



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ  
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ  
ΕΠΙΣΤΗΜΩΝ  
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

## Τοπολογική Ανάλυση Δεδομένων

Διπλωματική εργασία

Αλέξανδρος Βέτσικας

Επιβλέπουσα : Σοφία Λαμπροπούλου, Καθηγήτρια Ε.Μ.Π.

ΑΘΗΝΑ , ΙΟΥΛΙΟΣ 2023





ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ  
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ  
ΕΠΙΣΤΗΜΩΝ  
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

## Τοπολογική Ανάλυση Δεδομένων

Διπλωματική εργασία

Αλέξανδρος Βέτσικας

Εξεταστική Επιτροπή :

Σοφία Λαμπροπούλου, Καθηγήτρια Ε.Μ.Π.

Φιλία Βόντα, Καθηγήτρια Ε.Μ.Π.

Δημήτριος Κοντοκώστας, Επίκουρος Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ , ΙΟΥΛΙΟΣ 2023



## Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω όλους τους καθηγητές της σχολής Ε.Μ.Φ.Ε. για την επιμονή τους και την υπομονή τους όλα αυτά τα χρόνια, να μας κάνουν καλύτερους επιστήμονες. Πιο συγκεκριμένα τον κύριο Συμβώνη Αντώνιο για το εξαιρετικό στήσιμο των μαθημάτων πληροφορικής, δίνοντάς μας γερές βάσεις στον κλάδο αυτόν, και την κυρία Λαμπροπούλου Σοφία, για τις συμβουλές της και τη βοήθειά της στη μεταξύ μας συνεργασία. Επίσης, να ευχαριστήσω την κ. Βόντα Φιλία και τον κ. Κοντοκώστα Δημήτριο που δέχτηκαν να γίνουν μέλη της τριμελούς εξεταστικής επιτροπής. Τέλος, να ευχαριστήσω την οικογένεια μου και τους φίλους μου, που τόσα χρόνια με στηρίζουν σε αυτήν την μακράν πορεία.

## Περίληψη

Ο κλάδος της Τοπολογικής Ανάλυσης Δεδομένων, αποτελεί ένα σημαντικό παρακλάδι της επιστήμης των δεδομένων, καθώς τα τελευταία χρόνια έχουν επιτευχθεί σημαντικά αποτελέσματα με τεχνικές και εργαλεία που εισήχθησαν από την Αλγεβρική Τοπολογία. Πιο συγκεκριμένα, τα πιο βασικά εργαλεία είναι η Επιμένουσα Ομολογία, η οποία εξετάζει τα αλγεβρικά-τοπολογικά χαρακτηριστικά του χώρου στον οποίον ζουν τα δεδομένα, και ο αλγόριθμος Mapper που απεικονίζει τα δεδομένα που ζουν σε κάποιο χώρο τεράστιας διάστασης, σε ένα απλό και κατανοητό γράφημα. Ο σκοπός της παρούσας εργασίας είναι μια εισαγωγή στη θεωρητική θεμελίωση αυτών των εργαλείων.

**Λέξεις κλειδιά :** TDA, computational topology, simplicial complexes, Persistence Homology, Mapper Algorithm, Persistent Images, Barcodes

# Abstract

Topological Data Analysis (TDA) plays an important role in the field of data analysis science. In recent years there have been major discoveries through TDA, due to tools that have been inserted from Algebraic Topology. Specifically, the two main tools are Persistence Homology, which studies the topological features of the space where data live, and Mapper Algorithm, which visualizes data that live in large dimensions, to just a simple graph. The main purpose of this thesis is to do some introductory to the theoretical foundations of these tools.

**Λέξεις κλειδιά :** TDA,computational topology, simplicial complexes, Persistence Homology, Mapper Algorithm, Persistent Images, Barcodes

# Contents

<b>1</b>	<b>Εισαγωγή</b>	<b>9</b>
1.1	Τοπολογία . . . . .	9
1.2	Μετρικοί χώροι . . . . .	10
1.3	Απεικονίσεις, Ομοιομορφισμοί και Ομοτοπίες . . . . .	10
1.4	Άλγεβρα . . . . .	11
<b>2</b>	<b>Πλεγματικά Συμπλέγματα και Ομάδες Ομολογίας</b>	<b>14</b>
2.1	Πλεγματικά Συμπλέγματα . . . . .	14
2.2	Nerve, Čech and Rips complex . . . . .	16
2.3	Αλυσίδες, Κύκλοι, Σύνορα και Ομάδες Ομολογίας (Chains, Cycles, Boundaries and Homology Groups) . . . . .	19
2.3.1	Αλυσίδες . . . . .	19
2.3.2	Σύνορα και Κύκλοι . . . . .	19
2.3.3	Ομάδες Ομολογίας . . . . .	22
2.4	Επαγόμενες απεικονίσεις στην Ομολογία . . . . .	26
<b>3</b>	<b>Επιμένουσα Ομολογία και Επιμένουσες Εικόνες (Persistent Homology and Persistent Images)</b>	<b>28</b>
3.1	Επιμένουσα Ομολογία . . . . .	28
3.2	Διαγράμματα επιμονής και Barcodes . . . . .	30
3.3	Επιμένουσες Εικόνες . . . . .	32
<b>4</b>	<b>Mapper</b>	<b>34</b>
4.1	Ο Αλγόριθμος Mapper . . . . .	34



# 1 Εισαγωγή

Στον παρόν κεφάλαιο παρουσιάζονται βασικοί ορισμοί από την Τοπολογία και την Άλγεβρα, που θα χρησιμοποιηθούν στα επόμενα κεφάλαια.

## 1.1 Τοπολογία

**Ορισμός 1.1.1 :** Ως **τοπολογία** του χώρου  $\mathbb{X}$  ορίζεται οποιαδήποτε οικογένεια υποσυνόλων  $\tau$  του  $\mathbb{X}$ , η οποία ικανοποιεί τα παρακάτω αξιώματα:

1. Το κενό σύνολο  $\emptyset$  και ο  $\mathbb{X}$  ανήκουν στην  $\tau$ .
2. Η ένωση αριθμήσιμου ή υπεραριθμήσιμου το πλήθος στοιχείων της  $\tau$ , ανήκει στην  $\tau$ .
3. Η τομή πεπερασμένου το πλήθος στοιχείων της  $\tau$ , ανήκει στην  $\tau$ .

Τα στοιχεία της  $\tau$  καλούνται **ανοιχτά** σύνολα του  $\mathbb{X}$ . Το ζεύγος  $(\mathbb{X}, \tau)$  καλείται **τοπολογικός χώρος**. Ένα υποσύνολο  $C \subseteq \mathbb{X}$  καλείται **κλειστό** αν  $\mathbb{X} \setminus C \in \tau$ .

**Ορισμός 1.1.2 :** Έστω  $S \subseteq \mathbb{X}$ . Η **επαγόμενη τοπολογία** στον  $S$  ορίζεται ως  $\tau_S = \{U \cap S \mid U \in \tau\}$  και το ζεύγος  $(S, \tau_S)$  καλείται **τοπολογικός υπόχωρος** του  $(\mathbb{X}, \tau)$ .

**Ορισμός 1.1.3 :** Ένας τοπολογικός χώρος  $(\mathbb{X}, \tau)$  καλείται **μη συνεκτικός**, αν υπάρχουν δύο ξένα μεταξύ τους ανοιχτά σύνολα  $U, V \in \tau$ , έτσι ώστε  $\mathbb{X} = U \cup V$ . Σε αντίθετη περίπτωση, καλείται **συνεκτικός**.

**Ορισμός 1.1.4 :** Ένα ανοιχτό (ή κλειστό) **κάλυμμα** ενός τοπολογικού χώρου  $(\mathbb{X}, \tau)$  είναι μία συλλογή  $C$  από ανοιχτά (ή κλειστά) σύνολα, έτσι ώστε  $\mathbb{X} \subseteq \bigcup_{c \in C} c$ . Ο τοπολογικός χώρος  $(\mathbb{X}, \tau)$  καλείται **συμπαγής**, αν για κάθε ανοιχτό κάλυμμα  $C$  υπάρχει πεπερασμένο υποκάλυμμα  $C' \subseteq C$ , έτσι ώστε  $\mathbb{X} \subseteq \bigcup_{c \in C'} c$  και το  $C'$  έχει πεπερασμένο το πλήθος στοιχεία.

## 1.2 Μετρικοί χώροι

**Ορισμός 1.2.1 :** Ένας **μετρικός χώρος** είναι ένα ζεύγος  $(\mathbb{X}, d_{\mathbb{X}})$ , όπου  $\mathbb{X}$  είναι σύνολο και  $d_{\mathbb{X}}$  είναι μία συνάρτηση απόστασης  $d_{\mathbb{X}} : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}_{\geq 0}$  που ικανοποιεί τα παρακάτω αξιώματα:

1.  $d_{\mathbb{X}}(x, x) = 0, \forall x \in \mathbb{X}$
2.  $d_{\mathbb{X}}(x, y) = d_{\mathbb{X}}(y, x) > 0, \forall x, y \in \mathbb{X}, \text{ με } x \neq y$
3.  $d_{\mathbb{X}}(x, z) \leq d_{\mathbb{X}}(x, y) + d_{\mathbb{X}}(y, z) \quad \forall x, y, z \in \mathbb{X}$

**Ορισμός 1.2.2 :** Ως **ανοικτή μπάλα** στον  $(\mathbb{X}, d_{\mathbb{X}})$  με κέντρο  $c$  και ακτίνα  $r$  ορίζουμε το σύνολο  $B_r(c) = \{x \in \mathbb{X} \mid d_{\mathbb{X}}(c, x) < r\}$ .

**Ορισμός 1.2.3 :** Η μετρική σε ένα μετρικό χώρο  $(\mathbb{X}, d_{\mathbb{X}})$  επάγει τη λεγόμενη **μετρική τοπολογία**, η οποία αποτελείται από όλες τις ανοικτές μπάλες  $B_r(c)$  και όλες τις πιθανές εννώσεις τους. Θα συμβολίζεται ως  $\tau_{d_{\mathbb{X}}}$ .

**Ορισμός 1.2.4 :** Ως **νέφος σημειακών δεδομένων** (point cloud data) ορίζουμε κάθε πεπερασμένο σύνολο δεδομένων-σημείων εφοδιασμένο με κάποια μετρική  $d$ .

## 1.3 Απεικονίσεις, Ομοιομορφισμοί και Ομοτοπίες

**Ορισμός 1.3.1 :** Έστω  $\mathbb{T}, \mathbb{U}$  τοπολογικοί χώροι. Μία συνάρτηση  $f : \mathbb{T} \rightarrow \mathbb{U}$  είναι **συνεχής** αν και μόνο αν για κάθε ανοικτό  $U \subseteq \mathbb{U}$ , το  $f^{-1}(U)$  είναι ανοικτό στον  $\mathbb{T}$ . Αυτές τις συνεχείς συναρτήσεις συνήθως τις καλούμε απλά **απεικονίσεις** (maps).

**Ορισμός 1.3.2 :** Μία απεικόνιση  $g : \mathbb{T} \rightarrow \mathbb{U}$  είναι μία **εμφύτευση** (embedding) του  $\mathbb{T}$  στον  $\mathbb{U}$ , αν η  $g$  είναι μονοσήμαντη (injective).

**Ορισμός 1.3.3 :** Η απεικόνιση  $h : \mathbb{T} \rightarrow \mathbb{U}$  είναι **ομοιομορφισμός** αν είναι αμφι-μονοσήμαντη, με συνεχή αντίστροφη συνάρτηση. Δύο τοπολογικοί χώροι είναι **ομοιομορφικοί**, αν υπάρχει κάποιος ομοιομορφισμός μεταξύ τους.

**Ορισμός 1.3.4 :** Έστω οι απεικονίσεις  $g : \mathbb{X} \rightarrow \mathbb{U}$  και  $h : \mathbb{X} \rightarrow \mathbb{U}$ . Μία **ομοτοπία** (homotopy) είναι μια απεικόνιση  $H : \mathbb{X} \times [0, 1] \rightarrow \mathbb{U}$  έτσι ώστε  $H(\cdot, 0) = g$  και

$H(\cdot, 1) = h$ . Δύο απεικονίσεις καλούνται **ομοτοπικές** αν υπάρχει κάποια ομοτοπία που τις συνδέει.

**Ορισμός 1.3.5 :** Δύο τοπολογικοί χώροι  $\mathbb{T}, \mathbb{U}$  καλούνται **ομοτοπικά ισοδύναμοι**, αν υπάρχουν απεικονίσεις  $g : \mathbb{T} \rightarrow \mathbb{U}$  και  $h : \mathbb{U} \rightarrow \mathbb{T}$  έτσι ώστε η  $(h \circ g)$  να είναι ομοτοπική με την  $\iota_{\mathbb{T}} : \mathbb{T} \rightarrow \mathbb{T}$  και η  $(g \circ h)$  να είναι ομοτοπική με την  $\iota_{\mathbb{U}} : \mathbb{U} \rightarrow \mathbb{U}$ .

**Ορισμός 1.3.6 :** Έστω  $\mathbb{T}$  τοπολογικός χώρος, και έστω και ένας υπόχωρος  $\mathbb{U} \subset \mathbb{T}$ . Μία **retraction**  $r$  του  $\mathbb{T}$  στον  $\mathbb{U}$  είναι μία απεικόνιση από τον  $\mathbb{T}$  στον  $\mathbb{U}$  έτσι ώστε  $r(x) = x$  για κάθε  $x \in \mathbb{U}$ . Ο χώρος  $\mathbb{U}$  καλείται **deformation retract** του  $\mathbb{T}$ , αν η ταυτοτική απεικόνιση στον  $\mathbb{T}$  μπορεί να παραμορφωθεί με συνεχή τρόπο σε μία retraction χωρίς να μετακινούνται τα σημεία που ήδη ανήκουν στον  $\mathbb{U}$ . Πιο συγκεκριμένα, να υπάρχει μία ομοτοπία που καλείται **deformation retraction**  $R : \mathbb{T} \times [0, 1] \rightarrow \mathbb{T}$  έτσι ώστε η  $R(\cdot, 0)$  να είναι η ταυτοτική στον  $\mathbb{T}$ , η  $R(\cdot, 1)$  είναι μία retraction του  $\mathbb{T}$  στο  $\mathbb{U}$ , και  $R(x, t) = x$  για κάθε  $x \in \mathbb{U}$  και για κάθε  $t \in [0, 1]$ .

