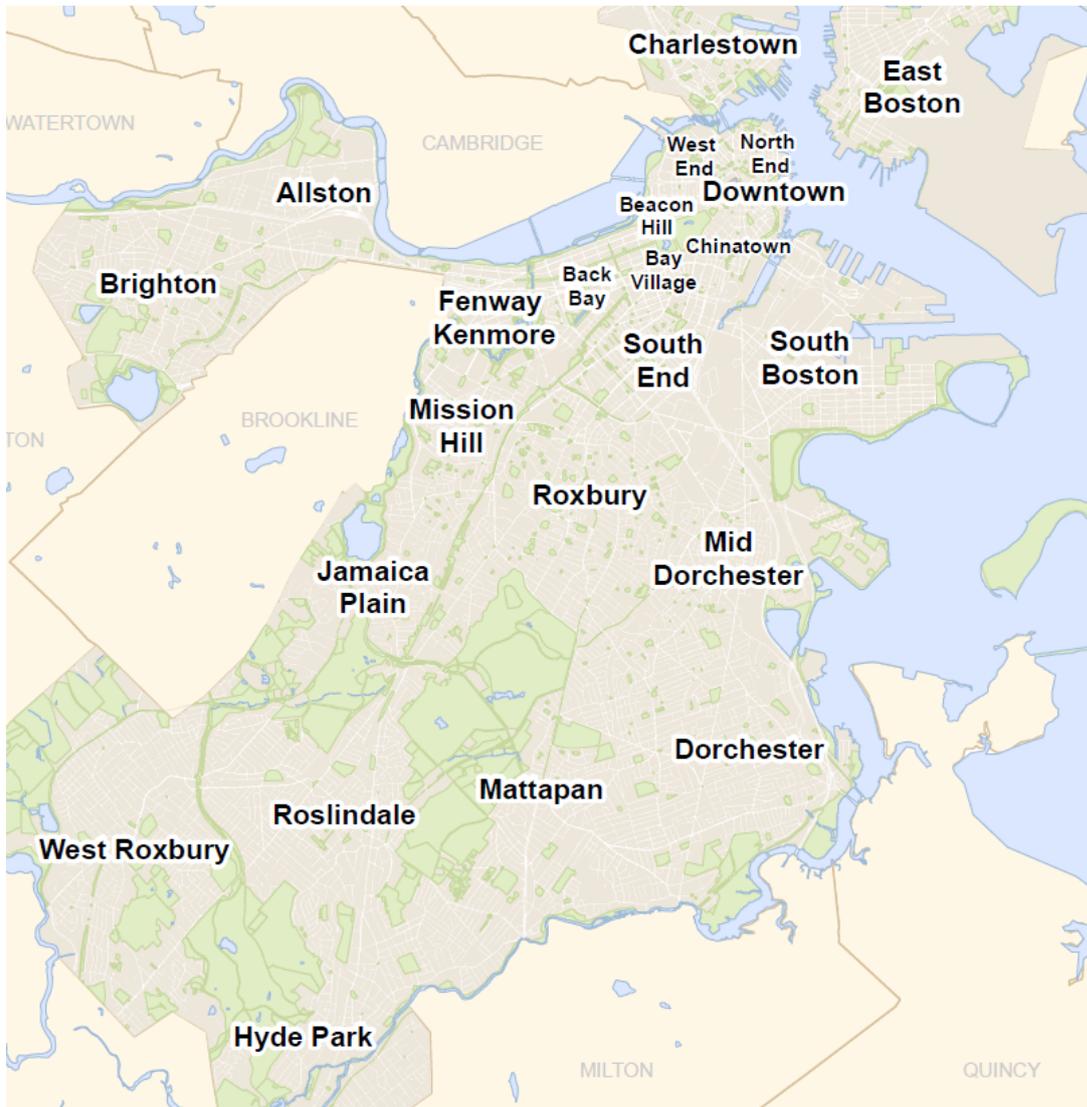


Figure 6.2: Boston neighborhoods [24].

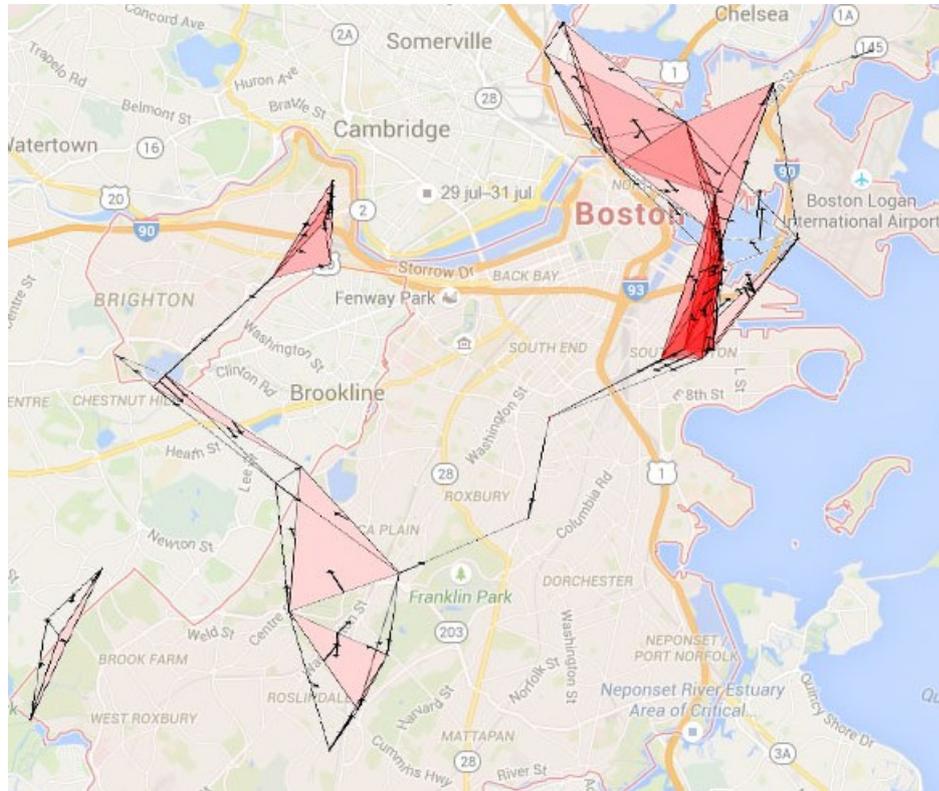


Figure 6.3: Boston.

6.1.1 The Vietoris-Rips Complex of Boston Using Geographic Distance

Our very first experiment was an attempt to build a Vietoris-Rips complex with vertices in all call locations in the whole Boston area. This problem turned out to be too complex to calculate because of the size. Because of that, calls were grouped by location using geographic proximity as the distance measure and the number of calls as function values on the vertices.

