CENTRO DE CIENCIAS MATEMÁTICAS

Aplicaciones relacionadas al análisis topológico de datos y aprendizaje geométrico profundo

10° Jornadas de Geometría, Topología y Dinámica

Adriana Haydeé Contreras Peruyero
Centro de Ciencias Matemáticas UNAM
haydeeperuyero@matmor.unam.mx

Supported by DGAPA research grant 2023

I. Análisis Topológico de Datos

- Motivación
- Complejos simpliciales y homología
- Aplicaciones
  - Pangenómica

2. Aprendizaje Geométrico Profundo

- Neurociencia

¿Qué es y por qué se usa?

¿Qué es y porqué se usa la topología para analizar datos?

¿Qué es y por qué se usa?

¿Qué es y porqué se usa la topología para analizar datos?

La topología es una rama de las matemáticas que es buena para extraer características cualitativas globales de estructuras geométricas complicadas.

¿Qué es y por qué se usa?

¿Qué es y porqué se usa la topología para analizar datos?

La topología es una rama de las matemáticas que es buena para extraer características cualitativas globales de estructuras geométricas complicadas.

El análisis topológico de datos:

usa la topología para resumir y aprender de la forma de los datos.

¿Por qué la topología es interesante para el análisis de datos?

- Invarianza del sistema de coordenadas: las características o invariantes topológicos no dependen del sistema coordenado que estemos usando. Lo que importa es la métrica (distancia/similitud entre los puntos de datos).
- Invarianza de deformación: las características topológicas son invariantes bajo homeomorfismos.
- Representación comprimida: La topología nos ofrece un conjunto de herramientas para resumir y representar los datos en formas compactas mientras se preserva su estructura topológica global.

Aplicaciones de TDA

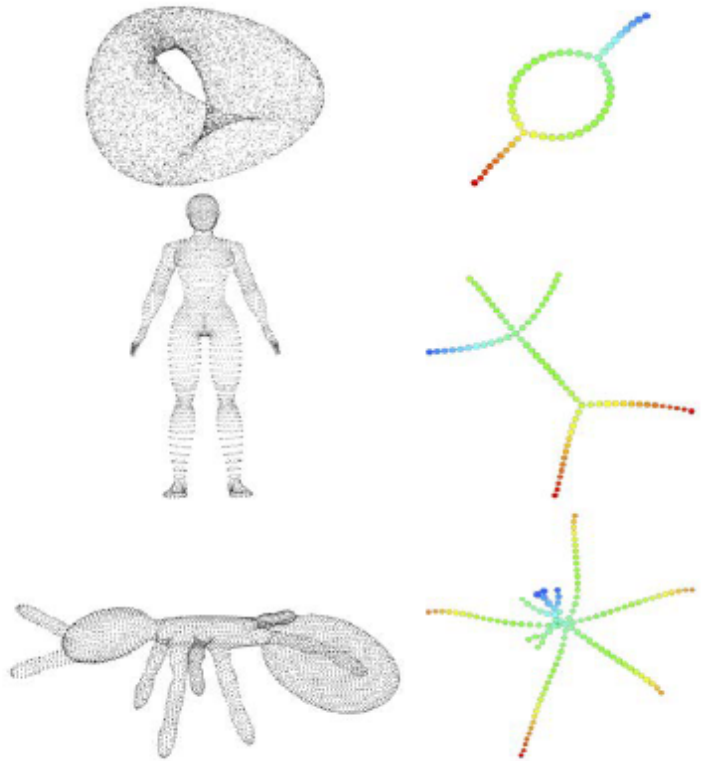


Figura: Mapper

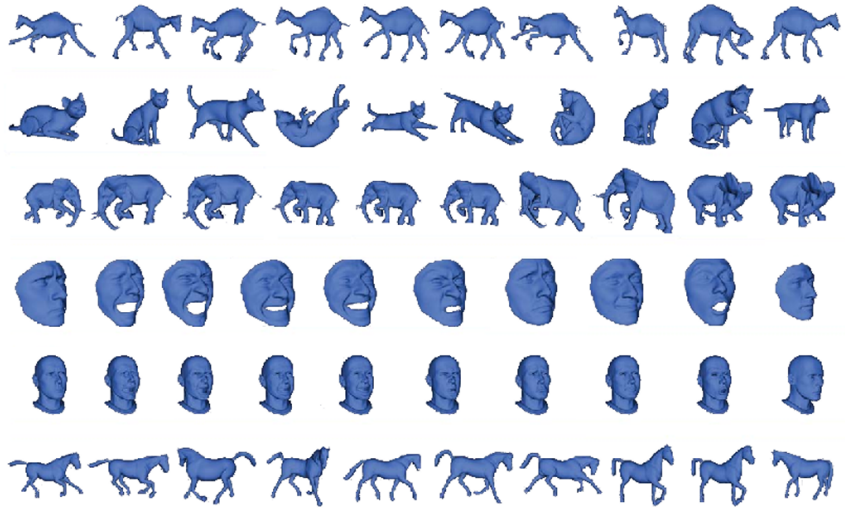


Figura: Reconocimiento de imágenes, [7, Figura 1]

Aplicaciones de TDA

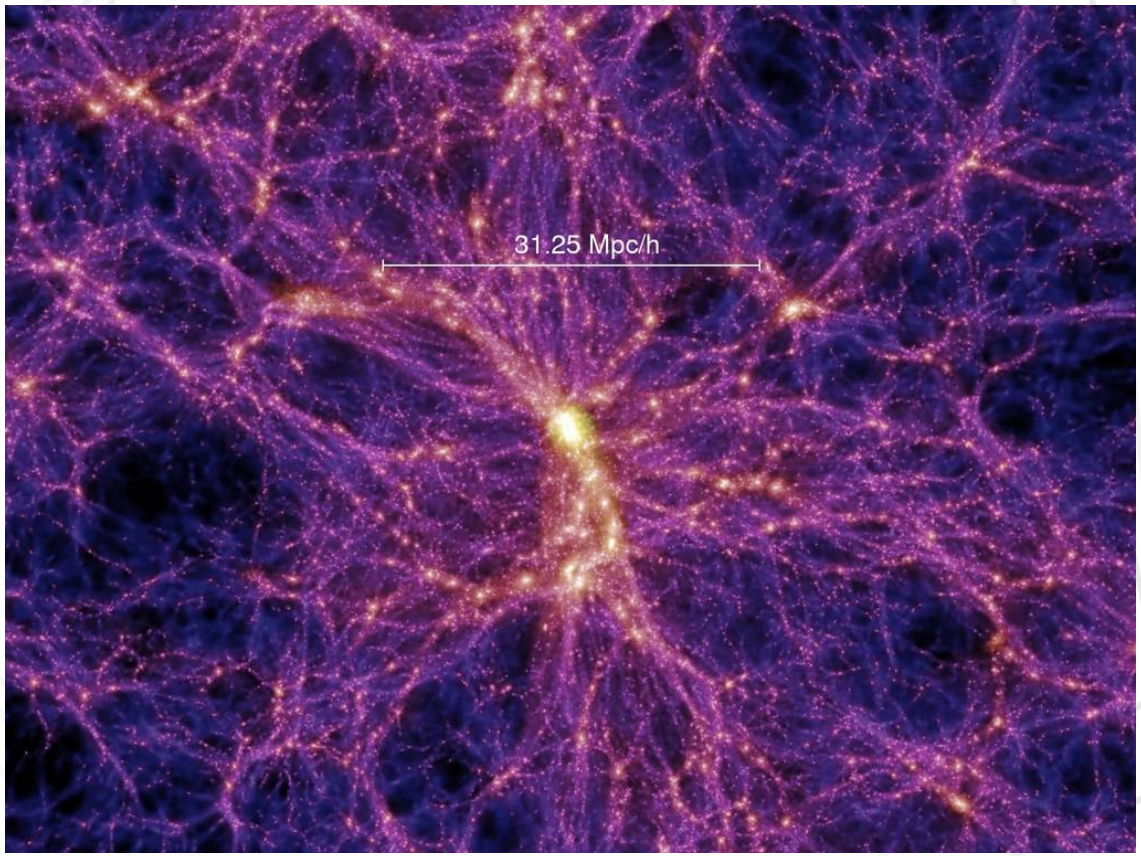


Figura: Estructura a gran escala del universo

Aplicaciones de TDA

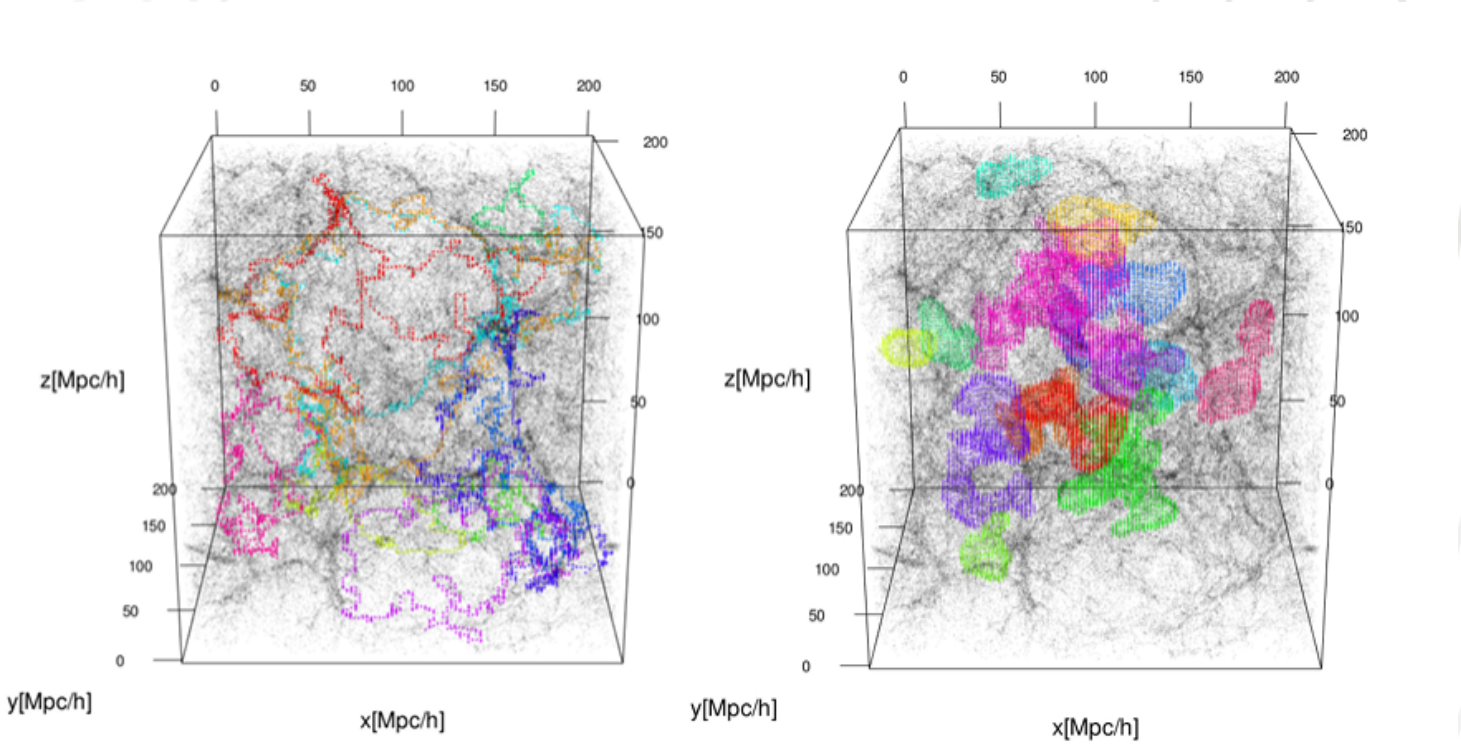


Figura: Vacíos cósmicos y filamentos, [24, Figura 12]

Aplicaciones de TDA

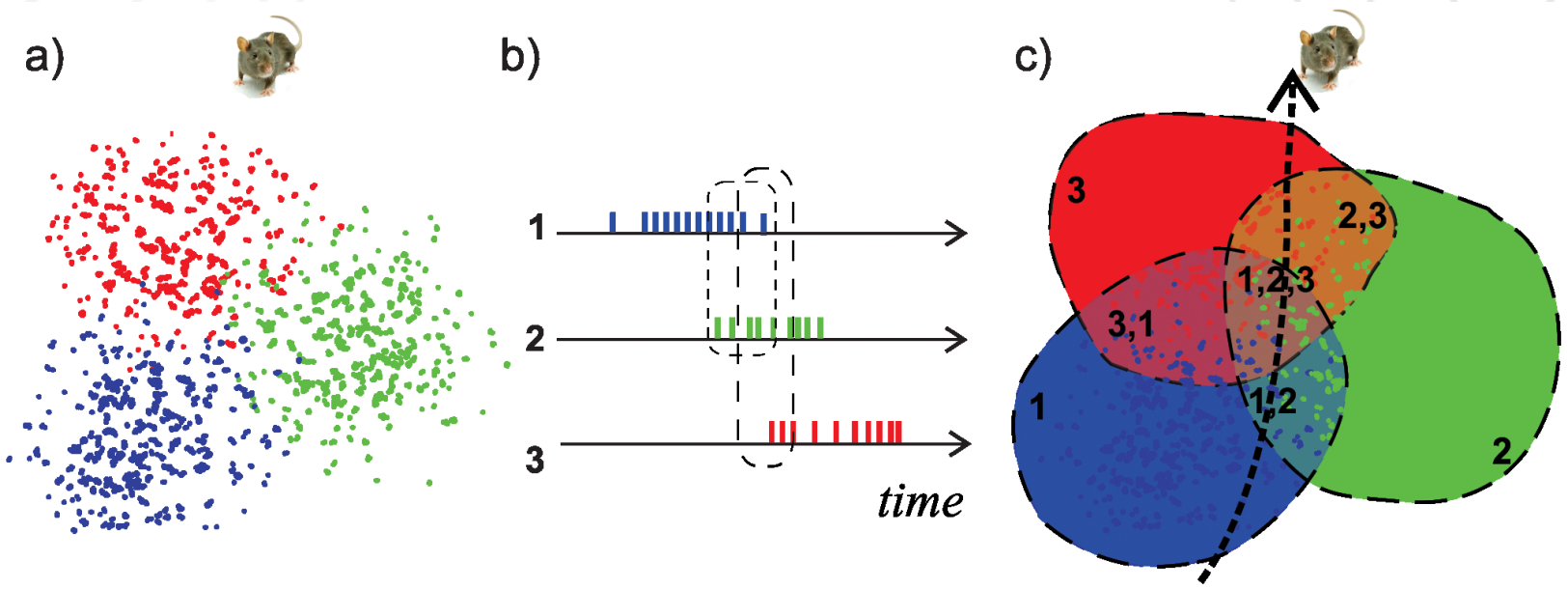


Figura: Células spike ayudan a revelar información del lugar [18, Figura 1]

Aplicaciones de TDA

Figura: Árbol cerebral de una persona de 24 años, [3, Figura 10]

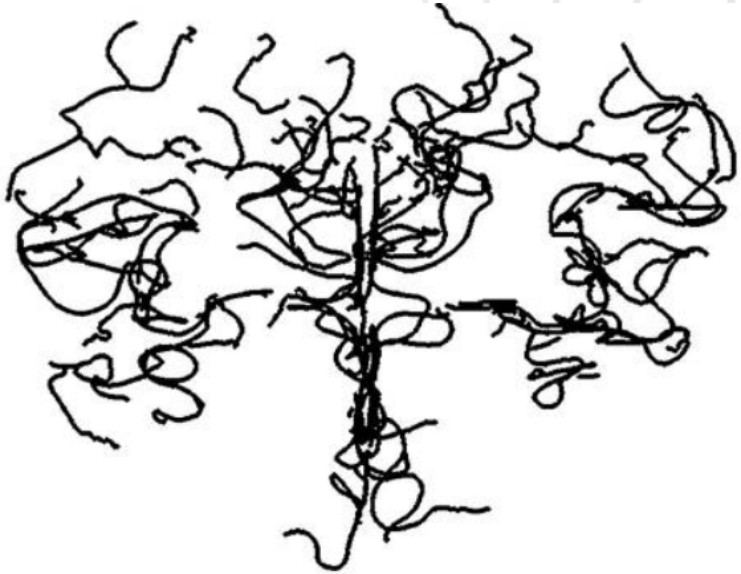


Figura: Árbol cerebral de una persona de 68 años, [3, Figura 11]

Aplicaciones de TDA

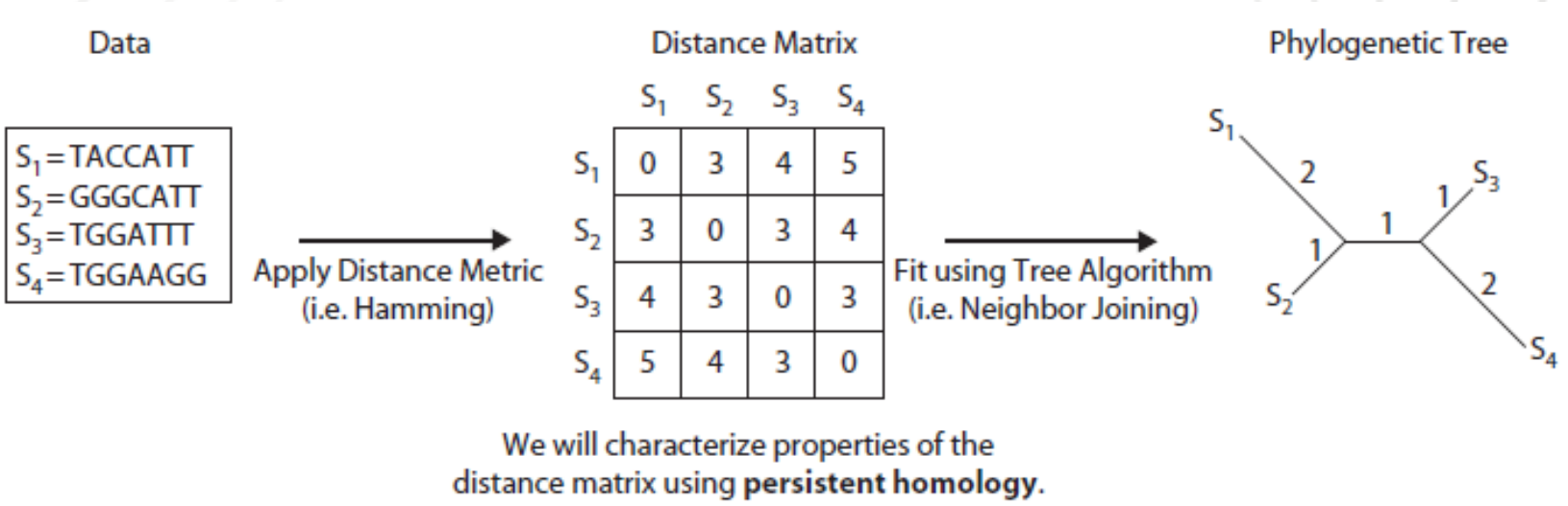


Figura: Análisis de datos genómicos usando TDA, [19, Figura 5.9]

Aplicaciones de TDA

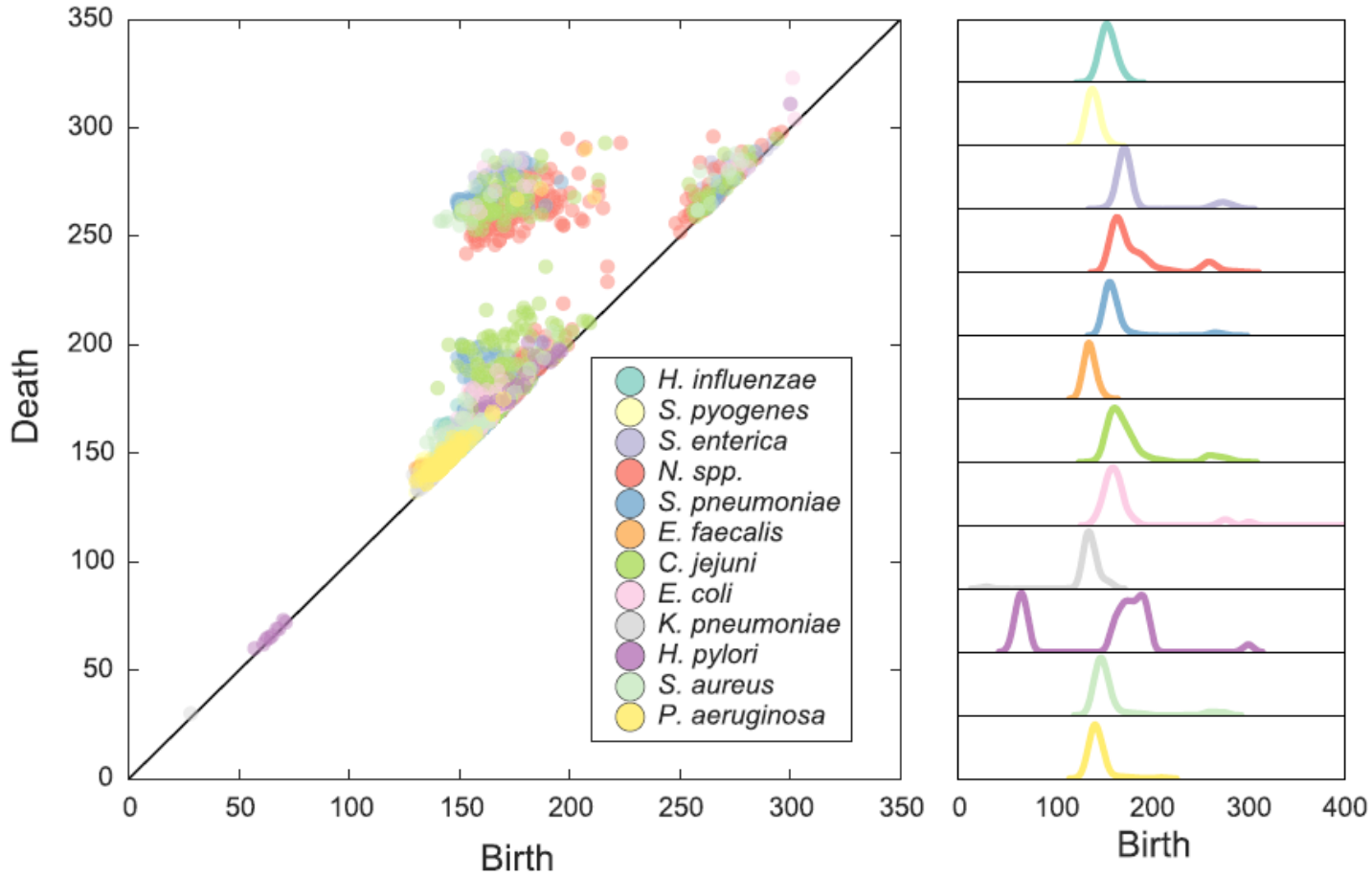


Figura: Diferencias entre 12 cepas de patógenos, [19, Figura 5.33]

¿Preguntas?

- ¿Cómo codificar espacios topológicos para fines computacionales?

¿Preguntas?

- ¿Cómo codificar espacios topológicos para fines computacionales?
- Buscar equivalencias homotópicas/homeomorfismos es "difícil". ¿Existen cantidades matemáticas que sean invariantes bajo equivalencias homotópicas **y** que sean fácil de calcular?