## 1.4 Άλγεβρα

**Ορισμός 1.4.1 :** Ομάδα  $(G, \cdot)$  είναι ένα σύνολο  $G$ , μαζί με μία διμελή πράξη  $\cdot$  στην  $G$ , η οποία ικανοποιεί τα παρακάτω αξιώματα:

1. Για κάθε  $a, b \in G$ ,  $a \cdot b \in G$ .
2. Για κάθε  $a, b, c \in G$ ,  $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ .
3. Υπάρχει ένα **ταυτοτικό στοιχείο**  $e \in G$  τέτοιο ώστε  $a \cdot e = e \cdot a = a$ , για κάθε  $a \in G$ .
4. Για κάθε  $a \in G$ , υπάρχει ένα **συμμετρικό στοιχείο**  $a' \in G$  τέτοιο ώστε  $a \cdot a' = a' \cdot a = e$ .

Αν επιπλέον  $a \cdot b = b \cdot a$  για κάθε  $a, b \in G$ , τότε λέμε ότι η  $(G, \cdot)$  είναι **αβελιανή**. Συνήθως τις αβελιανές ομάδες τις συμβολίζουμε με  $(G, +)$ , το ουδέτερο στοιχείο το συμβολίζουμε με 0 και το συμμετρικό στοιχείο ενός στοιχείου  $a \in G$  το καλούμε **αντίθετο** και το συμβολίζουμε με  $-a$ . Ένα υποσύνολο  $H \subseteq G$  λέγεται **υποομάδα** της  $G$ , αν είναι και το ίδιο ομάδα με την πράξη της  $G$ .

**Ορισμός 1.4.2 :** Έστω ένα μη κενό σύνολο γεννητόρων  $X$  μιας αβελιανής ομάδας  $G$ . Αν κάθε μη μηδενικό στοιχείο  $a \in G$  γράφεται με μοναδικό τρόπο στη μορφή :  
 $a = n_1x_1 + n_2x_2 + \dots + n_rx_r$ , όπου  $n_i \in \mathbb{Z}_{\neq 0}$  και τα  $x_i$  είναι ανά δύο διαφορετικά στοιχεία του  $X$ , τότε η  $G$  καλείται **ελεύθερη αβελιανή ομάδα** και το  $X$  καλείται **βάση** της ομάδας.

**Ορισμός 1.4.3 :** Για μία υποομάδα  $H \subseteq G$  και ένα στοιχείο  $a \in G$ , το αριστερό **σύνπλοκο** (coset) είναι το σύνολο  $aH = \{a * b \mid b \in H\}$  και το δεξιό σύνπλοκο είναι το σύνολο  $Ha = \{b * a \mid b \in H\}$ . Για τις αβελιανές ομάδες προφανώς αυτά τα σύνολα ταυτίζονται. Ως **ομάδα-πηλίκο** της  $G$  ως προς μία κανονική υποομάδα της  $H$  ορίζουμε το σύνολο  $G/H = \{aH \mid a \in G\}$  με την καλά ορισμένη πράξη που επάγεται από την  $G$  με τον εξής τρόπο :  $aH * bH = (a * b)H$ , για κάθε  $a, b \in G$ .

**Ορισμός 1.4.4 :** Μία απεικόνιση  $h : G \rightarrow H$  μεταξύ δύο ομάδων  $(G, +)$  και  $(H, *)$  καλείται **ομομορφισμός ομάδων** αν  $h(a + b) = h(a) * h(b)$  για κάθε  $a, b \in G$ . Αν επιπλέον η  $h$  είναι αμφιμονοσήμαντη, τότε την καλούμε **ισομορφισμό**. Δύο ομάδες  $G, H$  που συνδέονται με κάποιο ισομορφισμό, καλούνται **ισόμορφες** και συμβολίζουμε,  $G \cong H$ . Ως **πυρήνα** ενός ομομορφισμού  $h$  ορίζουμε το σύνολο :

$$\ker(h) = \{a \in G \mid h(a) = 0\}.$$

**Ορισμός 1.4.5 :** Ένα σύνολο  $R$  εφοδιασμένο με δύο διμελείς πράξεις, την “πρόσθεση”,  $+$ , και τον “πολλαπλασιασμό”,  $*$ , είναι ένας **δακτύλιος**, αν ικανοποιεί τα παρακάτω αξιώματα :

1. Το  $(R, +)$  είναι αβελιανή ομάδα.
2. Ο “πολλαπλασιασμός” είναι προσεταιριστικός, δηλαδή  $(a * b) * c = a * (b * c)$ .
3. Για κάθε  $a, b, c \in R$  ισχύει ο αριστερός και ο δεξιός επιμεριστικός νόμος, δηλαδή :  
 $a * (b + c) = a * b + a * c$  και  $(b + c) * a = b * a + c * a$ .

Εδώ θα θεωρούμε ότι ο δακτύλιος ως προς το πολλαπλασιασμό έχει ως ουδέτερο στοιχείο το  $\mathbf{1}$  και θα καλείται απλά δακτύλιος και όχι δακτύλιος με μοναδιαίο στοιχείο, για ευκολία. Αν επίσης και ο πολλαπλασιασμός είναι αντιμεταθετικός, τότε θα καλείται **αντιμεταθετικός δακτύλιος**. Τέλος, αν κάθε μη μηδενικό στοιχείο του αντιμεταθετικού δακτυλίου έχει πολλαπλασιαστικό αντίστροφο, τότε καλείται **σώμα** και το  $(R \setminus \{0\}, \cdot)$  είναι μία αβελιανή ομάδα.

**Ορισμός 1.4.6 :** Έστω ένα σώμα  $F$ . Ένας διανυσματικός χώρος πάνω από το  $F$  αποτελείται από μια αβελιανή ομάδα  $(V, +)$ , μαζί με μια πράξη βαθμωτού πολλαπλασιασμού των στοιχείων του  $V$  με στοιχεία του  $F$  από τα αριστερά, τέτοια ώστε για κάθε  $a, b \in F$  και  $\alpha, \beta \in V$  να ικανοποιούνται τα παρακάτω :

1.  $a\alpha \in V$

2.  $a(b\alpha) = (ab)\alpha$

3.  $(a + b)\alpha = (a\alpha) + (b\alpha)$

4.  $a(\alpha + \beta) = (a\alpha) + (a\beta)$

5.  $1\alpha = \alpha$

## 2 Πλεγματικά Συμπλέγματα και Ομάδες Ομολογίας

Σε αυτό το κεφάλαιο εισάγουμε τα δύο βασικά εργαλεία πάνω στα οποία στηρίζεται η Επιμένουσα Ομολογία (Persistent Homology), μία από τις μεθόδους που χρησιμοποιούνται στην Τοπολογική Ανάλυση Δεδομένων (TDA). Το ένα είναι τα πλεγματικά συμπλέγματα (simplicial complexes) και το άλλο είναι οι ομάδες ομολογίας (homology groups). Τα πρώτα, είναι κατά κάποιο τρόπο γεωμετρικοί σκελετοί που χτίζονται με κάποιους κανόνες πάνω στα δεδομένα-σημεία, μέσα σε κάποιο χώρο, δίνοντας την αφορμή στην τοπολογία να εμπλακεί στην μελέτη αυτών των γεωμετρικών αντικειμένων. Από την άλλη, οι ομάδες ομολογίας είναι τα αλγεβρικά εργαλεία που χρησιμοποιούνται στην επιμένουσα ομολογία για την μελέτη των δεδομένων.

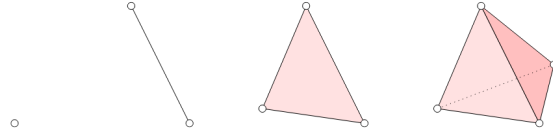
### 2.1 Πλεγματικά Συμπλέγματα

Τα πλεγματικά συμπλέγματα χρησιμοποιούνται ως η κύρια δομή δεδομένων για την αναπαράσταση των τοπολογικών χώρων.

**Ορισμός 2.1.1 :** Έστω τα σημεία  $u_0, u_1, u_2, \dots, u_k$  που ανήκουν στον χώρο  $\mathbb{R}^d$ . Αυτά τα σημεία είναι **αφινικά ανεξάρτητα** (affinely independent) αν τα  $k$  διανύσματα  $u_i - u_0$ ,  $1 \leq i \leq k$ , είναι γραμμικώς ανεξάρτητα. Στον  $\mathbb{R}^d$  έχουμε το πολύ  $d + 1$  αφινικά ανεξάρτητα σημεία. Ένα σημείο  $x = \sum_{i=0}^k \lambda_i u_i$  καλείται **αφινικός συνδυασμός** (affine combination) των  $u_i$ , αν  $\sum_{i=0}^k \lambda_i = 1$ .

**Ορισμός 2.1.2 :** Ένας αφινικός συνδυασμός  $x = \sum_{i=0}^k \lambda_i u_i$  είναι **κυρτός συνδυασμός**, αν τα  $\lambda_i \geq 0$ . Ως **κυρτή θήκη** (convex hull) των  $\{u_i\}_{i=0}^k$  ορίζεται το σύνολο όλων των κυρτών συνδυασμών των  $\{u_i\}_{i=0}^k$ .

**Ορισμός 2.1.3 :** Ορίζουμε ως  **$k$ -simplex**  $\sigma$  να είναι η κυρτή θήκη από  $k + 1$  αφινικά ανεξάρτητα σημεία  $\{u_i\}_{i=0}^k$  και συμβολίζουμε  $\sigma = [u_0, u_1, \dots, u_k]$ . Η **διάστασή** του είναι  $\dim(\sigma) = k$ . Ως **κορυφές** ορίζουμε τα 0-simplices, ως **αχμές** τα 1-simplices, ως **τρίγωνα** τα 2-simplices, και ως **τετράεδρα** τα 3-simplices (βλέπε Εικόνα 2.1). Επιπλέον, κάθε υποσύνολο των αφινικά ανεξάρτητων κορυφών, είναι πάλι αφινικά ανεξάρτητο και άρα πάλι ορίζει ένα simplex μικρότερης διάστασης.



Εικόνα 2.1 : Από αριστερά προς τα δεξιά: κορυφή, ακμή, τρίγωνο, τετράεδρο.

**Ορισμός 2.1.4 :** Ως **όψη** (face)  $\tau$  ενός simplex  $\sigma = [u_0, u_1, \dots, u_k]$  ορίζεται να είναι η κυρτή θήκη ενός μη κενού υποσυνόλου των  $\{u_i\}_{i=0}^k$  (συμβολισμός:  $\tau \leq \sigma$ ) και καλείται **γνήσια** (proper), αν το υποσύνολο δεν είναι ολόκληρο το σύνολο των  $\{u_i\}_{i=0}^k$  (συμβολισμός:  $\tau < \sigma$ ). Αντίστροφα, το  $\sigma$  καλείται **συν-όψη** (coface) του  $\tau$ , και αν  $\tau < \sigma$  το  $\sigma$  καλείται **γνήσια συν-όψη** (proper coface) του  $\tau$ .

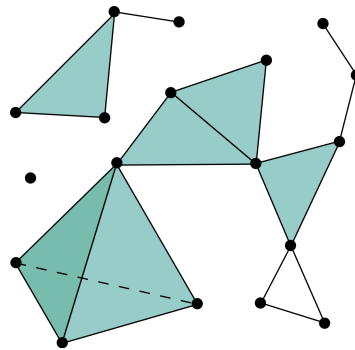
**Ορισμός 2.1.5 :** Έστω  $\sigma = [u_0, u_1, \dots, u_k]$  ένα  $k$ -simplex. Ως **σύνορο** του  $\sigma$ , ορίζεται να είναι η ένωση των  $(k - 1)$ -όψεων (ή αλλιώς facets) του  $\sigma$  και θα το συμβολίζουμε με  $bd(\sigma)$ . Για παράδειγμα οι κορυφές δεν έχουν σύνορο, οι ακμές έχουν τις αντίστοιχες κορυφές τους ως σύνορο, τα τρίγωνα τις ακμές που τα περιβάλλουν, και τα τετράεδρα τα τρίγωνα που τα περιβάλλουν. Ως **εσωτερικό** του  $\sigma$  ορίζεται το  $int(\sigma) = \sigma - bd(\sigma)$ .

**Ορισμός 2.1.6 :** Ένα **γεωμετρικό πλεγματοικό σύμπλεγμα** (geometric simplicial complex)  $K$  είναι μία πεπερασμένη συλλογή από simplices έτσι ώστε να ικανοποιούνται οι παρακάτω συνθήκες :

1. Αν  $\sigma \in K$  και  $\tau \leq \sigma$ , τότε  $\tau \in K$ , δηλαδή το  $K$  περιέχει κάθε όψη του  $\sigma$ .
2. Αν  $\sigma_1, \sigma_2 \in K$ , τότε η τομή  $\sigma_1 \cap \sigma_2$  είναι είτε κενή είτε κάποια κοινή τους όψη.

Ως **διάσταση** ενός γεωμετρικού πλεγματοικού συμπλέγματος  $K$  ορίζεται η μεγαλύτερη διάσταση από τα simplices που περιέχει.

Στην Εικόνα 2.2 παρουσιάζεται ένα παράδειγμα γεωμετρικού πλεγματοικού συμπλέγματος, με τα simplices να είναι κορυφές, ακμές, τρίγωνα και ένα τετράεδρο.



Εικόνα 2.2 : Παράδειγμα ενός γεωμετρικού πλεγματοικού συμπλέγματος.