Figure 6.3 shows the complex of Boston. We can see how it falls apart into different components that match the rough outline of the city and point to areas with most disturbances. The problem of this model is that we get simplices with great differences in values that are not representative enough. Some components have only a few calls and others have more than a few

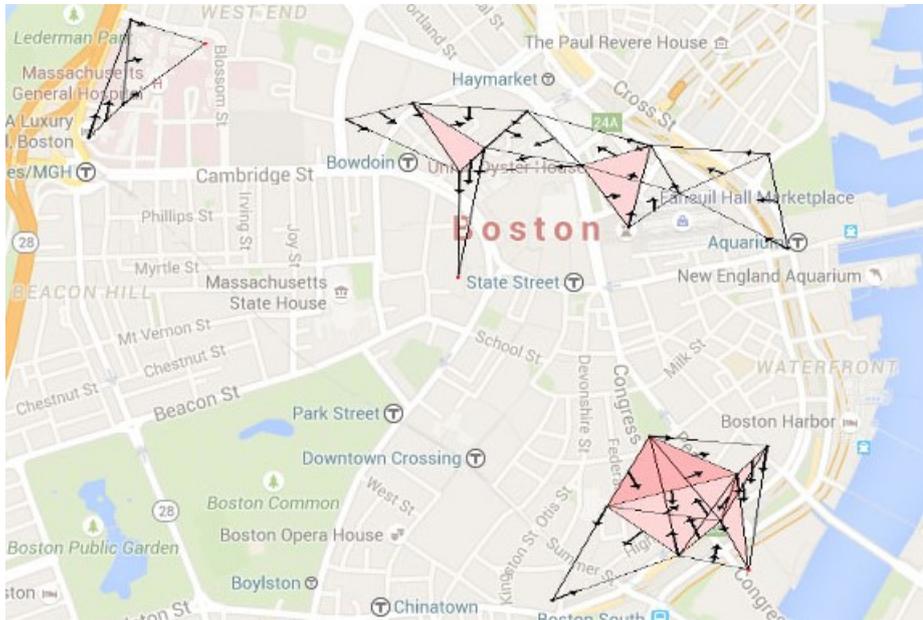


Figure 6.4: Complex of Downtown, where simplices are connected if they are close to each other.

hundred thousands of calls which only tell us big areas are represented by them. Some components are much more highly connected than others, which makes them impossible to connect.

One such interesting area lies roughly above Downtown. Most of the neighborhood areas seem to connect to it, so we take a closer look.

6.1.2 The Vietoris-Rips Complex of Downtown Using Geographic Distance

Figure 6.4 shows a complex of Downtown, Boston's central business district, also using geographic distance. Surprisingly, the calls are much more evenly distributed and the structure is clearer. This may be due to the fact that it is nearly impossible to find a quiet part of Downtown. It buzzes with activity day and night. There are a lot of business buildings where security is heightened, bars where people complain about the noise etc.

There are a few critical triangles in 6.4 and they all represent local maxima. If we wanted to analyze the part more closely with many critical simplices intertwined, we would again just change the resolution and only look at that part.

Looking at critical triangles, local maxima, they all have only one vertex with a much higher value, which means areas represented by other two vertices only seem prone to accidents, crime and other disturbances by association.

6.1.3 The Vietoris-Rips Complex of Downtown Using the Number of Calls as the Distance

The Vietoris-Rips complex was built on the calls made from the Downtown area grouped by location using the number of calls as the distance measure and as function values on the vertices.

Figure 6.5 also shows Downtown, but this time two simplices are said to be close if the difference between their numbers of calls is small. Because simplices can now be connected even if they lie on the other part of the neighborhood, we removed a few vertices that only had one or two calls made from it.

Here, the simplices connected in larger components all have from 50 to 400 calls, but there are also a few simplices by themselves and those have much bigger numbers, from 600 to 5000 calls. As we can see we can find those randomly through all Downtown. This might mean that areas with the most calls lie by themselves all around Downtown and are surrounded with areas with fewer disturbances (but not necessarily peaceful). These parts might be empty roads, parks and other less populated areas.

This distance would be the most useful at the areas with evenly distributed calls, so we could also analyze the paths between them.

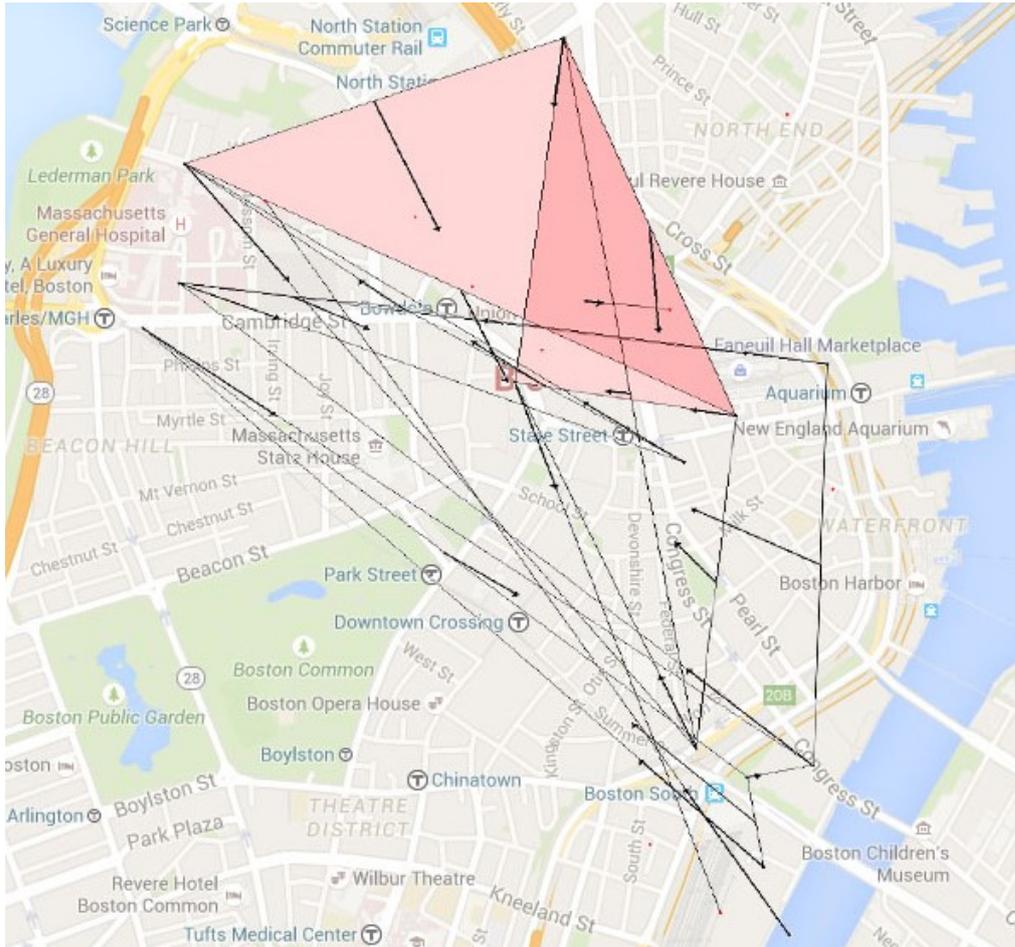


Figure 6.5: Complex of Downtown, where simplices are connected if they have a similar number of calls.

