Topology of Social and Managerial Networks

Martina Scolamiero

Dipartimento di Ingegneria Gestionale e della Produzione

Politecnico di Torino

Commission: Production Systems

PHD in Production Systems and Industrial Design XXV cycle

Prof. Francesco Vaccarino

Acknowledgements

I am very grateful to my advisor, prof. Francesco Vaccarino that with his enthusiasm made this Ph.D really valuable and enjoyable for me.
Thanks to my collaborators: professor Wojciech Chacholski, Giovanni Petri, Antonio Patriarca and Irene Donato for their patience and their kindness in sharing ideas. I also have to thank professor Sandra di Rocco, professor Bernd Sturmfels for precious suggestions and comments in these years. A special acknowledgement goes to professor Mario Rasetti who is a continuos source of inspiration for me. Working at the ISI foundation has been a beautiful experience. I had a great time with my collegues and friends at ISI that I all thank. Thanks also to the professors and coordinator of the Phd in Production Systems and Industrial Design at Politecnico di Torino. My Ph.D was funded by a grant within the Lagrange project on Complex Systems of Fondazione CRT.

I also have to thank my parents and my sister Giulia for encouraging me in this experience. Last but not least thanks to Andrea for special early morning support.

Contents

Chapter 1

Introduction

With the explosion of innovative technologies in recent years, organizational and managerial networks have reached high levels of intricacy. These are one of the many complex systems consisting of a large number of highly interconnected heterogeneous agents. The dominant paradigm in the representation of intricate relations between agents and their evolution (94)(7) is a network (graph). The study of network properties, and their implications on dynamical processes, up to now mostly focused on locally defined quantities of nodes and edges. These methods grounded in statistical mechanics gave deep insight and explanations on real world phenomena; however there is a strong need for a more versatile approach which would rely on new topological methods either separately or in combination with the classical techniques.

In this thesis we approach this problem introducing new topological methods for network analysis relying on persistent homology (29),(32). The results gained by the new methods apply both to weighted and unweighted networks; showing that classical connectivity measures on managerial and societal networks can be very imprecise and extending them to weighted networks with the aim of uncovering regions of weak connectivity.

In the first two chapters of the thesis we introduce the main instruments that will be used in the subsequent chapters, namely basic techniques from network theory and persistent homology from the field of computational algebraic topology. The third chapter of the thesis approaches social and organizational networks studying their connectivity in relation to the concept of social capital. Many sociological theories such as the theory of structural holes (21),(23), and of weak ties (68),(69) relate social capital, in terms of profitable managerial strategies and the chance of rewarding opportunities, to the topology of the underlying social structure. We review the known connectivity

measures for social networks, stressing the fact that they are all local measures, calculated on a node's Ego network, i.e considering a nodes direct contacts. By analyzing real cases it, nevertheless, turns out that the above measures can be very imprecise for strategical individuals in social networks, revealing fake brokerage opportunities. We, therefore, propose a new set of measures, complementary to the existing ones and focused on detecting the position of links, rather than their density, therefore extending the standard approach to a mesoscopic one. Widening the view from considering direct neighbors to considering also non-direct ones, using the "neighbor filtration", we give a measure of height and weight for structural holes (25), obtaining a more accurate description of a node's strategical position within its contacts. We also provide a refined version of the network efficiency measure (83), which collects in a compact form the height of all structural holes. The methods are implemented and have been tested on real world organizational and managerial networks. In pursuing the objective of improving the existing methods we faced some technical difficulties which obliged us to develop new mathematical tools.

The fourth chapter of the thesis deals with the general problem of detecting structural holes in weighted networks. We introduce thereby the weight clique rank filtration, to detect particular non-local structures, akin to weighted structural holes within the link-weight network fabric, which are invisible to existing methods. Their properties divide weighted networks in two broad classes: one is characterized by small hierarchically nested holes, while the second displays larger and longer living inhomogeneities. These classes cannot be reduced to known local or quasi local network properties, because of the intrinsic non-locality of homology, and thus yield a new classification built on high order coordination patterns. Our results show that topology can provide novel insights relevant for many-body interactions in social and spatial networks, (107),(108).

In the fifth chapter of the thesis, we develop new insights in the mathematical setting underlying multipersistent homology (35),(104). More specifically we calculate combinatorial resolutions and efficient Gröbner bases for multipersistence homology modules. In this new frontier of persistent homology, filtrations are parametrized by multiple elements (28),(30). Using multipersistent homology temporal networks can be studied and the weight filtration and neighbor filtration can be combined.

Chapter 2

Network Basics

Networks are very simple mathematical objects, defined by a set of vertices connected by edges that, in their simplicity, have the power to represent a wide variety of systems. Numerous are the examples of computer networks, as the internet network, biological networks, as protein interaction networks and social networks, as managerial and intra-organizational networks.

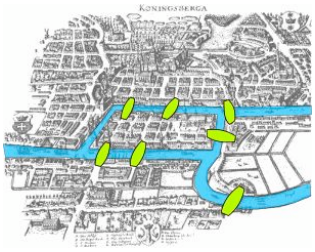
As network theory involves a wide range of disciplines: computer scientists, physicists, biologists, sociologists and mathematicians have over the years developed a rich set of tools to model and analyze networks to the aim of capturing their statistical properties (13),(4),(6),(7), (96),(52). The main approach from all disciplines, until recent years, was to deduce global properties of large systems and the evolution of dynamical processes over the system, from local properties, as degree distribution on clustering coefficient.

In this chapter we will introduce the basic concepts and definitions from network theory using the standard methods from statistical mechanics. Random networks introduced in section 2.4.2 are constructed to reproduce real world properties in a controlled way. For more information on network theory we remind the reader to (94),(95).