**Ορισμός 2.1.7 :** Δεδομένου ενός συνόλου  $V$ , ένα **αφηρημένο πλεγματοτικό σύμπλεγμα** (abstract simplicial complex) με σύνολο κορυφών  $V$  είναι ένα σύνολο  $K$  του οποίου τα στοιχεία είναι πεπερασμένα υποσύνολα του  $V$ , το οποίο έχει τις εξής ιδιότητες:

1. Τα στοιχεία του  $V$  ανήκουν στο  $K$ .
2. Για κάθε  $\sigma \in K$ , κάθε υποσύνολο του  $\sigma$  ανήκει στο  $K$ .

Τα στοιχεία του  $K$  καλούνται **simplices**. Επίσης, ως **διάσταση** ενός simplex  $\sigma$  του  $K$  ορίζεται η ποσότητα  $\dim(\sigma) = |\sigma| - 1$ , όπου  $|\sigma|$  = πλήθος στοιχείων του  $\sigma$ . Τέλος, ως **διάσταση του αφηρημένου πλεγματοτικού συμπλέγματος**  $K$  ορίζεται η μεγαλύτερη διάσταση από τα simplices που περιέχει.

**Ορισμός 2.1.8 :** Για κάθε  $0 \leq d \leq \dim(K)$ , ως  **$d$ -σκελετός** ( $d$ -skeleton) ενός πλεγματοτικού συμπλέγματος  $K$  ορίζεται το υποσύμπλεγμα του  $K$  που αποτελείται από τα simplices του  $K$  διάστασης το πολύ  $d$ . Συμβολίζεται με  $K^d$ .

**Ορισμός 2.1.9 :** Ο **υποκείμενος χώρος**  $|K|$  ενός αφηρημένου πλεγματοτικού συμπλέγματος  $K$  είναι η ταύτισή του με κάποιο γεωμετρικό πλεγματοτικό σύμπλεγμα. Ο **υποκείμενος χώρος**  $|K|$  ενός γεωμετρικού πλεγματοτικού συμπλέγματος  $K$  είναι ο εαυτός του.

## 2.2 Nerve, Čech and Rips complex

Παρακάτω παρουσιάζουμε κάποιους από τους “κανόνες” που χρησιμοποιούνται, για να χτίσουμε πλεγματοτικά συμπλέγματα πάνω στα δεδομένα-σημεία, σε κάποιο Ευκλείδιο χώρο  $\mathbb{R}^d$ .

**Ορισμός 2.2.1 :** Έστω μία πεπερασμένη συλλογή από μη κενά σύνολα  $\mathcal{U} = \{U_a\}_{a \in A}$ . Ορίζουμε ως **νεύρο (nerve)** του συνόλου  $\mathcal{U}$  το αφηρημένο πλεγματοτικό σύμπλεγμα  $N(\mathcal{U})$  που έχει ως σύνολο κορυφών το σύνολο δεικτών  $A$ , και όπου ένα υποσύνολο  $\{a_0, a_1, \dots, a_m\} \subseteq A$  παράγει ένα  $m$ -simplex στο  $N(\mathcal{U})$  αν και μόνο αν  $U_{a_0} \cap U_{a_1} \cap \dots \cap U_{a_m} \neq \emptyset$ .

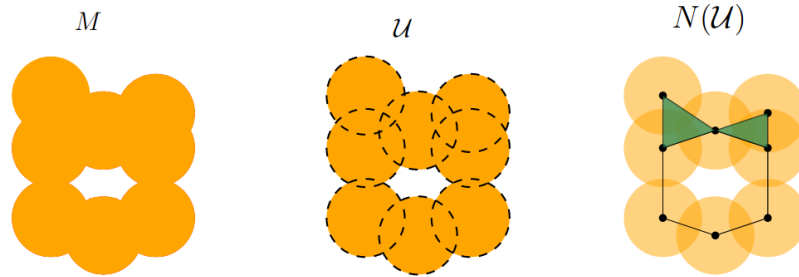
**Θεώρημα 2.2.1 :** (Nerve Theorem) Έστω ένα πεπερασμένο κάλυμμα  $\mathcal{U}$  (ανοιχτό ή κλειστό) ενός μετρικού χώρου  $M$ . Ο υποκείμενος χώρος  $|N(\mathcal{U})|$  είναι **ομοτοπικά ισοδύναμος** με τον  $M$  αν κάθε μη κενή τομή  $\bigcap_{i=0}^k U_{a_i}$  από στοιχεία του καλύμματος, είναι ομοτοπικά ισοδύναμη με ένα σημείο, δηλαδή **συσταλή**.



**Ορισμός 2.2.2 :** (Čech complex) Έστω  $(M, d)$  μετρικός χώρος και  $P = \{p_1, p_2, \dots, p_q\}$  ένα πεπερασμένο υποσύνολο του  $M$ . Δεδομένου  $r \in \mathbb{R}^+$ , το **Čech complex**  $\mathbb{C}^r(P)$  ορίζεται να είναι το νεύρο από την οικογένεια των κλειστών μπαλών  $\{B(p_i, r)\}_{1 \leq i \leq q}$ , δηλαδή

$$\mathbb{C}^r(P) = \{\sigma \subseteq P \mid \bigcap_{p_i \in \sigma} B(p_i, r) \neq \emptyset\}.$$

Στην Εικόνα 2.3 παρουσιάζεται το νεύρο  $N(\mathcal{U})$  του  $\mathcal{U}$ , το οποίο είναι ένα Čech complex.

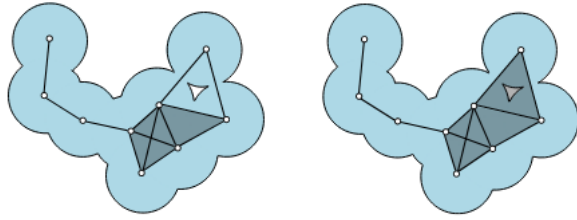


Εικόνα 2.3 : Αριστερά είναι ο μετρικός χώρος  $M$ , στη μέση είναι το κάλυμμα  $\mathcal{U}$  από ανοικτές μπάλες και δεξιά το νεύρο  $N(\mathcal{U})$ .

Επειδή τα Čech complexes έχουν υπολογιστική δυσκολία στην κατασκευή τους, συνήθως στην πράξη χρησιμοποιούνται τα Vietoris-Rips complexes που υπολογίζονται πιο εύκολα.

**Ορισμός 2.2.3 :** (Vietoris-Rips complex) Έστω  $(P, d)$  πεπερασμένος μετρικός χώρος. Δεδομένου  $r > 0$ , το **Vietoris-Rips complex** είναι το πλεγματοικό σύμπλεγμα  $\mathbb{VR}^r(P)$ , όπου ένα simplex  $\sigma \in \mathbb{VR}^r(P)$  αν και μόνο αν  $d(p, q) \leq 2r, \forall p, q \in \sigma$ .

Στην Εικόνα 2.4 βλέπουμε στα αριστερά ένα Čech complex και στα δεξιά ένα Vietoris-Rips complex. Στο Čech complex παρατηρούμε ότι το πάνω δεξιά τρίγωνο δεν περιέχει το εσωτερικό του, ενώ στο Vietoris-Rips complex το αντίστοιχο τρίγωνο περιέχει το εσωτερικό του. Η διαφορά αυτή έγκειται στο ότι στο Čech complex δεν ικανοποιείται η συνθήκη της τριπλής μη κενής τομής των αντίστοιχων μπαλών, ώστε να έχουμε ένα 2-simplex, δηλαδή και το εσωτερικό του τριγώνου. Από την άλλη στο Vietoris-Rips complex ικανοποιείται η συνθήκη των αποστάσεων για τις τρεις κορυφές του τριγώνου, για αυτό περιλαμβάνεται και το εσωτερικό.



Εικόνα 2.4 : Αριστερά είναι το Čech complex και δεξιά το Vietoris-Rips complex.

## 2.3 Αλυσίδες, Κύκλοι, Σύνορα και Ομάδες Ομολογίας (Chains, Cycles, Boundaries and Homology Groups)

Οι ομάδες ομολογίας παρέχουν πληροφορίες για τις τρύπες, τα κενά και τις συνεκτικές συνιστώσες ενός τοπολογικού χώρου. Προκειμένου όμως να τις ορίσουμε, πρέπει πρώτα να ορίσουμε τι είναι αλυσίδες, κύκλοι και σύνορα.

### 2.3.1 Αλυσίδες

Έστω  $K$  ένα  $k$ -πλεγματοικό σύμπλεγμα με  $m_p$  το πλήθος  $p$ -simplices,  $0 \leq p \leq k$ . Μια  $p$ -**αλυσίδα** ( $p$ -chain)  $c$  είναι ένα τυπικό άθροισμα (formal sum) από  $p$ -simplices του  $K$ . Ο συνήθης συμβολισμός είναι ο εξής:

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

όπου τα  $\sigma_i$  είναι τα  $p$ -simplices και τα  $a_i$  είναι σταθερές που γενικά ανήκουν σε κάποιο σώμα, όμως στην TDA συνήθως το σώμα αυτό είναι το  $\mathbb{Z}_2 = \{0, 1\}$ . Όταν μιλάμε για αφηρημένα πλεγματοικά συμπλέγματα, τα simplices τα βλέπουμε ως σύνολα από κορυφές. Τότε, οι αλυσίδες στην ουσία είναι η συμμετρική διαφορά των συνόλων που τις αποτελούν. Ως **μηδενική  $p$ -αλυσίδα** ορίζεται η  $0 = \sum_{i=1}^{m_p} 0a_i$ .

Δύο  $p$ -αλυσίδες  $c = \sum_{i=1}^{m_p} a_i \sigma_i$  και  $c' = \sum_{i=1}^{m_p} b_i \sigma_i$  προστίθενται όπως τα πολυώνυμα, δηλαδή :

$$c + c' = \sum_{i=1}^{m_p} (a_i + b_i) \sigma_i$$

με την πράξη  $a_i + b_i$  να γίνεται στο σώμα  $\mathbb{Z}_2$  όπως είπαμε. Επίσης, ισχύει ότι  $c + c = \sum_{i=1}^{m_p} (a_i + a_i) \sigma_i = 0$ , αφού  $a_i + a_i = 0$  στον  $\mathbb{Z}_2$ , άρα  $c = -c$ . Έτσι μπορούμε να ορίσουμε την αβελιανή  $p$ -**ομάδα αλυσίδων** ( $p$ -th chain group)  $C_p(K)$ , με μηνικό στοιχείο το  $0 = \sum_{i=1}^{m_p} 0a_i$ , αφού εύκολα δείχνεται ότι ικανοποιούνται όλα τα αξιώματα μιας αβελιανής ομάδας. Άρα για κάθε  $p$  έχουμε και μία ομάδα αλυσίδων.

### 2.3.2 Σύνορα και Κύκλοι

Οι ομάδες αλυσίδων σχετίζονται μεταξύ τους μέσω του **συνοριακού τελεστή** (boundary operator)  $\partial_p$ . Σύμφωνα με αυτόν, θα ορίσουμε τις υποομάδες  $Z_p(K)$  και  $B_p(K)$  των ομάδων  $C_p(K)$ , ώστε να καταλήξουμε στις ομάδες ομολογίας  $H_p(K)$ .

**Ορισμός 2.3.1 :** Έστω ένα  $p$ -simplex  $\sigma = [u_0, u_1, \dots, u_p]$ . Όπως ορίσαμε και πριν, το **σύνορο** του  $\sigma$  μέσω του συνοριακού τελεστή  $\partial_p$  ορίζεται να είναι:

$$\partial_p \sigma = \sum_{i=0}^p [u_0, u_1, \dots, \tilde{u}_i, \dots, u_p]$$

όπου  $\tilde{u}_i$  εννοούμε ότι το  $u_i$  δεν ανήκει στο αντίστοιχο  $(p-1)$ -simplex, ή αλλιώς  $(p-1)$ -facet όπως τα καλούμε. Στην ουσία το σύνορο ενός  $p$ -simplex μπορούμε να το δούμε ως μια αλυσίδα από  $(p-1)$ -simplices, η οποία έχει μη μηδενικούς όρους  $a_i$  μόνο στα facets του  $p$ -simplex.

Στην συνέχεια επεκτείνουμε τον ορισμό του συνοριακού τελεστή και στις αλυσίδες. Αρχικά, ορίζουμε το σύνορο μιας κορυφής  $u$  να είναι:  $\partial_0 u = 0$ , όπου  $0$  είναι το ταυτοτικό στοιχείο της ομάδας  $C_{-1} = \{0\}$ , την οποία συμπεριλαμβάνουμε για τεχνικούς λόγους. Το **σύνορο μιας  $p$ -αλυσίδας**  $c = \sum_{i=1}^{m_p} a_i \sigma_i$  ορίζουμε να είναι :

$$\partial_p c = \sum_{i=1}^{m_p} a_i \partial_p \sigma_i$$

όπου  $m_p$  είναι το πλήθος των  $p$ -simplices ενός πλεγματοειδούς συμπλέγματος  $K$ . Αυτό που διαπιστώνουμε είναι ότι μια  $p$ -αλυσίδα απεικονίζεται μέσω του  $\partial_p$  σε μια  $(p-1)$ -αλυσίδα, αφού τα  $\partial_p \sigma_i$  είναι  $(p-1)$ -αλυσίδες. Άρα ο **συνοριακός τελεστής** :

$$\partial_p : C_p \longrightarrow C_{p-1}$$

είναι καλά ορισμένος. Αυτό που επίσης παρατηρούμε είναι ότι η απεικόνιση  $\partial_p$  είναι στην ουσία ομομορφισμός ομάδων, αφού εναλλάσσεται με την πράξη της πρόσθεσης αλυσίδων. Δηλαδή :

$$\partial_p(c + c') = \partial_p c + \partial_p c'.$$

**Ορισμός 2.3.2 :** Ως **αλυσιδωτό σύμπλεγμα** (chain complex) ορίζουμε την παρακάτω ακολουθία από ομάδες αλυσίδων και συνοριακών τελεστών ενός  $k$ -πλεγματοειδούς συμπλέγματος  $K$ :

$$\{0\} = C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} C_{k-2} \quad \dots \quad C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} C_{-1} = \{0\}.$$

**Θεμελιώδες Λήμμα της Ομολογίας :** Η σχέση  $\partial_p \partial_{p+1} \sigma = 0$  ισχύει για κάθε ακέραιο  $p$  και κάθε  $(p+1)$ -αλυσίδα  $\sigma$ , όπου  $\partial_p \partial_{p+1}$  εννοούμε  $\partial_p \circ \partial_{p+1}$ .