Complejos simpliciales

Complejos simpliciales

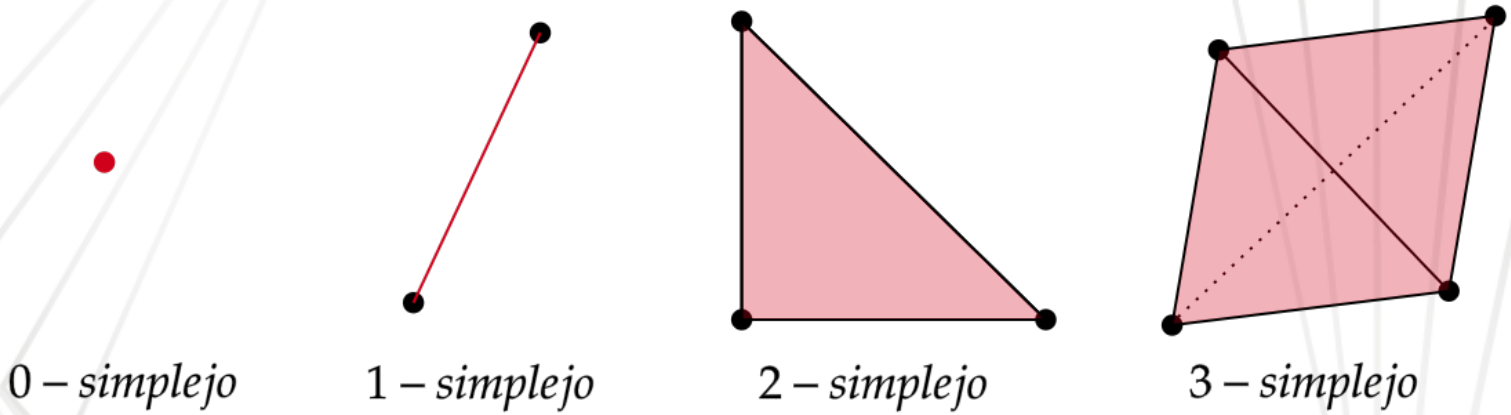


Figura: Complejos simpliciales

$$\{\text{Datos}\} \longrightarrow \{\text{Complejos simpliciales}\}$$

Definición

Un complejo simplicial (finito) K en $\mathbb{R}^d$ es una colección (finita) de simplejos (simplices) tal que:

1. *Cualquier cara de un simplejo de K es un simplejo de K,*
2. *La intersección de cualesquiera dos simplejos de K es o el vació o una cara común de ambos.*

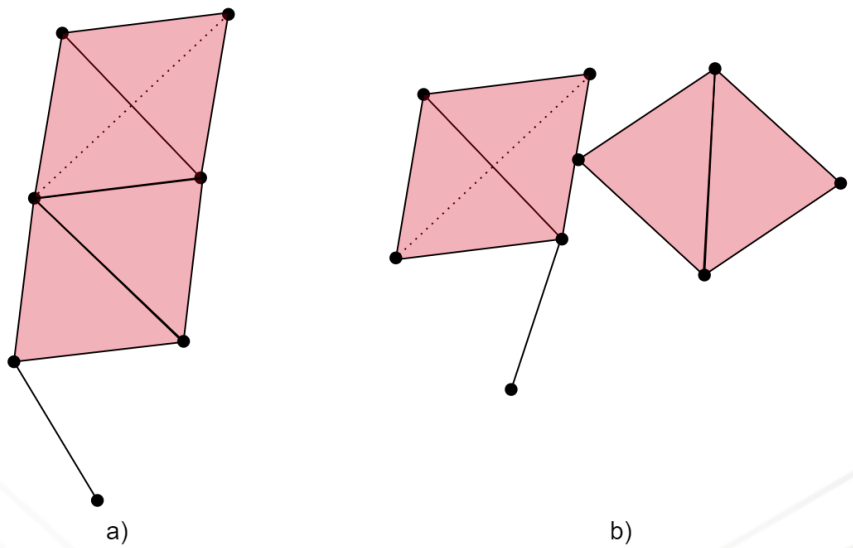


Figura: a) Complejo simplicial b) No es un complejo simplicial

IMPORTANTE

Los complejos simpliciales pueden ser vistos al mismo tiempo como espacios geométricos/topológicos (buenos para la inferencia geométrica y topológica) y como objetos combinatorios (complejos simpliciales abstractos, buenos para cálculos computacionales)

$$\{\text{Datos}\} \longrightarrow \{\text{Complejos simpliciales}\} \longrightarrow \{\text{Invariantes algebraicos}\}$$

Complejo de Čech

Definición

Dada una colección de puntos $X = \{x_i\} \subset \mathbb{R}^n$, el complejo de Čech, C_r, es un complejo simplicial abstracto que tiene como vértices a los puntos de X y cuyos k-simplejos $\{x_i\}_0^k$ son tales que $\cap_{j=0}^{k} B_{\frac{r}{2}}(x_{i,j}) \neq 0$.

Desde el punto de vista computacional, el complejo de Čeck es muy costoso.

Ejemplo: Complejo de Čech

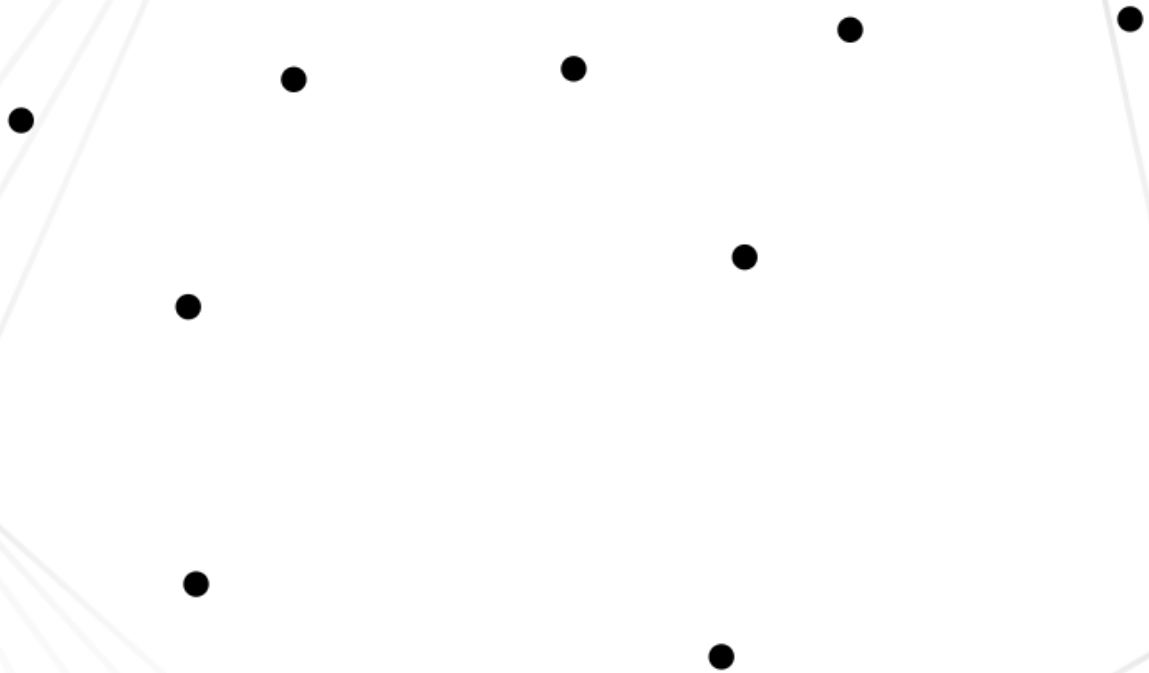


Figura: Colección de puntos $X = \{x_i\}$.

Ejemplo: Complejo de Čech

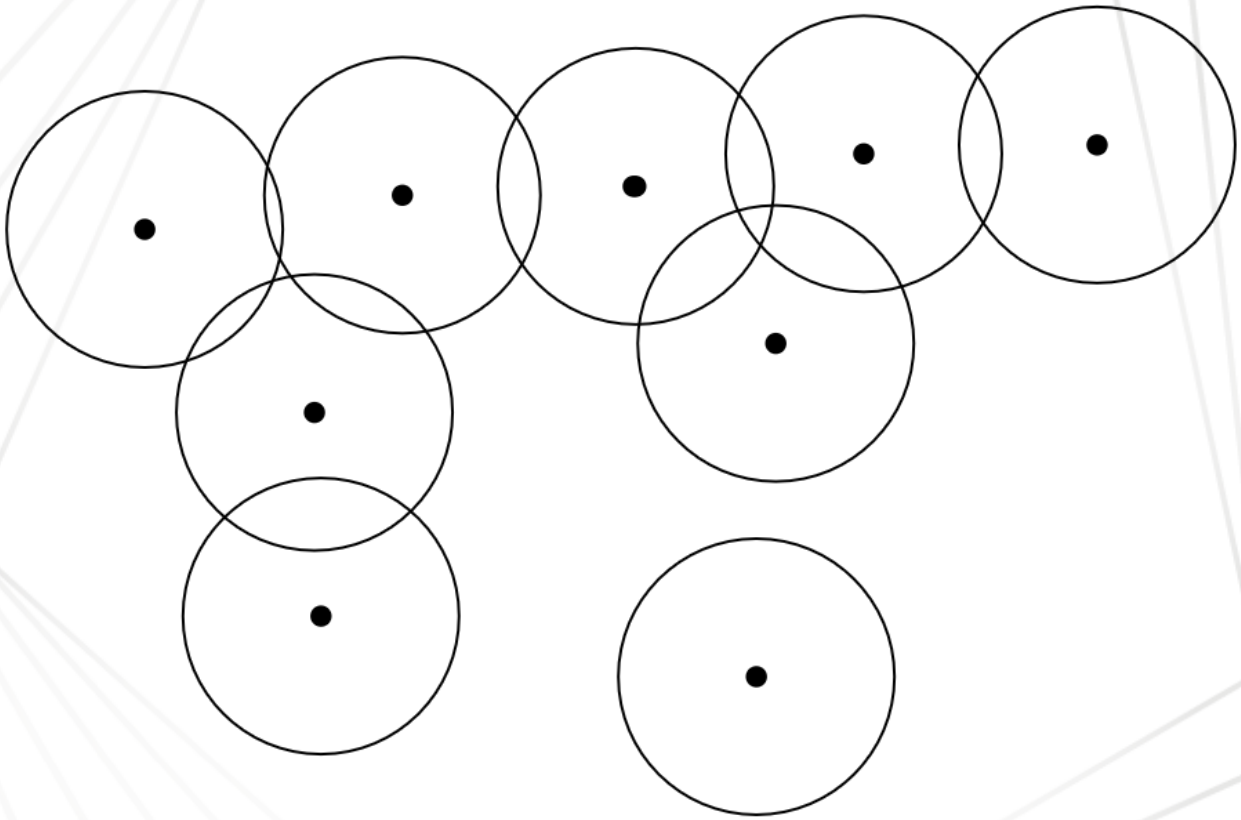


Figura: Bolas de radio $r/2$ centradas en los puntos x_i.

Ejemplo: Complejo de Čech

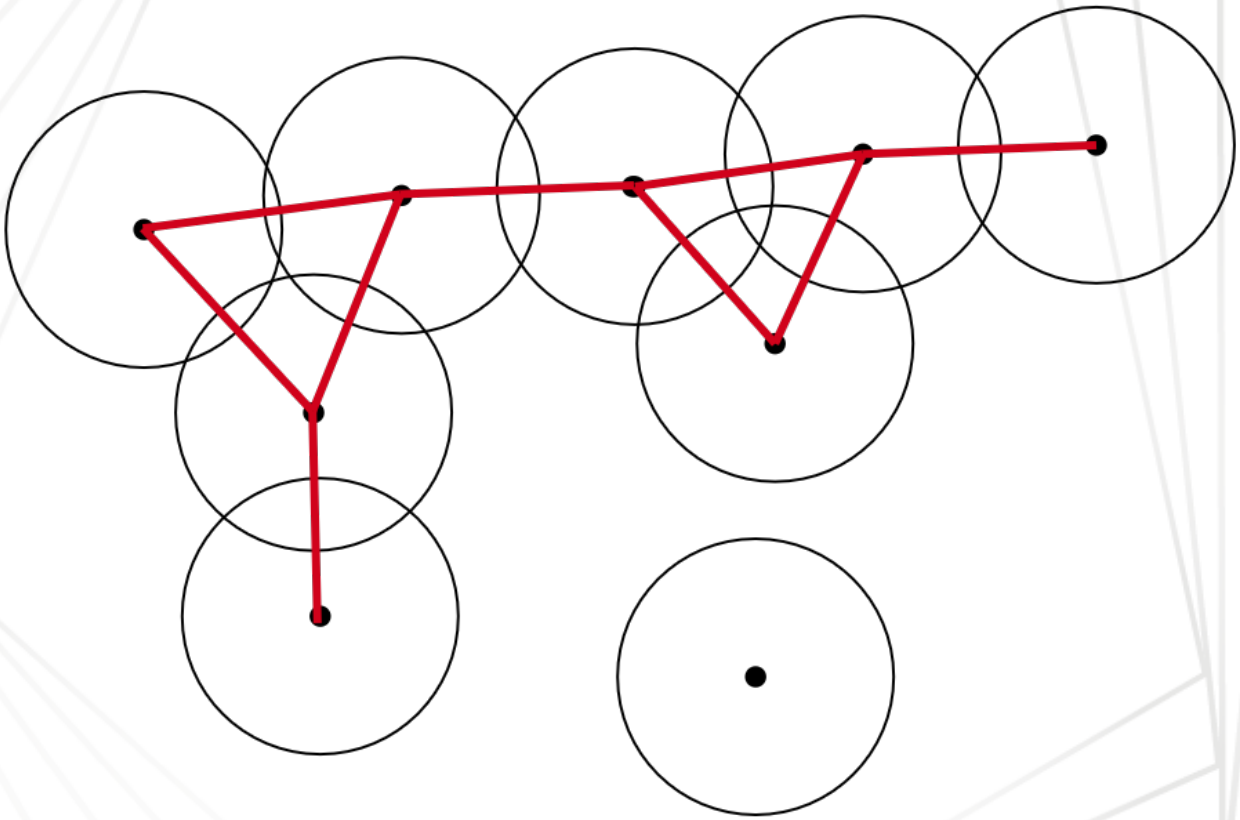


Figura: Agregamos 1–simplejos si dos bolas se intersecan.

Ejemplo: Complejo de Čech

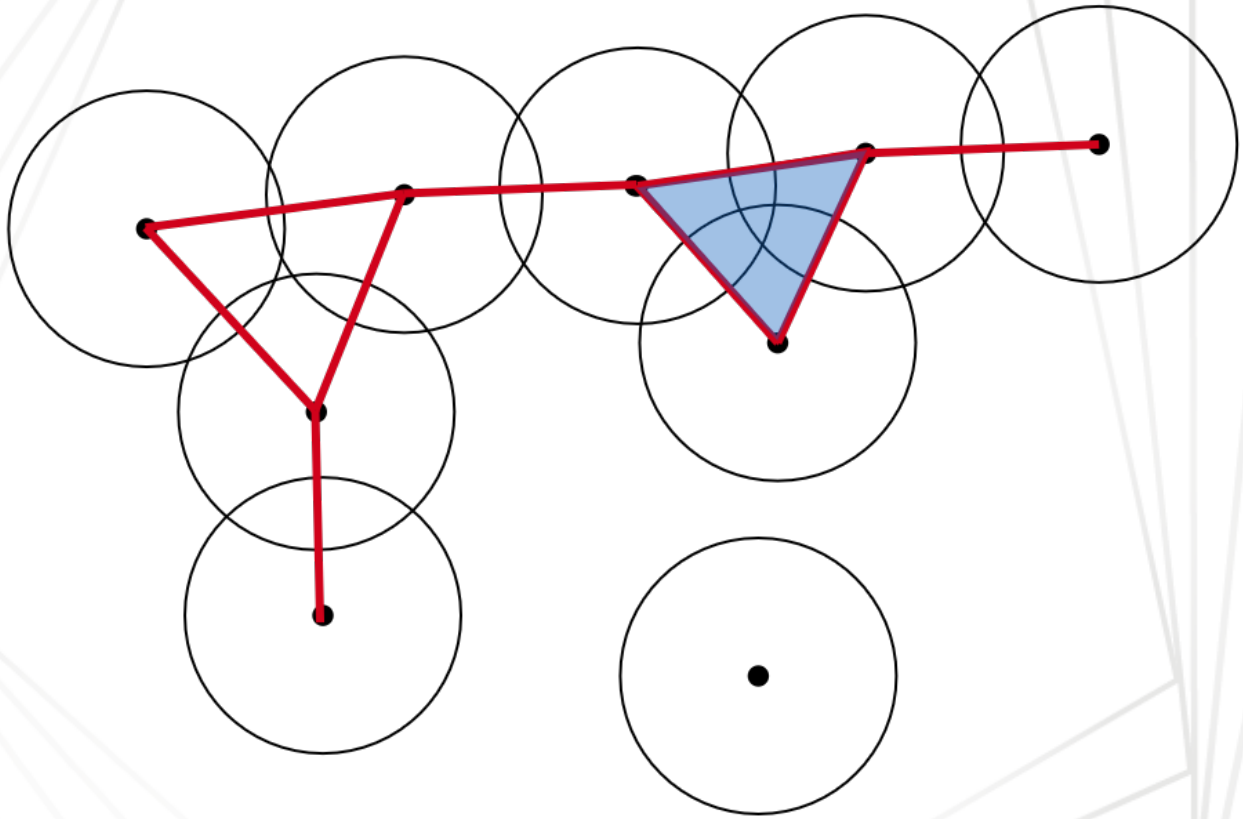


Figura: Agregamos 2–simplejos si tres bolas se intersecan.

Ejemplo: Complejo de Čech

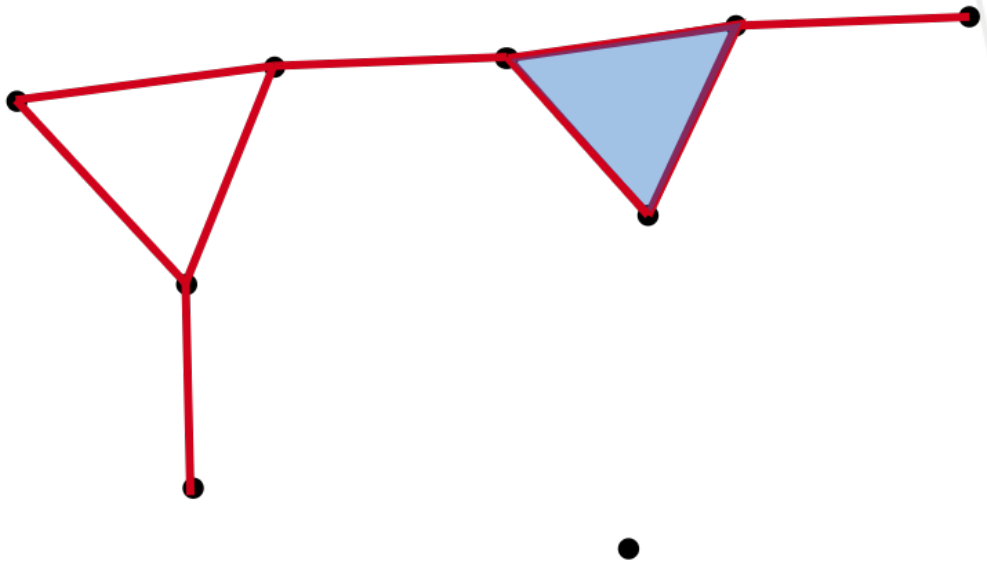


Figura: Complejo de Čech resultante.

Complejo de Vietoris-Rips

Definición

Dada una colección de puntos $\{x_i\}$ en un espacio métrico (X, d_X) el complejo de Vietoris-Rips, $\mathcal{VR}_r$, es un complejo simplicial abstracto que tiene como vértices a los puntos de X y cuyos k-simplejos $\{x_i\}_0^k$ son tales que $d(x_{i,j}, x_{i,l}) \leq 2r$ para toda pareja $1 \leq j, l \leq k$.

El complejo de Vietoris-Rips es menos costoso computacionalmente que el complejo de Čech aunque tenga más simplejos.

Ejemplo: Complejo de Vietoris-Rips

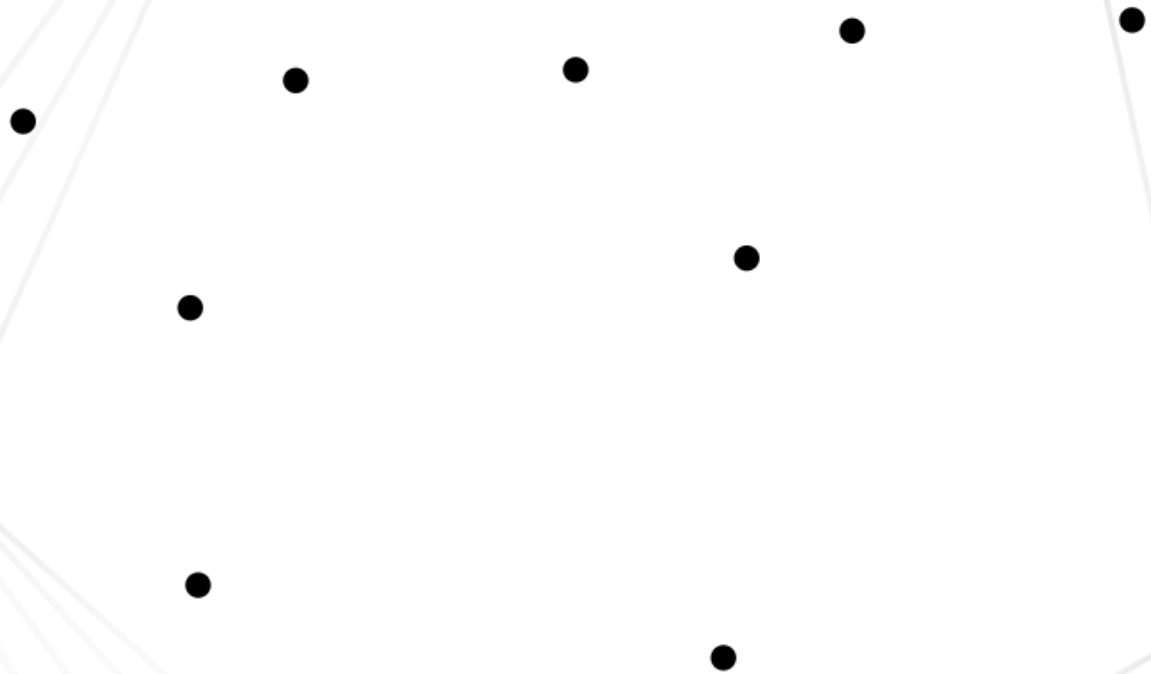


Figura: Colección de puntos $X = \{x_i\}$.