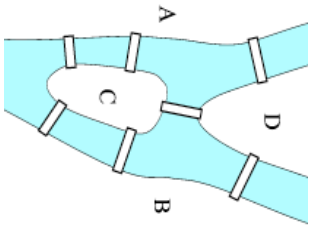
Vertices and edges in a network can be labelled with additional information such as the names for vertices, representing the identity of agents, of weights on links, representing the strength of an interaction between agents.
Our approach in this thesis is to give new insight on the connectivity of weighted and unweighted networks using methods based on geometry and algebraic topology, as will be explained in chapters 4 and 5.

2.1 Graph definitions

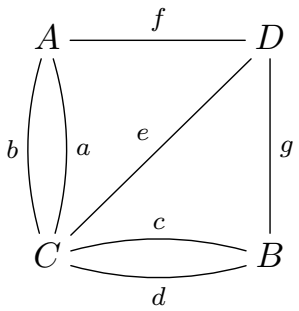
Graphs appeared for the first time in literature in 1735 in the paper "Solutio problematis ad geometriam situs pertinentis", written by Leonhard Euler (60). The question addressed in the paper was a practical question that the inhabitants of the city of Konisberg asked Euler. The city of Konisberg, in figure 2.1 (a) and (b), is divided in four parts by seven bridges, comprehending an islet surrounded by the river. The question was: *is it possible to trace a path that starting from one of the four areas traverses every bridge once and returns to the starting point?*



(a)



(b)



(c)

Figure 2.1: Konisberg city in figure (a), a schematization of the seven bridges in figure (b), a graph with nodes the four regions in the city of Konisberg and edges the bridges connecting the regions (c).

The abstraction with which Euler interpreted the problem, lead to a graph, 2.1 (c). In terms of the graph the problem was if it was possible to construct a closed path passing for each vertex in the graph once.
Graphs are also traceable in the fist steps of topology, Euler's formula relating the number of edges, vertices, and faces of a convex polyhedron was in fact studied and generalized by Cauchy (34) and L'Huillier (78) and sets the basis for the combinatorics

underlying simplicial homology, see section 3.3.

In chemistry there has been a large use and development of graph theory, the origin of a part of the standard terminology of graph theory in fact comes from this fusion of fields. In particular, the term graph was introduced by Sylvester in (110) where he draws an analogy between "quantic invariants" and "co-variants" of algebra and molecular diagrams. Famous problems in graph theory are mostly regard the combinatorics of graphs, we mention "The Four Color Problem" just to give an example. The introduction of probabilistic methods in graph theory, especially in the study of Erdos and Renyi of the asymptotic probability of graph connectivity, gave then rise to random graph theory, see subsection 2.4.2. Using methods from statistical mechanics and topology (lately) (119), graph theory has been successfully applied to the study of natural, economics and societal phenomena in the field of complex networks. We will now give the basic definitions about direct and indirect graphs.

Definition 2.1.1. *An indirect graph $G = (V, E)$ is defined by a set of vertices V and a set $E \subseteq (V \times V)/\sim$, of equivalence classes of pairs of vertices called edges. Where $\sim$ is the equivalence relation $(a, b) \sim (b, a)$ for a and b in V.*

The dimension of a graph is the number of its vertices. A vertex $a \in V$ is **connected** or **adjacent** to vertex $b \in V$ if there is an edge between them, i.e if $(a, b) \in E$. In this case a and b are called **neighbors**.
The number of neighbors of node v is the node's **degree**, denoted by k_v.

Definition 2.1.2. *A direct graph $G = (V, E)$ is defined by a set of vertices V and a set $E \subseteq (V \times V)$, of pairs of vertices called edges.*

In this case an edge from a to b is not identified with the edge from b to a. For directed graphs, we say a is a successor of b and b is a predecessor of a if the edge $e = (a, b)$ connects a to b. Shifting the attention from vertices to edges, we say that a is the source of the edge e, $a = s(e)$, and b is the target of the edge e, $b = t(e)$. In this case, the notion of degree is not well defined and we need to distinguish between the *in*-degree y_v, that is the number of predecessors of v and the *out*-degree z_v, the number of successors of v.
If in the definition of a direct 2.1.2 or indirect graph 2.1.1, we consider also the pairs (a, a) for $a \in V$ in the edge set, then the graph has **self loops**. If multiple edges are admitted between two vertices, the graph is called a **multigraph**. If we impose instead that at most one edge can connect two vertices, this is called a **simple graph**.

Definition 2.1.3. *A graph with a weight function $w : E \to \mathbb{R}$ defined on edges is called a weighted graph.*

A common way to represent a graph is through the **adjacency matrix**. This matrix shows which vertices are adjacent to which others. For unweighted N-dimensional graphs this is the $N \times N$ matrix $A = \{a_{i,j}\}$ with entries

$$a_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases}$$

The sum of elements in the rows (or equivalently in the columns) of the matrix gives the degrees of nodes in the graph.

Indeed, in weighted graphs, each edge carries a weight $w_{i,j}$, so that we can define $a_{i,j} = w_{i,j}$ if i and j are connected by and edge and $a_{i,j} = 0$ otherwise. For weighted graphs the adjacency matrix is most commonly called incidence matrix.

Remark 2.1.4. *The adjacency matrix of a graph (weighted or unweighted) is symmetric if and only if the graph is indirect.*

A **quiver** is a directed multigraph with self loops. This is a very general object and often in applications simpler mathematical objects are preferred. In this thesis we will deal mostly with simple undirected graphs, both weighted and unweighted. In particular, the topology of weighted graphs will be taken into account in chapter 5.
The following theorem due to Euler relates the number of vertices and edges in a simple graph.

Theorem 2.1.5. *For every simple graph $G = (V, E)$ we have:*

$$\sum_{v \in V} k_v = 2|E|. \tag{2.1.1}$$

Proof. Every edge (a, b) contributes two times in the sum on the right, once in k_a and once in k_b. □

Theorem 2.1.5 implies that the sum of the degrees of the vertices in the graph is even and that the number of vertices with odd degree is even. The version of this theorem for directed graphs is:

$$\sum_{v \in V} y_v = \sum_{v \in V} z_v = |E|. \tag{2.1.2}$$

We will now introduce the definition of morphism between graphs as function between the vertices that preserves the edge relations.