**Απόδειξη :** Αυτό που αρκεί να δείξουμε είναι ότι για κάθε  $(p+1)$ -simplex  $\tau$ ,  $\partial_p \partial_{p+1} \tau = 0$ . Το σύνορο  $\partial_{p+1} \tau$  αποτελείται από τις  $p$ -όψεις του  $\tau$ . Κάθε  $(p-1)$ -όψη του  $\tau$  ανήκει σε ακριβώς δύο  $p$ -όψεις, άρα με συντελεστές στο  $\mathbb{Z}_2$ ,  $\partial_p \partial_{p+1} \tau = 0 \in C_{p-1}$ .  $\square$

Τώρα μπορούμε να ορίσουμε τις υποομάδες  $Z_p(K)$  και  $B_p(K)$  των  $C_p(K)$ , ενός  $k$ -πλεγματού συμπλέγματος  $K$ .

**Ορισμός 2.3.3 :** Ως  $p$ -κύκλος ( $p$ -cycles)  $c$  ορίζεται η  $p$ -αλυσίδα που ικανοποιεί τη σχέση:

$$\partial_p c = 0.$$

Επειδή το  $\partial_p$  εναλλάσσεται με την πράξη της πρόσθεσης, τότε το σύνολο των  $p$ -κύκλων του  $K$ ,  $Z_p(K)$ , είναι υποομάδα της  $C_p(K)$ , και την καλούμε **ομάδα των  $p$ -κύκλων** ( $p$ -cycle group). Με άλλα λόγια, η  $Z_p(K)$  είναι ο πυρήνας του συνοριακού τελεστή  $\partial_p$  :

$$Z_p(K) = \text{Ker } \partial_p.$$

**Ορισμός 2.3.4 :** Ένα  $p$ -σύνορο ( $p$ -boundary)  $c$  είναι μια  $p$ -αλυσίδα, η οποία είναι το σύνορο από κάποια  $(p+1)$ -αλυσίδα, δηλαδή :

$$c = \partial_{p+1} d, \text{ όπου } d \in C_{p+1}.$$

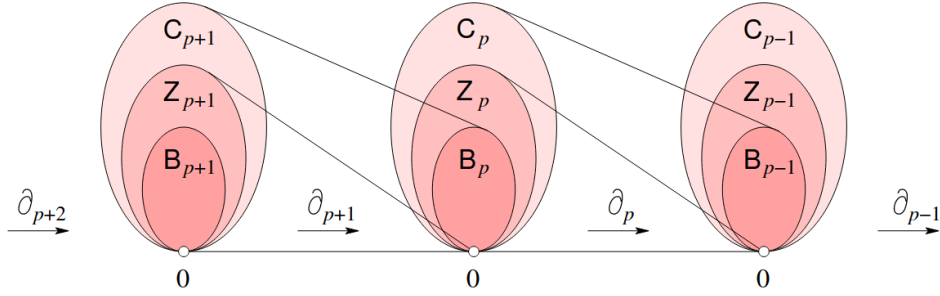
Επειδή το  $\partial_p$  εναλλάσσεται με την πράξη της πρόσθεσης, τότε το σύνολο των  $p$ -συνόρων του πλεγματού συμπλέγματος  $K$ ,  $B_p(K)$ , είναι υποομάδα της  $C_p(K)$  και την καλούμε **ομάδα  $p$ -συνόρων** ( $p$ -boundary group). Με άλλα λόγια, η  $B_p(K)$  είναι η εικόνα του συνοριακού τελεστή  $\partial_{p+1}$  :

$$B_p(K) = \text{Im } \partial_{p+1}.$$

Λόγω του θεμελιώδους λήμματος της ομολογίας, έχουμε ότι  $\partial_p B_p(K) = 0$ . Άρα :

$$B_p(K) \subseteq Z_p(K).$$

Τέλος, εύκολα αντιλαμβάνεται κανείς ότι για κάθε  $p$  η ομάδα  $C_p(K)$  αποτελεί διανυσματικό χώρο, με μία βάση του να αποτελείται από όλα τα  $p$ -simplices του  $k$ -πλεγματού συμπλέγματος  $K$ . Επίσης, οι υποομάδες  $Z_p(K), B_p(K)$  είναι γραμμικοί υπόχωροί του. Στην Εικόνα 2.5 παρουσιάζονται σχηματικά οι σχέσεις μεταξύ των ομάδων  $C_i$ , των υποομάδων τους  $Z_i, B_i$  και των συνοριακών τελεστών  $\partial_i$ .



Εικόνα 2.5 : Βλέπουμε σχηματικά όλα αυτά που αναπτύχθηκαν μέχρι στιγμής στην Ενότητα 2.3 .

Με βάση τα προηγούμενα, μπορούμε τώρα να ορίσουμε τις ομάδες ομολογίας.

### 2.3.3 Ομάδες Ομολογίας

Αφού οι ομάδες των συνόρων αποτελούν υποομάδες των ομάδων των κύκλων, μπορούμε να δημιουργήσουμε ομάδες-πηλίκα. Με άλλα λόγια, μπορούμε να διαμερίσουμε κάθε ομάδα κύκλων σε κλάσεις από κύκλους, οι οποίοι διαφέρουν μεταξύ τους ως προς κάποιο σύνορο. Συνεπώς, θα ορίσουμε τις ομάδες ομολογίας και πιο συγκεκριμένα τις πλεγματικές ομάδες ομολογίας (simplicial homology groups), αφού οι τοπολογικοί χώροι που μελετάμε είναι πλεγματικά συμπλέγματα.

**Ορισμός 2.3.5 :** Η  $p$ -πλεγματική ομάδα ομολογίας ( $p$ -th simplicial homology group) ενός  $k$ -πλεγματικού συμπλέγματος  $K$  ορίζεται να είναι η ομάδα-πηλίκο :

$$H_p(K) = Z_p(K)/B_p(K).$$

Για ευκολία απλά θα τις καλούμε  $p$ -ομάδες ομολογίας. Τα στοιχεία της  $H_p(K)$  τα καλούμε **κλάσεις ομολογίας** (homology classes) και τα συμβολίζουμε με  $[c]$ . Δύο κύκλοι στην ίδια κλάση ομολογίας καλούνται **ομόλογοι** (homologous). Η πράξη στην  $H_p(K)$  ορίζεται ως:

$$[c] + [c'] = [c + c'].$$

Τέλος, ως πηλίκο διανυσματικών χώρων, ο  $H_p(K)$  είναι διανυσματικός χώρος, με :

$$\dim H_p(K) = \dim Z_p(K) - \dim B_p(K).$$

**Ορισμός 2.3.6 :** Ως  $p$ -οστό αριθμό Betti  $\beta_p$  ορίζουμε :  $\beta_p = \dim H_p(K)$ . Το  $\beta_0$  μετράει το πλήθος των συνεκτικών συνιστωσών του  $k$ -πλεγματικού συμπλέγματος  $K$ , το  $\beta_1$  το πλήθος των οπών, το  $\beta_2$  το πλήθος των δισδιάστατων οπών (ή αλλιώς κενών), και ούτω καθεξής.

Ο ορισμός των αριθμών Betti ενός πλεγματοειδούς συμπλέγματος  $K$  μέσω της αντιστοίχισής τους με κάποια γεωμετρικά χαρακτηριστικά τού  $K$ , δεν είναι καθόλου τυχαίος.

Προτού προχωρήσουμε στην ανάλυση του ορισμού των αριθμών Betti, θα κάνουμε κάποιες συμβάσεις ως προς τους συμβολισμούς για αυτήν και μόνο την παράγραφο. Για το συμβολισμό των simplices θα χρησιμοποιήσουμε τον συμβολισμό των συνόλων, δηλαδή αν  $\sigma$  είναι ένα  $p$ -simplex, τότε  $\sigma = \{u_0, u_1, \dots, u_p\}$ . Στη συνέχεια, αν  $c$  είναι μια  $p$ -αλυσίδα, με  $c = \sum_{i=1}^{m_p} a_i \sigma_i$ , θα την συμβολίζουμε ως εξής:

$$c = \{ \{r_0, \dots, r_p\} + \{s_0, \dots, s_p\} + \dots + \{t_0, \dots, t_p\} \}$$

όπου οι προσθετέοι όροι είναι τα  $p$ -simplices που την αποτελούν και τα  $r_i, s_j, t_k$ , εν γένει όχι διαφορετικά μεταξύ τους.

Έπειτα, όσον αναφορά την πρόσθεση αλυσίδων, αν :

$$c = \{ \{r_0, \dots, r_p\} + \{s_0, \dots, s_p\} + \dots + \{t_0, \dots, t_p\} \}$$

και

$$c' = \{ \{r'_0, \dots, r'_p\} + \{s'_0, \dots, s'_p\} + \dots + \{t'_0, \dots, t'_p\} \}$$

τότε :

$$c + c' =$$

$$\{ \{r_0, \dots, r_p\} + \{s_0, \dots, s_p\} + \dots + \{t_0, \dots, t_p\} + \{r'_0, \dots, r'_p\} + \{s'_0, \dots, s'_p\} + \dots + \{t'_0, \dots, t'_p\} \}.$$

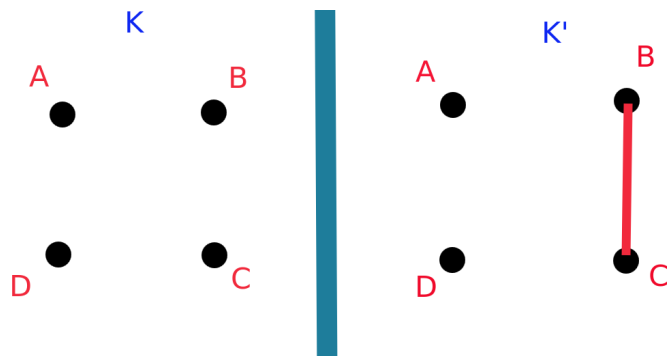
Στην περίπτωση που δρα ο συνοριακός τελεστής  $\partial_p$  στην παραπάνω  $p$ -αλυσίδα  $c$ , τότε απλά αντικαθιστούμε το κάθε  $p$ -simplex που εμφανίζεται στην αλυσίδα  $c$ , με το σύνορό του, που είναι μια  $(p-1)$ -αλυσίδα. Για παράδειγμα, ας πάρουμε το πρώτο  $p$ -simplex  $\{r_0, \dots, r_p\}$ . Τότε,  $\partial_p(\{r_0, \dots, r_p\}) = \left\{ \sum_{i=0}^p \{r_0, r_1, \dots, \tilde{r}_i, \dots, r_p\} \right\}$ , όπου  $\tilde{r}_i$  εννοούμε ότι το  $r_i$  δεν ανήκει στο αντίστοιχο  $(p-1)$ -simplex. Άρα :

$$\partial_p c = \left\{ \sum_{i=0}^p \{r_0, r_1, \dots, \tilde{r}_i, \dots, r_p\} + \sum_{i=0}^p \{s_0, s_1, \dots, \tilde{s}_i, \dots, s_p\} + \dots + \sum_{i=0}^p \{t_0, t_1, \dots, \tilde{t}_i, \dots, t_p\} \right\}.$$

Αφού κάναμε όλες αυτές τις συμβάσεις, στη συνέχεια θα αναλύσουμε τον ορισμό των αριθμών Betti.

Ας ξεκινήσουμε από την διάσταση 0 και το  $\beta_0 = \dim Z_0 - \dim B_0$ . Έστω τα πλεγματικά συμπλέγματα  $K = \{\{A\}, \{B\}, \{C\}, \{D\}\}$ ,  $K' = \{\{A\}, \{B\}, \{C\}, \{D\}, \{B, C\}\}$  που παρουσιάζονται στην Εικόνα 2.6. Αυτά τα δύο πλεγματικά συμπλέγματα έχουν μία διαφορά. Το  $K$  δεν περιέχει την ακμή  $\{B, C\}$ . Αυτό που τελικά θα δούμε είναι ότι  $\beta_0(K) = 4$  και  $\beta_0(K') = 3$ . Δηλαδή, το  $K$  έχει 4 συνεκτικές συνιστώσες και το  $K'$  έχει 3. Επειδή  $\partial_0\{u\} = 0$  για κάθε κορυφή  $u$ , τότε  $C_0(K) = Z_0(K)$  και  $C_0(K') = Z_0(K')$ . Άρα  $\dim Z_0(K) = 4$  και  $\dim Z_0(K') = 4$ , αφού μια βάση της  $C_0(K)$  είναι το σύνολο  $\{\{A\}, \{B\}, \{C\}, \{D\}\}$  και μια βάση της  $C_0(K')$  είναι το σύνολο  $\{\{A\}, \{B\}, \{C\}, \{D\}\}$ . Στο  $K$ , έχουμε ότι  $B_0(K) = \{0\}$ , αφού δεν έχουμε κάποια ακμή ώστε κάποια 0-αλυσίδα να είναι σύνορο. Άρα  $\dim B_0(K) = 0$  και συνεπώς  $\beta_0(K) = 4$ . Όμως, στην περίπτωση του  $K'$  έχουμε ένα 1-simplex και ισχύει ότι  $\partial_1\{B, C\} = \{\{B\} + \{C\}\}$ , δηλαδή η αλυσίδα  $\{\{B\} + \{C\}\} \in B_0(K')$ . Άρα,  $\dim B_0(K') = 1$  και συνεπώς  $\beta_0(K') = 3$ .