Ejemplo: Complejo de Vietoris-Rips

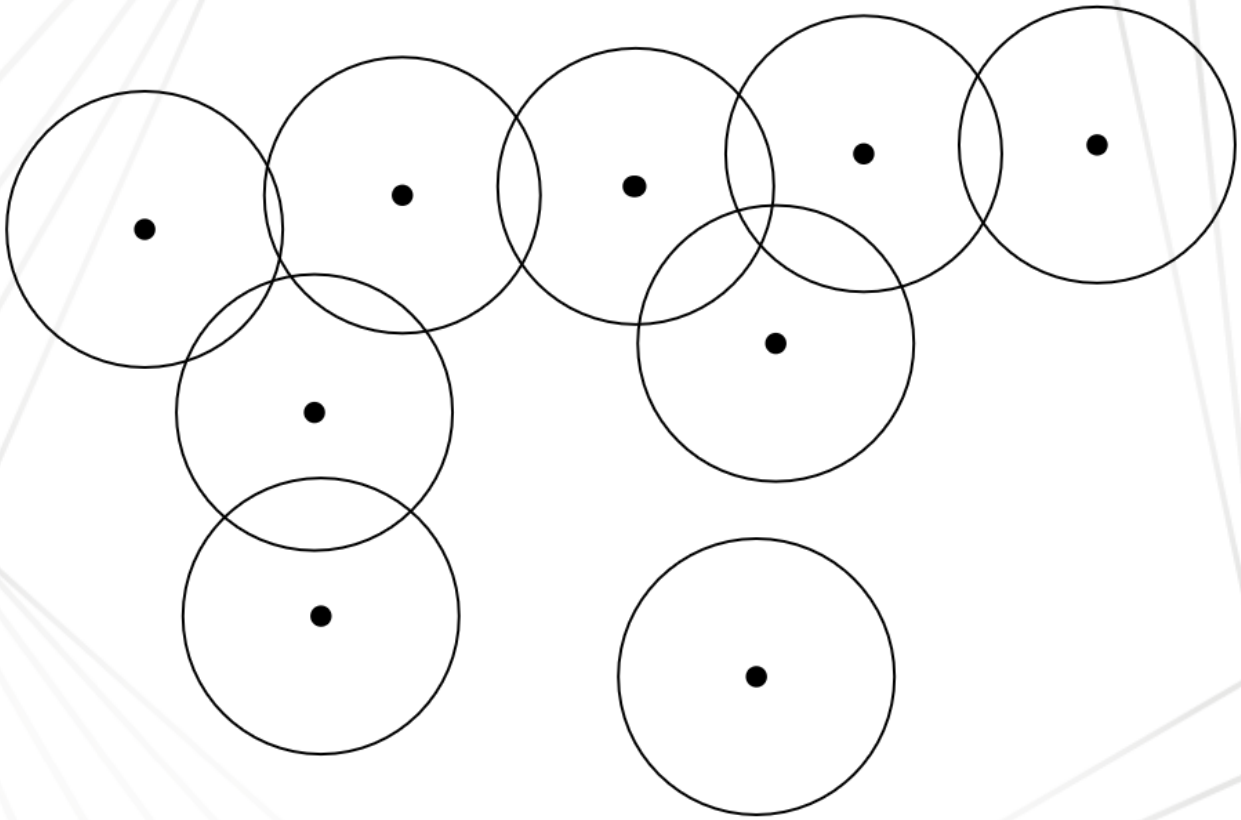


Figura: Bolas de radio $r/2$ centradas en los puntos x_i.

Ejemplo: Complejo de Vietoris-Rips

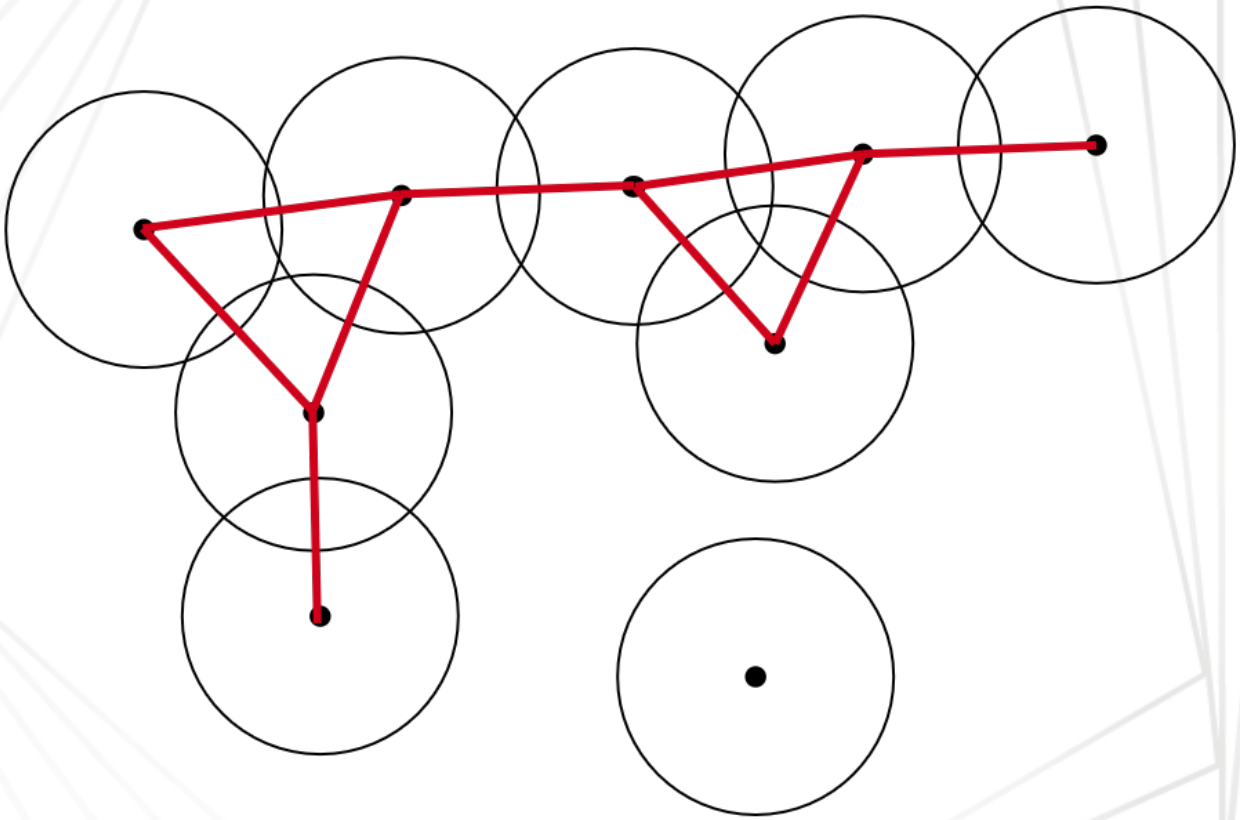


Figura: Agregamos 1–simplejos si dos bolas se intersecan.

Ejemplo: Complejo de Vietoris-Rips

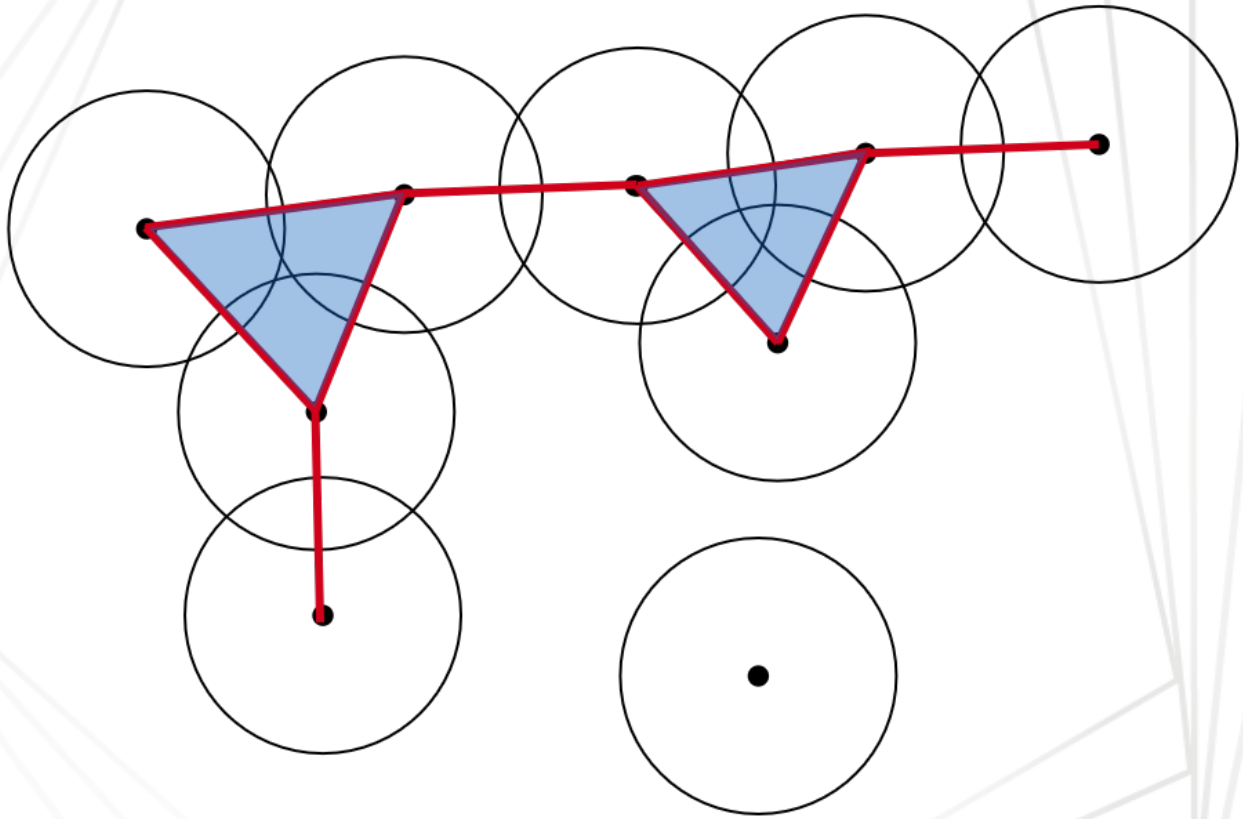


Figura: Agregamos 2—simplejos.

Ejemplo: Complejo de Vietoris-Rips

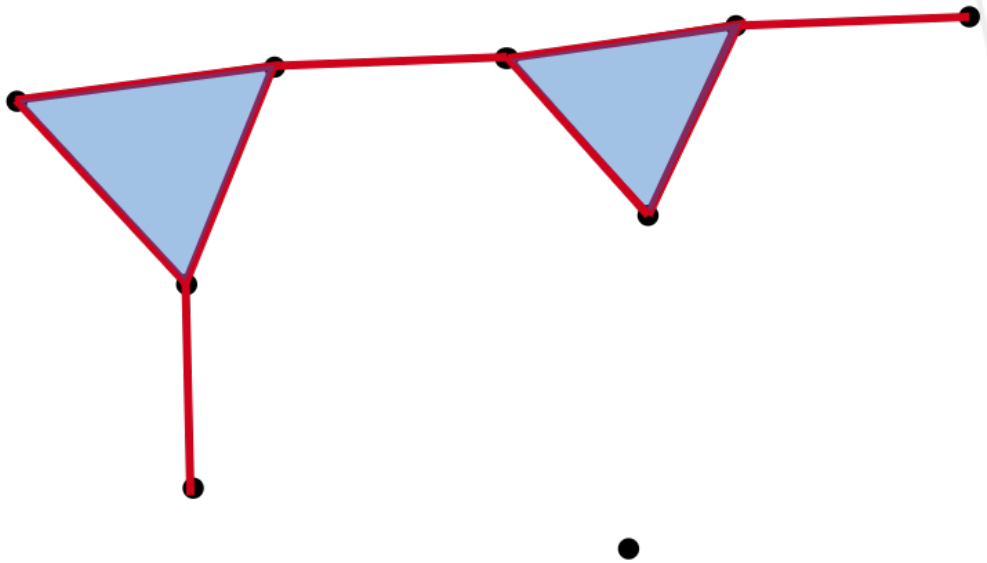


Figura: Complejo de Vietoris-Rips resultante.

Otros complejos

- Alpha
- Whitness

¿Cuál es el parámetro (radio) correcto para construir el complejo de Čech?

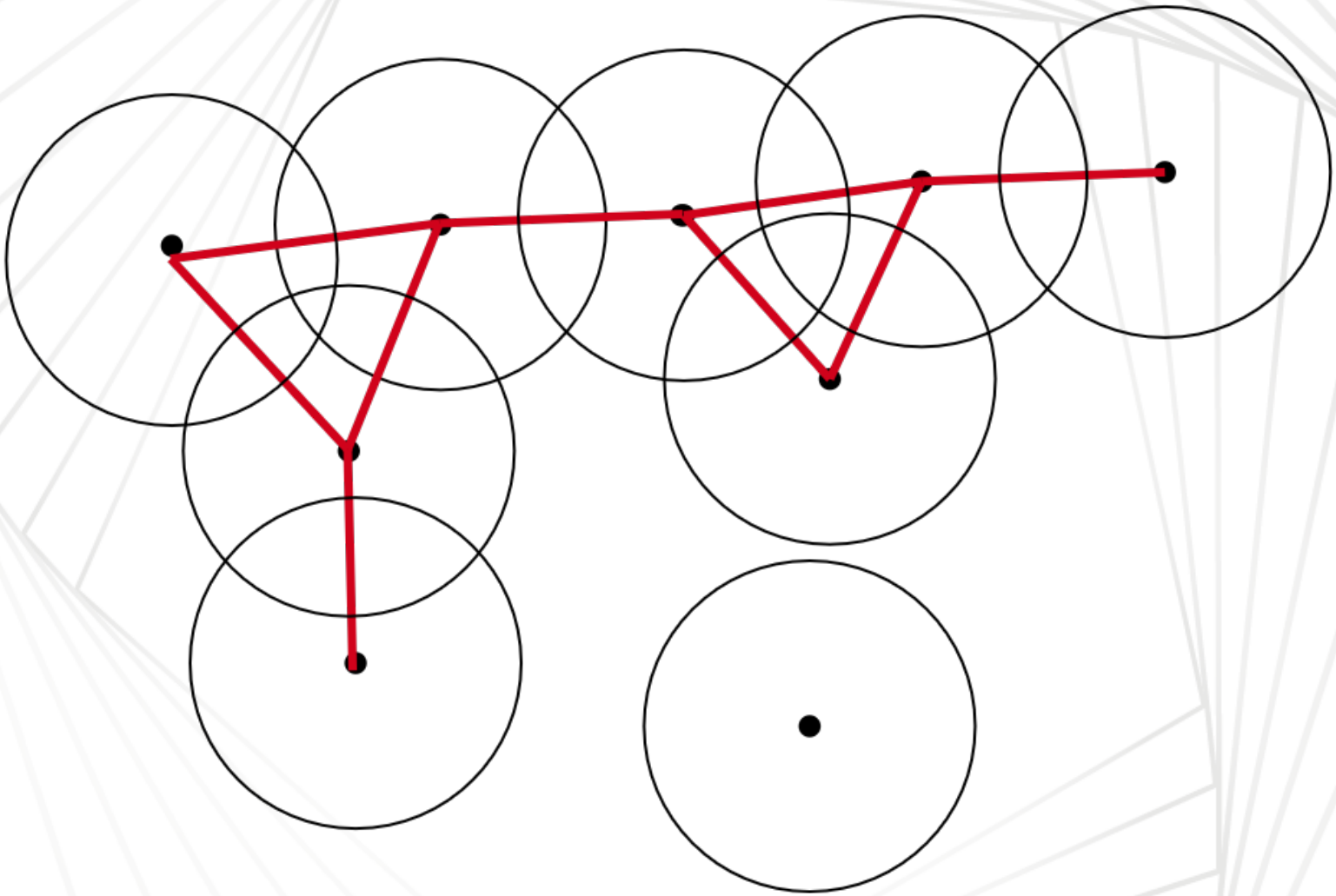


Figura: Complejo de Čech con otro parámetro.

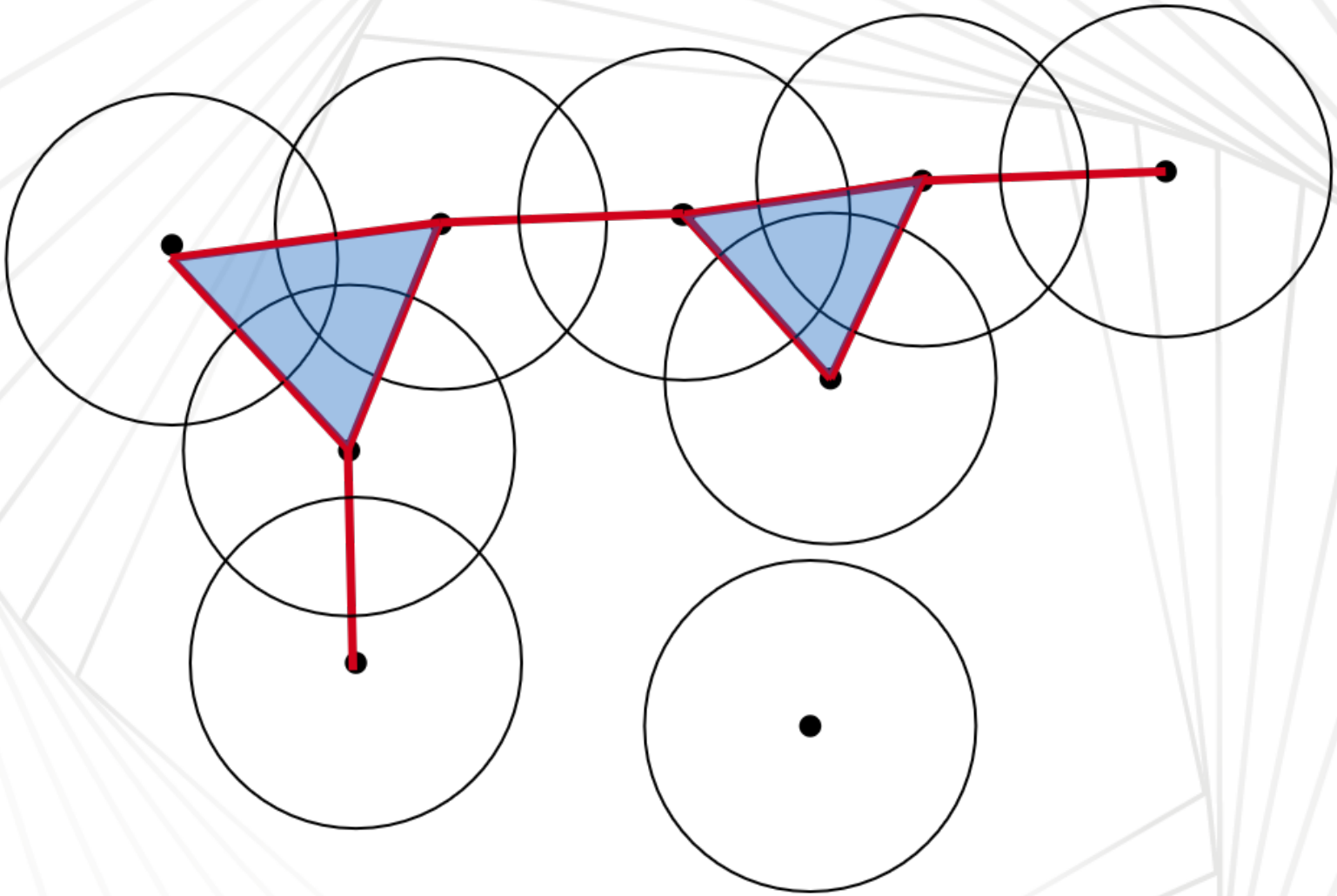


Figura: Complejo de Čech con otro parámetro.

Idea de Persistencia

- Variar el parámetro y llevar un “registro” de cuando aparecen y desaparecen las características topológicas a estudiar.
- Al variar el radio de las bolas en la construcción de los complejos, obtenemos una colección creciente de complejos simpliciales.

Filtración

- Una filtración de un complejo simplicial K es una colección $\{K_2, K_2, ..., K_N\}$ de subcomplejos de K tal que

$$\emptyset \subset K_1 \subset K_2 \subset \cdots \subset K_N = K.$$

Filtración

- Una filtración de un complejo simplicial K es una colección $\{K_2, K_2, ..., K_N\}$ de subcomplejos de K tal que

$$\emptyset \subset K_1 \subset K_2 \subset \cdots \subset K_N = K.$$

- Todo complejo simplicial abstracto tiene asociada una filtración dada por sus $l-$esqueletos.

Filtración

- Una filtración de un complejo simplicial K es una colección $\{K_2, K_2, ..., K_N\}$ de subcomplejos de K tal que

$$\emptyset \subset K_1 \subset K_2 \subset \cdots \subset K_N = K.$$

- Todo complejo simplicial abstracto tiene asociada una filtración dada por sus $l-$esqueletos.
- La filtración asociada a un complejo simplicial abstracto NO es única.