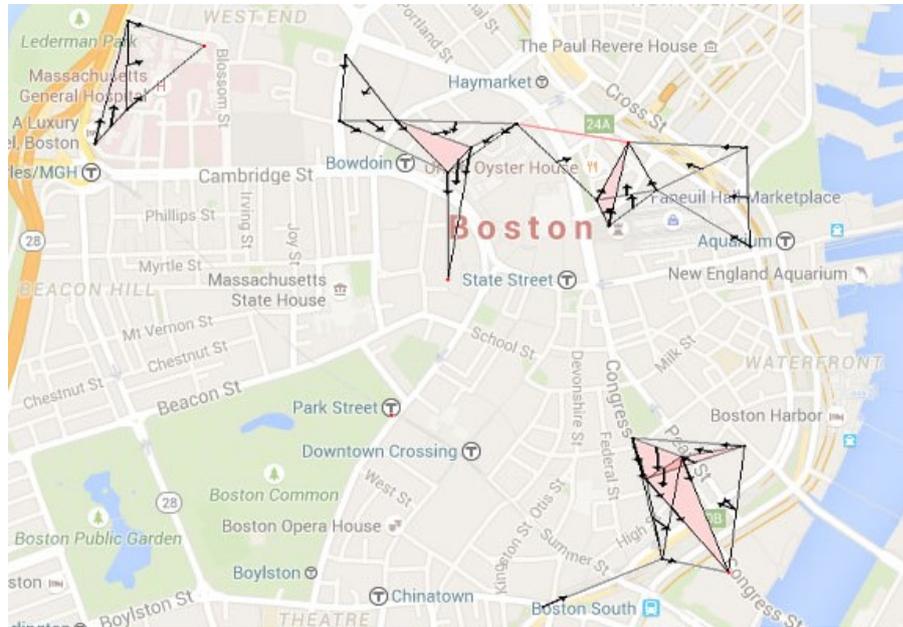


Figure 6.6: Complex of Downtown with a critical edge.

6.1.4 The Vietoris-Rips Complex of Downtown using the Geographic Distance and the Changing Parameter for Cancelling Critical Simplices

The algorithm uses the parameter that determines when two critical simplices will be paired. By increasing the threshold we can see new critical simplices appearing, especially critical edges. Edges represent areas that split the local surrounding areas into areas with more and with fewer disturbances.

Figure 6.6 shows a complex where we only pair two simplices if their values of Morse function are similar. Geographic distance is used here.

We got a new critical edge. Its boundary vertices are connected by other non-critical paths, which means this saddle is not that significant but still points to a bigger value difference. This might be an area between a park that would represent the area with fewer disturbances, and a shopping center with a larger number of people on the area of the same size and a higher probability of accidents and crime.

6.2 Results Based on Persistence Diagrams

We were interested in comparing different neighborhoods and we tried analyzing and comparing them with the use of persistence.

Persistence parameters can be changed in multiple ways, which gives us many possibilities for the analysis. It illustrates how a complex is built by changing the parameter values. We can see how long the components live for. The ones that have long lives point to the areas that represent focal points in the neighborhood, and those with short lives point to details or even just to the noise in the data.

In the next sections, we compared the characteristics of the calls made in different seasons. We also wanted to see what happens if we only look at some categories of calls. Do shootings, for instance, make a complex with different characteristics than medical emergencies.

Different shades of green represent a different component size - the darker the color, the bigger the component.

6.2.1 Summer vs Winter

We analyzed data from two winters and two summers in Roxbury, South End and South Boston. We wanted to see if there are any similarities between winter and summer diagrams or not.

Diagrams 6.7 and 6.8 show us how the complex is built using the number of calls as its parameter.

The left column is for winters and the right column for summers. Diagrams 6.7 are diagrams for Roxbury and diagrams 6.8 for South End and South Boston.

Roxbury is populated largely by African Americans, Caribbean Americans, and Latinos and is historically the center of Boston's black community.

The South End is the center of the city's LGBT population and also populated by artists and young professionals as well as a vibrant African American community, while South Boston is a predominantly Irish-American

neighborhood.

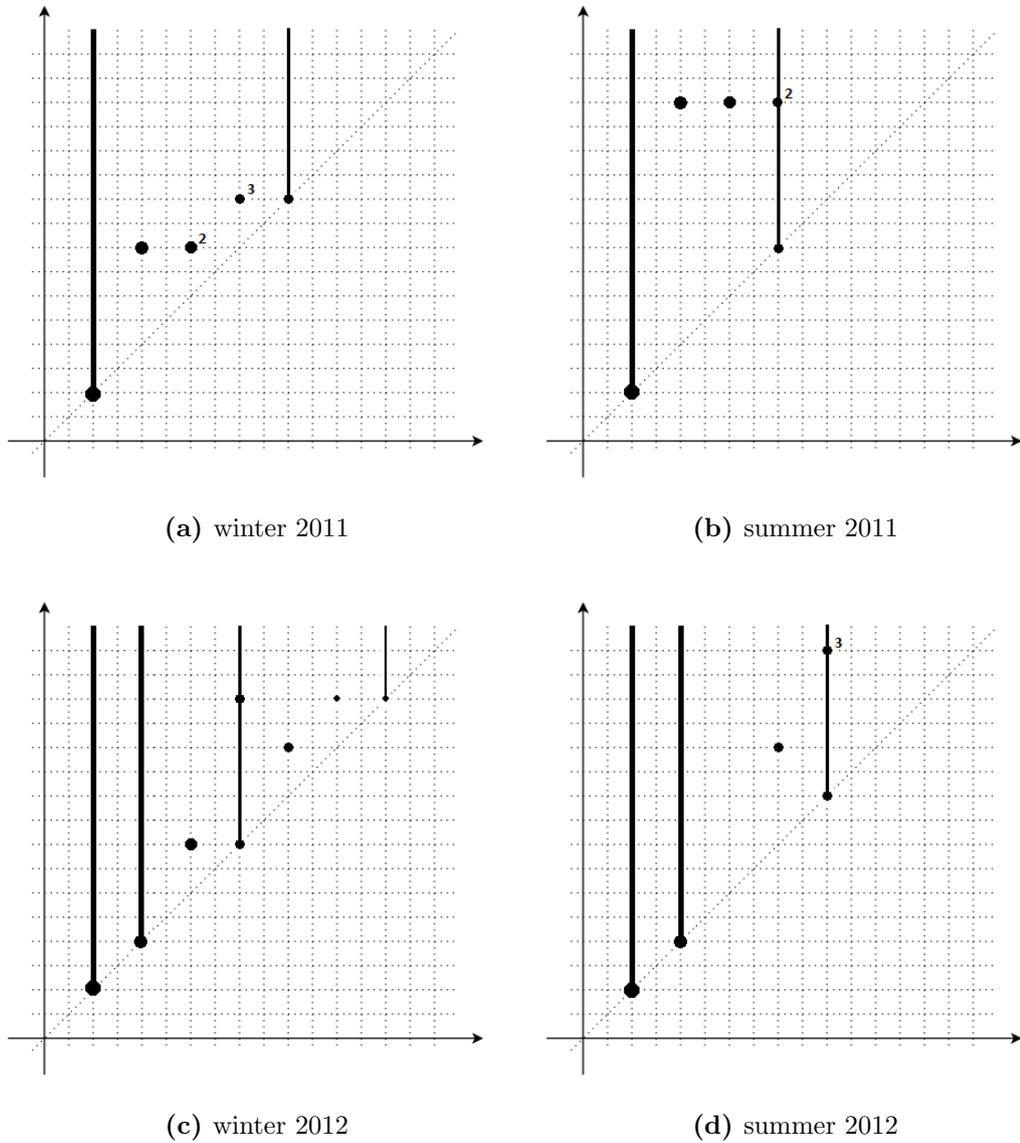


Figure 6.7: Roxbury

Since the parameter used is the number of calls, the components will be born in the same order as their sizes from bigger to smaller.