Στο παραπάνω παράδειγμα είδαμε πώς υπολογίζεται το  $\beta_0$ . Ο μέγιστος αριθμός συνεκτικών συνιστωσών που μπορεί να έχει ένα πλεγματικό σύμπλεγμα, προφανώς είναι όσος ο αριθμός των κορυφών του. Έστω ένα πλεγματικό σύμπλεγμα  $K$ , που έχει 0-simplices, 1-simplices, 2-simplices κ.ο.κ. Ας πάρουμε τον 0-σκελετό  $K^0$  του  $K$ . Τότε  $Z_0(K^0) = \#\text{κορυφών}$  και  $B_0(K^0) = \{0\}$ . Βάζοντας την πρώτη ακμή  $\{u, v\}$  του  $K$  στο  $K_0$ , το πλήθος των συνεκτικών συνιστωσών του  $K^0$  μειώνεται κατά 1. Αυτό που συμβαίνει παράλληλα στο χώρο των κύκλων  $Z_0(K^0)$  και των συνόρων  $B_0(K^0)$  είναι ότι μία 0-αλυσίδα, συγκεκριμένα η  $\{\{u\} + \{v\}\}$ , γίνεται πλέον σύνορο. Οπότε  $B_0(K^0) = \{0, \{\{u\} + \{v\}\}\}$ , δηλαδή η διάσταση του  $B_0(K^0)$  αυξήθηκε κατά ένα, αφού η βάση του πλέον είναι το  $\{\{u\} + \{v\}\}$ , οπότε και μειώθηκε τελικά το  $\beta_0(K^0)$  κατά ένα. Επαναληπτικά, είναι εύκολο να δει κανείς ότι όσο η νέα ακμή, που βάζουμε από το  $K$  στο  $K^0$ , δεν έχει σύνορο το οποίο ανήκει στο  $B_0(K^0)$ , τότε οι συνεκτικές συνιστώσες θα μειώνονται. Διαφορετικά, αν ανήκει ήδη στο  $B_0(K^0)$ , τότε παραμένει ίδιος ο αριθμός  $\beta_0(K^0)$ .



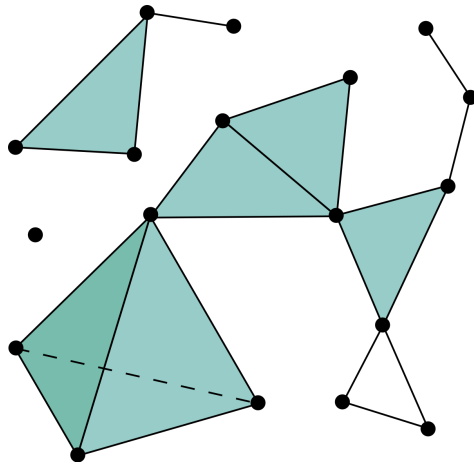
Εικόνα 2.6 : Στα αριστερά το πλεγματικό σύμπλεγμα  $K$  και στα δεξιά το πλεγματικό σύμπλεγμα  $K'$ .



Στη συνέχεια θα αναλύσουμε τους 1-κύκλους, οι οποίοι είναι κλειστές πολυγωνικές γραμμές, ενός πλεγματοειδούς συμπλέγματος  $K$  με  $\dim K \geq 1$ . Αυτή τη φορά θα πάρουμε τον 1-σκελετό  $K^1$  του  $K$  και θα εξετάσουμε τι συμβαίνει με τους 1-κύκλους που ανήκουν στον  $Z_1(K^1)$ , καθώς βάζουμε 2-simplices (τριγώνια) από το  $K$  στο  $K^1$ . Αρχικά, έχουμε ότι  $B_1(K^1) = \{0\}$  και άρα  $\beta_1(K^1) = \dim Z_1(K^1)$ . Βάζουμε ένα 2-simplex, έστω το  $\{u, v, w\}$ , από το  $K$  στο  $K^1$ . Τότε το σύνορο του  $\{u, v, w\}$ , που είναι  $\partial_1\{u, v, w\} = \{\{u, v\} + \{v, w\} + \{w, u\}\}$ , είναι 1-κύκλος, ο οποίος τελικά γίνεται 1-σύνορο εξαιτίας της τοποθέτησης αυτού του τριγώνου. Άρα το  $B_1(K^1) = \{0, \{\{u, v\} + \{v, w\} + \{w, u\}\}\}$  και η βάση του είναι η  $\{\{u, v\} + \{v, w\} + \{w, u\}\}$ , που έχει 1 στοιχείο. Συνεπώς  $\dim B_1(K^1) = 1$  κι άρα το  $\beta_1(K^1)$  μειώνεται κατά ένα. Επαναληπτικά, βάζοντας τριγώνια από το  $K$  στο  $K^1$ , ελέγχουμε αν το σύνορό τους ανήκει στο  $B_1(K^1)$ . Αν ανήκει ήδη, τότε δεν αυξάνεται η διάσταση  $\dim B_1(K^1)$ , αλλιώς αυξάνεται κατά 1. Αυτό που κάνουν στην ουσία τα 2-simplices είναι να γεμίζουν το εσωτερικό-οπή κάθε 1-κύκλου, δηλαδή κάθε κλειστής πολυγωνικής γραμμής. Όταν επιτυγχάνεται το γέμισμα της οπής κάποιας κλειστής πολυγωνικής γραμμής, τότε αυξάνεται κατά ένα η διάσταση του  $B_1(K^1)$ .

Τέλος, όσον αναφορά το  $\beta_2$ , το οποίο μετράει το πλήθος των διδιάστατων κενών, δηλαδή κενών σαν το εσωτερικό κενό μιας σφαίρας, αρκεί κανείς να συνειδητοποιήσει ότι οι 2-κύκλοι είναι κλειστές επιφάνειες, που προκύπτουν από συγκόλληση 2-simplices (τριγώνων) με τρόπο ώστε να μην παραβιάζεται ο ορισμός του πλεγματοειδούς συμπλέγματος, οι οποίες γεμίζουν από 3-simplices (τετράεδρα).

**Παράδειγμα:** Το πλεγματοειδές σύμπλεγμα της Εικόνας 2.7 αποτελεί ένα απλό παράδειγμα για να διαπιστώσει κανείς όλα αυτά που αναπτύχθηκαν για τους αριθμούς Betti.



Εικόνα 2.7 : Στην παραπάνω εικόνα βλέπουμε ότι  $\beta_0 = 3$ ,  $\beta_1 = 1$ . Αν το τετράεδρο είναι κενό στο εσωτερικό του, τότε  $\beta_2 = 1$ , διαφορετικά  $\beta_2 = 0$ .

## 2.4 Επαγόμενες απεικονίσεις στην Ομολογία

Προτού προχωρήσουμε στο επόμενο κεφάλαιο και τον ορισμό της επιμένουσας ομολογίας (persistence homology), θα χρειαστεί να ορίσουμε πρώτα κάποιες απεικονίσεις και να μελετήσουμε μερικές ιδιότητές τους.

**Ορισμός 2.4.1 :** Έστω  $K$  και  $K'$  δύο αφηρημένα πλεγματικά συμπλέγματα. Μία απεικόνιση  $f : K \rightarrow K'$  λέγεται **πλεγματική απεικόνιση** (simplicial map) αν έχει την παρακάτω ιδιότητα: όποτε το σύνολο  $\{u_0, \dots, u_m\}$  είναι simplex στο  $K$ , τότε το σύνολο  $\{f(u_0), \dots, f(u_m)\}$  είναι simplex στο  $K'$ .

**Ορισμός 2.4.2 :** Έστω τώρα τα πλεγματικά συμπλέγματα  $K, K'$  και μία πλεγματική απεικόνιση  $f : K \rightarrow K'$ . Αυτή η απεικόνιση επάγει την **επαγόμενη απεικόνιση  $p$ -αλυσίδων** (induced  $p$ -chain map):

$$\begin{aligned} \tilde{f}_p : C_p(K) &\rightarrow C_p(K') \\ \sum a_i \sigma &\mapsto \sum a_i f(\sigma) \end{aligned}$$

Με βάση τις επαγόμενες απεικονίσεις μπορούμε να εξάγουμε τα παρακάτω λήμματα.

**Λήμμα 2.4.1** Έστω μία πλεγματική απεικόνιση  $f : K \rightarrow K'$ . Τότε ισχύει η σχέση  $\partial'_p \circ \tilde{f}_p = \tilde{f}_{p-1} \circ \partial_p$ , για κάθε  $p \geq 0$ . Δηλαδή, το παρακάτω διάγραμμα είναι μεταθετικό:

$$\begin{array}{ccc} C_p(K) & \xrightarrow{\partial_p} & C_{p-1}(K) \\ \tilde{f}_p \downarrow & & \downarrow \tilde{f}_{p-1} \\ C_p(K') & \xrightarrow{\partial'_p} & C_{p-1}(K') \end{array}$$

**Απόδειξη :** Έστω ένα  $p$ -simplex  $\sigma \in K$ . Εξ ορισμού της ομάδας  $C_p(K)$ , το  $\sigma$  είναι γεννήτορας της. Τώρα έχουμε ότι:

$$\begin{aligned} (\partial'_p \circ \tilde{f}_p)(\sigma) &= \partial'_p(\tilde{f}_p(\sigma)) = \sum_{f(v) \in f(\sigma)} (f(\sigma))^{(f(v))} = \sum_{v \in \sigma} f(\sigma^{(v)}) = \\ &= \tilde{f}_{p-1} \left( \sum_{v \in \sigma} \sigma^{(v)} \right) = \tilde{f}_{p-1}(\partial_p(\sigma)) = (\tilde{f}_{p-1} \circ \partial_p)(\sigma). \end{aligned}$$

Αφού η σχέση  $\partial'_p \circ \tilde{f}_p = \tilde{f}_{p-1} \circ \partial_p$  ισχύει για κάθε γεννήτορα της  $C_p(K)$  και επειδή οι απεικονίσεις  $\partial'_p, \tilde{f}_p, \tilde{f}_{p-1}, \partial_p$  είναι γραμμικές, έχουμε ότι ισχύει :

$$\partial'_p \circ \tilde{f}_p = \tilde{f}_{p-1} \circ \partial_p, \text{ για κάθε } p \geq 0.$$

Να θυμίσουμε ότι αν  $\sigma = [u_0, u_1, \dots, u_p]$ , με τον συμβολισμό  $\sigma^{(u_i)}$  εννοούμε την αφαίρεση της κορυφής  $u_i$  από το σύνολο  $[u_0, u_1, \dots, \tilde{u}_i, \dots, u_p]$ .  $\square$

**Λήμμα 2.4.2** Έστω μία πλεγματική απεικόνιση  $f : K \rightarrow K'$ . Τότε για κάθε  $p \geq 0$ :

1.  $\tilde{f}_p(Z_p(K)) \subset Z_p(K')$ .
2.  $\tilde{f}_p(B_p(K)) \subset B_p(K')$ .

**Απόδειξη :**

Για το 1 : Έστω  $c \in Z_p(K)$ . Τότε  $\partial_p(c) = 0$ . Επομένως από το Λήμμα 2.4.1 έχουμε :  $\partial'_p(\tilde{f}_p(c)) = \tilde{f}_{p-1}(\partial_p(c)) = \tilde{f}_{p-1}(0) = 0$ . Άρα  $\tilde{f}_p(c) \in Z_p(K')$ .

Για το 2 : Έστω  $c' \in B_p(K)$ . Θέλουμε να δείξουμε ότι  $\tilde{f}_p(c') \in B_p(K')$ , δηλαδή, εξ'ορισμού της ομάδας των συνόρων, ότι  $\tilde{f}_p(c') = \partial_{p+1}(\sigma)$ , για κάποιο  $\sigma \in C_{p+1}(K')$ . Γνωρίζουμε ότι υπάρχει  $\beta \in C_{p+1}(K)$  τέτοιο ώστε  $c' = \partial_{p+1}(\beta)$ . Επομένως :  $\tilde{f}_p(c') = \tilde{f}_p(\partial_{p+1}(\beta)) = \partial'_{p+1}(\tilde{f}_{p+1}(\beta))$ .

**Λήμμα 2.4.3** Η παρακάτω γραμμική απεικόνιση είναι καλά ορισμένη.

$$f_p : H_p(K) \rightarrow H_p(K')$$

$$[c] \mapsto [\tilde{f}_p(c)]$$

**Απόδειξη :** Η απεικόνιση είναι καλά ορισμένη, αφού το  $c$  είναι κύκλος και από το προηγούμενο Λήμμα 2.4.2 το  $\tilde{f}_p(c)$  είναι πάλι κύκλος. Επίσης, η επιλογή εκπροσώπου δεν επηρεάζει το αποτέλεσμα. Αν το  $c$  είναι εκπρόσωπος κάποιου συμπλόκου τότε και το  $c + b, b \in B_p(K)$  είναι εκπρόσωπος του ίδιου συμπλόκου. Το  $\tilde{f}_p(c + b) + B_p(K')$  είναι σύμπλοκο της  $H_p(K')$ . Από την άλλη  $\tilde{f}_p(c + b) + B_p(K') = \tilde{f}_p(c) + \tilde{f}_p(b) + B_p(K')$ , λόγω της γραμμικότητας της  $\tilde{f}_p$ . Τέλος,  $\tilde{f}_p(c) + \tilde{f}_p(b) + B_p(K') = \tilde{f}_p(c) + B_p(K')$ , διότι  $\tilde{f}_p(B_p(K)) \subset B_p(K')$ .  $\square$

**Θεώρημα 2.4.1 :** Έστω  $K, K'$  και  $K''$  πλεγματικά συμπλέγματα. Έστω επίσης και οι πλεγματικές απεικονίσεις  $f : K \rightarrow K'$  και  $g : K' \rightarrow K''$ . Τότε :

$$(g \circ f)_p = g_p \circ f_p.$$

Δηλαδή, η γραμμική απεικόνιση του λήμματος 2.4.3 που αντιστοιχεί στη σύνθεση  $g \circ f$  είναι η ίδια με την σύνθεση των γραμμικών απεικονίσεων  $f_p, g_p$  του Λήμματος 2.4.3 που αντιστοιχούν στις  $f$  και  $g$ .