Complejo simplicial filtrado

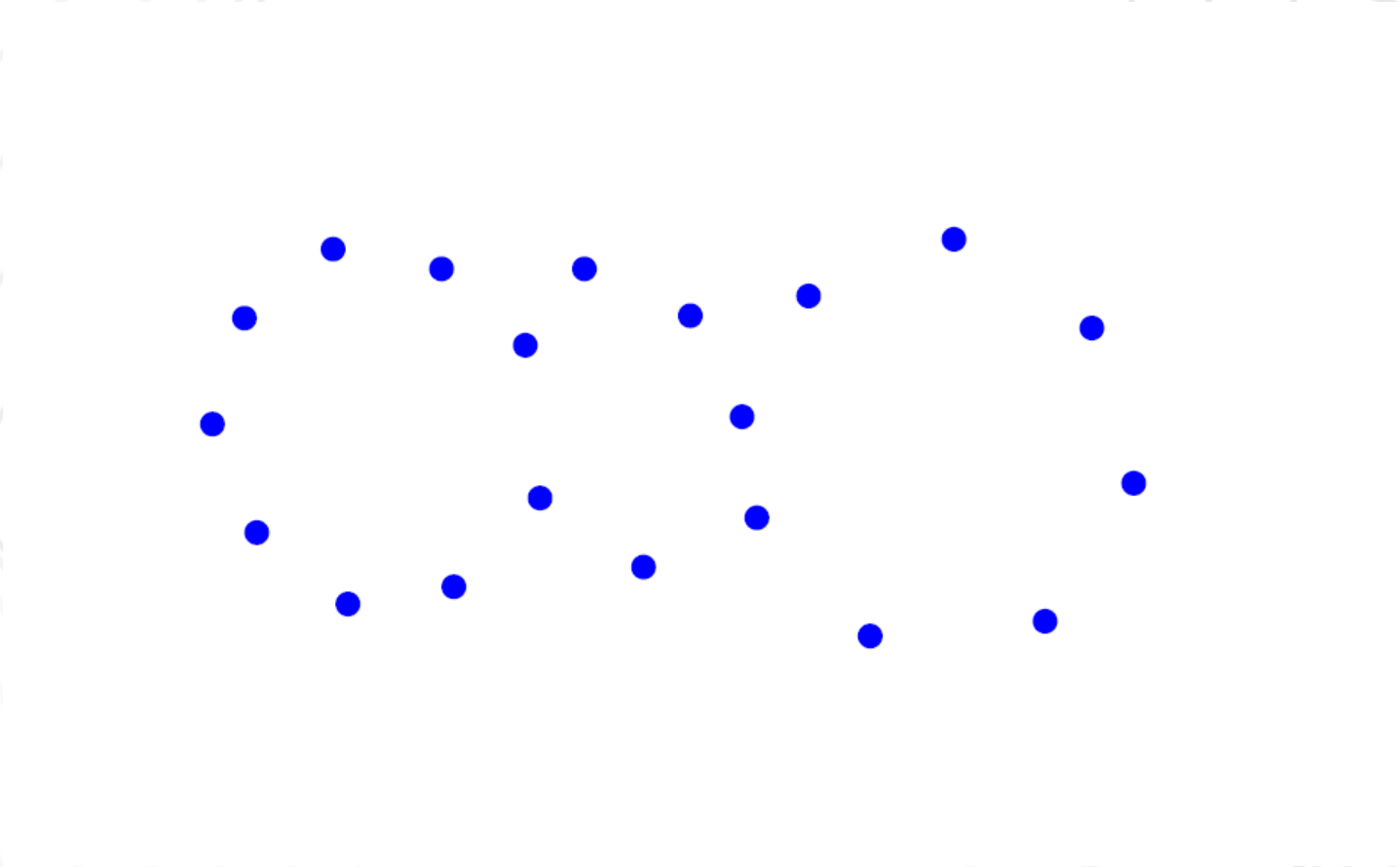


Figura: Radio=0

Complejo simplicial filtrado

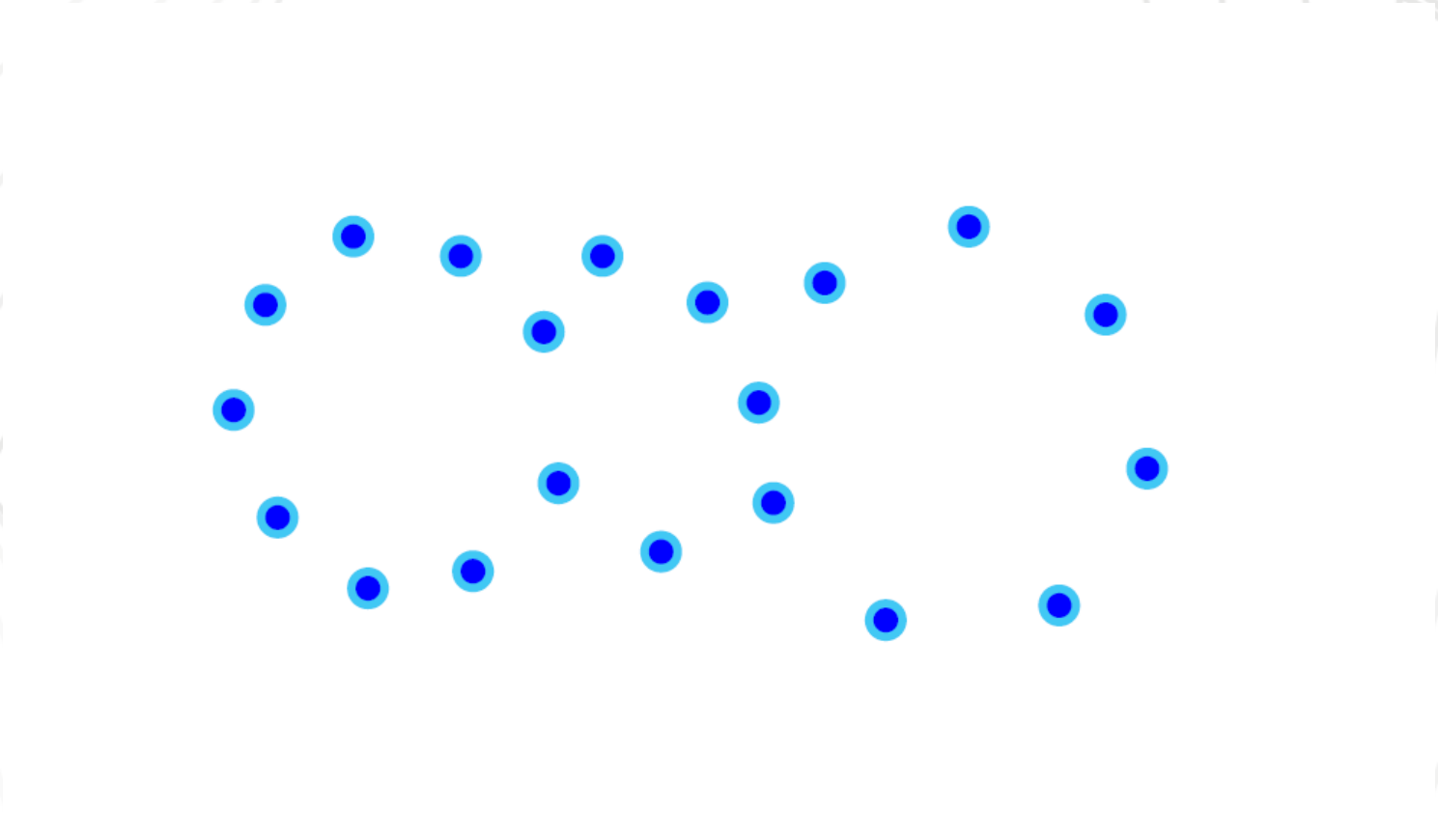


Figura: Radio=1

Complejo simplicial filtrado

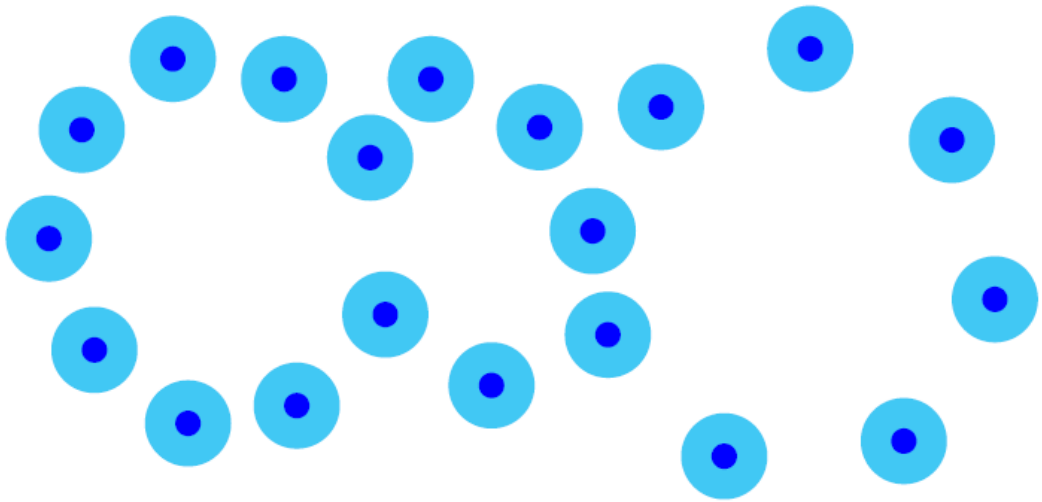


Figura: Radio=2

Complejo simplicial filtrado

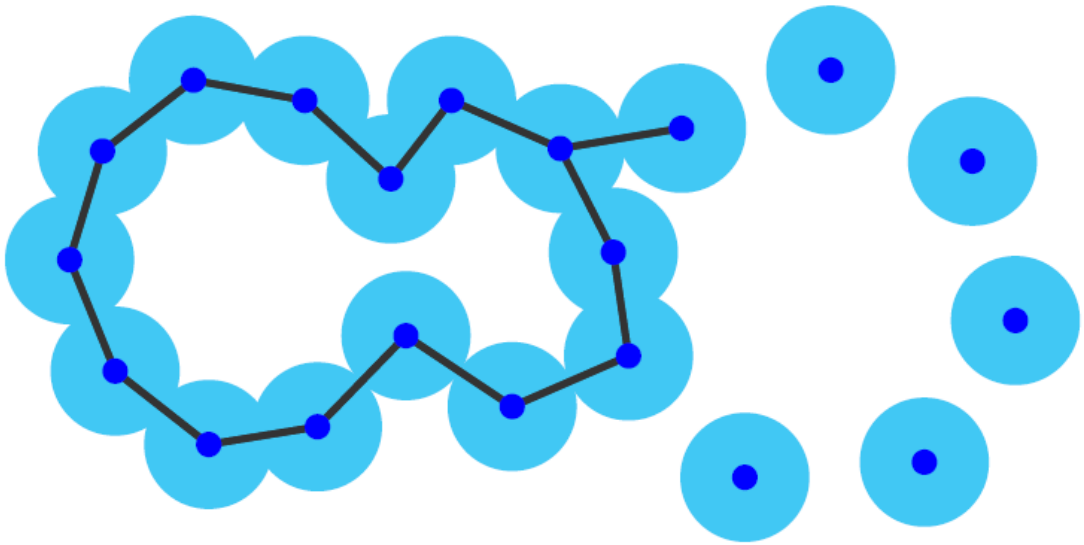


Figura: Radio=3

Complejo simplicial filtrado

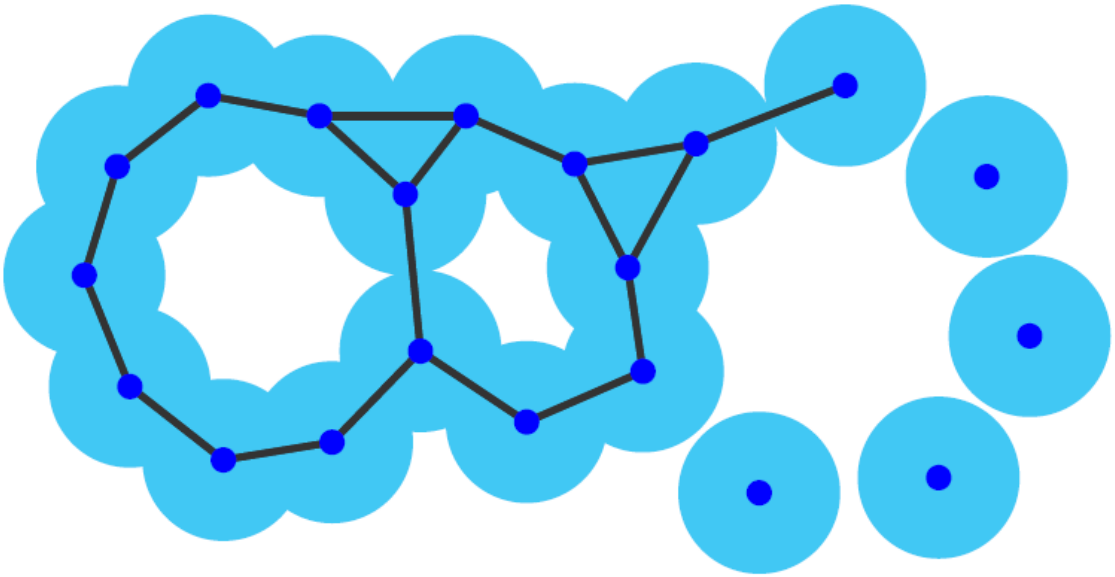


Figura: Radio=4

Complejo simplicial filtrado

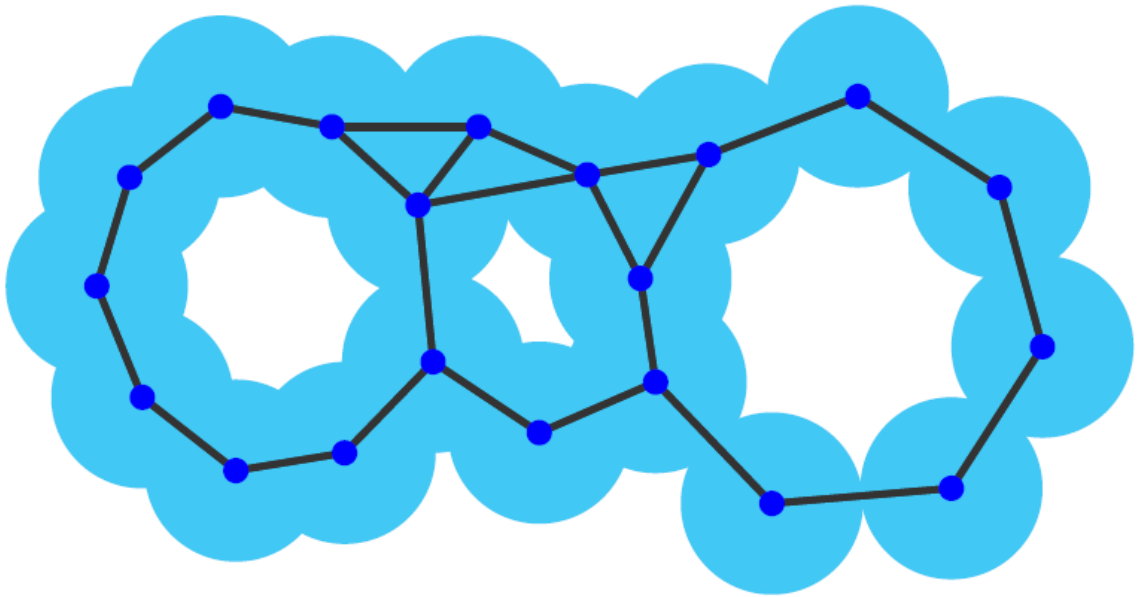


Figura: Radio=5

Complejo simplicial filtrado

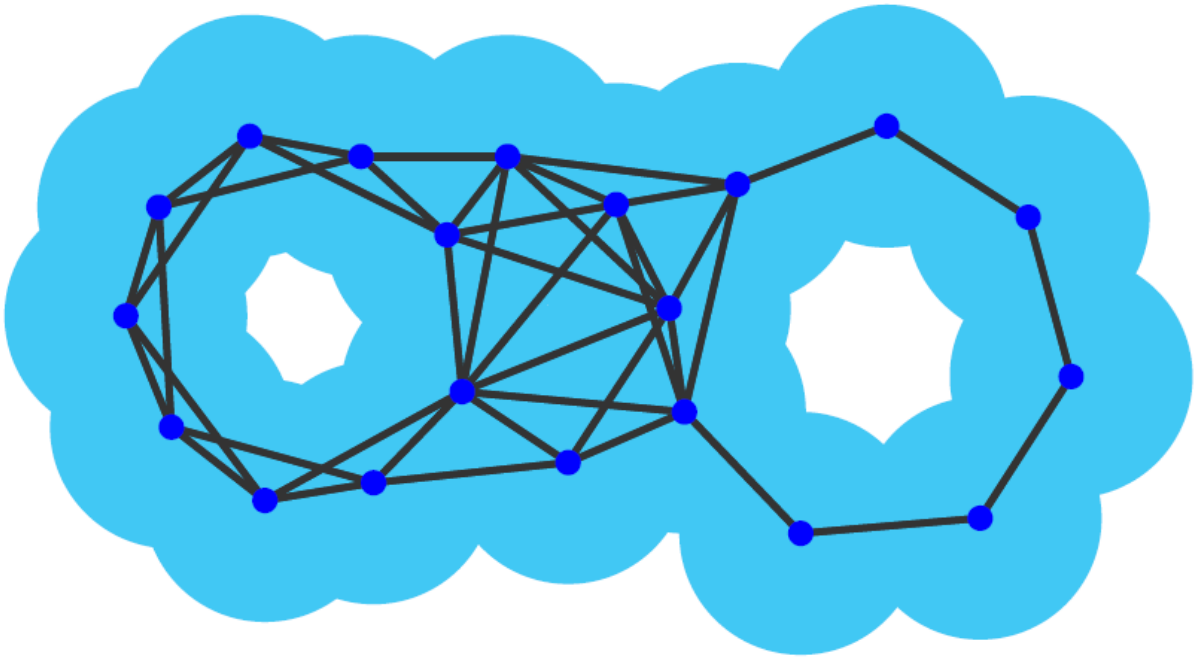


Figura: Radio=6

Complejo simplicial filtrado

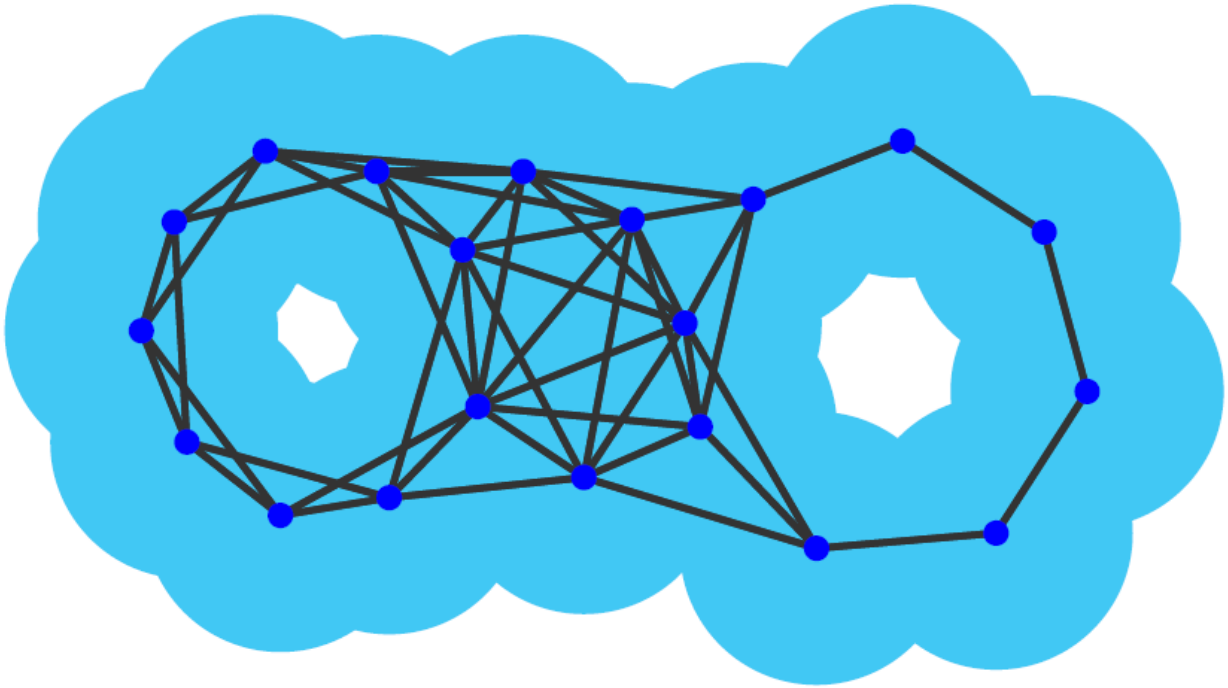


Figura: Radio=7

Complejo simplicial filtrado

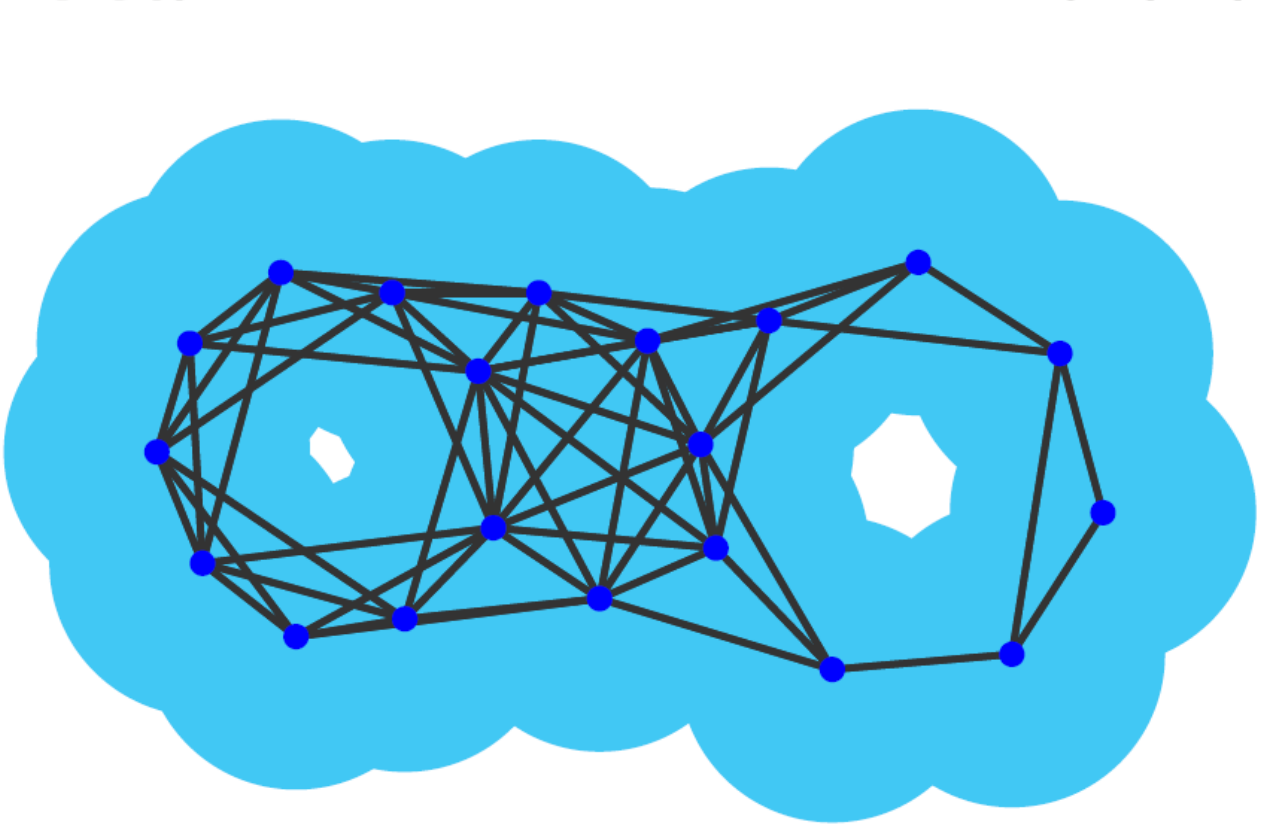


Figura: Radio=8

Complejo simplicial filtrado

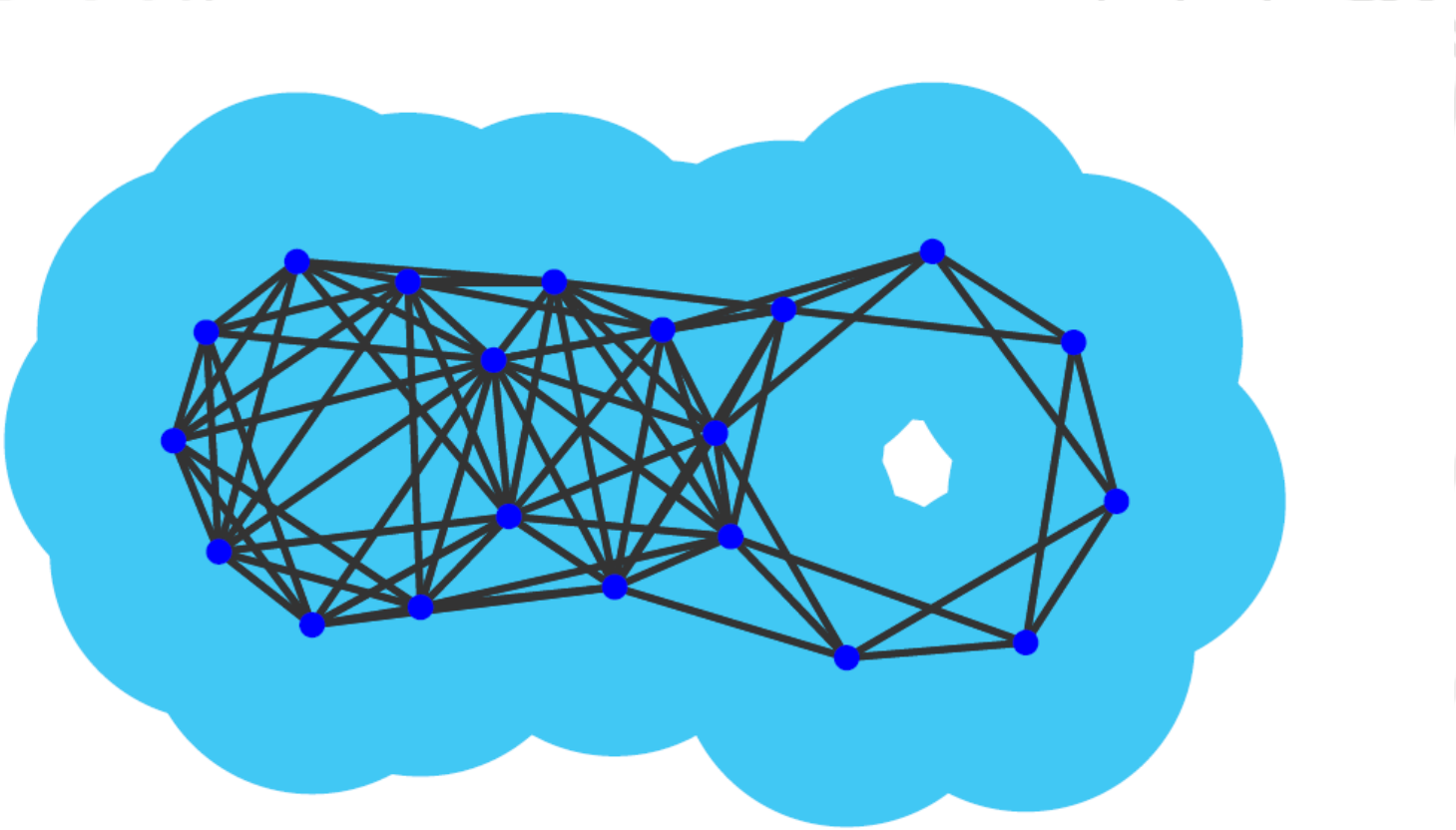


Figura: Radio=9

Complejo simplicial filtrado

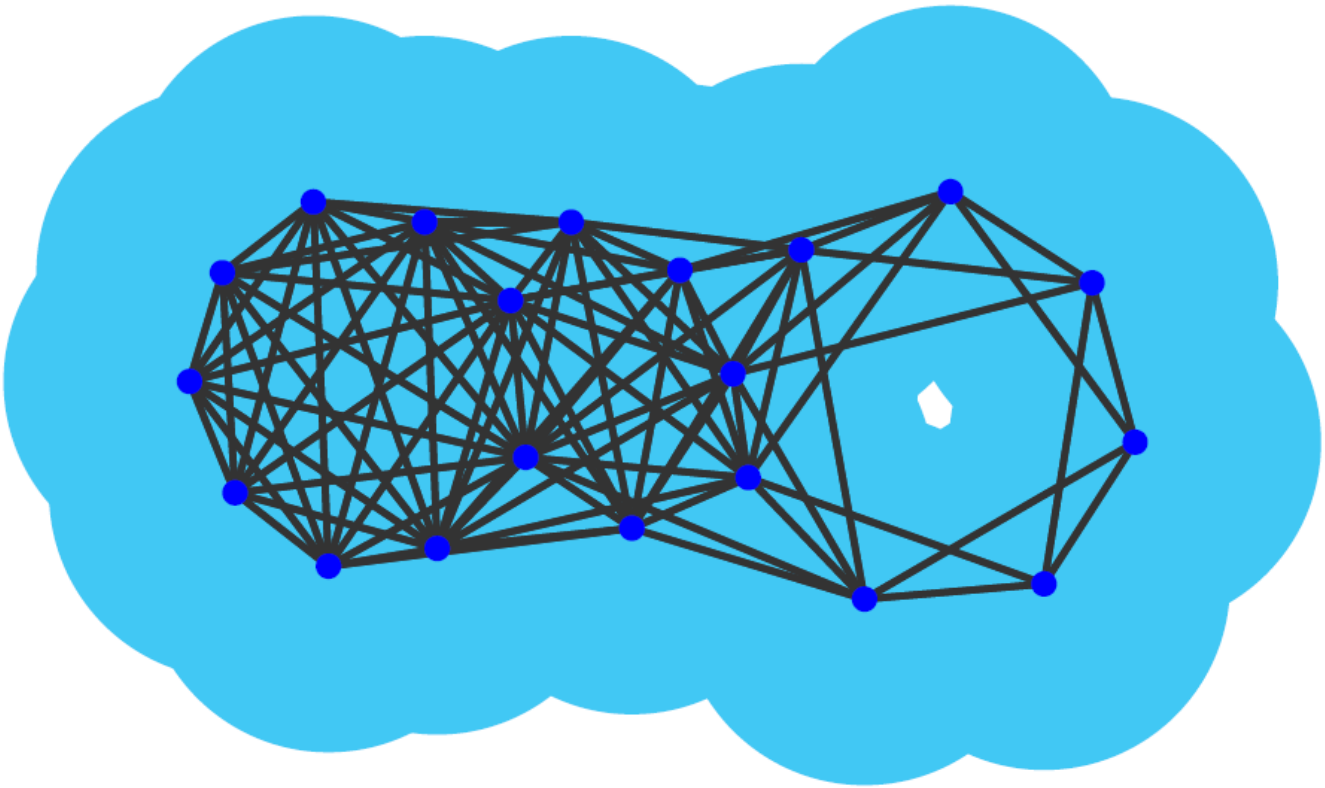


Figura: Radio=10

Complejo simplicial filtrado

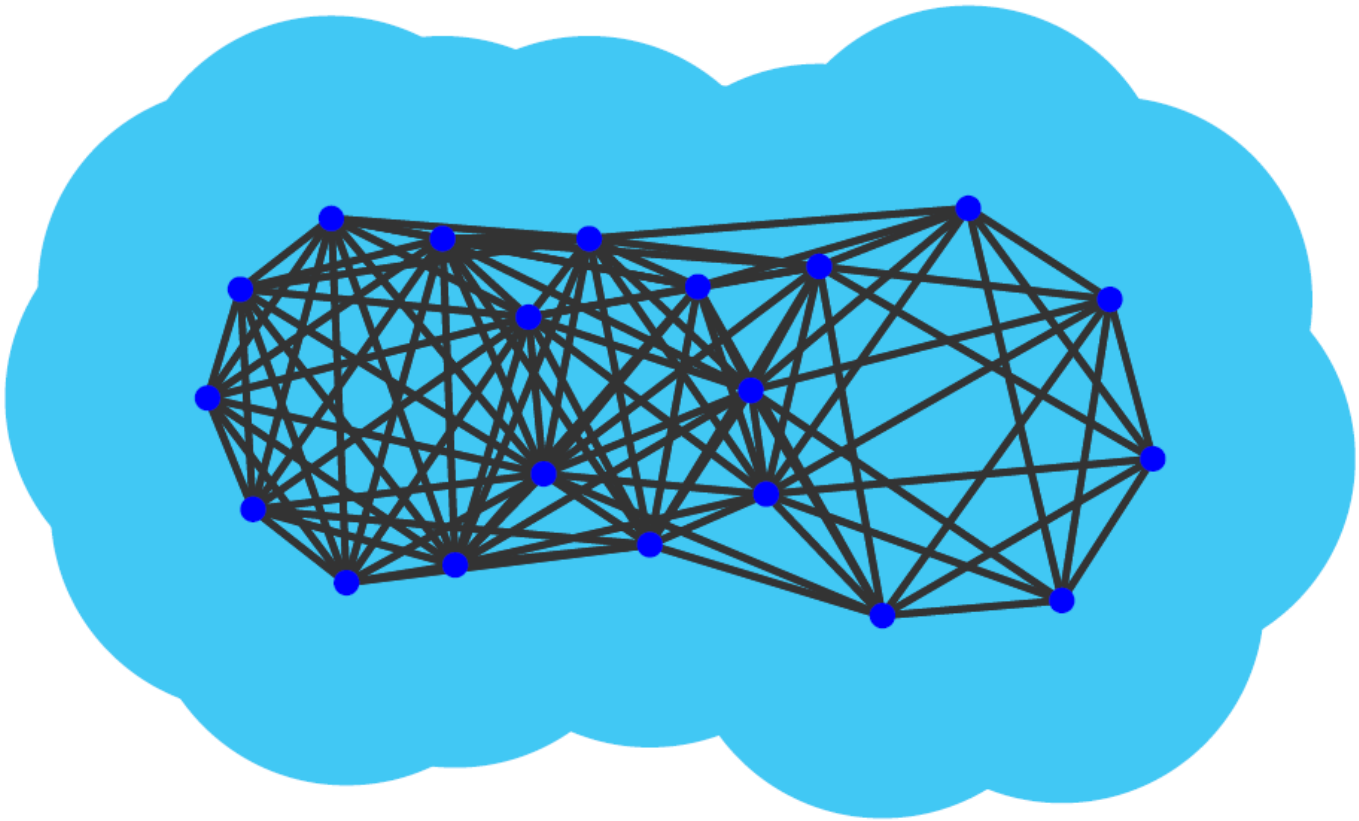


Figura: Radio=11

Homología persistente

- Consideremos la filtración de un complejo simplicial abstracto K, $K_0 \subset K_1 \subset \cdots \subset K_m$.

Homología persistente

- Consideremos la filtración de un complejo simplicial abstracto K, $K_0 \subset K_1 \subset \cdots \subset K_m$.
- Denotemos por $\iota_{i,j} : K_i \longrightarrow K_j$ la inclusión canónica.

Homología persistente

- Consideremos la filtración de un complejo simplicial abstracto K, $K_0 \subset K_1 \subset \cdots \subset K_m$.
- Denotemos por $\iota_{i,j} : K_i \longrightarrow K_j$ la inclusión canónica.
- Estas inclusiones inducen morfismos en homología para cada n:

$$\iota_{i,j_*} : H_n(K_i) \longrightarrow H_n(K_j).$$

Homología persistente

- Consideremos la filtración de un complejo simplicial abstracto K, $K_0 \subset K_1 \subset \cdots \subset K_m$.
- Denotemos por $\iota_{i,j} : K_i \longrightarrow K_j$ la inclusión canónica.
- Estas inclusiones inducen morfismos en homología para cada n:

$$\iota_{i,j_*} : H_n(K_i) \longrightarrow H_n(K_j).$$

- El n-ésimo grupo de homología persistente de nivel i, j se define como:

$$H_{i,j:n}(K) := \mathrm{Im}(\iota_{i,j_*}).$$

Homología persistente

- Consideremos la filtración de un complejo simplicial abstracto K, $K_0 \subset K_1 \subset \cdots \subset K_m$.
- Denotemos por $\iota_{i,j} : K_i \longrightarrow K_j$ la inclusión canónica.
- Estas inclusiones inducen morfismos en homología para cada n:

$$\iota_{i,j_*} : H_n(K_i) \longrightarrow H_n(K_j).$$

- El n-ésimo grupo de homología persistente de nivel i, j se define como:

$$H_{i,j:n}(K) := \mathrm{Im}(\iota_{i,j_*}).$$

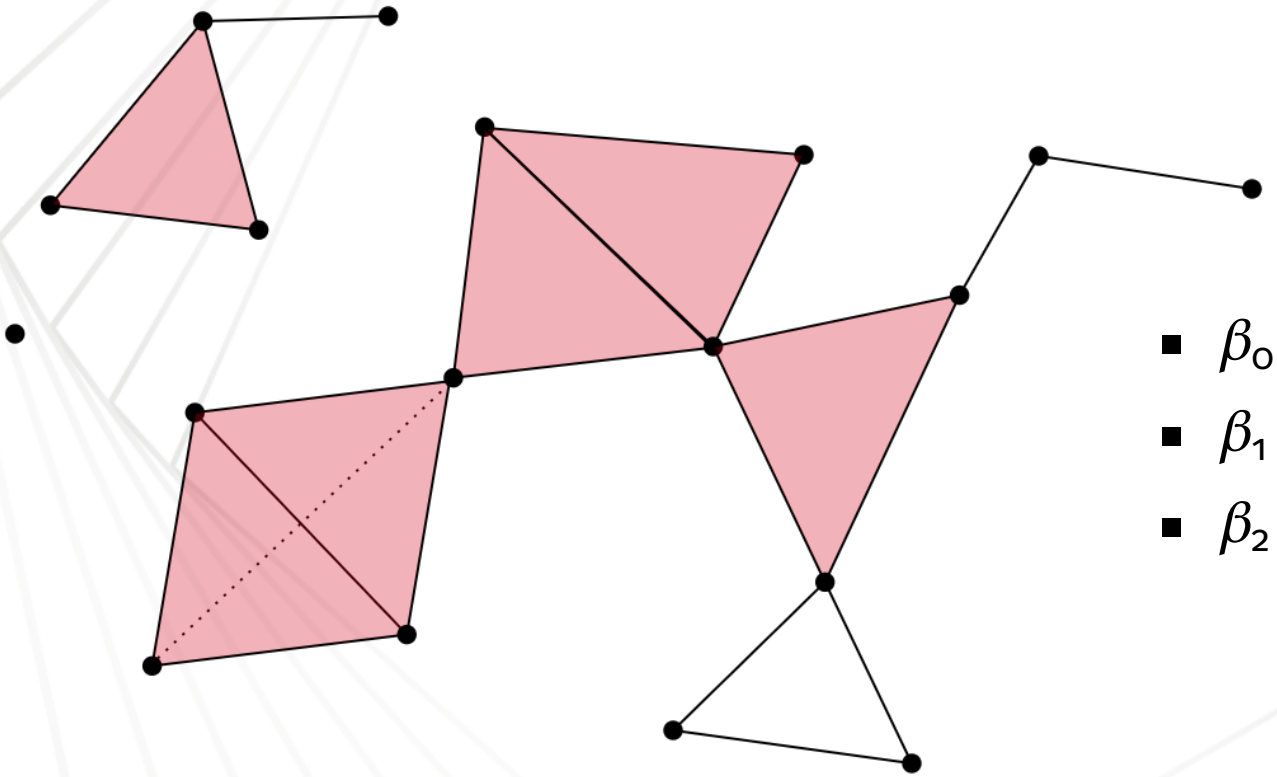
- A la dimensión de este grupo la denotamos por $\beta_{i,j;n}$ y se llama el n-ésimo número de Betti persistente de nivel i, j.

Números de Betti

Números de Betti

- H_k es un grupo (espacio vectorial) en el cual cada elemento es una clase de equivalencia de ciclos asociados al mismo "agujero".
- La dimensión de H_k es llamada el β_k- número de Betti.
  - $\beta_0 =$ Número de componentes conexas.
  - $\beta_1 =$ Número de "agujeros".
  - $\beta_2 =$ Número de "vacíos/huecos".

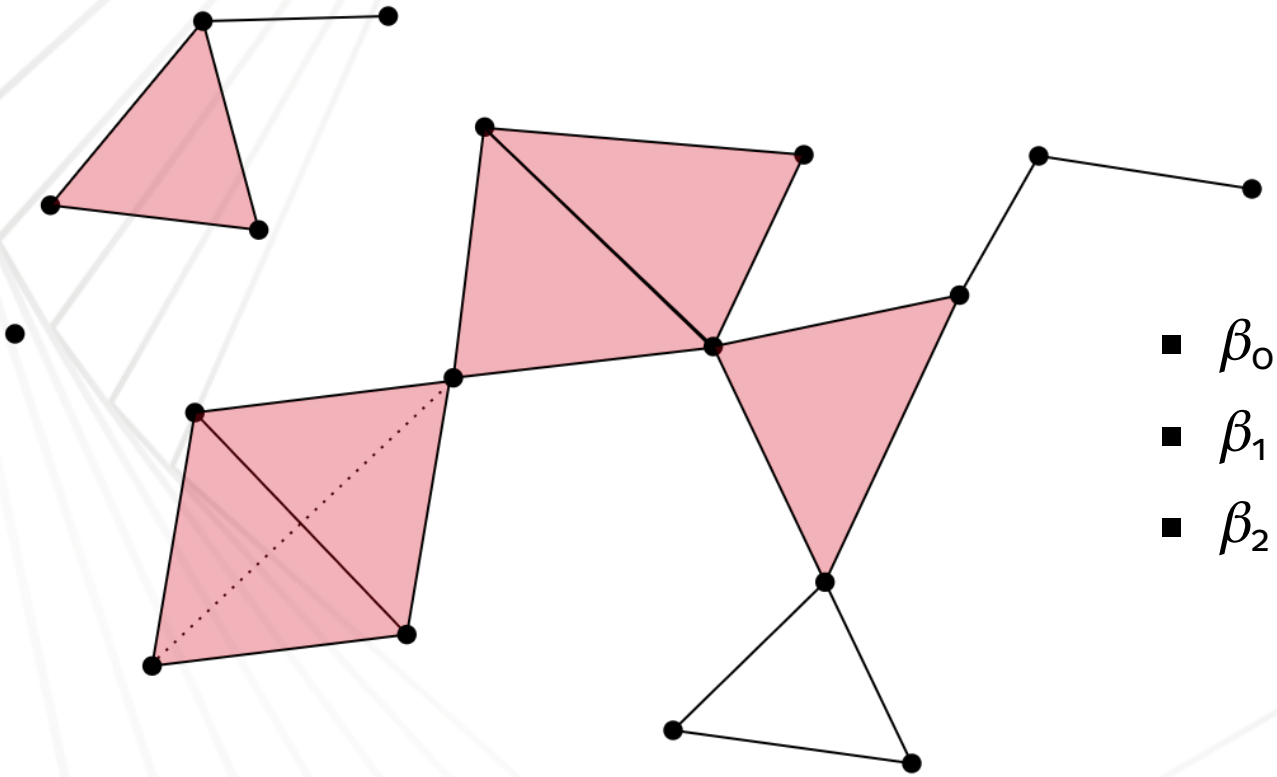
Ejemplo



- $\beta_0 =$
- $\beta_1 =$
- $\beta_2 =$

Figura: Complejo simplicial

Ejemplo



- $\beta_0 = 3$
- $\beta_1 = 1$
- $\beta_2 = 1$

Figura: Complejo simplicial

Códigos de barras

Un código de barras es una representación gráfica de $H_{i,j;n}(K)$. Consiste de una colección de segmentos de líneas horizontales en un plano, cuyos ejes corresponden al intervalo de persistencia (eje horizontal) y a un orden, que puede ser arbitrario, de los generadores de homología (eje vertical).

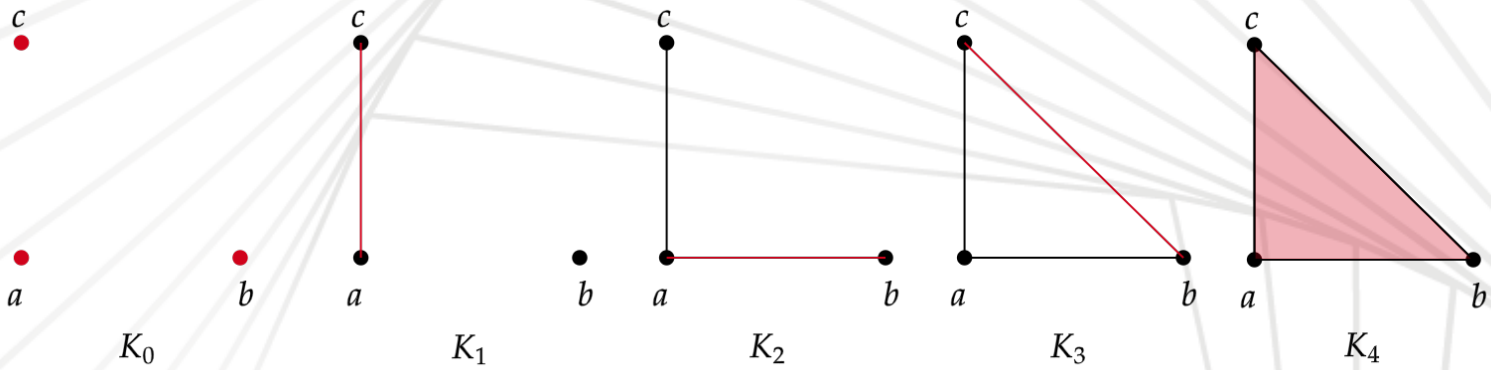


Figura: Ejemplo: Filtración

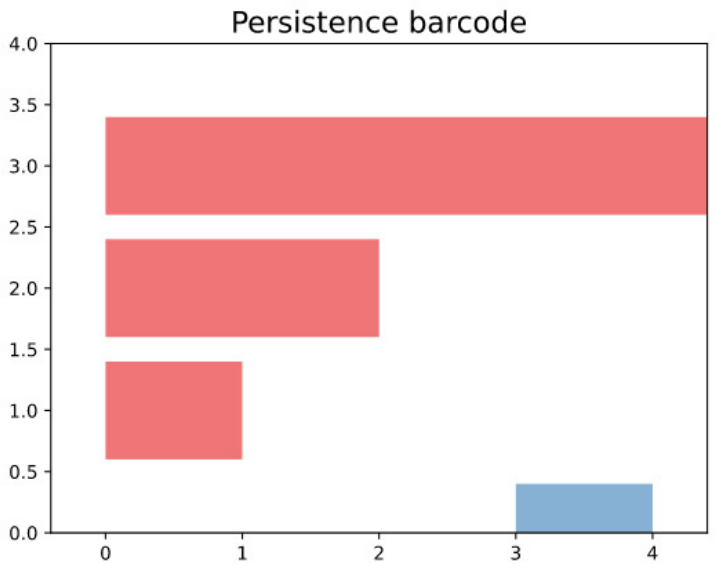


Figura: Código de barras

Aplicaciones

Colaboradores

- Shaday Guerrero Flores - CCM
- Nelly Selem Mojica - CCM
- José María Ibarra Rodríguez - C3
- Noé Bárcenas - CCM
- Jesús Hernández Hernández - CCM

¿Qué es la Pangenómica?

En la genómica, el término pangenoma se refiere al conjunto completo de genes de todas las cepas que forman un clado filogenético, es decir, es la unión de todos los genomas de los organismos que pertenecen a un mismo clado.
El campo de estudio del pangenoma se llama Pangenoma.