Definition 2.1.6. *Given two graphs $G_0 = (V_0, E_0)$ and $G_1 = (V_1, E_1)$ a graph map is a function $f : V_0 \to V_1$ such that the incidence relations are preserved. This means that if $(a, b) \in E_0$ then $(f(a), f(g)) \in E_1$.*

Remark 2.1.7. *The definition of simplicial map 3.2.4 that will be given in the next chapter, is a generalization of this definition 2.1.6 of graph map.*

Two graphs are said isomorphic if there is a bijective function between them. By the definition nodes in isomorphic graphs have the same degrees.

Example 2.1.8. *The two graphs in figure 2.2 are not isomorphic because in the first graph there are two vertices of degree three, namely 2 and 5, while in the second graph there are three vertices of degree three that are e, b and d.*

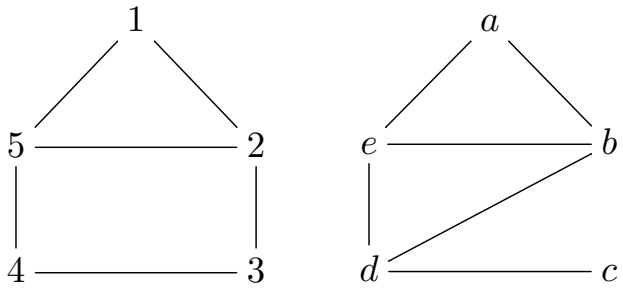


Figure 2.2: Two non isomorphic graphs

Being isomorphic is an equivalence relation on the set of all graphs. A fundamental branch of graph theory is the research of properties of graphs invariant for isomorphism. Isomorphism classes of graphs are anyway a fine classification; two real world networks for example are very rarely isomorphic but can share the sam type of degree distribution. In the following of this chapter we will go through some methods for distinguishing different types of networks and studying their structure.

Remark 2.1.9. *Note that in network theory vertices are often referred to as nodes and edges are referred to as links, both terms will be used in the following of this thesis.*

2.2 Connectivity

The topological notion of connectedness is translated in graph theory as follows: a graph is said to be **connected** if there is a path between any pair of nodes. In general graphs

are not connected, but they have several *connected components*, which are maximal connected subgraphs. All the connected components of G, provide a partition of the graph's nodes.
The natural generalization of an edge is a **path**, as the word suggests this is a way through the graph, paved by edges, connecting two non necessarily adjacent vertices in a connected component.

Definition 2.2.1. *A sequence of edges* $1, \ldots, n$ *such that* $t(i) = s(i+1)$ *for* $i \in \{1, \ldots, n-1\}$ *is a path from* $s(1)$ *to* $t(n)$ *we denote with* $\sigma_{1,n}$.

The length of a path $l(\sigma)$ is given by the number of edges in the path. A closed path, meaning that the first and the last vertex coincide, is called *loop*, for example a loop of length three is a *triangle*.
There can be many paths between two nodes but the shortest path measure is unique and defines a distance between the nodes on a network.

Definition 2.2.2. *The shortest path between two nodes* $s, t \in V$ *is the minimum length of a path connecting them.*

$$d(s,t) := min_{\sigma_{s,t}} l(\sigma_{s,t}).$$

Definition 2.2.3. *The diameter of a graph* G *is the maximum distance between two nodes in the graph*

$$D(G) := max_{(s,t)} d(s,t).$$

The diameter shows how compact the graph is, the smaller the diameter, more compact is the graph. A related measure is the average shortest path among all pairs of nodes. The most compact graph is a clique.

Definition 2.2.4. *The vertices* $v_0 \ldots v_n$ *compose a* $n+1$ *clique if and only if there is an edge between every pair of them.*

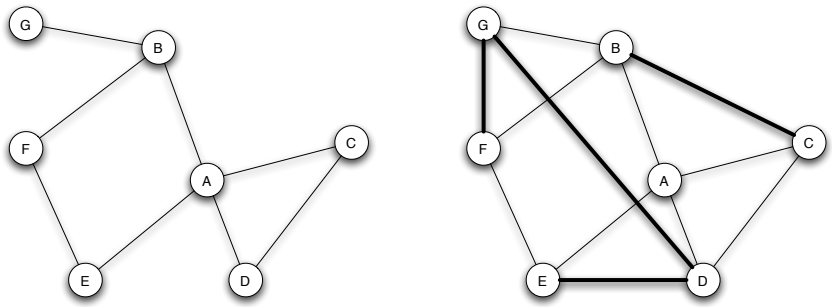
The definition is equivalent to saying that the nodes determine a complete subgraph. In a clique both the diameter and the shortest path between any two nodes in one. The diameter can also be defined as the maximum over all the eccentricities calculated for nodes in the network. The **small world** property is associated to networks with a small diameter (95).

Definition 2.2.5. *The eccentricity is the distance from a given node to the furthest node from it in the network:*

$$E(v) = max_s d(v,s).$$

In the study of the connections in a social network, a very studied property is **transitivity**, (94).
Many possible relations can be defined between couples of vertices of a graph, we will consider the relation "connected by an edge ". If this relation is transitive, the network is said to be transitive itself. This means that it is always the case that the friend of an individual is also his friend. By definition, in transitive networks every connected component is a clique, i.e a complete subgraph, this makes the topology of transitive networks trivial. Usually networks are not transitive but it is interesting to understand the level of transitivity of a network, that is to which extent, open triangles tend to close. In the evolution of a graph over time, in particular we are interested in the mechanisms by which nodes arrive and depart and by which edges form and vanish. For social networks it is very common that if two people have a common acquaintance, they know each other, (53),(68). This idea is stated in the **triadic closure principle**:

If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.