**Απόδειξη :** Έστω  $[c] \in H_p(K)$ . Από Λήμμα 2.4.3 έχουμε ότι  $(g \circ f)_p([c]) = [(g \circ f)_p(c)]$ . Όμως,  $(g \circ f)_p(c) = \sum a_i (g \circ f)(c) = \sum a_i g(f(c)) = \tilde{g}_p(\sum a_i f(c)) = \tilde{g}_p \circ \tilde{f}_p$ , το οποίο προκύπτει από τον Ορισμό 2.4.2 των  $\tilde{g}_p, \tilde{f}_p$ . Άρα :

$$(g \circ f)_p([c]) = [\widetilde{(g \circ f)_p(c)}] = [(\tilde{g}_p \circ \tilde{f}_p)(c)] = [\tilde{g}_p(\tilde{f}_p(c))] = g_p([\tilde{f}_p(c)]) = g_p(f_p([c])).$$

το οποίο είναι και το ζητούμενο.  $\square$

### 3 Επιμένουσα Ομολογία και Επιμένουσες Εικόνες (Persistent Homology and Persistent Images)

Έπειτα από όλη αυτή την εισαγωγή εργαλείων και εννοιών από την τοπολογία και την άλγεβρα, έχουμε ό,τι είναι απαραίτητο για να μπορέσουμε να ορίσουμε την Επιμένουσα Ομολογία, που είναι μια από τις κύριες μεθόδους που χρησιμοποιούνται στην Τοπολογική Ανάλυση Δεδομένων. Πρώτα από όλα, όμως, να πούμε τι κάνει στην ουσία αυτή η Επιμένουσα Ομολογία. Δεδομένου ενός νέφους σημειακών δεδομένων και χρησιμοποιώντας κάποιον αλγόριθμο κατασκευής πλεγματού συμπλέγματος πάνω στα δεδομένα, για παράδειγμα Vietoris-Rips Complex ή Čech Complex (βλ. ορισμός 2.2.2, ορισμός 2.2.3), καθώς μεταβάλλεται μία συγκεκριμένη παράμετρος (π.χ. η ακτίνα στα προαναφερθέντα συμπλέγματα), μεταβάλλεται και το ίδιο το πλεγματού σύμπλεγμα και συνεπώς και τα αλγεβρικά/τοπολογικά χαρακτηριστικά του. Η δουλειά της επιμένουσας ομολογίας είναι να καταγράφει τη μεταβολή αυτών των αλγεβρικών/τοπολογικών χαρακτηριστικών του πλεγματού συμπλέγματος καθώς αυτό μεταμορφώνεται ταυτόχρονα με τη μεταβολή της παραμέτρου μέσα σε κάποιο διάστημα.

Τέλος, αφού παρουσιάσουμε την Επιμένουσα Ομολογία, ορίζουμε τις Επιμένουσες Εικόνες (Persistent Images-PI), οι οποίες μετασχηματίζουν τα αποτελέσματα της Επιμένουσας Ομολογίας σε διανύσματα, ώστε αυτά να μελετηθούν με αλγορίθμους της Μηχανικής Μάθησης.

#### 3.1 Επιμένουσα Ομολογία

**Ορισμός 3.1.1 :** Έστω ένα πλεγματού σύμπλεγμα  $K$  και μία συνάρτηση  $\rho : K \rightarrow \mathbb{R}$  με την ιδιότητα:

$$\text{Αν το } \sigma \text{ είναι μία όψη του } \tau, \text{ τότε } \rho(\sigma) \leq \rho(\tau).$$

Τότε λέμε ότι η  $\rho$  είναι **μονοτονική**.

Λόγω αυτής της ιδιότητας της  $\rho$ , το σύνολο  $K(\alpha) = \rho^{-1}((-\infty, \alpha])$  είναι υποσύμπλεγμα (subcomplex) του  $K$  για κάθε  $\alpha \in \mathbb{R}$ .

**Ορισμός 3.1.2 :** Ένα **φιλτράρισμα** (filtration) :

$$\mathcal{F}_\rho : \emptyset = K_{-1} \subset K_0 \subset K_1 \subset \dots \subset K_n = K,$$

είναι μια ακολουθία από υποσυμπλέγματα του  $K$ , που παράγεται από τις τιμές  $-\infty = \alpha_{-1} < \alpha_0 < \alpha_1 < \dots < \alpha_n$ , όπου  $K_i = K(\alpha_i) = \rho^{-1}((-\infty, \alpha_i])$  για κάθε  $i \in \{-1, \dots, n\}$ . Αφού  $K_i \subset K_{i+1}$  για κάθε  $i$ , τότε έχουμε μια ακολουθία από εμφυτεύσεις :

$$f^{i,j} : K_i \hookrightarrow K_j .$$

Να σημειώσουμε ότι αυτές οι εμφυτεύσεις  $f^{i,j}$  από το  $K_i$  στο  $K_j$  είναι πλεγματού απεικονίσεις. Επομένως η καθεμιά τους επάγει, σύμφωνα με τα προηγούμενα, τις παρακάτω γραμμικές απεικονίσεις :

$$f_p^{i,j} : H_p(K_i) \rightarrow H_p(K_j).$$

Επίσης, σύμφωνα με το Θεώρημα 2.4.1 έχουμε ότι :

$$f_p^{k,j} \circ f_p^{i,k} = f_p^{i,j} \text{ για κάθε } i \leq k \leq j.$$

**Ορισμός 3.1.2 :** Έστω ότι έχουμε ένα φιλτράρισμα  $\emptyset = K_{-1} \subset K_0 \subset K_1 \subset \dots \subset K_n = K$  και τις γραμμικές απεικονίσεις  $f_p^{i,j} : H_p(K_i) \rightarrow H_p(K_j)$  που επάγονται από τις εμφυτεύσεις από το  $K_i$  στο  $K_j$  για κάθε  $-1 \leq i \leq j \leq n$ . Ορίζουμε ως  **$p$ -ομολογικό πρότυπο** ( $p$ -homology module) του  $K$  την ακολουθία :

$$(\{H_p(K_i)\}_{-1 \leq i \leq n}, \{f_p^{i,j}\}_{-1 \leq i \leq j \leq n})$$

Δηλαδή, σχηματικά για κάθε  $p$  η ακολουθία είναι :

$$0 = H_p(K_0) \rightarrow H_p(K_1) \rightarrow \dots \rightarrow H_p(K_i) \xrightarrow{f_p^{i,j}} \dots \rightarrow H_p(K_j) \rightarrow \dots \rightarrow H_p(K_n) = H_p(K).$$

**Ορισμός 3.1.3 :** Οι  **$p$ -επιμένουσες ομάδες ομολογίας** ( $p$ -th persistent homology groups) είναι οι εικόνες των ομομορφισμών, οι οποίοι επάγονται από τις εμφυτεύσεις, δηλαδή  $H_p^{i,j} = \text{im } f_p^{i,j}$ , για κάθε  $0 \leq i \leq j \leq n$ . Οι αντίστοιχοι  **$p$ -οστοί επιμένοντες αριθμοί Betti** ( $p$ -th persistent Betti numbers) είναι τα  $\beta_p^{i,j} = \text{rank } H_p^{i,j}$ .

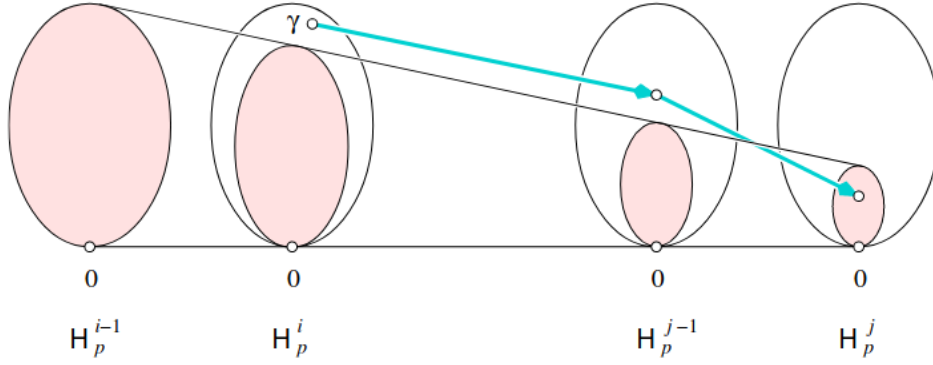
Να σημειώσουμε ότι  $H_p^{i,i} = H_p(K_i)$ , εφόσον  $f_p^{i,i} = \text{id}_{H_p(K_i)}$ . Οι  $p$ -επιμένουσες ομάδες ομολογίας αποτελούνται από τις κλάσεις ομολογίας του  $K_i$  που συνεχίζουν να υπάρχουν στο  $K_j$ , δηλαδή  $H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i))$ .

**Ορισμός 3.1.4 :** Μία μη τετριμμένη κλάση ομολογίας  $\gamma \in H_p(K_i)$  λέμε ότι **γεννιέται** στο  $K_i$ , στο χρόνο  $i = \text{birth}(\gamma)$ , αν  $\gamma \notin H_p^{i-1,i}$ . Ομοίως, μία μη τετριμμένη κλάση ομολογίας  $\gamma \in H_p(K_i)$  **πεθαίνει** καθώς εισέρχεται στο  $K_j$ ,  $i < j$ , στο χρόνο  $j = \text{death}(\gamma)$ , αν η  $f_p^{i,j-1}(\gamma)$  είναι μη μηδενική μη-τετριμμένα (δηλαδή όχι όταν  $f_p^{i,j-1} = 0$ ) αλλά  $f_p^{i,j}(\gamma) = 0$ . Ως **επιμονή** (persistence) της κλάσης  $\gamma$  ορίζουμε :

$$\text{death}(\gamma) - \text{birth}(\gamma) = j - i.$$

Στην Εικόνα 3.1 παρουσιάζεται η γέννηση της κλάσης  $\gamma$  στο  $K_i$  και ο θάνατός της στο  $K_j$ .

**Ορισμός 3.1.5 :** Οι κλάσεις ομολογίας  $\alpha \in H_p(K_i)$  και  $\beta \in H_p(K_j)$ , με  $i < j$ , λέμε ότι **συγχωνεύονται** στην  $H_p(K_m)$ , με  $m > j$ , αν  $f_p^{i,m}(\alpha) = f_p^{j,m}(\beta)$ . Τότε, λέμε ότι το  $\alpha$  είναι **γρηραιότερο** του  $\beta$ .



Εικόνα 3.1

### 3.2 Διαγράμματα επιμονής και Barcodes

Απεικονίζοντας σημεία στις δύο διαστάσεις, μπορούμε να οπτικοποιήσουμε τη συλλογή από επιμένοντες αριθμούς Betti. Κάποια σημεία μπορεί να έχουν κάποια συντεταγμένη άπειρη και κάποια μπορεί να ταυτίζονται μεταξύ τους. Για αυτό θα μιλάμε για πολυσύνολα (multisets) από σημεία στον επεκταμένο Ευκλείδιο χώρο  $\overline{\mathbb{R}^2}$ .

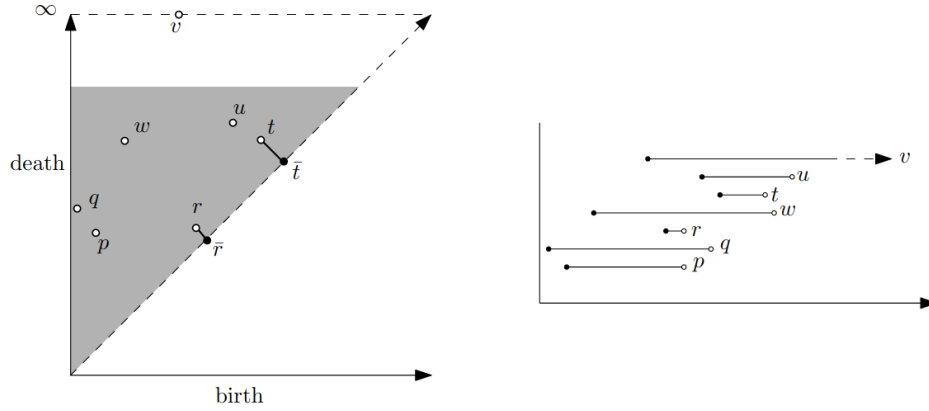
Αν  $\mu_p^{i,j}$  είναι ο αριθμός των  $p$ -διάστατων κλάσεων που γεννιούνται στο  $K_i$  και πεθαίνουν στο  $K_j$  έχουμε ότι :

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$$

για κάθε  $p$  και  $i < j$ . Πράγματι, η πρώτη διαφορά μετρά τις κλάσεις που γεννιούνται πριν από το  $i$  ή στο  $K_i$  και πεθαίνουν στο  $K_j$ , ενώ η δεύτερη διαφορά μετρά τις κλάσεις που γεννιούνται πριν από το  $i-1$  ή στο  $K_{i-1}$  και πεθαίνουν στο  $K_j$ . Σημειώνοντας κάθε ως άνω σημείο  $(i, j)$  με πολλαπλότητα του  $\mu_p^{i,j}$  παίρνουμε το  $p$ -**Διάγραμμα Επιμονής** ( $p$ -Persistent Diagram-PD) για τη συγκεκριμένη συνάρτηση  $\rho$  που μας δίνει το παραπάνω φιλτράρισμα και το συμβολίζουμε με  $Dgm_p(\mathcal{F}_\rho)$ . Στο διάγραμμα αυτό λαμβάνουμε υπόψιν για τεχνικούς λόγους, όπως θα δούμε παρακάτω, και την κύρια διαγώνιο, όπου κάθε σημείο της θεωρείται ότι έχει άπειρη πολλαπλότητα. Τα σημεία που βρίσκονται πολύ κοντά στη διαγώνιο ονομάζονται **θόρυβος** (noise).

Από την άλλη τα **Barcodes** παρέχουν τις ίδιες πληροφορίες, μόνο που για σημεία απεικονίζουμε ευθύγραμμα τμήματα με μήκος όσο η διάρκεια ζωής-επιμονή  $j - i$ .

Στην Εικόνα 3.2 παρουσιάζεται κάποιο διάγραμμα επιμονής και το αντίστοιχό του Barcode. Κάποια σημεία έχουνε ιδιαίτερη σημασία, όπως το σημείο  $v$  που έχει άπειρη επιμονή, καθώς επίσης και τα σημεία θορύβου κοντά στη διαγώνιο  $r$  και  $t$ .