Comparing winter and summer diagrams, we can notice a slight change in the duration of components' lives. In winter, it seems that new components

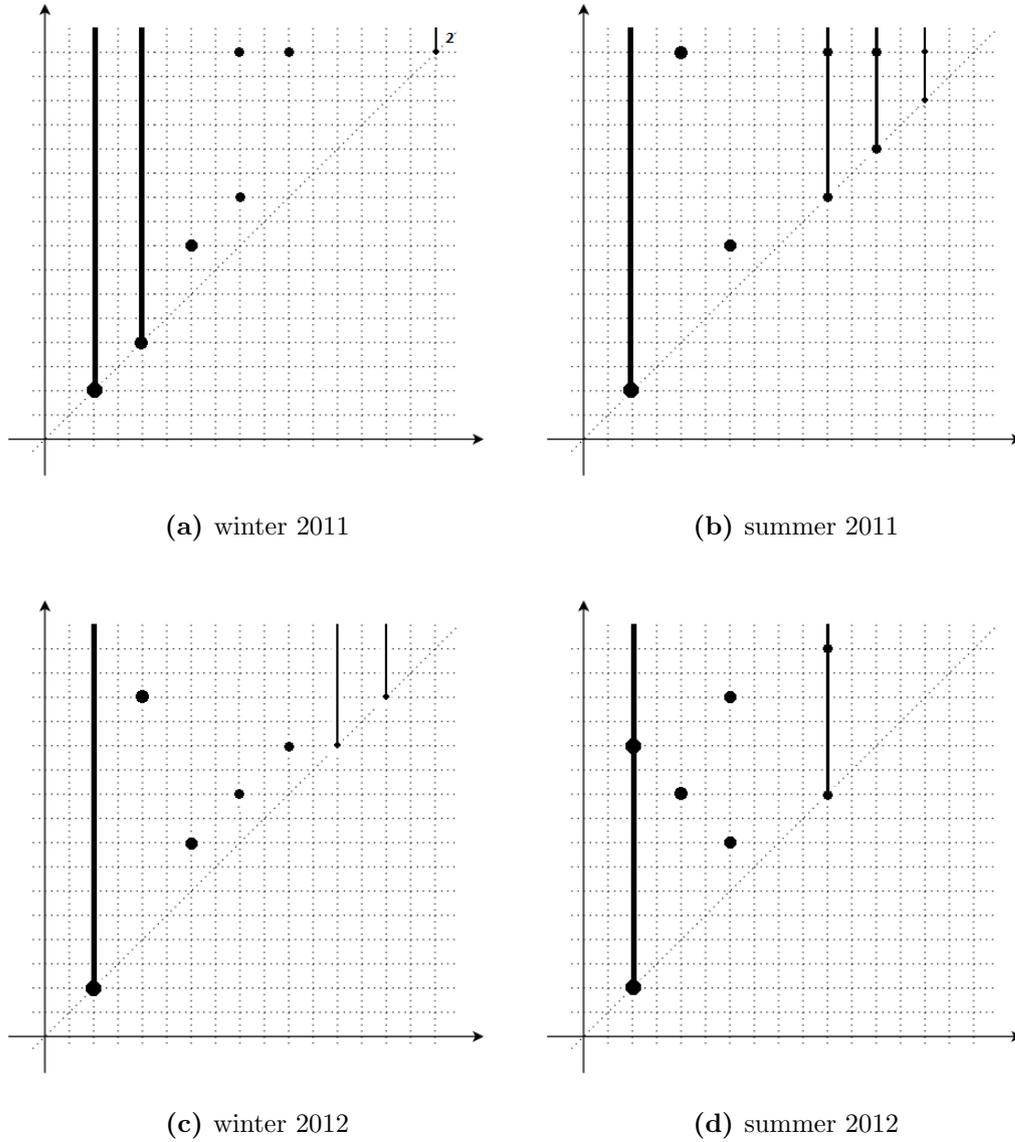


Figure 6.8: South End and South Boston

die quicker than in summer. This means that the calls were made more randomly and we have less obvious focal points, while in winter we have a few focal points to which other components are quickly connected.

This might happen, for instance, because people tend to stay at a limited number of places in winter due to colder temperatures, while in summer,

they are much more active.

6.2.2 Call Categories

We would imagine different types of calls produce different diagrams since the connection between parties that are too loud and seizures does not seem obvious. We wanted to see if the diagrams for the same category of calls but from different areas have any similarities.

Every call has a type that tells us what kind of situation was the reason behind the call. Maybe it was an armed robbery, an illness, panhandling etc. Each call only belongs to one type but can be assigned to multiple (or none) categories. We analyzed three categories - major medical emergencies, private conflict and guns.

Four neighborhoods analyzed in these parts are:

- Brighton, a neighborhood, heavily populated by students from nearby universities as well as recent graduates.
- Hyde Park, a neighborhood with a distinct suburban feel,
- Jamaica, a community of white professionals and Latinos,
- Mattapan, a neighborhood with Boston's highest concentrations of African Americans.

We have built diagrams using two parameters, the number of calls and the times at which the calls occurred. If a component contains a call that was made early on, this component will also be born early on. This also means that components with a bigger number of calls have a better chance of being born earlier than components with fewer calls.

Major Medical Emergencies

Major medical emergencies indicates events that reflect major medical emergencies (e.g., stroke). This category contains calls with these types:

- ANAPHX - Anaphylaxis,
- CARST - Cardiac Arrest,
- CARDIS - Cardiac Disorder,
- DIABET - Diabetic,
- DIFFBR - Difficulty Breathing,
- ILL1 - Illness 1,
- ILL23RATE - Illness,
- INJ23 - Injury,
- ODRATE - Drug Overdose,
- REACT - Reaction to Rx or Sting,
- SEIZRATE - Seizure,
- STROKE - Stroke.

Diagrams 6.9 show the persistence in each city where the number of calls determines the births of the simplices.

We can see that in all neighborhoods we got a relatively large number of components and that most of them did not die early. This means we have multiple focal points around the neighborhood where medical emergencies occur, with Brighton having the most of such points. Focal points show areas where accidents and other medical emergencies are more likely to occur.

On the other hand, if we look at the persistence based on time, we get diagrams 6.10 that show a totally different story.

Most of the components are born in the first stages and there are a lot of them. In some neighborhoods, they last for a long time (Hyde Park, Mattapan), which points to more randomness in the data, while in Jamaica they merge much quicker. This can point to more connected areas or more evenly distributed places with emergencies more likely to happen.

In the later stages, not many independent simplices are born. This means that we get a clear picture of the structure very quickly in time and later on only fill out the details.

Private Conflict

Private Conflict indicates the prevalence of events that reflect interpersonal conflict in the neighborhood (e.g., domestic violence). Types:

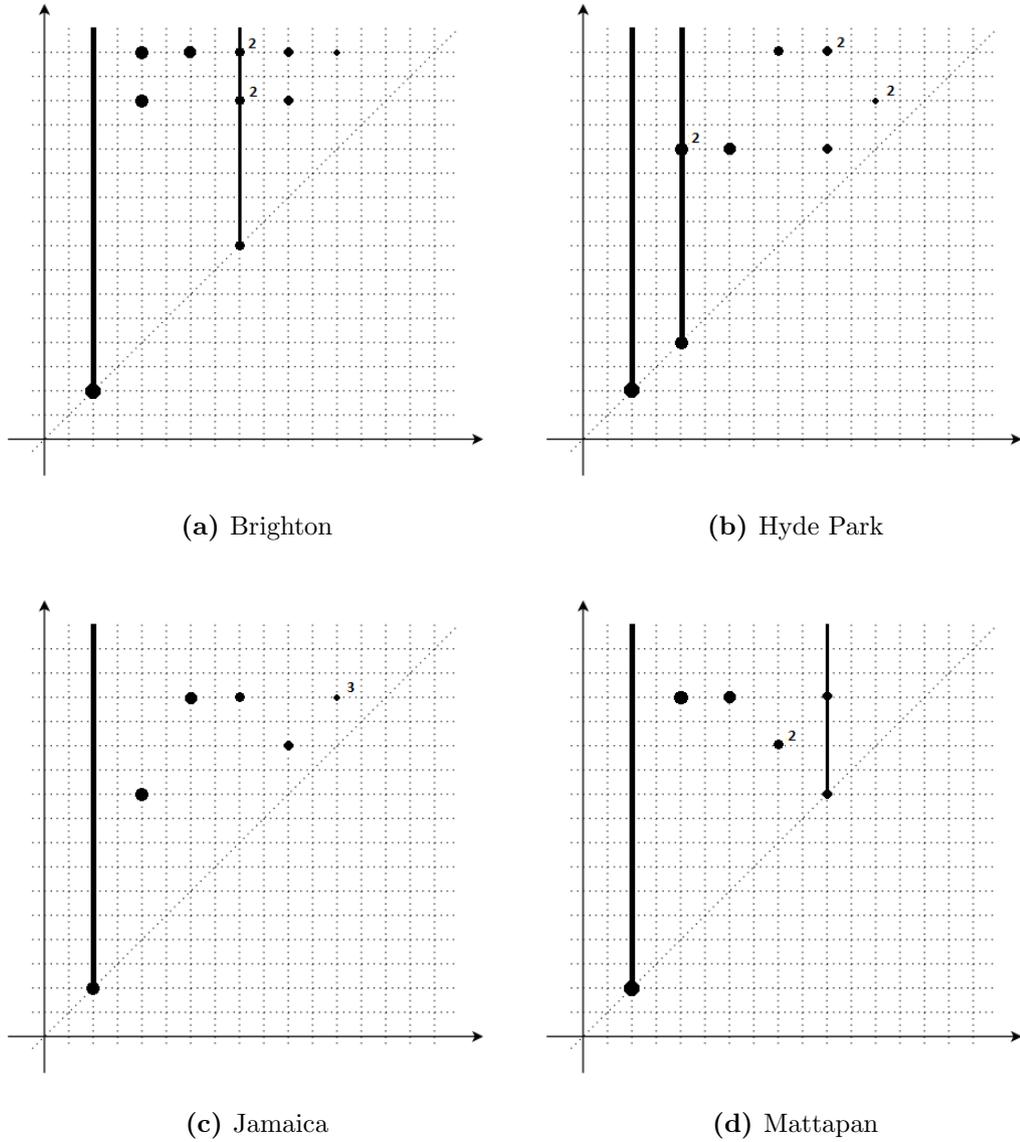


Figure 6.9: Major Medical Emergencies persistent diagrams based on the number of calls.

- BEIP - Breaking/Entering in Progress,
- DISTRB - Disturbance,
- DVIP - Domestic Violence,
- IVDRUG - Investigate Drug Location,

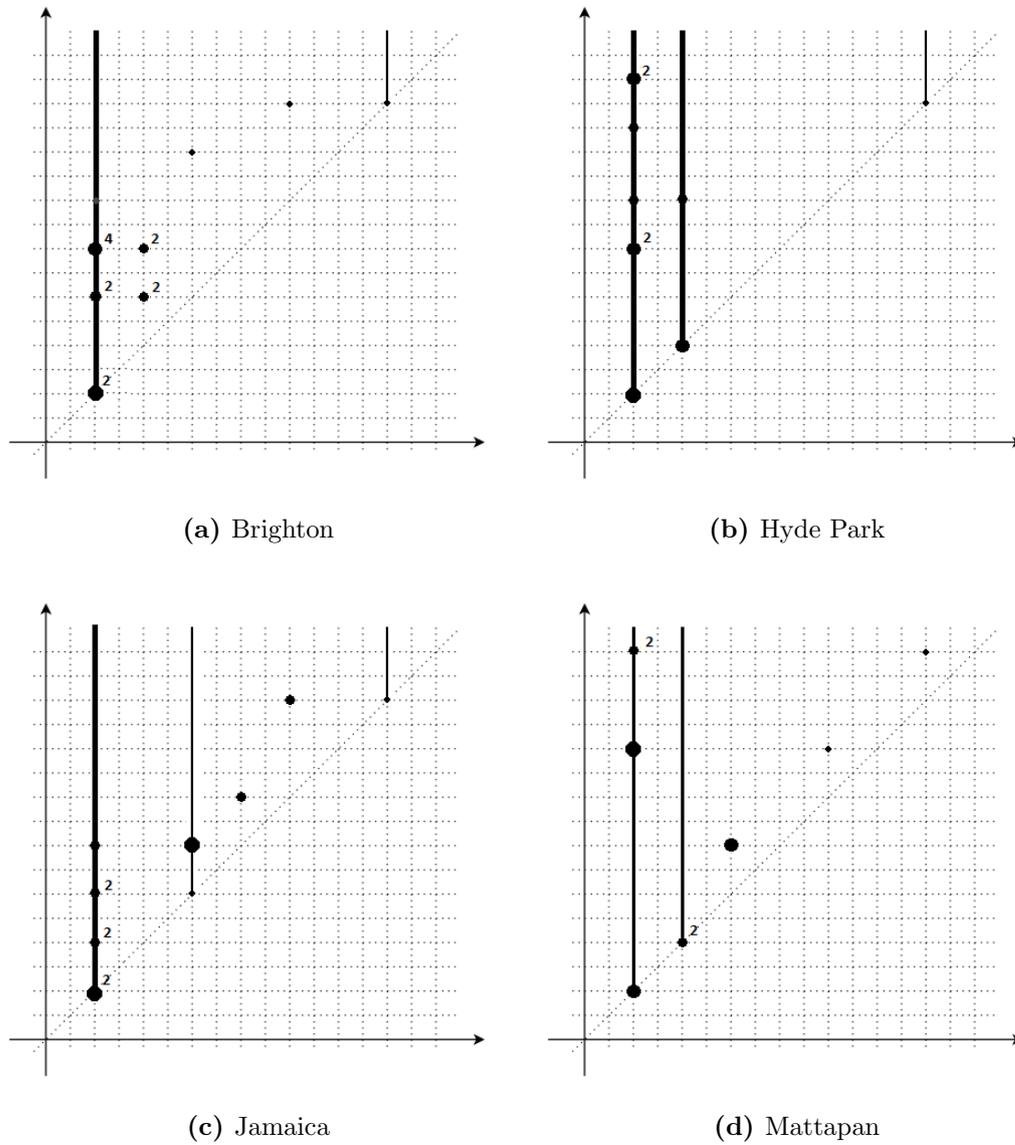


Figure 6.10: Major Medical Emergencies persistence diagrams based on time.

- LANTEN - Landlord/Tenant Trouble,
- PARTYRATE - Party, Loud Music,
- VANRPT - Vandalism Report,
- VIORDR - Violent Restraining Order.

Diagrams 6.11 show a difference between Hyde Park, the suburban neighborhood and other neighborhoods.

What they all have in common is that there are not many components appearing. They are limited to residential areas. However, Hyde Park has only one big component, which means that simplices are connected as soon as they are born. This might be because of a certain area in which parties and/or acts of vandalism are much more common than in other parts of the neighborhood. It pulls other components to it.