(a) *Before new edges form.* (b) *After new edges form.*

Figure 2.3: Effects of triadic closure of the connectivity of a network, (53)

The triadic closure principle was deeply investigated by M.Granovetter in the seminal paper " The Strength of Weak Ties", (68). In this study, weighted network are taken into account and the weights correspond to the strength of a relation between individuals. The triadic closure principle, in this case takes the form a the following hypothesis:

The stronger the tie between A and B, the larger the proportion of individuals to whom

they will be both tied, this is connected by a weak or a strong tie. This overlap in their friendship circles is predicted to be least when their tie is absent, most when it is strong, and intermediate when it is weak.

This hypothesis is based on the evidence that the stronger is a tie between two individuals, more similar they are in various ways. This evidence also supports the statement that if strong ties connect A to B and A to C, both B and C being similar to A are probably similar to one another increasing the likelihood of friendship (a link between them). Reasonings relating local properties to larger structures like the triadic closure, are the mile stone, in Granovetter's article, in the study of how information propagates between and within groups . Resulting in the evidence that weak ties (not frequent and superficial contacts) are the ones responsible for the spreading of information within the network.
The triadic closure can be quantitatively measured using the **clustering coefficient**,(94). We call **connected triple** , a triple of vertices with at least two edges connecting them; a closed connected triple is a triangle.

Definition 2.2.6. *The global clustering coefficient of a graph is*

$$C = \frac{3 \cdot \#(\ triangles)}{\#connected\ triples}.$$

This is an numerical invariant varying from 0 to 1. When $C = 1$, the graph is transitive on the other side if $C = 0$ then there are no triangles as for example in a tree or a square lattice.
In many real networks, especially social ones, the tendency of clustering can be clearly observed (121) as we will explain in chapter 4 section 4.3. For example the network of film actor collaborations has been found to have $C = 0.20$; a network of collaborations between biologists has been found to have $C = 0.09$, a network of e-mail communication has $C = 0.16$. These are typical values for social networks. Some denser networks have even higher values as 0.5 or 0.6.
Technological and biological networks by contrast tend to have somewhat lower values. The internet at the autonomus system level, for instance, has a clustering coefficient of only about 0.01.

The value of the clustering coefficient of a network must not be valued in an absolute sense but in comparison to the one of a graph in which nodes are not linked according to a rule but randomly (a null model). For simplicity, let's assume that the degree of every node is constant, $k_v = c$ for all $v \in V$ and that neighborhood relations are

completely random. In this network the clustering coefficient is approximately c/N, being $N = |V|$.

For the networks considered above, the value of c/N is 0.0003 for the film actor network, 0.00001 for the biology collaborations and 0.00002 for the e-mail network. Thus the measured clustering coefficients are much larger than this estimate based on the assumption of random network connections. This is the evidence of a behavior in social networks: there is much a greater chance that two people will be acquainted if they have another common friend than if they don't, as predicted by Granovetter in (68).

The **local clustering coefficient**, (94), is the local version of the clustering coefficient, introduced by Duncan J. Watts and Steven Strogatz in 1998 to determine if a network is a small-world (121). The clustering coefficient of node v is defined as the number of triangles which include v normalized dividing by the couple of vertices connected to v, that are $k_v \times (k_v - 1)/2$. This measure quantifies to which extend nodes tend to form links to the neighbors of their neighbors. The following is an equivalent definition:

Definition 2.2.7. *Let k_v the number of neighbors of $v \in V$ and e_v the number of links between them. The local clustering coefficient of v is*

$$C_v = \frac{e_v}{1/2(k_v(k_v - 1))}$$

In words this is the probability that a pair of neighbors of v are neighbors themselves. It is empirically shown that there is a rough dependence between local clustering coefficient and degree. Vertices with higher degree having a lower local clustering coefficient on average. The local clustering coefficient can also be used as an indicator of so called "structural holes" in a network, as we will see in chapter 4 section 4.3.

2.3 Centrality measures

Centrality measures address the problem of understanding which are the most important or central nodes in a network. Of course there are many possible definitions of importance and hence ways of defining a centrality measures. In this section we will go through some of the most used centrality measures on networks.

The most simple centrality measure of a node v, is given by its degree k_v. This measure is called **degree centrality** and mostly used in the study of social networks

as for example in (116), how we will see in chapter 4 subsection 4.2.2. In the degree centrality measure all neighbors are considered equivalent, each contributing with value one to the centrality measure. In real world networks, anyway some neighbors are more powerful or more connected and thus should contribute with a major value when measuring centrality. This feature is included in the **eigenvector centrality measure**.

The eigenvector centrality measure begins with an initial guess x_i of the centralities for each node $i \in V$. Then the values are added according to the incidences of vertices encoded in the adjacency matrix $A = \{a_{i,j}\}$, as in the following formula:

$$x'_i = \sum_j a_{i,j} x_j. \tag{2.3.1}$$

If for example the initial value for all nodes is one, then the initial eigenvector centrality measure coincides with the degree centrality. The formula 2.3.1 can also be written in matrix form as $x' = Ax$. Iterating the process, we obtain more accurate estimates for the centrality and the formula at the t-th iteration is $x(t) = A^t x(0)$. In the limit for $t \to \infty$ the limiting vector of centralities is proportional to the leading eigenvector k_1 of the adjacency matrix. Using this reasoning we can define the eigenvector centrality in the following way:

Definition 2.3.1. *Given a graph $G = (V, E)$ with adjacency matrix $A = \{a_{i,j}\}$ and dimension d, denoting with x a vector in $\mathbb{N}^d$ we define the eigenvector centrality as the vector that satisfies*

$$Ax = k_1 x.$$

This is the greatest eigenvector of the adjacency matrix.

The last two centrality measures we are going to introduce depend on the length of paths connecting nodes in the network. While the eccentricity measures the maximum distance from a node, the average distance from that node is calculated with the closeness centrality.

Definition 2.3.2. *The closeness centrality is the average distance from a given starting node to all the other nodes in the network.*

The node's centrality measure that is more significant for applications is anyway the betweenness centrality.

Definition 2.3.3. *Betweenness centrality of node v measures how often node v appears on the shortest path between nodes in the network.*

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$$

Betweenness centrality is evidently a measure of importance according to the centrality of a node in the flow of information in the context of a social network. Other centrality measures more specific to the field social networks will be introduced in section 4.3.