Microorganismo

DNA

A ⇨ T

C ⇨ G

Nucleótidos

```
>AL645882.2 Streptomyces
CCCGCGGAGCGGGTACCACATCGCT
TCCCGCTCCGCGGGAGCGCTGGCGG
ACGCTCCGTCCGCTGCGCTTCCGGA
CCTTGGGCCTGCGGCGGGTCCGCTG
AACCGCTTGCGCGGGCCTGGCTGGC
GGCTGCCTCAGATAGATGCAGCATG
TCCCAAGTTTCGAGAGGATGGCCAG
AGCGATTCGACGTGCCGGGTGACGC
TCCTGGAGCGCCGGGCGTGCCCGCG
GAGCTGGACGGCATGGCGCCGGCTT
GGCGGTGAGCGAGGCCCTGGGGGGC
CGACTGAGCTGGCCGTACATTGCGC
```

Genoma

Figura: Los organismos almacenan información en el DNA

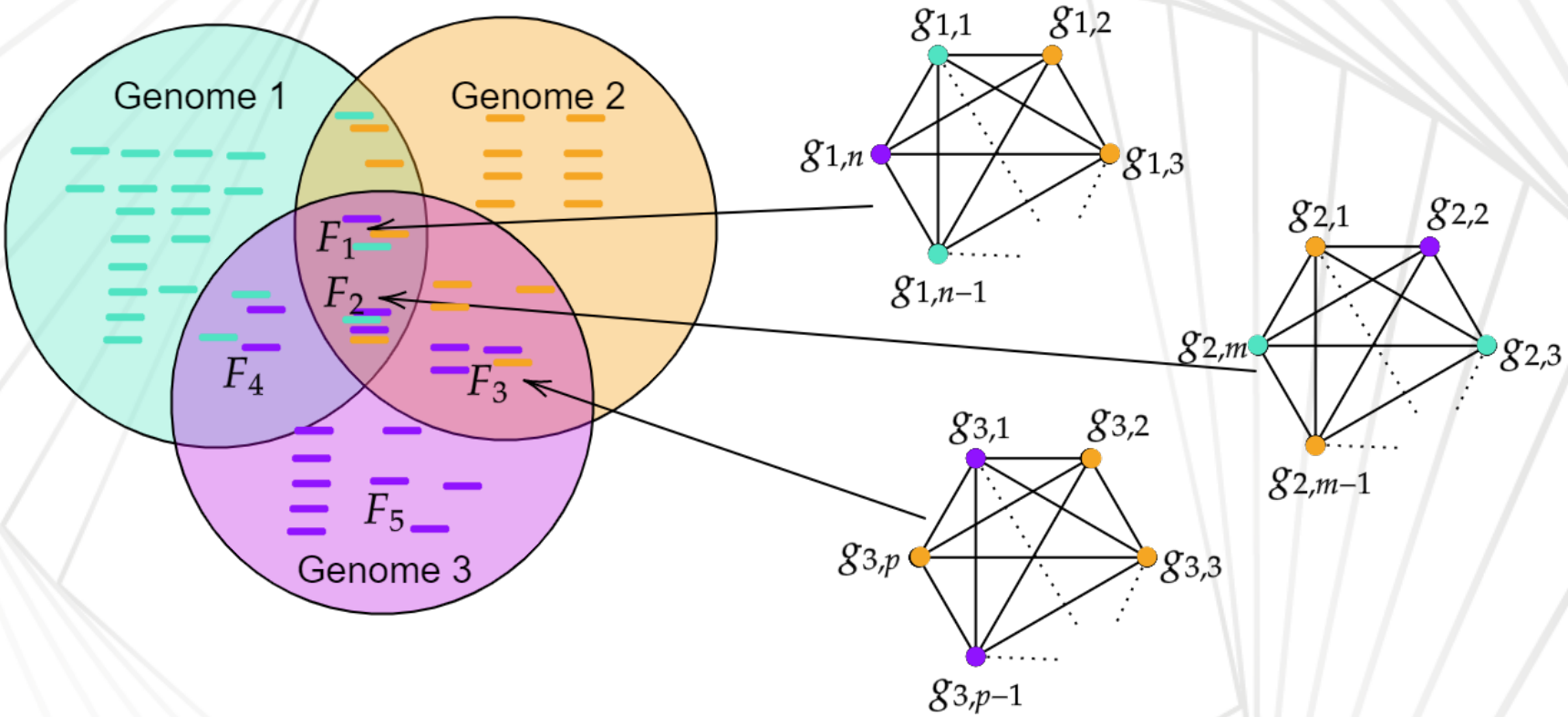


Figura: Para un parámetro t de agrupamiento obtenemos una colección de familias de genes.

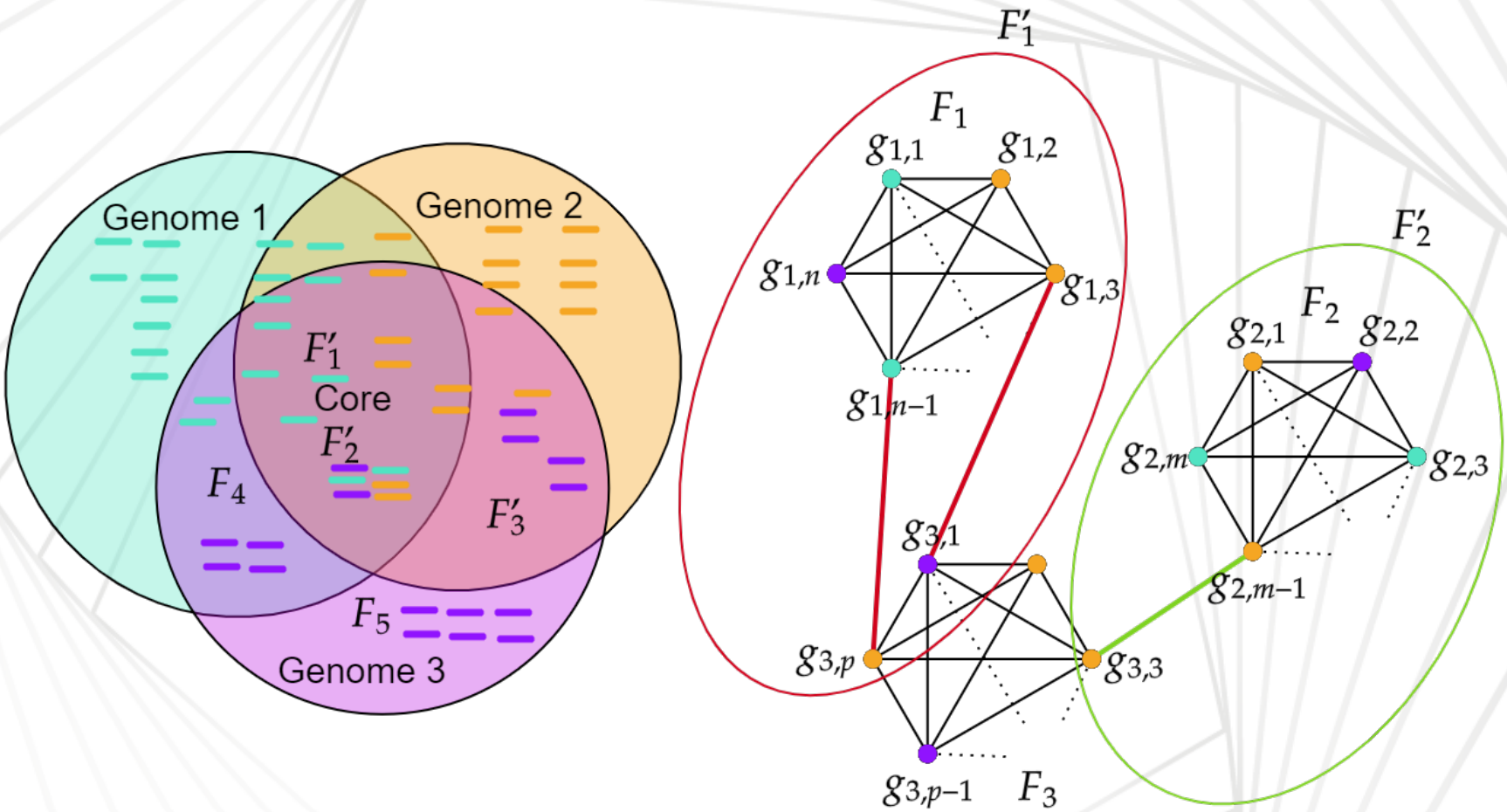


Figura: Al variar el parámetro de agrupamiento obtenemos las familias cambian.

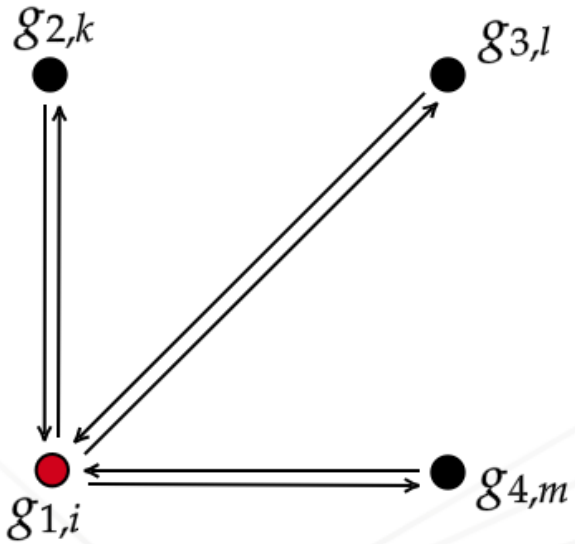
Objetivos

Explorar el pangenoma usando TDA:

- Detectar conjuntos de genes que persisten.
- Clasificar familias del genoma central, genoma periférico y genoma diferencial.
- ¿Cuántas familias quedan al final?
- Comparar pangenomas de diferentes familias o especies (en proceso).
- ¿Se puede detectar transferencia horizontal?

1er método

- Usar el evalue para construir una matriz de distancias entre los genes de varios genomas.
- ¡No es simétrica!
- El evalue de BLAST es el número de hits esperados de calidad similar (puntuación) que podrían encontrarse por casualidad.

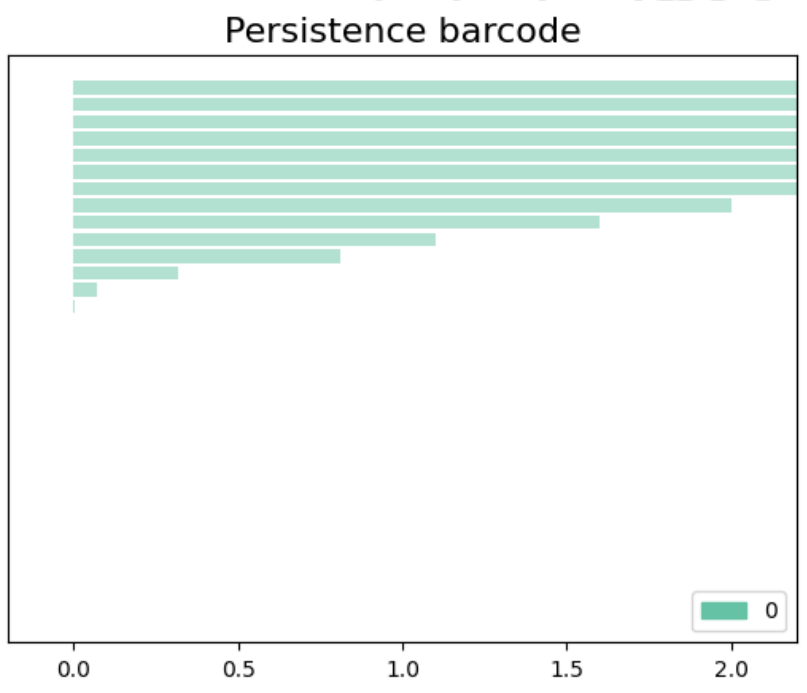


	qseqid	sseqid	evalue
0	2603V|GBPINHCM_01420	NEM316|AOGPFIKH_01528	4.110000e-67
1	2603V|GBPINHCM_01420	A909|MGIDGNCP_01408	4.110000e-67
2	2603V|GBPINHCM_01420	515|LHMFJANI_01310	4.110000e-67
3	2603V|GBPINHCM_01420	2603V|GBPINHCM_01420	4.110000e-67
4	2603V|GBPINHCM_01420	A909|MGIDGNCP_01082	1.600000e+00

sseqid qseqid	2603V|GBPINHCM_00065	2603V|GBPINHCM_00097	2603V|GBPINHCM_00348	2603V|GBPINHCM_00401
2603V|GBPINHCM_00065	1.240000e-174	NaN	NaN	NaN
2603V|GBPINHCM_00097	NaN	9.580000e-100	NaN	NaN
2603V|GBPINHCM_00348	NaN	NaN	0.0	NaN
2603V|GBPINHCM_00401	NaN	NaN	NaN	2.560000e-135

sseqid qseqid	2603V|GBPINHCM_00065	2603V|GBPINHCM_00097	2603V|GBPINHCM_00348	2603V|GBPINHCM_00401
2603V|GBPINHCM_00065	1.240000e-174	5.000000e+00	5.0	5.000000e+00
2603V|GBPINHCM_00097	5.000000e+00	9.580000e-100	5.0	5.000000e+00
2603V|GBPINHCM_00348	5.000000e+00	5.000000e+00	0.0	5.000000e+00
2603V|GBPINHCM_00401	5.000000e+00	5.000000e+00	5.0	2.560000e-135

```
([4, 15, 25, 30, 37, 39], 0.001)
([18, 25, 30, 37, 39], 0.001)
([4, 18, 25, 30, 37, 39], 0.001)
([15, 18, 25, 30, 37, 39], 0.001)
([4, 15, 18, 25, 30, 37, 39], 0.001)
([2, 39], 0.003)
([13, 39], 0.003)
([2, 13, 39], 0.003)
([23, 39], 0.003)
([2, 23, 39], 0.003)
([13, 23, 39], 0.003)
([2, 13, 23, 39], 0.003)
([35, 39], 0.003)
```

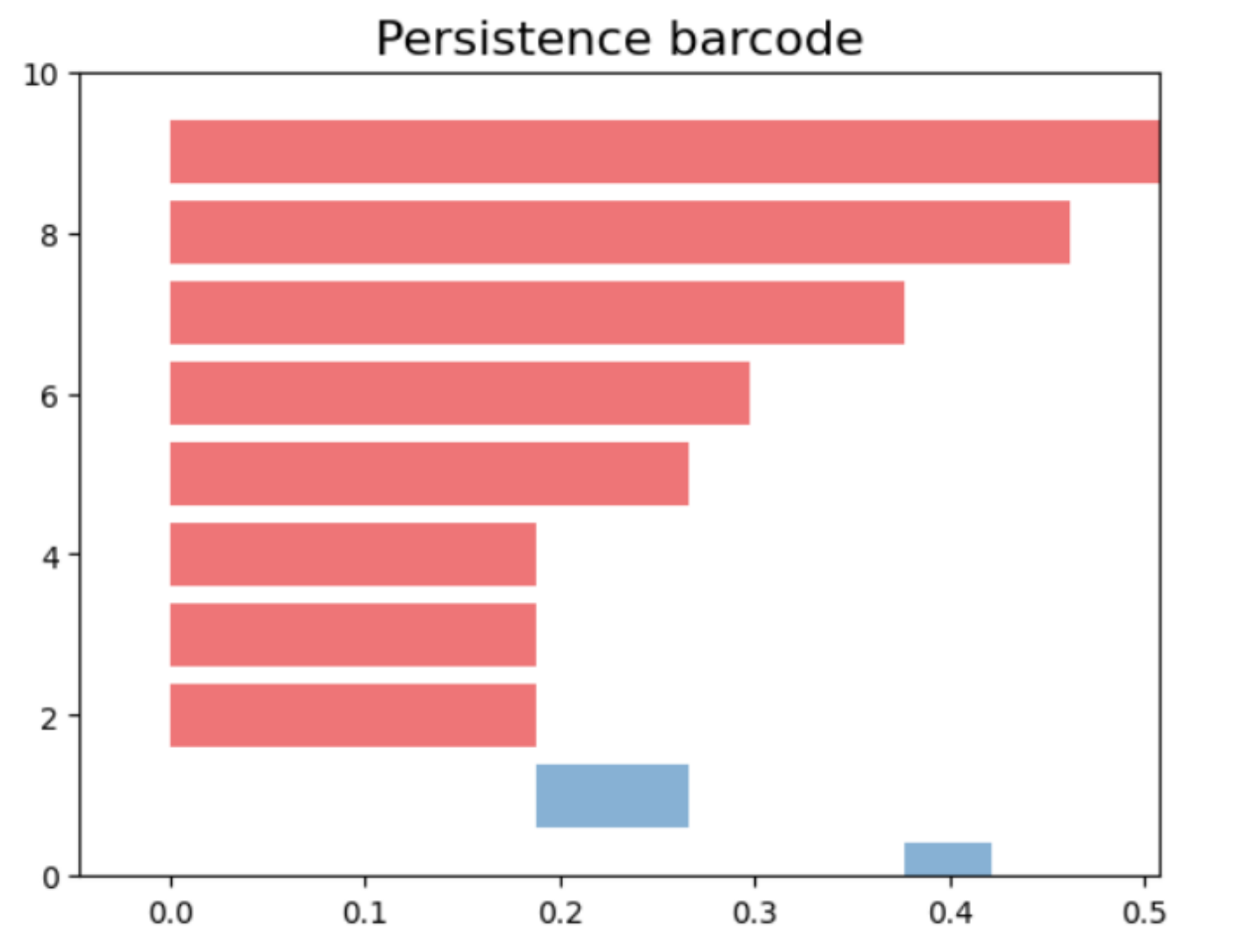


Columna1	t_birth	t_death	persistence	2603V	515	A909	N
('2603V|GBPINHCM_00554', '2603V|GBPINHCM_01231', '515|LHMFJANI_00548', '515|LHMFJANI_01178', 'A909|MGIDGNCP_00580', 'A909|MGIDGNCP_01268', 'NEM316|AOGPFIKH_00621', 'NEM316|AOGPFIKH_01341')	0.014	2	1.986	1	1	1	
('2603V|GBPINHCM_00401', '515|LHMFJANI_00394', '515|LHMFJANI_01625', 'A909|MGIDGNCP_00405', 'NEM316|AOGPFIKH_00403', 'NEM316|AOGPFIKH_01842')	1.3	2	0.7	1	1	1	
('2603V|GBPINHCM_01042', '2603V|GBPINHCM_01420', '515|LHMFJANI_01310', 'A909|MGIDGNCP_01408', 'NEM316|AOGPFIKH_01528')	1.6	2	0.4	1	1	1	
('2603V|GBPINHCM_00065', '515|LHMFJANI_00064', 'A909|MGIDGNCP_00064', 'A909|MGIDGNCP_00627', 'NEM316|AOGPFIKH_00065')	0.086	2	1.914	1	1	1	
('2603V|GBPINHCM_00348', '515|LHMFJANI_00342', 'A909|MGIDGNCP_00352', 'NEM316|AOGPFIKH_00350', 'NEM316|AOGPFIKH_01341')	0.003	2	1.997	1	1	1	
('2603V|GBPINHCM_01042', 'A909|MGIDGNCP_01082', 'A909|MGIDGNCP_01408', 'NEM316|AOGPFIKH_01528')	1.6	2	0.4	1	0	1	
('2603V|GBPINHCM_00748', '515|LHMFJANI_00064', 'A909|MGIDGNCP_00064', 'NEM316|AOGPFIKH_00065')	0.83	2	1.17	1	1	1	
('2603V|GBPINHCM_00097', '515|LHMFJANI_00097', 'A909|MGIDGNCP_00096', 'NEM316|AOGPFIKH_00098')	9.58E-100	2	2	1	1	1	
('2603V|GBPINHCM_00815', '515|LHMFJANI_00781', 'A909|MGIDGNCP_00877', 'NEM316|AOGPFIKH_00855')	0	2	2	1	1	1	
('2603V|GBPINHCM_00748', '2603V|GBPINHCM_01042', 'A909|MGIDGNCP_01082')	2	2	0	1	0	1	
('2603V|GBPINHCM_00748', '2603V|GBPINHCM_01042')	2	2	0	1	0	0	
('2603V|GBPINHCM_00748', 'A909|MGIDGNCP_01082')	2	2	0	1	0	1	
('515|LHMFJANI_01625', 'A909|MGIDGNCP_01221')	1.1	2	0.9	0	1	1	
('515|LHMFJANI_01130', 'A909|MGIDGNCP_01221')	1.31E-85	2	2	0	1	1	
('A909|MGIDGNCP_01343', 'NEM316|AOGPFIKH_01415')	7.89E-143	2	2	0	0	1	
('2603V|GBPINHCM_01226',)	0	2	2	1	0	0	

2do método

- Usar la distancia de Hamming.
- Asociar el complejo simplicial.
- Obtener los diagramas de persistencia.
- Detecta transferencia horizontal.
- Agregar puntos medios.

	g_A909	g_2603V	g_515	g_NEM316
A909|MGIDGNCP_01408	1	1	1	1
A909|MGIDGNCP_00096	1	1	1	1
A909|MGIDGNCP_01343	1	0	0	1
A909|MGIDGNCP_01221	1	0	1	0
A909|MGIDGNCP_01268	1	1	1	1
A909|MGIDGNCP_00580	1	1	1	1
A909|MGIDGNCP_00352	1	1	1	1
A909|MGIDGNCP_00064	1	1	1	1
A909|MGIDGNCP_00627	1	0	0	0
A909|MGIDGNCP_01082	1	1	0	0
A909|MGIDGNCP_00877	1	1	1	1
A909|MGIDGNCP_00405	1	1	1	1
2603V|GBPINHCM_00748	0	1	0	0
2603V|GBPINHCM_01226	0	1	0	0
515|LHMFJANI_01625	0	0	1	1


Persistence barcode
10
8
6
4
2
0
0.0
0.1
0.2
0.3
0.4
0.5

Aprendizaje Geométrico Profundo

Aprendizaje Geométrico Profundo

- El término aprendizaje geométrico profundo fue dado por Michael Bronstein.
- Involucra la codificación de una comprensión geométrica de los datos como un sesgo inductivo para ayudar a los modelos de aprendizaje profundo.
- El entendimiento de la geometría esta codificado generalmente en: la simetría e invarianza, la estabilidad y la representación multiescala.
- Le proporcionamos esta información a los algoritmos de aprendizaje profundo para que no tenga que aprender esto.