Persistence diagrams based on time 6.12 are much more similar to each other than those based on the number of calls. This again points to having a basic structure very quickly. We can see that these structures slightly vary from neighborhood to neighborhood, but are actually very similar to the structures of major medical emergencies diagrams of the same district. These diagrams are a lot more dependent on the structure of the area than on particular types of calls.

Guns

Guns, indicates the prevalence of events that involve the use of guns (e.g., shooting). Types:

- ABDWIP - Assault/Battery with a Dangerous Weapon in Progress,
- RATESHOT - Person Shot,
- PERGUN - Person with Gun,
- SHOTS - Shots Fired.

Here it needs to be noted that the sizes of components are much smaller than in the previous two categories. Events involving the use of guns are much rarer. This means that an area represented by a single point usually has 30 calls at the most made from it in the past two years, while in the other two categories the numbers are in hundreds and thousands of calls.

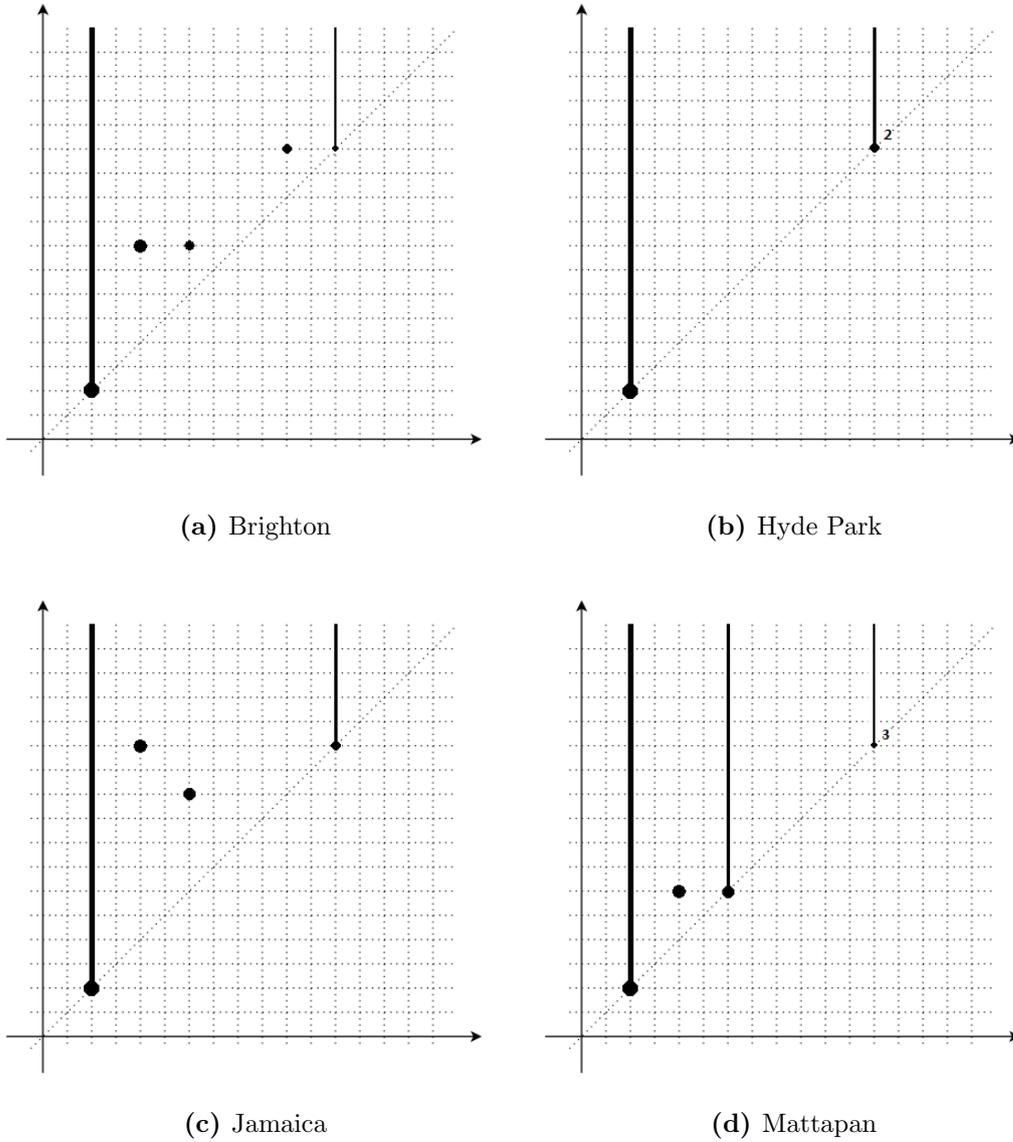


Figure 6.11: Private Conflict persistent diagrams based on the number of calls.

Analyzing data about guns, we tried to distinguish between random occurrences involving guns and dangerous areas of the district.

When looking at how simplices are born and merged by the components' sizes in diagrams 6.13, Jamaica's structure is the most random with more