2.4 Relevant constructions

In this section we will report some fundamental graph constructions that will be used in the thesis. In particular the subgraphs are considered to represent and distinguish relevant properties of real world networks. Such characteristics are not traceable in the whole network because of the entanglement of links within the network. The random graphs are instead constructed to understand the relevance of the traced properties in comparison to null models.

2.4.1 Subgraph constructions

Definition 2.4.1. *Given a subset of vertices $S \subseteq V$, the subgraph induced by S is the graph with vertex set S and whose edges are all the edges of G which only connect elements of S.*

We will now introduce the subgraph constructions used in the following of the thesis to construct the neighbor and the weight clique rank filtration in subsection 3.4.1.

Definition 2.4.2. *Fixed a vertex $v \in V$, the closed neighborhood graph of v is the subgraph of G induced by v and its neighbors.*

We will denote the closed neighborhood graph of v in G with $N_G[v]$ or with $N[v]$ if the graph is obvious from the context. Iterating this construction we define the neighborhood of radius k.

Definition 2.4.3. *The closed neighborhood of radius k of node v in G, denoted by $N^k_G[v]$, is the subgraph of G induced by v and it's k-neighbors, i.e the nodes in G connected with v by a path of length $\leq k$.*

In the context of social networks, the closed neighborhood of a node is ofter referred to as it's **Ego network** and the closed neighborhood of radius k as the **radius k ego network**.

Remark 2.4.4. *It will be useful to consider the graph induced only by v's neighbors (not v itself). This is the so called open Ego network and denoted with $N_G(v)$. The radius k open Ego network is defined in analogy to definition 2.4.3.*

In weighted complex networks, filterings are commonly used to isolate statistically relevant structures and give a meaningful, even if reduced, representation of the network. Filtering techniques can be global, defined by a global threshold, or local as for the disparity filter method, (114).

The **global threshold method** is very simple, we choose a threshold parameter ω and consider the links in the network with weight $\leq \omega$. This is the graph induced by couples of nodes linked by edges of weight smaller or equal than the threshold parameter. This type of approach will not represent multi-scale patterns, which instead can be unearthed using the disparity filter method.

The **disparity filter method** offers a practical procedure to extract the relevant connectivity backbone in complex multi scale networks, preserving the edges that are significant with respect to a randomized model for the assignment of weights and connections. The weights in the null model (i.e the randomized model) associated to the edges incident to a certain node of degree k are produced by a random assignment from a uniform distribution. The links that are kept in the filtration are the ones with weight major than in the null hypothesis. This method permits to keep edges with low weight that anyway have a local relevance. In figure 2.4 we confront the number of nodes kept in the backbones with the global threshold method and the disparity filter method, as a function of the fraction of weight (left panels) and edges (right panels) retained by the filters. Note that he disparity filter reduces the number of edges significantly keeping a large number of nodes and variety of weights.

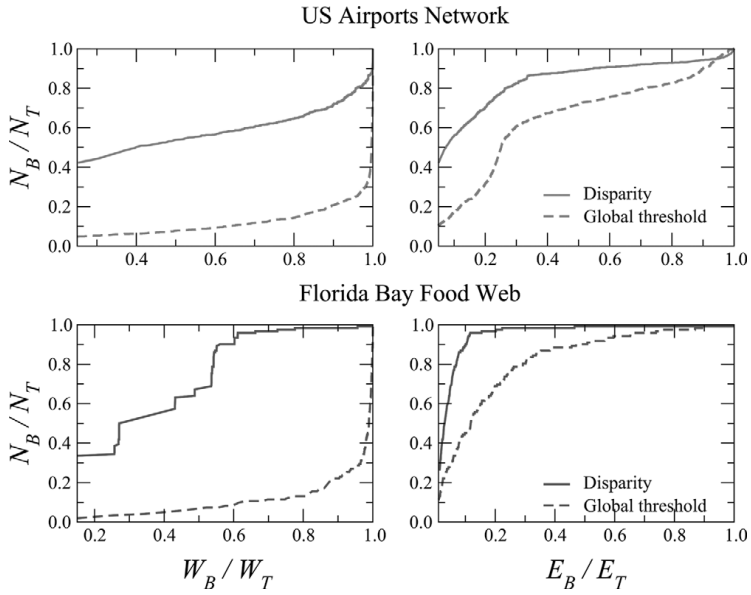


Figure 2.4: Comparison of the number of nodes preserved by global thresholding and disparity filter method in the US airport network and the Florida bay food web, in function of weights (left) and number of edges (right),(114).

2.4.2 Random graphs

For the disparity filter explained above, the topology of the graph is preserved while the weights of links are chosen randomly from a normal distribution. In this section we will explain methods to construct the topology of a network with some random process, maintaining some properties decided a priori.

Definition 2.4.5. *A random graph is a graph generated by a random process.*

We will begin this section with the simplest example of random graph.
Fixed a number n of nodes and a number m of edges, this model proposed by Edgar Gilbert and denoted by $\mathbf{G}(\mathbf{n}, \mathbf{m})$ is built by placing m edges within the nodes at random. Generally the graph is required to be simple, so the position of each edge should be chosen between the couples of nodes that are distinct and not already connected. This is equivalent to choosing m couples between the $\binom{n}{2}$ possible couples of vertices that can be considered. Another equivalent definition of the model is to say that the network is created by choosing uniformly at random among the set of all simple graphs with exactly n vertices and m edges. some properties of this model are very easy to calculate; just to give an example the average number of edges is obviously m and the average

degree is $< k >= 2m/n$. Other properties are less immediate to calculate and more work has been done of a slight variation of this model, namely the **Erdos Renyi model**, $\mathbf{G(n, p)}$.
The Erdos Renyi model is specified by two parameters: the number of vertices in the graph, n, and the probability of an edge, p. Fixed the two parameters n and p the Erdos Reny graph is built including an edge between each pair of vertices with probability p, independently from the chosen pair. Equivalently, $G(n, p)$ is an ensemble of graphs with n vertices, in which every single graph appears with probability

$$P(G) = p^m(1-p)^{\binom{n}{2}-m}, \tag{2.4.1}$$

where m is the number of vertices in the graph and we assume non simple graphs to have probability zero.