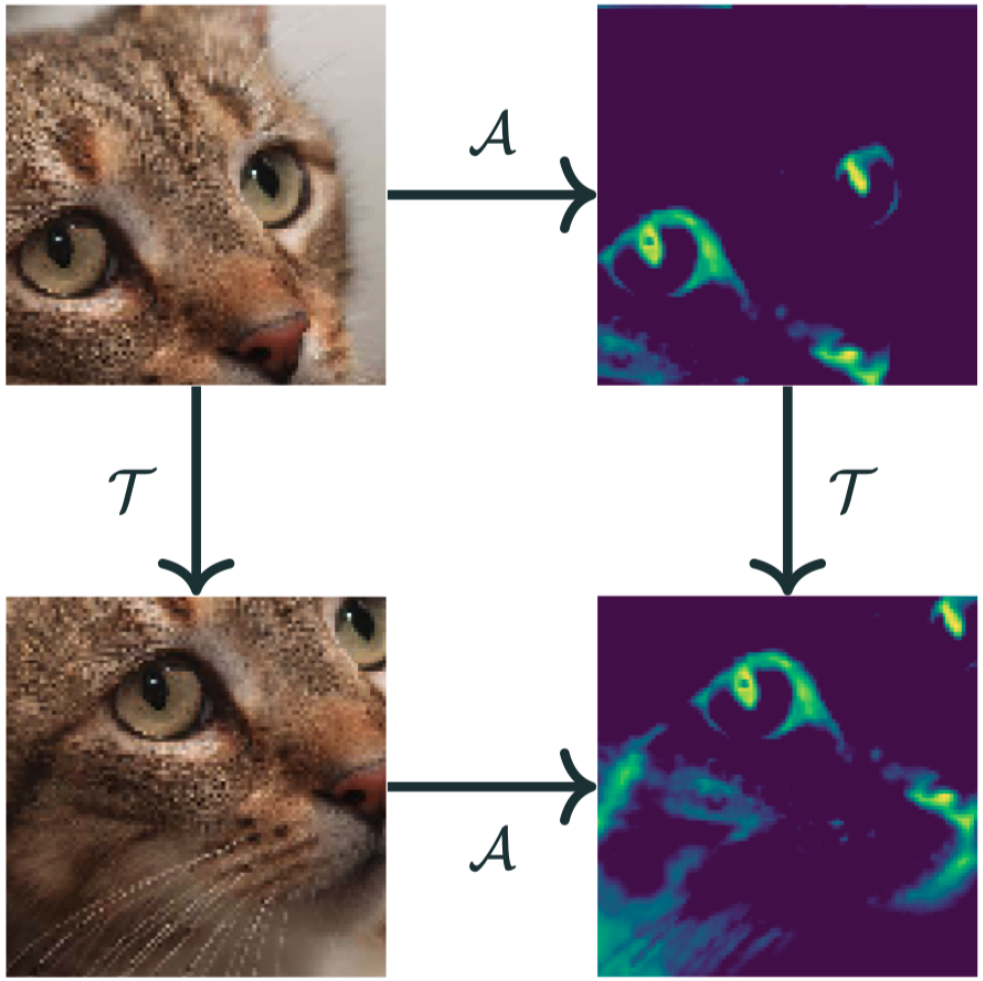


Figura: Equivarianza de la traslación, [14]

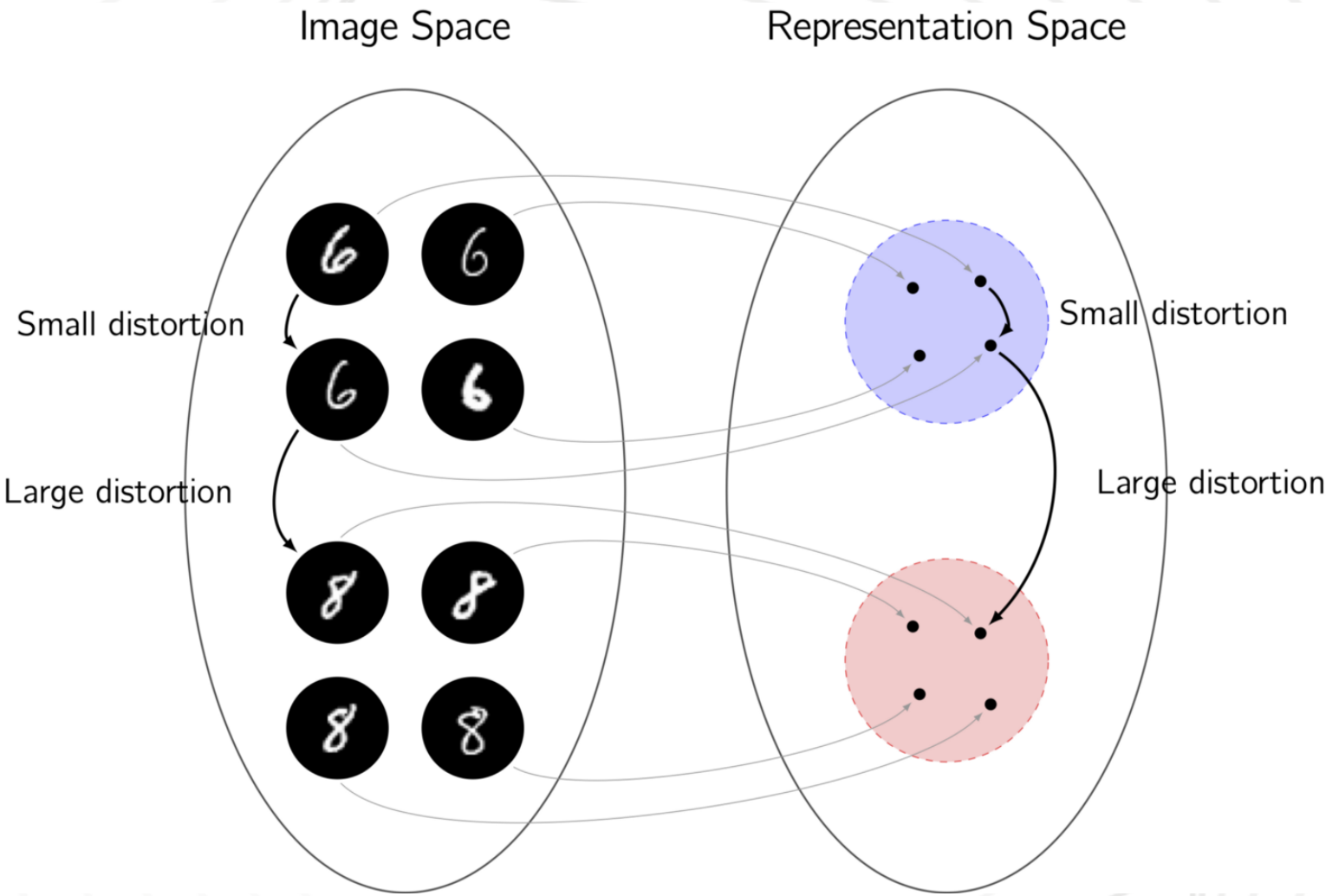


Figura: Estabilidad, [14]

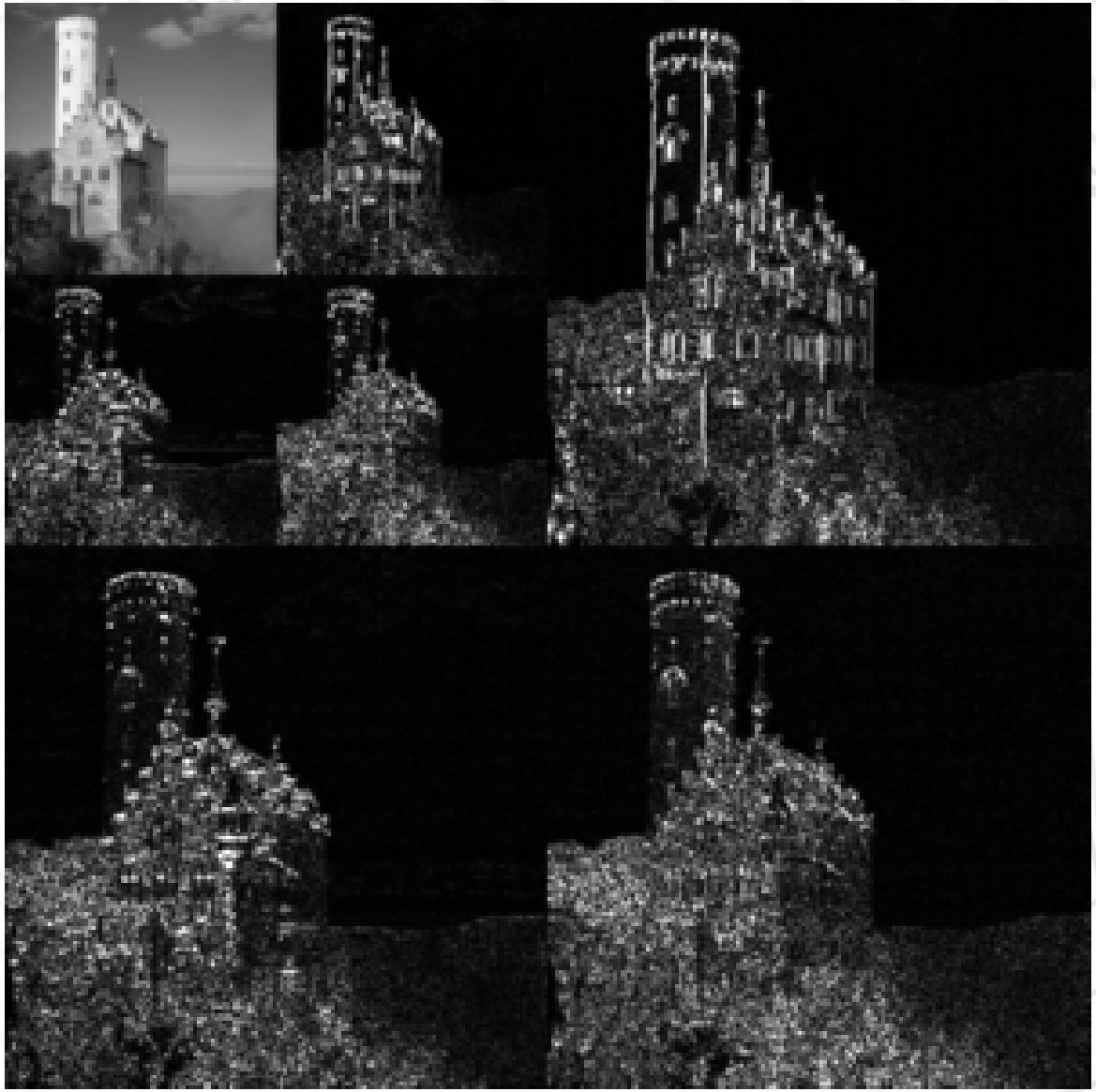


Figura: Multiescala, [14]

Ejemplos de aplicaciones de GDL

Decidir cuando dos 3-variedades dadas como “plumbing graphs” dan 3-variedades homeomorfas, [17].

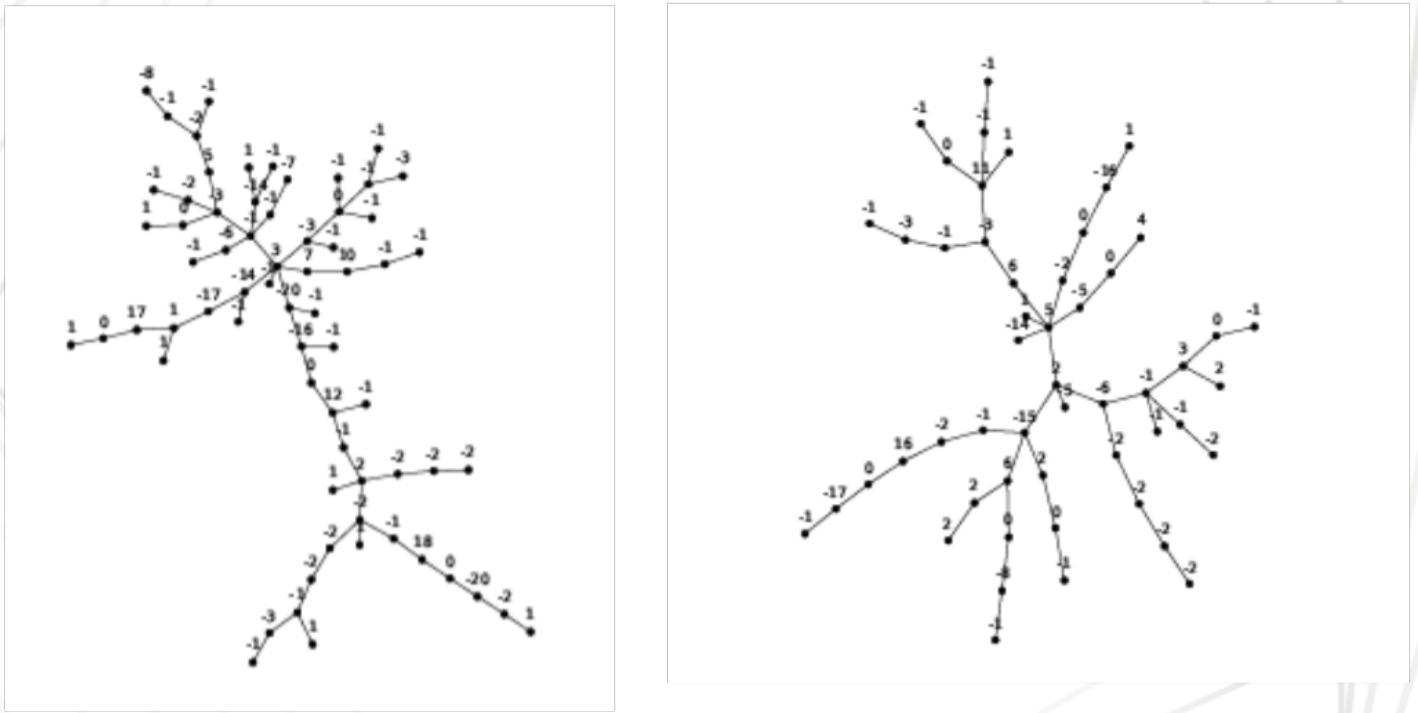


Figura: Par de plumbing graphs equivalentes reconocidas por los algoritmos.

Ejemplos de aplicaciones de GDL

Construir un funtor que mande nudos a gráficas y aplicar GDL para predecir invariantes de nudos, [13]

Reducción de dimensiones

- PCA:Dado un conjunto de puntos $\{x_1, ..., x_m\} \in \mathbb{R}^n$ encontrar una proyección a $\mathbb{R}^k$ tal que el error de hacer esto sea mínimo.

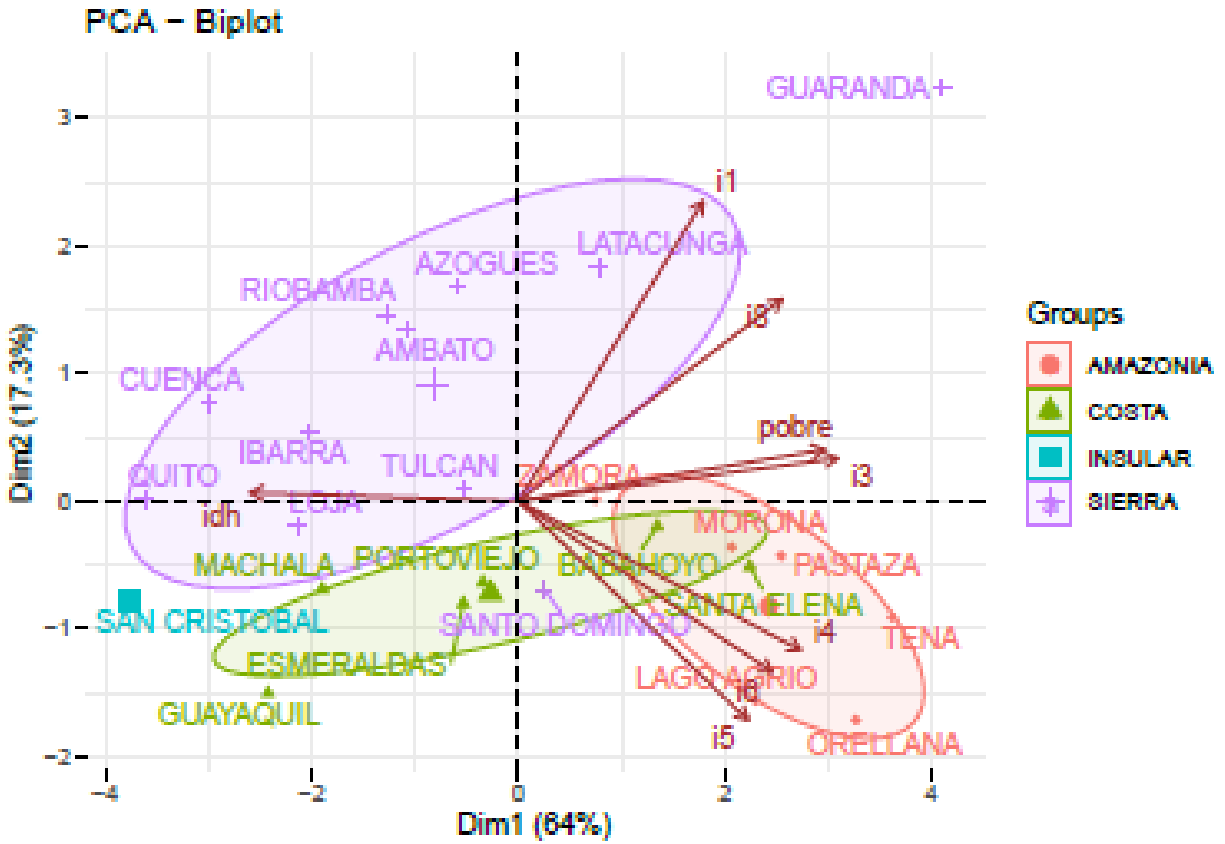


Figura: Primera y segunda componente

Reducción de dimensiones

- MDS: Dado un espacio métrico finito (X, d_X), encontrar un encaje óptimo de X en $\mathbb{R}^k$ de tal forma que se preserve lo mejor posible la métrica original de los datos.

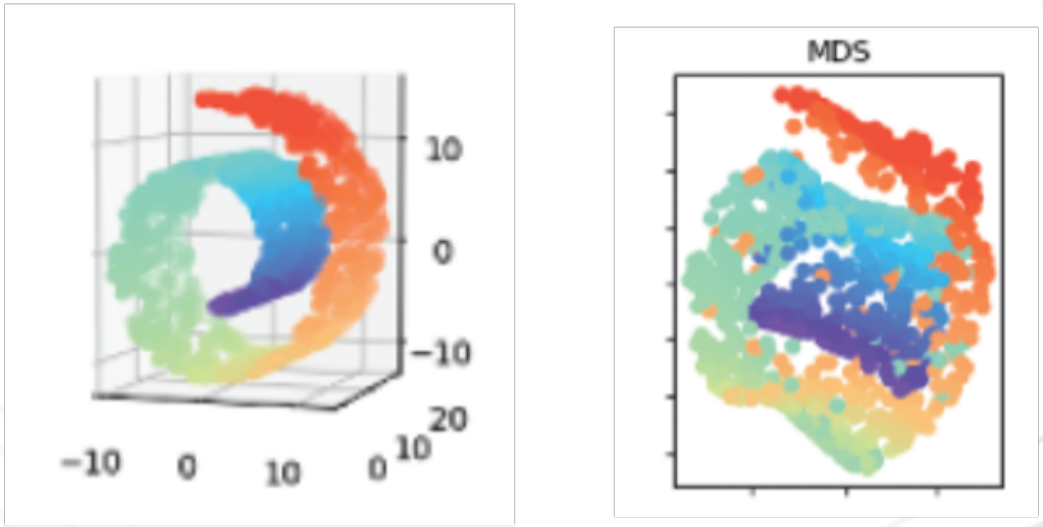


Figura: [19, Figura 4.2]

Reducción de dimensiones

- Si el conjunto de datos es un subconjunto de $\mathbb{R}^n$ entonces PCA y MDS métrico coincide.
- En estos métodos uno intenta conservar en lo posible la geometría intrínseca de los datos.
- Si los datos no pueden ser encajados en un $\mathbb{R}^n$ entonces estos métodos no capturan la geometría intrínseca.

Aprendizaje de Variedades

Idea

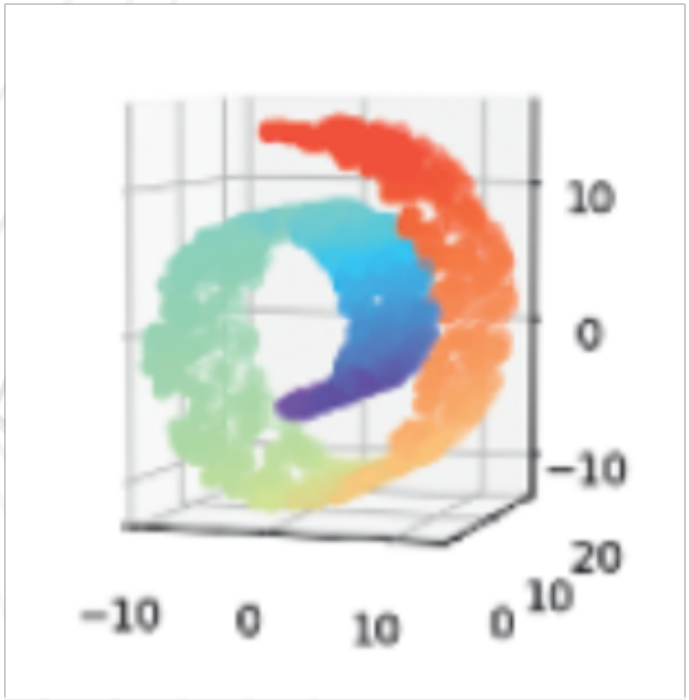
La estructura de la variedad puede ser reconstruida considerando distancias peqeñas como un indicador adecuado de la distancia intrínseca e ignorar las distancias grandes.