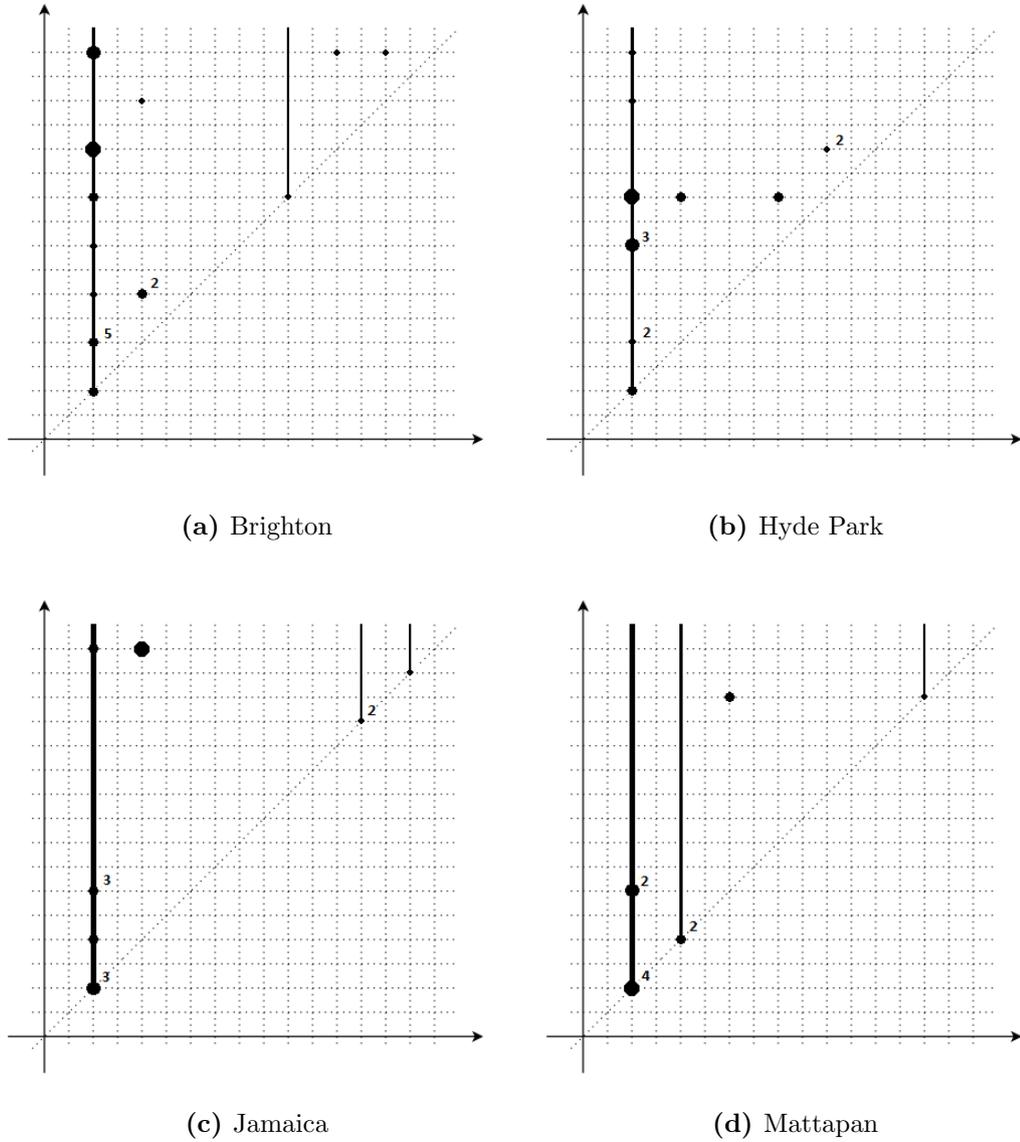


Figure 6.12: Private Conflict persistence diagrams based on time.

components than the rest, while Mattapan's first component is a lot larger than all the rest, which could point to a dangerous area.

The diagrams 6.14 also show that Hyde Park does not have a strong center.

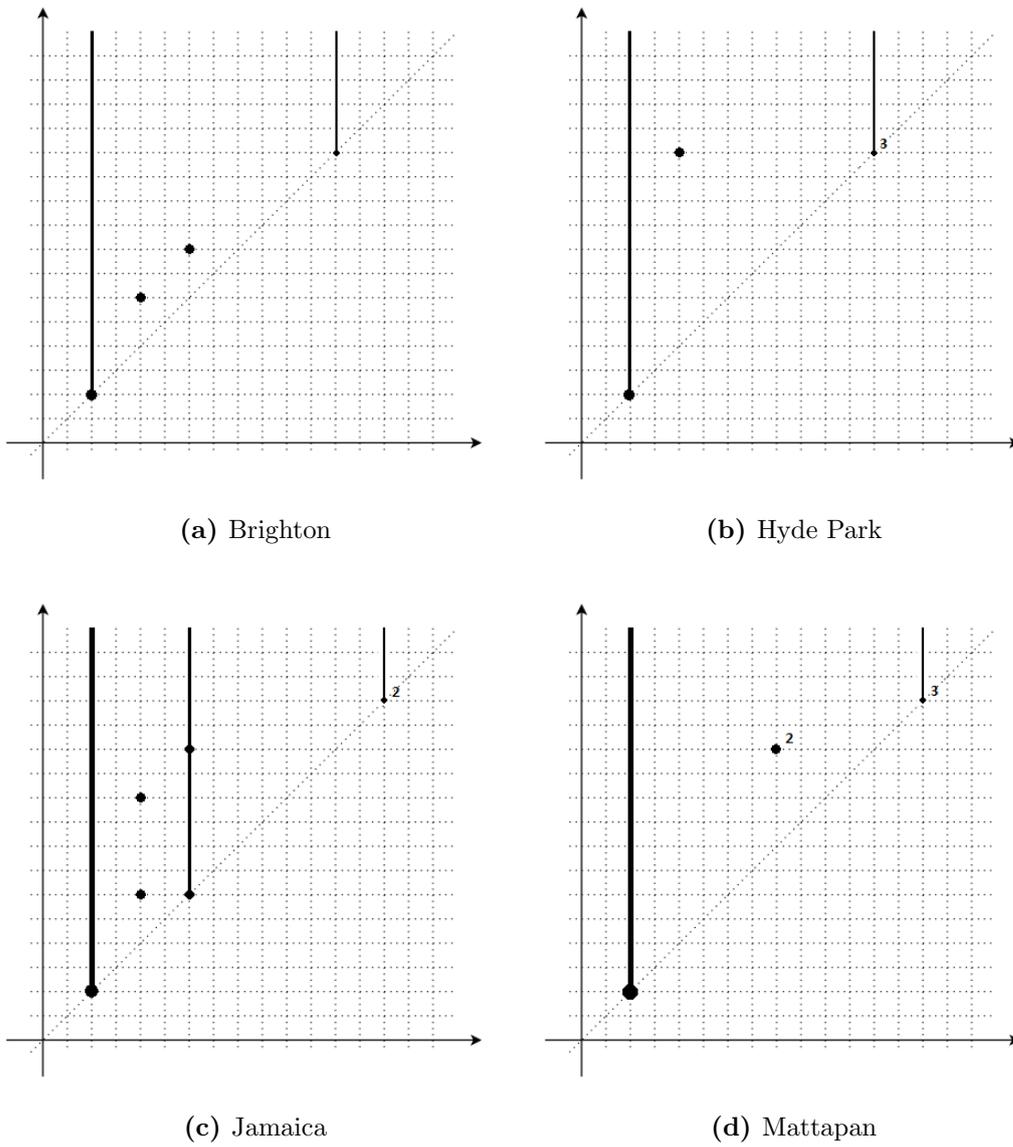


Figure 6.13: Guns persistent diagrams based on the number of calls.