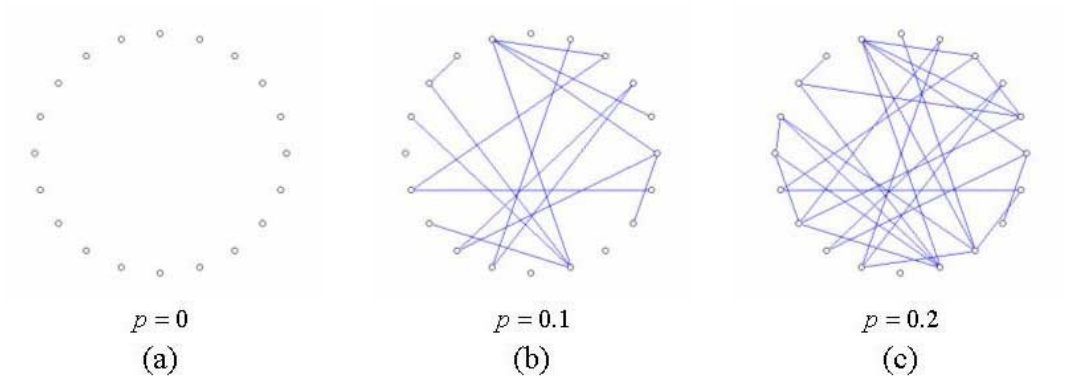


Figure 2.5: Erdos Renyi model construction for different values of the parameter p.

In this model, the probability of a graph to have m edges is

$$P(m) = \binom{\binom{n}{2}}{m} p^m(1-p)^{\binom{n}{2}-m}. \tag{2.4.2}$$

The probability of a graph with n vertices and m edges is in fact the probability of a graph with m edges that is $P(G)$ multiplied by the number of ways that the m edges can be chosen from the $\frac{n}{2}$ possible edges. The latter are combinations of class k form a set with $\frac{n}{2}$ elements, calculated with the binomial coefficient. From the formula 2.4.1 we recover the mean value of edges that is

$$< m >= \binom{n}{2} p. \tag{2.4.3}$$

Exploiting the formula for the mean degree written above for random graphs, we deduce that the mean degree of an Erdos Renyi graph is:

$$< k >= (n-1)p. \tag{2.4.4}$$

This expression completely follows intuition as it says that the mean number of edges linked to a node v is the probability of such node to be adjacent to another vertex multiplied by the number of vertices in the graph excluding v. With a similar reasoning it is possible to calculate the degree distribution in this model.
We have said that the probability of a node of being connected to one of the other $n-1$ nodes is p, the probability of being connected to k nodes is then $p^k \cdot (1-p)^{n-1-k}$. There are $\binom{n-1}{k}$ ways of choosing k neighbors from the possible $(n-1)$ and hence the total probability for a node to have degree k is given by the formula:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}. \tag{2.4.5}$$

A major interest is for large networks because these random models are constructed to reproduce properties of real world networks. A few calculations reveal that when n grows, $n \longrightarrow \infty$, the degree distribution tends to a be a Poissonian distribution

$$p_k \to e^{-c} \frac{c^k}{k!}. \tag{2.4.6}$$

Equation 2.4.6 is the reason why this model is sometimes called Poissonian random graph. How the size of components in the Erdos Renyi model change according to the parameters n and p has been deeply investigated .
The main results in this direction are:

- if the product $np < 1$, then a graph in $G(n,p)$ will almost surely have no connected components of size larger than $O(log(n))$,

- if $np = 1$, then a graph in $G(n,p)$ will almost surely have a largest component whose size is of order $n2/3$,

- if instead $np \to c > 1$, where c is a constant, then a graph in $G(n,p)$ will almost surely have a unique giant component containing a positive fraction of the vertices. No other component will contain more than $O(log(n))$ vertices.

In terms of p a sharp threshold for the connectedness of graphs in $G(n,p)$ is $p > \frac{\log_e(n)}{n}$.

Although the Erdos Renyi graph is one of the best studied models of networks, this model presents however some limitations in representing real world networks. The first problem that in this model is that for large n the clustering coefficient tends to zero whereas, how we have seen in the previous section, this is not the case in real world networks, especially social ones. This random graph also differs form real world networks because there is no relations between the degrees of adjacent vertices. The degrees in real world networks, by contrast, are usually correlated because of intrinsic coordination patterns. One other problem is that this random graph there is no community structure while in real world networks vertices group into communities, there are sets of highly connected vertices. The main problem anyway in modeling real structures with this random graph is anyway in the degree distribution. Most real world networks do not have a gaussian distribution but a power law have a power law distribution

$$p_k = Ck^{-\alpha}, \tag{2.4.7}$$

where α is a positive constant and C is a normalization constant.

To the aim of representing real world networks with major precision, the random graph models described up to now have been generalized so as to give the random network any degree distribution we please. We will only explain a few famous of such generalizations: the configuration model.

In a **Configuration Model** the degree sequence is fixed, hence it is possible to reproduce a power law distribution. This means that the number of neighbors of each node is known and in turn also the number of edges in the network, since the number of edges is given by $m = \frac{1}{2}\sum_i k_i$.
Given a graph $G = (V, E)$, the construction of the configuration model can be summarized in the following steps:

- specify the degree k_i of each node $i \in V$,
- give each vertex i a total number of k_i stubs of edges. This produces a total of $2m$ stubs where m is the total number of edges,
- choose two of the stubs uniformly at random and connect them to one another, creating an edge.

Iterating the last step until all the stubs are used up, we obtain a network in which every vertex has the desired degree.