Εικόνα 3.2 : Αριστερά το διάγραμμα επιμονής και δεξιά το barcode, για κάποια διάσταση  $p$ .

Έστω τώρα δύο διαγράμματα επιμονής  $Dgm_p(\mathcal{F}_f)$  και  $Dgm_p(\mathcal{F}_g)$  για δύο συναρτήσεις  $f, g$ . Τώρα θέλουμε να εξετάσουμε όλες τις αμφιμονοσήμαντες απεικονίσεις από τα σημεία του  $Dgm_p(\mathcal{F}_f)$  στα σημεία του  $Dgm_p(\mathcal{F}_g)$ . Επειδή μπορεί να διαφέρουν ως προς το πλήθος των μη διαγώνιων στοιχείων, μπορούμε να δανειζόμαστε σημεία από τη διαγώνιο, τα οποία έχουν άπειρη πολλαπλότητα όπως είπαμε, για να φτιάχνουμε τις αμφιμονοσήμαντες απεικονίσεις. Αυτός είναι και ο λόγος που συμπεριλάβαμε τις διαγωνίους στα διαγράμματα επιμονής.

**Ορισμός 3.2.1 :** Έστω το σύνολο :

$$\Pi = \{\pi : Dgm_p(\mathcal{F}_f) \rightarrow Dgm_p(\mathcal{F}_g) \mid \pi \text{ αμφιμονοσήμαντη}\}$$

με όλες τις αμφιμονοσήμαντες απεικονίσεις των διαγραμμάτων. Έστω και η νόρμα άπειρο  $\|\cdot\|_\infty$  μεταξύ δύο σημείων του  $\mathbb{R}^2$ . Ορίζουμε ως **Bottleneck distance** την παρακάτω ποσότητα :

$$d_B(Dgm_p(\mathcal{F}_f), Dgm_p(\mathcal{F}_g)) := \inf_{\pi \in \Pi} \sup_{x \in Dgm_p(\mathcal{F}_f)} \|x - \pi(x)\|_\infty$$

Μέσω αυτής τη απόστασης διαπιστώνεται το παρακάτω σημαντικό θεώρημα ευστάθειας των φιλτραρισμάτων [4], μέσω του οποίου μας δίνεται η δυνατότητα να “πειράζουμε” τις συναρτήσεις που τα παράγουν όσο θέλουμε, χωρίς να χάσουμε σημαντικές πληροφορίες.

**Θεώρημα 3.2.1 :** (Θεώρημα Ευστάθειας για Φιλτραρίσματα) Έστω  $K$  ένα πλεγματοειδές σύμπλεγμα και  $f, g : K \rightarrow \mathbb{R}$  δύο μονοτονικές συναρτήσεις. Για κάθε διάσταση  $p$ , η bottleneck distance μεταξύ των διαγραμμάτων  $Dgm_p(\mathcal{F}_f), Dgm_p(\mathcal{F}_g)$  φράσσεται από πάνω από την  $L_\infty$ -νόρμα της διαφοράς των συναρτήσεων  $f, g$ , δηλαδή:

$$d_B(Dgm_p(\mathcal{F}_f), Dgm_p(\mathcal{F}_g)) \leq \|f - g\|_\infty.$$

Σημειώνεται ότι  $\|f - g\|_\infty := \sup_{\sigma \in K} |f(\sigma) - g(\sigma)|$ .

### 3.3 Επιμένουσες Εικόνες

Μία από τις γέφυρες που ενώνει τη Μηχανική Μάθηση με την Τοπολογική Ανάλυση Δεδομένων είναι οι Επιμένουσες Εικόνες. Αυτές είναι το αποτέλεσμα του μετασχηματισμού των διαγραμμάτων επιμονής σε διανύσματα, με σκοπό να αξιοποιηθούν από αλγόριθμους της μηχανικής μάθησης.

Έστω  $B$  ένα PD με συντεταγμένες γέννησης-θανάτου. Έστω ένας γραμμικός μετασχηματισμός  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  με τύπο:

$$T(x, y) = (x, y - x)$$

και έστω  $T(B)$  το μετασχηματισμένο πολυσύνολο των γεννήσεων-θανάτων. Έστω επίσης  $\varphi_u : \mathbb{R}^2 \rightarrow \mathbb{R}$  παραγωγίσιμη κατανομή πιθανότητας με μέσο  $u = (u_x, u_y) \in \mathbb{R}^2$ . Στην πλειοψηφία των περιπτώσεων, επιλέγεται ως  $\varphi_u$  η κανονικοποιημένη συμμετρική Γκαουσιανή,  $g_u(x, y)$ , με μέσο  $u$  και διασπορά  $\sigma^2$  που έχει τύπο :

$$g_u(x, y) = \frac{1}{2\pi\sigma^2} e^{-[(x-u_x)^2 + (y-u_y)^2]/2\sigma^2}$$

Επιλέγουμε επίσης μία συνάρτηση βάρους  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  η οποία είναι μηδενική στον οριζόντιο άξονα (για τεχνικούς λόγους πάλι, που δεν θα επεκταθούμε σε αυτήν την εργασία), συνεχής και κατά τμήματα παραγωγίσιμη. Με όλα αυτά τα στοιχεία, μετατρέπουμε το PD σε μία βαθμωτή συνάρτηση πάνω από το  $\mathbb{R}^2$ .

**Ορισμός 3.3.1 :** Για ένα PD  $B$ , η αντίστοιχη **επιμένουσα επιφάνεια** (persistence surface)  $\rho_B : \mathbb{R}^2 \rightarrow \mathbb{R}$  είναι η συνάρτηση :

$$\rho_B(z) = \sum_{u \in T(B)} f(u)\varphi_u(z).$$

Η επιλογή της  $f$  είναι σημαντική για να εξασφαλίσουμε ότι ο μετασχηματισμός από το PD στην επιμένουσα επιφάνεια είναι ευσταθής.

Τέλος, η επιφάνεια  $\rho_B$  περιορίζεται σε ένα πεπερασμένο διάστασης διάνυσμα τιμών, μέσω της διακριτοποίησης μιας σχετικής με το PD  $B$  υποπεριοχής του  $\mathbb{R}^2$  και ολοκλήρωσης της  $\rho_B$  πάνω από κάθε περιοχή της διακριτοποιημένης υποπεριοχής. Δηλαδή, σταθεροποιούμε ένα πλέγμα στο επίπεδο με  $n$  κελιά (pixels) και αντιστοιχούμε στο κάθε κελί το ολοκλήρωμα της  $\rho_B$  πάνω από το αντίστοιχο κελί.



**Ορισμός 3.3.2 :** Για ένα PD  $B$ , η **Επιμένουσα Εικόνα** (Persistent Image-PI) του είναι η διατεταγμένη συλλογή από  $\{I(\rho_B)_{q_i}\}_{1 \leq i \leq n}$ , όπου  $I(\rho_B)_{q_i} = \iint_{q_i} \rho_B dydx$  και την ολοκλήρωση να γίνεται στο ορθογώνιο  $[x'_{q_i}, x_{q_i}] \times [y'_{q_i}, y_{q_i}]$  του κελιού  $q_i$ .

Οι PIs παρέχουν ένα βολικό τρόπο για να συνδυαστούν διάφορα PDs από διαφορετικές ομολογικές διαστάσεις σε ένα και μόνο διάνυσμα. Πράγματι, αν έχουμε υπολογίσει τα PDs των  $H_0, H_1, \dots, H_k$ , τότε μπορούμε να παραθέσουμε τα PI διανύσματα των  $H_0, H_1, \dots, H_k$  σε ένα και μόνο διάνυσμα που αντιπροσωπεύει όλες τις ομολογικές διαστάσεις ταυτόχρονα, και το οποίο αξιοποιούμε με αλγορίθμους Μηχανικής Μάθησης.

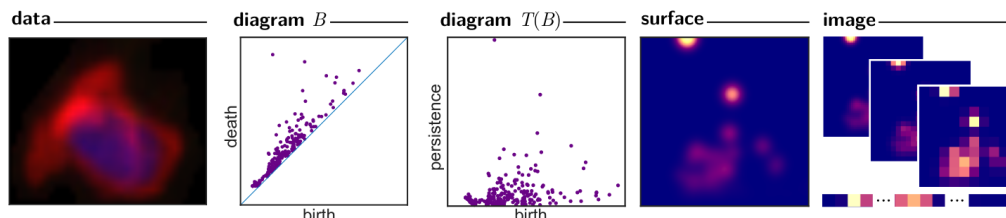
Για την δημιουργία των PIs ο χρήστης κάνει τρεις επιλογές : την ανάλυση (resolution), την κατανομή (distribution) μαζί με τις αντίστοιχες παραμέτρους της, και τη συνάρτηση βάρους. Ένα πλεονέκτημα των PIs είναι η ευελιξία τους, ενώ ένα μειονέκτημά τους είναι ότι δεν υπάρχει συνταγή για την κατάλληλη επιλογή των παραπάνω χαρακτηριστικών για την κάθε υπό-μελέτη περίπτωση.

**Η ανάλυση της εικόνας :** Η ανάλυση μιας PI αντιστοιχεί στο πλέγμα που απλώνεται πάνω στο PD.

**Η συνάρτηση κατανομής :** Η μέθοδος αυτή απαιτεί την επιλογή μιας συνάρτησης κατανομής πιθανότητας συσχετισμένη με κάθε σημείο του PD. Όπως αναφέραμε, η επιλογή της διασποράς, για παράδειγμα στη Γκαουσιανή, είναι πάλι στο χέρι του χρήστη και όπως έχει δείχθει στην [9], δεν επιφέρει σημαντικές αλλαγές στα αποτελέσματα.

**Η συνάρτηση βάρους :** Προκειμένου να ισχύουν τα αποτελέσματα στην εργασία [5] για την ευστάθεια της μετάβασης από τα PDs στα PIs, πρέπει η  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  να είναι μηδενική πάνω στον οριζόντιο άξονα, συνεχής και κατά τμήματα παραγωγίσιμη. Μια από τις επιλογές είναι η συνάρτηση βάρους, που εξαρτάται μόνο από την επιμένουσα συνιστώσα  $y$  στο διάγραμμα  $T(B)$ . Υπάρχουν όμως και περιπτώσεις όπου τα σημεία με μικρή ή μεσαία επιμονή, έχουν καλύτερη επιρροή στα αποτελέσματα, οπότε και δίνεται διαφορετική βαρύτητα σε αυτά.

Η Εικόνα 3.3 απεικονίζει την παραπάνω διαδικασία. Ξεκινάει από τα δεδομένα, από τα οποία υπολογίζεται κάποιο  $p$ -PD  $B$ , στη συνέχεια μετασχηματίζεται το  $p$ -PD  $B$  στο διάγραμμα  $T(B)$  και έπειτα ολοκληρώνεται η  $\rho_B$  για 3 τιμές της ανάλυσης της εικόνας (τα τρία εικονίδια). Τέλος, το μπροστά εικονίδιο μετατρέπεται σε ένα διατεταγμένο διάνυσμα, ώστε να αξιοποιηθεί από αλγορίθμους της Μηχανικής Μάθησης.



Εικόνα 3.3 : Σχηματικά ο αλγόριθμος που μετασχηματίζει τα δεδομένα σε PI.

## 4 Mapper

Ένα άλλο επίσης σπουδαίο εργαλείο είναι ο αλγόριθμος Mapper. Τα τελευταία χρόνια έχει χρησιμοποιηθεί κατα κύριο λόγο στην Ιατρική, επιτυγχάνοντας σημαντικά αποτελέσματα, όπως η ανάδειξη υποτύπων του καρκίνου του μαστού και υποτύπων του διαβήτη τύπου 2 (βλ. Εικόνα 4.1 και Εικόνα 4.2), αλλά και σε άλλους τομείς, όπως το μπάσκετ, όπου ανέδειξε 12 διαφορετικούς ρόλους παικτών έναντι των κλασικών 5. Όσον αφορά την Ιατρική, στην ουσία δείχνει την κατεύθυνση και όχι ακριβώς την ανακάλυψη, δηλαδή κατά κάποιο τρόπο ενθαρρύνει τους ερευνητές να κινηθούν προς μια κατεύθυνση, γιατί υπάρχουν ενδείξεις ότι κάτι ενδιαφέρον υπάρχει προς τα εκεί.

### 4.1 Ο Αλγόριθμος Mapper

**Ορισμός 4.1.1 :** Έστω  $f : X \rightarrow \mathbb{R}^d$ ,  $d \geq 1$ , μία συνεχής πραγματική διανυσματική συνάρτηση και έστω  $\mathcal{U} = \{U_a\}_{a \in A}$  κάλυμμα του  $im(f)$ . Έστω επίσης και το κάλυμμα pull back του  $X$  που ορίζεται να είναι το  $\mathcal{V} = \{f^{-1}(U_a) \mid U_a \in \mathcal{U}\}$ , δηλαδή  $X \subseteq \bigcup_{a \in A} f^{-1}(U_a)$ .

Τώρα κάθε σύνολο  $f^{-1}(U_a)$  το γράφουμε ως την ένωση των συνεκτικών του συνιστωσών, δηλαδή  $f^{-1}(U_a) = \bigcup_{j=1}^{j_a} V_{a,j}$ , όπου  $j_a$  είναι το πλήθος των συνεκτικών συνιστωσών του  $f^{-1}(U_a)$ . Αφού κάνουμε αυτή τη διαμέριση, ορίζουμε ως **εκλεπτυσμένο** (refined) κάλυμμα pull back του  $X$  το  $\tilde{\mathcal{V}} := \bigcup_{a \in A} \bigcup_{j=1}^{j_a} V_{a,j}$ . Τέλος, ορίζουμε ως **Mapper**  $M_f(X, \mathcal{U})$  να είναι το νεύρο του καλύμματος  $\tilde{\mathcal{V}}$ , δηλαδή :

$$M_f(X, \mathcal{U}) := \mathcal{N}(\tilde{\mathcal{V}}).$$

Στην περίπτωση που ο  $X$  είναι ένα νέφος σημειακών δεδομένων, δηλαδή ένα πεπερασμένο σύνολο σημείων εφοδιασμένο με κάποια μετρική, τότε η διάσπαση των  $f^{-1}(U_a)$  σε συνεκτικές συνιστώσες γίνεται εφαρμόζοντας τον ίδιο clustering αλγόριθμο σε κάθε τέτοιο σύνολο. Δηλαδή, τα  $V_{a,j}$  είναι σύνολα κορυφών, με  $V_{a,j} \cap V_{a,k} = \emptyset$  για κάθε  $j \neq k$ , με  $j, k \in \{1, \dots, j_a\}$ . Κάθε τέτοιο σύνολο  $V_{a,j}$  θα αποτελεί και ένα κόμβο στο τελικό γράφημα. Επιπλέον, μια ακμή θα ενώνει δύο κόμβους του τελικού γραφήματος, αν τα αντίστοιχα σύνολά τους έχουν μη κενή τομή.