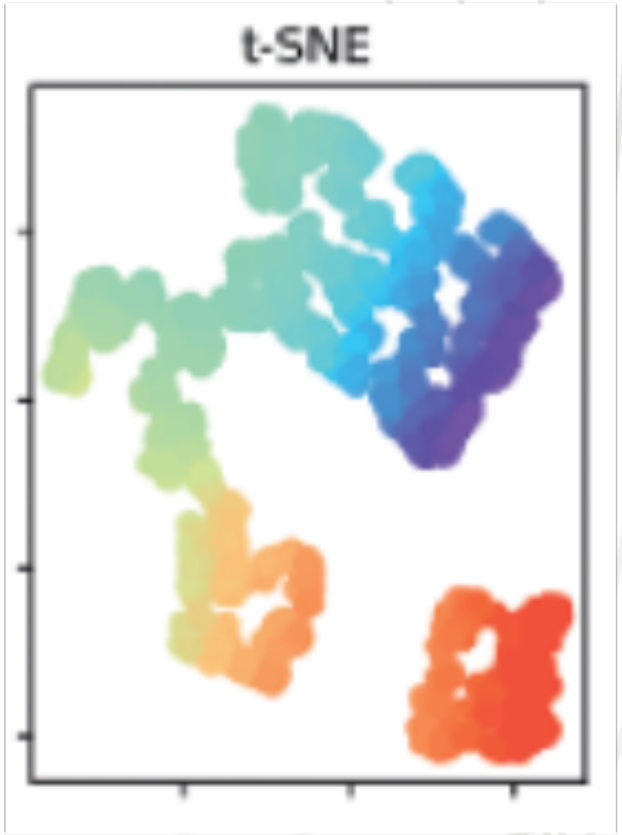
Las técnicas de aprendizaje de variedades se basan en tomar los k vecinos más cercanos a un punto x y con estos se busca aproximar el plano tangente a x.

t-SNE

- Es un método estadístico basado en el encaje de vecinos estocástico.
- Es una técnica no lineal no supervisada utilizada para la exploración de datos y la visualización de datos de alta dimensión.
- La distancia entre los puntos se convierte en probabilidades.
- Se calcula la similitud entre los puntos en la dimensión inicial.
- Se crea un espacio de dimensión baja en el cual se representa los puntos.
- Se calcula la medida de similitud entre el espacio original y el final.
- Usualmente se proyecta a $\mathbb{R}^2$ y se realiza un algoritmo de clustering.

t-SNE





UMAP

- Primer paso: crear una representación topológica de los datos usando la gráfica del k–vecino más cercano.
- Segundo paso: encontrar una representación de baja dimensión que maximice la entropía cruzada al gráfico de alta dimensión.
- Tiene 3 parámetros básicos:
  1. `n_neighs`: número de vecinos a considerar para la aproximación de la métrica local.
  2. `min_dist`: la distancia mínima entre los puntos en el espacio de baja dimensión.
  3. La métrica a utilizar para crear la representación topológica de los datos.

t-SNE vs UMAP

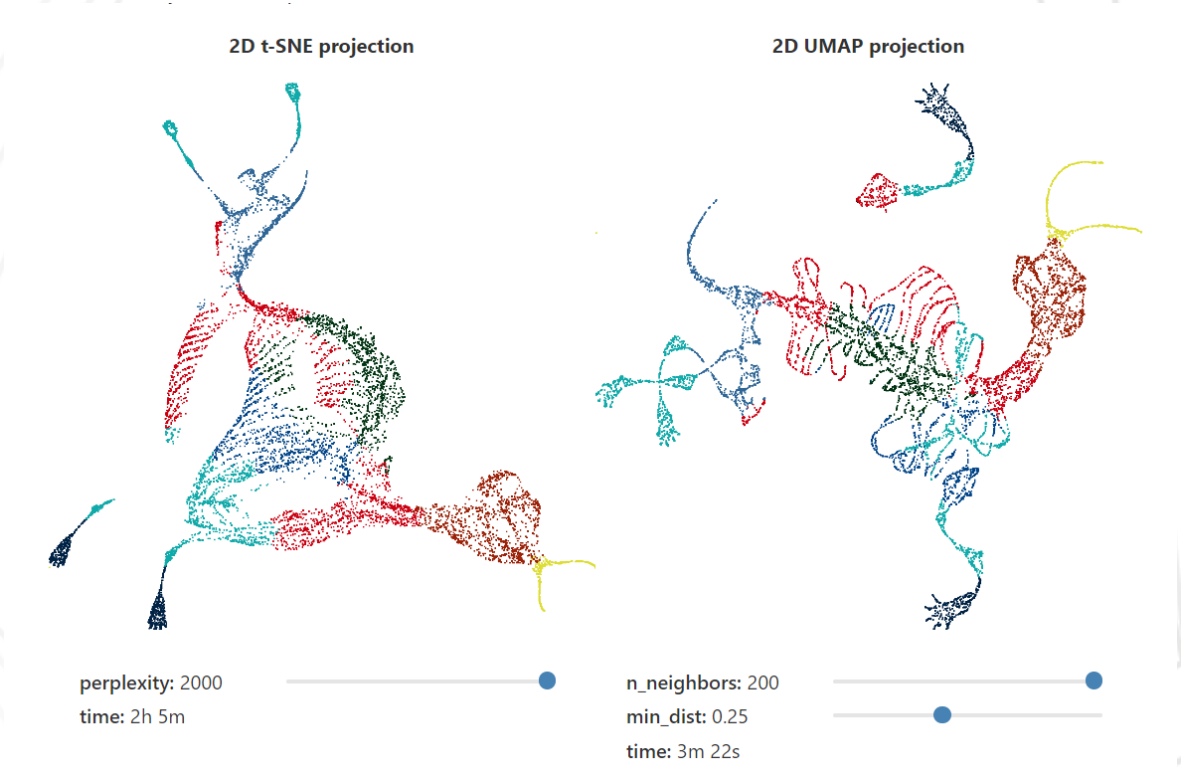


Figura: Ejemplo de: https://pair-code.github.io/understanding-umap/

TriMap

- Compara distancias entre tripletas de puntos.
- Una tripleta (i, j, k) es tal que i es más cercano a j que a k.
- Cada tripleta se asocia a un peso donde valores grandes implican que el par (i, k) esta más lejos que el par (i, j).
- El encaje se construye con un subconjunto de tripletas donde el punto j está en el conjunto de los vecinos más cercanos a i y k está en el conjunto de los puntos más lejanos a i y j.

TriMap vs UMAP vs t-SNE

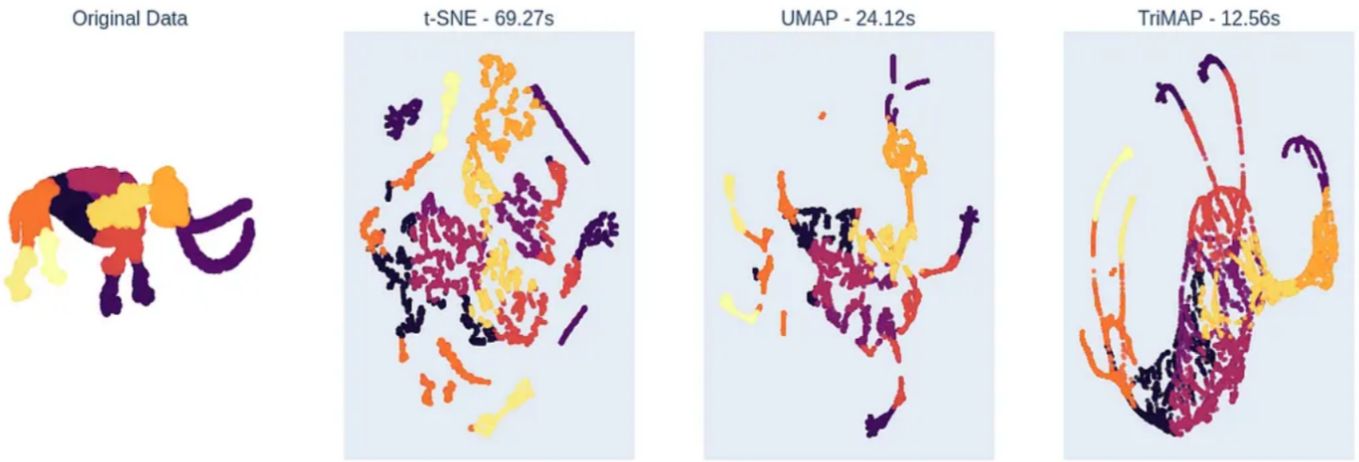


Figura: Ejemplo de: https://towardsdatascience.com/why-you-should-not-rely-on-t-sne-umap-or-trimap-f8f5dc333e59

Colaboradores

- Sebastian Pujalte Ojeda - Instituto de Fisiología Celular UNAM
- Sophie Achard - Centre de recherche Inria Grenoble Rhone-Alpes
- Pablo Suárez Serrato - Instituto de Matemáticas UNAM

Objetivo

¿El espectro del autismo puede ser detectado por medio de las características topológicas y geométricas de gráficas inferidas de imágenes fMRI?

- Autism Brain Imaging Data Exchange (ABIDE) ([8]) tiene datos de 539 individuos con autismo y 573 controles.
- Estimar las conexiones funcionales entre las 417 parcelaciones del cerebro usando análisis de ondas.
- La base de datos con la que trabajamos consta de 866 individuos, 402 pacientes y 464 controles.

Longitud del espectro marcado

- La longitud del espectro marcado es una función en una gráfica que asigna a cada clase de homotopía de un lazo el ínfimo de las longitudes sobre todos los representantes de sus clases de conjugación.
- Para una gráfica G, su parte convexa, $Conv(G)$, es la subgráfica de G que se encuentra al remover iterativamente todos los vértices de grado 1.
- Constantine y Lafont [9] dan una forma de comparar la similitud entre dos gráficas usando la longitud del espectro marcado, la cual solo depende de $Conv(G)$.
- Al remover vértices de grado 1 se puede perder información.

Cono de una gráfica

- El cono de una gráfica consiste en añadir un vértice y unirlo con todos los vértices de la gráfica.
- Para una gráfica G, añadir un cono a G nos deja una nueva gráfica $\widehat{G}$ tal que $\widehat{G} = Conv(\widehat{G})$.
- Calcular el espectro de longitud marcad es difícil, entonces se usa como aproximación los eigenvalores de la matriz de ciclos sin retroceso.

Matriz de Ciclos sin Retroceso

- Un ciclo sin retroceso es una trayectoria cerrada tal que $e_{i+1} \neq e_i^{-1}$ para todo i.
- La matriz de ciclos sin retroceso B es una matriz de tamaño $2m \times 2m$ que representa todos los pares de aristas que están en trayectorias de longitud 2 sin retroceso.
- Cada fila y columna representa una arista $e \in E$ en una de sus dos orientaciones.
- Los ciclos sin retroceso son topológicamente relevantes porque aristas con retroceso son homotópicamente triviales.
- Los ciclos sin retroceso son los represenantes más pequeños en sus clases de homotopía.

Distancia de ciclos sin retroceso

Definición

Sean G, H dos gráficas y sean $\{\lambda_k\}_{k=1}^{r}$, $\{\mu_k\}_{k=1}^{r}$, los r eigenvalores sin retroceso, la distancia de ciclos sin retroceso ([22]) entre G y H se define como $NB_d(G, H) = d(\{\lambda_k\}, \{\mu_k\})$.

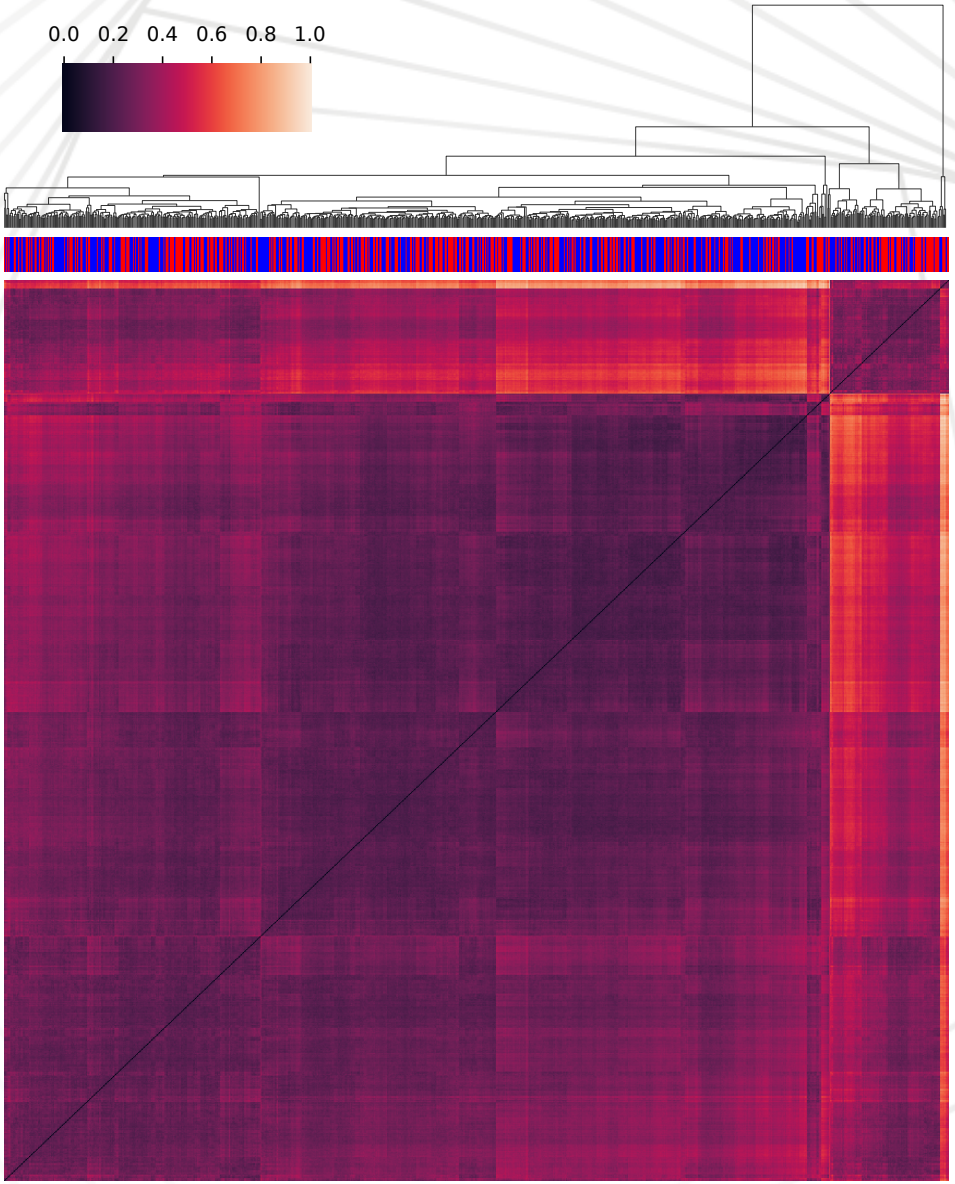


Figura: Clustermap

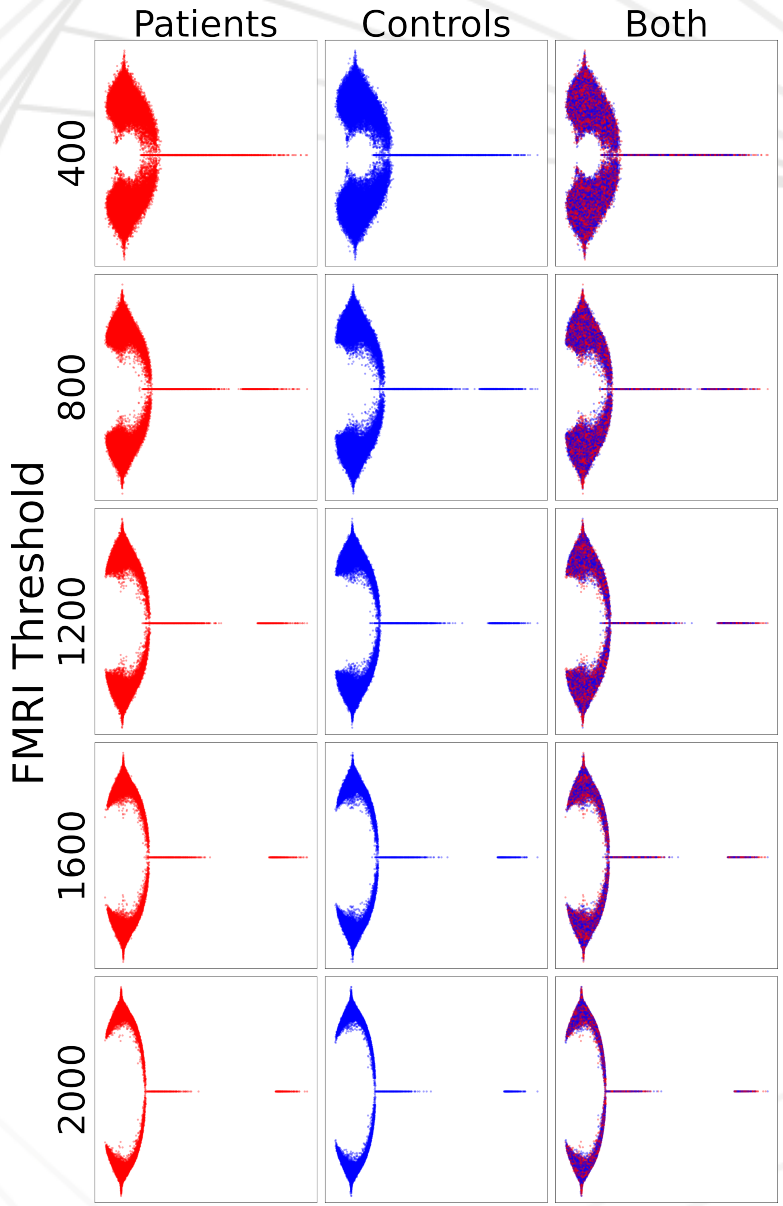


Figura: Distribución de los eigenvalores

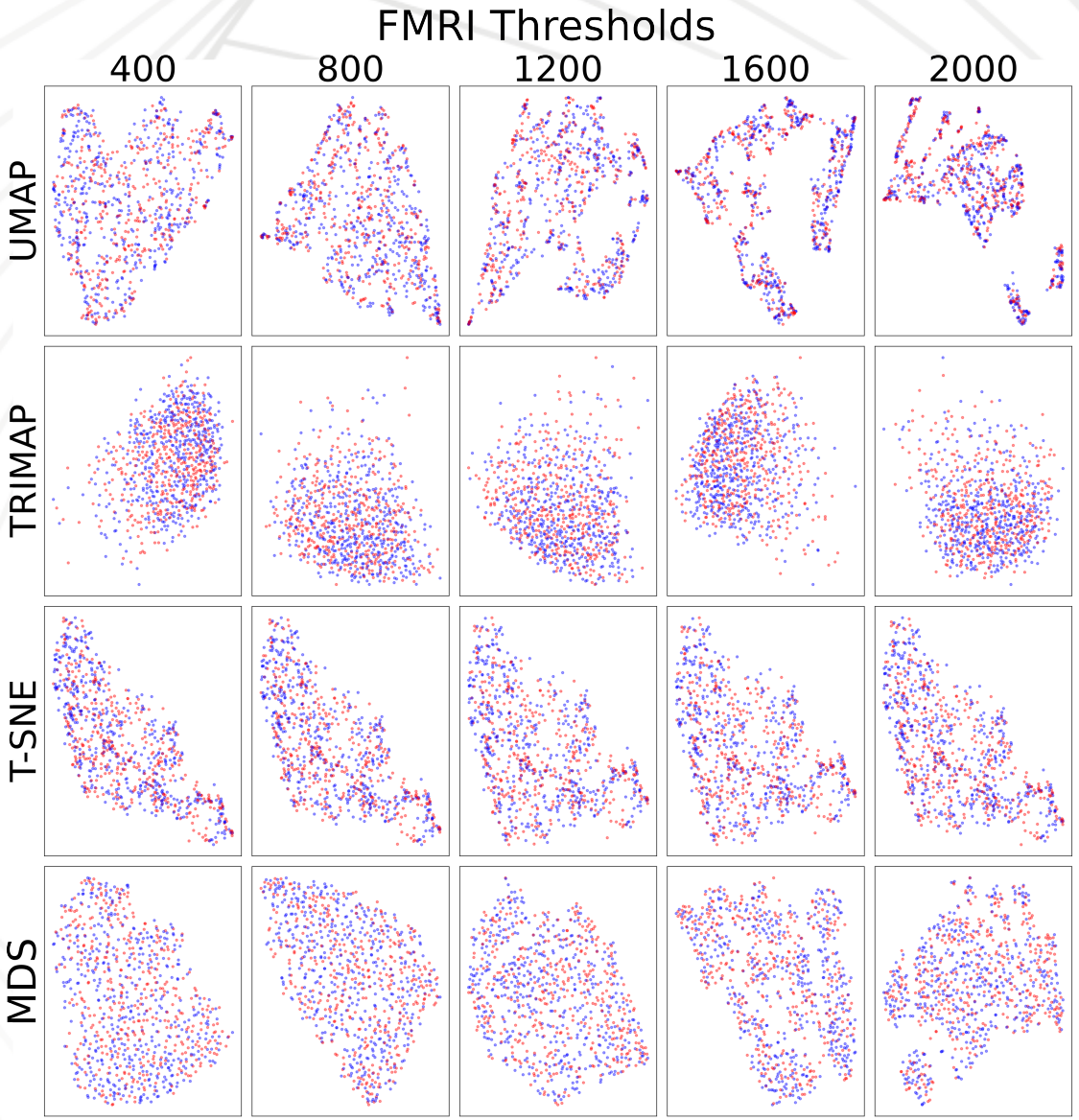


Figura: Diferentes métodos de aprendizaje de variedades usando la matriz de distancia de ciclos sin retroceso.

Bibliografía I

Ahn, Y-Y, et al. *Flavor network and the principles of food pairing.* Scientific reports 1 (2011): 196.

Achard, S. and Bullmore, Ed. *Efficiency and cost of economical brain functional networks.* PLoS computational biology, 3-2, 2007. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030017

Bendich, P., et al. *Persistent homology analysis of brain artery trees.* The annals of applied statistics 10.1 (2016): 198.

Bronstein, M., *Geometric foundations of Deep Learning*, Towards Data Science. https://towardsdatascience.com/geometric-foundations-of-deep-learning-94cdd45b451d

Bubenik, P. *Topology for Data Science 1: An Introduction to Topological Data Analysis.* Tercera Escuela de Análisis Topológico de Datos y Topología Estocástica ABACUS, Estado de México (2017).

Lum, P. Y., et al. *Extracting insights from the shape of complex data using topology.* Scientific reports 3 (2013): 1236.

Chazal, F., et al. *Gromov-Hausdorff stable signatures for shapes using persistence.* Computer Graphics Forum. Vol. 28. No. 5. Oxford, UK: Blackwell Publishing Ltd, 2009.

Craddock, C. and Benhajali, Y. and Chu, C. and Chouinard, F. and Evans, A. and Jakab, A. and Khundrakpam, B. Singh and Lewis, J.D. and Li, Q. and Milham, M. et al *The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives.* Frontiers in Neuroinformatics, 7 (2013). http://fcon_1000.projects.nitrc.org/indi/abide/

Constantine, David, and Jean-François Lafont. *Marked length rigidity for one-dimensional spaces.* Journal of Topology and Analysis 11.03 (2019): 585-621.

Bibliografía II

Dabaghian, Y., et al. *A topological paradigm for hippocampal spatial map formation using persistent homology.* PLOS Computational Topology, (2012): e1002581.https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002581

Eren, A. M. and Kiefl, E. et al *Community-led, integrated, reproducible multi-omics with anvi'o*. Nature microbiology, 6-1 (2021).

Guzmán-Sáenz, A., et al *Signal enrichment with strain-level resolution in metagenomes using topological data analysis*. BMC genomics, Springer, 20-2 (2019).

Jaretzki, Lennart. *Geometric deep learning approach to knot theory.* arXiv preprint arXiv:2305.16808 (2023).

McEwen, J. *A brief introduction to Geometric Deep Learning*, Towards Data Science. https://towardsdatascience.com/a-brief-introduction-to-geometric-deep-learning-dae114923ddb

Michael, B. *Introduction to Topological Data Analysis* http://bertrand.michel.perso.math.cnrs.fr/Enseignements/TDA/Etics-2021-compresse.pdf

Otter, N., et al. *A roadmap for the computation of persistent homology.* EPJ Data Science 6.1 (2017): 17.

Putrov, Pavel, and Song Jin Ri. *Graph Neural Networks and 3-Dimensional Topology.* arXiv preprint arXiv:2305.05966 (2023).

Singh, G., Mémoli, F. and Carlsson, G. *Topological methods for the analysis of high dimensional data sets and 3d object recognition.* Eurographics Symposium on Point-Based Graphics 2 (2007).

Rabadán, Raúl, and Andrew J. Blumberg. *Topological data analysis for genomics and evolution: topology in biology.* Cambridge University Press (2019).

Suárez-Serrato, P., Yazdani, M. *A first stab at Topological Gastronomy.* IPAM Cultural Analytics, White Papers, UCLA 2016.

Bibliografía III

Dey, T. K., and Yusu W.. *Computational topology for data analysis.* Cambridge University Press, 2022.

Torres, L. and Suárez-Serrato, P. and Eliassi-Rad, T. *Non-backtracking cycles: length spectrum theory and graph mining applications.* Applied Network Science, Springer. 4-1, 2019.

Torres, L. *SuNBEaM: Spectral Non-Backtracking Embedding And pseudo-Metric*. 2018. `https://github.com/leotrs/sunbeam`.

Xu, X., et al. *Finding cosmic voids and filament loops using topological data analysis.* Astronomy and Computing 27 (2019): 34-52.

Gracias