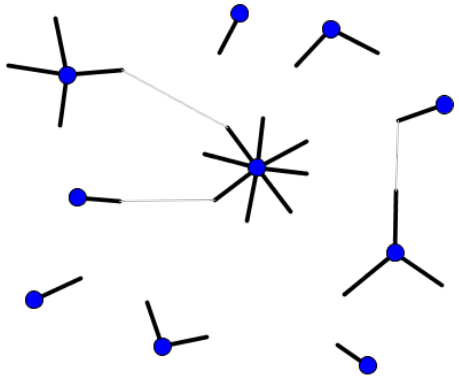


Figure 2.6: Construction of the configuration model.

Remark 2.4.6. *Note that for the configuration model the sum of all degrees must be a pair number otherwise the construction cannot work and that self loops are admitted in the construction rule.*

A further issue in the configuration model is that while all matchings of stabs are performed with uniform probability, not all networks appear with the same probability because different configurations can give rise to the same network.
The choice of a random rule for attachment is not very realistic, in real world networks higher coordination rules determine preferential attachments.

The **Preferential Attachment** model by Barabasi and Albert reflects this coordination property that is present in many real world networks (1),(2). The Barabasi Albert model is one of several proposed models that generates scale free networks. It incorporates two important general concepts: growth and preferential attachment. Both growth and preferential attachment exist widely in real networks. Preferential attachment means that a node in most probably linked to a node with a high degree. In a social context this can be translated as: the more friends I have, the more is the probability to make new friends. We will now concisely present the steps in the construction of this model:

- consider a network with N nodes, for $N \geq 2$, and such that the degree of each node in the initial network is at least one, if not the node will remain isolated,

- new nodes are added to the network one at a time. Each new node is connected to existing nodes with a probability that is proportional to the number of links that the existing nodes already have.

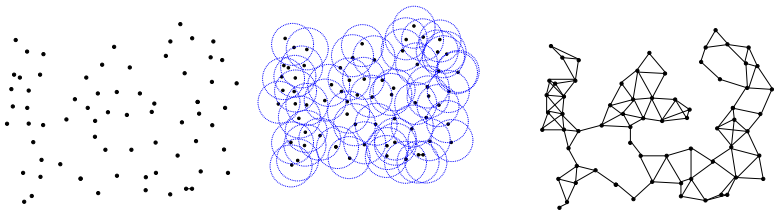
The probability p_i that node a node i is connected to a node of degree k_i is

$$p_i = \frac{p_i}{\sum_j p_j}. \tag{2.4.8}$$

The degree distribution of a Barabasi Albert model is a power law with exponent $\alpha = 3$. Note that differently from the previous cases, correlations between the degrees of connected nodes develop spontaneously. Although there is no direct calculation of the clustering coefficient in the Barabasi Albert model, the empirically determined clustering coefficients are generally significantly higher for this model than for random networks. The clustering coefficient also scales with network size following approximately a power law.

The last class of random graphs we are going to introduce are **Random Geometric Graphs**, $G(n, r)$. These graphs are used in chapter 5, subsection 5.4.1 to understand the classification of weighted networks in two broad classes according to their homological properties. M.Penrose, in (105) introduce this model that gives a way to construct a graph from a random set of points in a plane according to the mutual distance between couples of points. The parameter in the model are two, namely the number of points in the graph, n, and a radius $r(n) > 0$.
We will summarize this construction in the following steps:

- choose a sequence V_n of independent and uniformly distributed points in $[0, 1]^d$,
- given the radius parameter $r(n) > 0$, connect two points if their l_p distance is at most r.



(a) **(b)** **(c)**

Figure 2.7: Construction of the random geometric graph, (49)

Remark 2.4.7. *The RGG is a distance graph of n random points in the metric space* $[0, 1]^d$*, using* l_p *distance. Distance graphs are at the basis of the Rips-Vietoris complexes, as we will see is chapter 3 section 3.2.1.*

Chapter 3

Persistent Homology

The theory of persistent homology builds a bridge between computational algebraic topology and data analysis using homology as an effective tool to associate a computable invariant, the barcode, to a point cloud, (54),(55),(32), (56). The theory has a wide range of applications: classical ones varying from biological and medical data analysis (39), (46) to shape recognition or coverage problems in sensor networks, (31),(47). The main idea of persistent homology is to approximate a point cloud embedded in a metric space with an increasing sequence of simplicial complexes (filtration), see (29, 55). By analyzing the persistent, i.e. long living, homological features in the filtration, the shape of the point cloud can be approximatelly inferred (38). In this chapter we will introduce the standard construction of persistent homology and two new filtrations built on networks. In the following of the thesis we will in fact develop two methods of applying persistent homology to complex networks to the aim of understanding the topology of weighted and unweighted networks. Different applications of persistent homology to networks can be found in (57), (77), (84), (51).

3.1 Homology of a Chain Complex

In this section we will give the very general definition of homology. This will be specialized to simplicial homology in section 3.3. For background notions on homological algebra, we refer the reader to (122), (82).

We recall that a **ring** is a triple $(A, +, \cdot)$ consisting of a set A, two binary operations: addition$+ : A \times A \to A\,, (a, b) \to a + b$; and multiplication $\cdot : A \times A \to A\,, (a, b) \to ab$,

such that $(A, +)$ is an abelian group, with zero element 0 , and the following conditions are satisfied:

- $(ab)c = a(bc)$,
- $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$

for all a, b, $c \in A$

The conditions tell us that multiplication is associative and both right and left distributive over addition. We will consider commutative rings with identity, i.e $ab = ba$ for all $a, b \in A$ and $1 \in A$.
Fixed a field k, we can define a **k-algebra**; this is a ring A, such that A has a k-vector space structure compatible with the multiplication of the ring, that is, such that

$$\lambda(ab) = (a\lambda)b = a(\lambda b) = (ab)\lambda \qquad \forall \lambda \in K, \text{and } a, b \in A \tag{3.1.1}$$

Definition 3.1.1. *Let A be a k-algebra. A right A-**module** is a pair $(M, \cdot)$, where M is a k-vector space and $\cdot : M \times A \to M, \quad (m, a) \to ma$, is a binary operation satisfying the following conditions:*

- $(x + y)a = xa + ya$;
- $x(a + b) = xa + xb$;
- $x(ab) = (xa)b$;
- $x1 = x$;
- $(x\lambda)a = x(a\lambda) = (xa)\lambda$

for all $x, y \in M \quad a, b \in A$ and $\lambda \in K$.