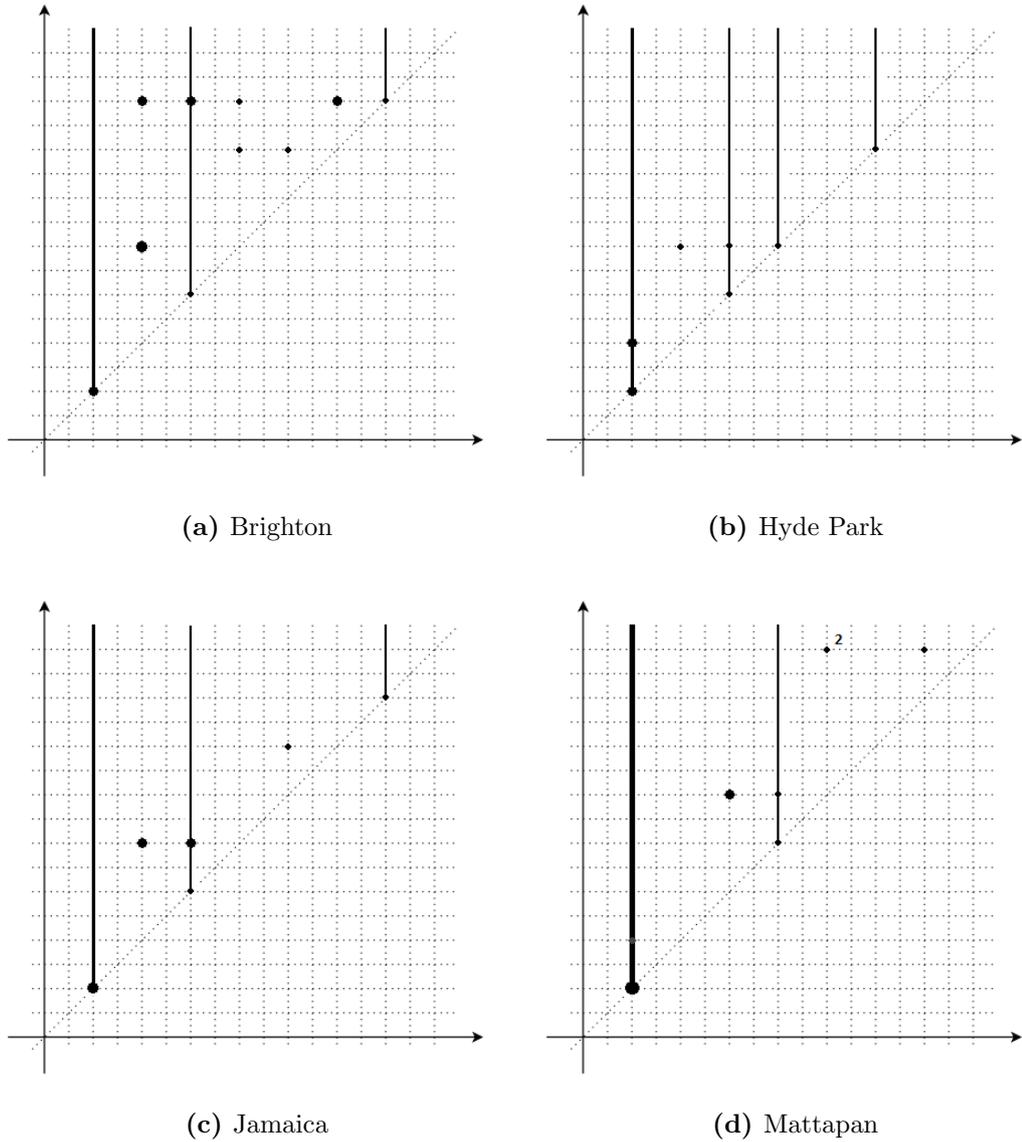


Figure 6.14: Guns persistence diagrams based on time.

Chapter 7

Conclusions

In this thesis, topology methods were used to analyze the data of 911 calls from the Boston area. The basic model used was the Vietoris-Rips complex. Various parameters can be used to build the complex. While constructing the complex, we also add function values to vertices that are later used to calculate the Morse field. Various combinations of call parameters can be used for the function values. From the complex, the Morse field is calculated that produces critical simplices. These carry information about the data. We can look at how the complex is built depending on some parameter. We can illustrate the process by drawing persistence diagrams. The diagrams tell us how persistent data is, and we can compare the diagrams with data from different neighborhoods, times, types of calls, etc.

The results show that we can draw some conclusions about the calls. Critical simplices tell us where the places with the most and the least disturbances are and also where the areas that represent the saddles are, that is, the areas that separate these extremes. We can see that some neighborhoods have equally distributed calls for certain types of calls and others do not.

The persistent diagrams of an area illustrate its structure. We can compare these structures while changing parameters and see if they change or not. The structure also shows if there are multiple focal point which point to the dispersion of the calls. Having only one focal point would mean there

is a center where most of the calls origin from and other calls quickly connect to it. Persistence analysis revealed that the time of year can affect the complex characteristics and that for some parameters the structure of the neighborhood dictates the results.

Topological data analysis allows us to explore and gain insights at varying levels of granularity. Each level gives us some information about the data. There are a lot of possible distances that can be used, and we have only used a few different parameters, so there are still a lot of possibilities for future work. The parameter for setting the threshold of cancelling can also be adjusted for further possibilities.

We only built the simplices to dimension 2, which means that we only analyzed the persistence diagrams for dimension 0. We assume higher dimensions could hold much more information about the data.

Another possibility is calculating persistence with multiple parameters as described in [17].

Bibliography

- [1] G. Carlsson, Topology and data, *Bulletin of the American Mathematical Society* 46 (2) (2009) 255 – 308.
- [2] V. de Silva, R. Ghrist, Coordinate-free coverage in sensor networks with controlled boundaries via homology, *Int. Journal of Robotics Research* 25 (12) (2006) 1205 – 1222.
- [3] M. Nicolau, A. J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, *Proceedings of the National Academy of Sciences of the United States of America* 108 (17) (2011) 7265 – 7270.
- [4] V. Robins, P. J. Wood, A. P. Sheppard, Theory and algorithms for constructing discrete morse complexes from grayscale digital images, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (8) (2011) 1646 – 1658.
- [5] R. Forman, A user’s guide to discrete morse theory, in: *Proc. of the 2001 Internat. Conf. on Formal Power Series and Algebraic Combinatorics*, A special volume of *Advances in Applied Mathematics*, 2001, p. 48.
- [6] G. Carlsson, A. Zomorodian, A. Collins, L. Guibas, Persistence barcodes for shapes, in: *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, SGP ’04*, ACM, 2004, pp. 124–135.
- [7] K. Mischaikow, V. Nanda, Morse theory for filtrations and efficient computation of persistent homology, *Discrete & Computational Geometry* 50 (2) (2013) 330 – 353.

-
- [8] S. L. Morgan, C. Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd Edition, Cambridge University Press, 2015.
- [9] C. Winship, *Econometrics in the age of big data: Measuring and assessing 'broken windows' using large scale administrative records*, *Sociological Methodology*.
- [10] C. Winship, C. Knight, *The Causal Implications of Mechanistic Thinking: Identification Using Directed Acyclic Graphs (DAGs)*, Vol. 2013, Springer, 2013.
- [11] E. W. Chambers, V. de Silva, J. Erickson, R. Ghrist, Vietorisrips complexes of planar point sets, *Discrete & Computational Geometry* 44 (1) (2010) 75 – 90.
- [12] A. Zomorodian, Fast construction of the vietoris-rips complex, *Computer and Graphics* (2010) 263–271.
- [13] A. J. Zomorodian, *Topology for Computing*, The Press Syndicate of the University of Cambridge, 2005.
- [14] H. Edelsbrunner, J. L. Harer, *Computational Topology: An Introduction*, The American Mathematical Society, 2010.
- [15] E. Čech, *Théorie générale de l'homologie dans un espace quelconque*.
- [16] H. King, K. Knudson, N. Mramor, Generating discrete morse functions from point data, *Experimental Mathematics* 14 (4) (2005) 435 – 444.
- [17] M. Allili, T. Kaczynski, C. Lapid, Reducing complexes in multidimensional persistent homology theory, *ArXiv e-prints*.
- [18] L. Vietoris, Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen, *Mathematische Annalen* 97 (1) (1927) 454–472.

-
- [19] M. Gromov, Hyperbolic groups, in: *Essays in Group Theory*, Vol. 8 of *Mathematical Sciences Research Institute Publications*, Springer New York, 1987, pp. 75–263.
- [20] A. Björner, *Handbook of combinatorics* (vol. 2), MIT Press, Cambridge, MA, USA, 1995, Ch. *Topological Methods*, pp. 1819–1872.
- [21] E. Welzl, Smallest enclosing disks (balls and ellipsoids), in: *Results and New Trends in Computer Science*, Springer-Verlag, 1991, pp. 359–370.
- [22] Wikipedia, Boston — wikipedia, the free encyclopedia, [Online; accessed 14-June-2015] (2015).
URL <https://en.wikipedia.org/w/index.php?title=Boston&oldid=666788558>
- [23] Wikipedia, Neighborhoods in boston — wikipedia, the free encyclopedia, [Online; accessed 14-June-2015] (2015).
URL https://en.wikipedia.org/w/index.php?title=Neighborhoods_in_Boston&oldid=655910895
- [24] City of boston neighborhoods.
URL http://www.cityofboston.gov/images_documents/Neighborhoods_tcm3-8205.pdf