---

**Algorithm 1** Mapper Algorithm

---

**Είσοδος :**

- 1) Ένα πεπερασμένο σύνολο σημείων  $\mathbb{X}$ , εφοδιασμένο με μία μετρική ή ένα μέτρο ανομοιογένειας μεταξύ των δεδομένων.
- 2) Μία συνάρτηση  $f : \mathbb{X} \rightarrow \mathbb{R}^d$ .
- 3) Ένα κάλυμμα  $\mathcal{U}$  του  $f(\mathbb{X})$ .
- 4) Ένας αλγόριθμος clustering ώστε, για κάθε  $U_a \in \mathcal{U}$ , να διαμερίσει το  $f^{-1}(U_a)$  σε σύνολα  $V_{a,j}$ .

**Υπολογισμός :** Υπολογίζει το νεύρο  $M_f(\mathbb{X}, \mathcal{U})$  του καλύμματος  $\bigcup_{a \in A} \bigcup_{j=1}^{j_a} V_{a,j}$  του  $\mathbb{X}$ .

**Έξοδος :** Ένα πλεγματοικό σύμπλεγμα, δηλαδή το νεύρο  $M_f(\mathbb{X}, \mathcal{U})$ , όπου :

- μία κορυφή  $v_{a,j}$ , για κάθε cluster  $V_{a,j}$ .
  - μια ακμή μεταξύ  $v_{a_1,j}$  και  $v_{a_2,k}$  αν και μόνο αν  $V_{a_1,j} \cap V_{a_2,k} \neq \emptyset$
- 

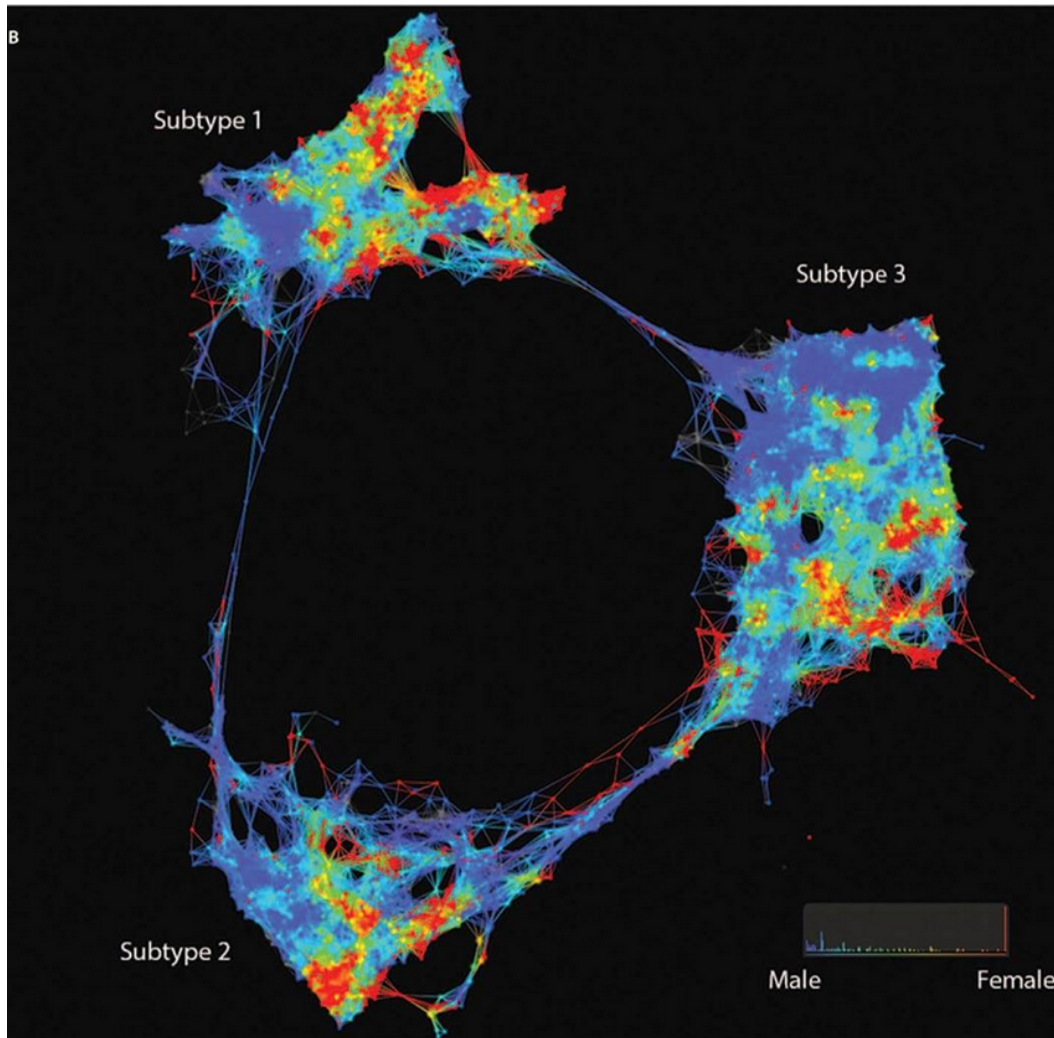
**Η επιλογή της  $f$ .** Η επιλογή της συνάρτησης  $f$ , η οποία καλείται συνήθως *filter*, εξαρτάται σημαντικά από τα δεδομένα που θέλει κανείς να δώσει περισσότερη σημασία. Μερικές από αυτές είναι οι παρακάτω :

- Density Estimates: Ο Mapper μπορεί να βοηθήσει να κατανοήσουμε την δομή και την συνεκτικότητα περιοχών με μεγάλη πυκνότητα.
- Η Centrality Function  $f(x) = \sum_{y \in \mathbb{X}} d(x, y)$  και η Συνάρτηση Εκκεντρότητας  $f(x) = \max_{y \in \mathbb{X}} d(x, y)$  είναι σε μερικές περιπτώσεις καλές επιλογές, που δεν απαιτούν την γνώση των δεδομένων.
- Η Kernel Density Function  $f_\varepsilon(x) = C_\varepsilon \sum_{y \in \mathbb{X}} e^{-d_{\mathbb{X}}^2(x,y)\varepsilon^{-1}}$ .

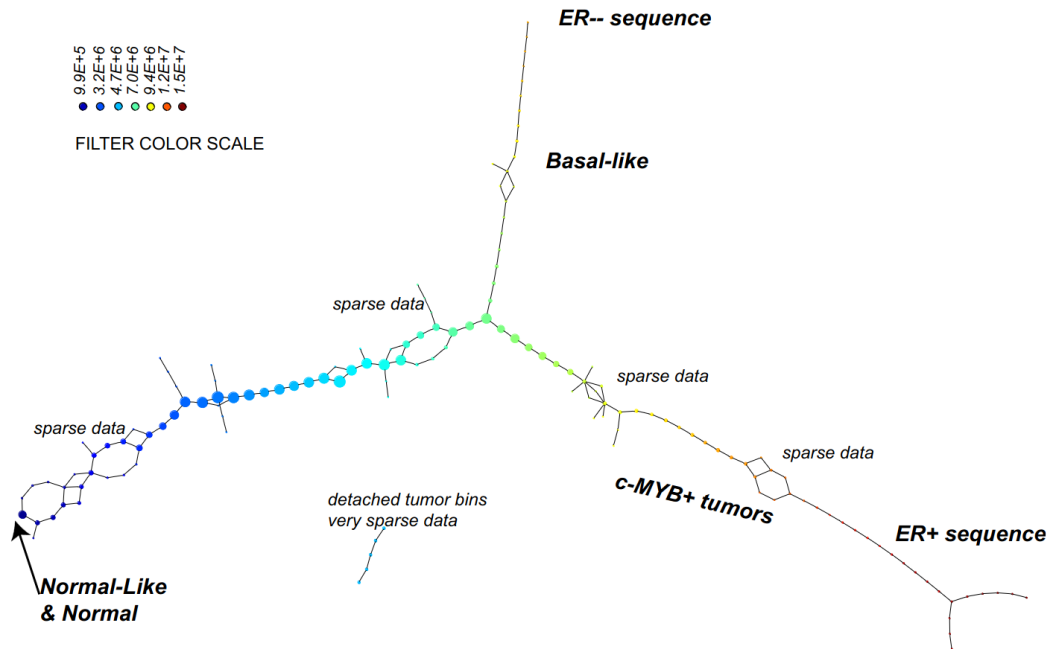
**Η επιλογή του καλύμματος  $\mathcal{U}$ .** Όταν η  $f$  είναι πραγματική συνάρτηση, μια συνήθης επιλογή για το κάλυμμα  $\mathcal{U}$  είναι ένα σύνολο από διαστήματα ίσου μήκους  $r > 0$ , που καλύπτει τον  $f(\mathbb{X})$ . Το μήκος  $r$  συνήθως καλείται ανάλυση (resolution) του καλύμματος και το ποσοστό επικάλυψης  $g$  μεταξύ δύο διαδοχικών διαστημάτων καλείται gain του καλύμματος. Είναι σημαντικό να τονίσουμε ότι το αποτέλεσμα του Mapper είναι πολύ ευαίσθητο στην επιλογή του καλύμματος  $\mathcal{U}$  και μικρές αλλαγές στην ανάλυση και το gain μπορεί να επιφέρει τεράστιες αλλαγές στο τελικό αποτέλεσμα, δηλαδή η μέθοδος δεν είναι ευσταθής. Συνήθως εκτελείται ο Mapper για διάφορες τιμές των παραμέτρων και επιλέγονται εκείνες που φαίνεται να προσφέρουν τις περισσότερες πληροφορίες σύμφωνα με την οπτική του χρήστη.

**Η επιλογή των clusters.** Ο αλγόριθμος Mapper απαιτεί την ομαδοποίηση της αντίστροφης εικόνας των ανοικτών συνόλων  $U \in \mathcal{U}$ . Μία μέθοδος είναι να εφαρμόσουμε σε κάθε σύνολο  $f^{-1}(U)$  κάποιο αλγόριθμο ομαδοποίησης. Μια άλλη είναι να κατασκευάσουμε ένα γράφημα γειτνίασης πάνω στα δεδομένα του  $\mathbb{X}$ , όπως για παράδειγμα το γράφημα k-NN. Έτσι για κάθε  $U \in \mathcal{U}$ , παίρνουμε τις συνεκτικές συνιστώσες του υπογραφήματος με κορυφές στο  $f^{-1}(U)$ .

Τέλος, κλείνουμε την εργασία παραθέτοντας δύο παραδείγματα όπου χρησιμοποιήθηκε ο Mapper και είχε εξαιρετική επιτυχία.



Εικόνα 4.1 : Στην παραπάνω εικόνα βλέπουμε τους 3 υποτύπους του διαβήτη τύπου 2. Κάθε κόμβος είναι είτε κάποιος ασθενής είτε κάποια ομάδα ασθενών με σημαντική ομοιότητα ως προς τα κλινικά χαρακτηριστικά τους. Δύο κόμβοι ενώνονται με κάποια ακμή γιατί έχουν τουλάχιστον έναν κοινό ασθενή. Τέλος, με κόκκινο χρώμα είναι οι ομάδες που αποτελούνται αμιγώς από γυναίκες ενώ με μπλε οι ομάδες που αποτελούνται αμιγώς από άνδρες. Τα ενδιάμεσα χρώματα προφανώς είναι για την μίξη αντρών και γυναικών, ανάλογα με το ποσοστό που συμβάλει το κάθε φύλο στον κόμβο (βλ. [7] απ'όπου πάρθηκε η εικόνα).



Εικόνα 4.2 : Progression Analysis of Disease (PAD) ανάλυση των δεδομένων του Netherlands Kanker Instituut (NKI). Το αποτέλεσμα (output) έχει τρία progression arms, γιατί οι όγκοι (που είναι τα δεδομένα-σημεία) ταξινομούνται με βάση το βαθμό απόκλισης από το υγιές (Healthy State model-HSM). Κάθε κάδος χρωματίζεται με την μέση τιμή των τιμών του φίλτρου  $f$  στα σημεία. Οι μπλε κάδοι περιέχουν όγκους που η απόκλισή τους από το HSM είναι μικρή (Normal και Normal-like tumors). Οι κόκκινοι κάδοι περιέχουν όγκους που απόκλισή τους είναι μεγάλη από το HSM. Η εικόνα του φίλτρου  $f$  υποδιαιρέστηκε σε 15 διαστήματα με 80% επικάλυψη (βλ. [8] απ'όπου πάρθηκε η εικόνα).

## References

- [1] Dey, Tamal, and Yusu Wang. Computational Topology for Data Analysis. Cambridge University Press, 2022.
- [2] Edelsbrunner, Herbert, and J. Harer. Computational Topology: An Introduction. American Mathematical Society, 2010.
- [3] Karvonen, Elli. “Persistent Homology and Its Applications.” The University of Helsinki, 2021.
- [4] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, Jan 2007.
- [5] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: a stable vector representation of persistent homology. *J. Machine Learning Research*, 18:218–252, 2017
- [6] Chazal, F., Michel, B. (2021). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, 4, 108.
- [7] Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, 7(311), 311ra174
- [8] Nicolau, M., Levine, A. J., amp; Carlsson, G. (2011). 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*, 108(17), 7265–7270.
- [9] Paul Bendich, James S. Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.*, 10(1):198–218, 2016