For us an A-module we will mean a right A-module.
If M and N are A-modules, then a map $\phi : M \to N$ is a homomorphism of A-modules if, for any m, $n \in M$ and r, $s \in A$, $f(rm + sn) = rf(m) + sf(n)$. This, like any homomorphism of mathematical objects, is just a mapping which preserves the structure of the objects. We will denote the **category of A-modules and A-module homomorphisms** with $mod(A)$.

Definition 3.1.2. *A **chain complex** $\{C, \partial\}$ of A-modules is a family $\{C_n\}_{n\in\mathbb{Z}}$ of A-modules, together with left A-module maps $\partial = \partial_n : C_n \to C_{n-1}$ such that each composite $\partial \circ \partial : C_n \to C_{n-2}$ is zero.*

The maps ∂_n are called the differentials of C. Sometimes we will refer to $\{C, \partial\}$ with C for short.

Remark 3.1.3. *Note that by definition* $Im(\partial_{n+1}) \subseteq \ker(\partial_n)$.

A morphisms between two chain complexes is defined in the following way:

Definition 3.1.4. *A* **chain map** f *between two chain complexes* $\{C, \delta\}$ *and* $\{D, \varphi\}$ *is given by a family of module homomorphisms* $f_n : C_n \to D_n$ *that commute with the differentials* $f_n \circ \delta_{n+1} = \varphi_{n+1} \circ f_{n+1}$ *for all* n.

Chain maps can be composed element-wise, $(f \circ g)_n := f_n \circ g_n$. It is easy to check that composition is associative and has the identity element $1_n = id$. This defines the **category** Ch_A **of chain complexes of** A**-modules and chain maps** between them. Given a chain complex, the first question is usually about exactness: one wants to check if the kernel of the differential at one step coincides with the image of the differential at the previous step. If this is not the case, the homology of the chain complex measures the lack of exactness.

Fixed a chain complex C the kernel of $\partial_n : C_n \longrightarrow C_{n-1}$ is the module of n**-cycles** of C, denoted $Z_n = Z_n(C)$; the image of $\partial_{n+1} : C_{n+1} \longrightarrow C_n$ is the module of n**-boundaries** of C, denoted $B_n = B_n(C)$.

Definition 3.1.5. *The* n**-th homology** *module of the chain complex* C *is the quotient*

$$H_n(C) = Z_n / B_n.$$

The quotient is well defined because $B_n \subseteq Z_n$ by remark 3.1.3.
The homology class of a cycle $z \in Z_n$ will be denoted by $\overline{z} \in H_n$.

Definition 3.1.6. *A chain complex* C *is exact if* $H_n(C) = 0$ *for all* n.

The image through the chain map $f : C \to D$ of a cycle in C is a cycle in D, the same is true for boundaries. This property ensures that the following homomorphism is well defined.

$$\begin{array}{rcl} H_n(f) : H_n(C) & \longrightarrow & H_n(D) \\ \overline{z} & \mapsto & [f(z)] \end{array}$$

The homomorphism $H_n(f)$, induced in homology by f shares the properties:

- if $f\,, g : C \to D$, then $H_n(f \circ g) = H_n(f) \circ H_n(g)$,

- $H_n(id_C) = id_D$.

We can then claim that the n-th homology is a covariant functor from the category of chain complexes of A-modules to the category of A-modules.

$$\begin{aligned} H_n : Ch_A &\longrightarrow mod(A) \\ C &\mapsto H_n(C) \\ f : C \to D &\mapsto H_n(f) : H_n(C) \to H_n(D). \end{aligned}$$

3.1.1 Homotopy Invariance

Given two chain maps f, $g : (C, \delta) \to (D, \varphi)$, a **chain homotopy** from f to g is a chain map $h : (C, \delta) \to (D, \varphi)$ such that :

$$h(C_n) \subset D_{n+1} \quad \text{and} \quad f - g = \varphi h + h\delta. \tag{3.1.2}$$

If there is a chain homotopy between two chain maps f and g, we will write $f \simeq g$. The relation of chain homotopy on chain maps is an equivalence relation:

- reflexive, $0 : f \simeq f$

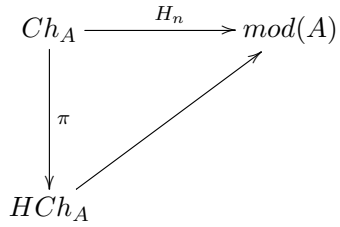
- symmetric, if $h : f \simeq g$, then $-h : f \simeq g$.

- transitive, if $h : e \simeq f$ and $l : f \simeq b$, then $h + l : e \simeq g$.

We denote with $[f]$ the equivalence class of a chain map. The homotopy relation is compatible with composition.

Proposition 3.1.7. *Given three chain complexes $C\,, D$, E and chain maps $f, g : C \to D$ and $f', g' : D \to E$ if $f \simeq g$ and $f' \simeq g'$ then $f'f \simeq g'g$.*

Thanks to this proposition we can define a law of composition for homotopy classes of chain maps as: $[f'] \cdot [f] := [f' \cdot f]$. The homotopy classes of chain maps are associative by definition, and have an identity element $[id]$. A new category is defined by chain complexes and homotopy classes of chain maps, denoted by HCh_A. We will now prove that the homology functor factors through HCh_A



Proposition 3.1.8. *If $f \simeq g : C \to D$ are two homotopically equivalent chain maps, then $H_n(f) = H_n(g)$ for all n.*

Proof. It is a simple calculation to show that the two functions coincide on every element, $H_n(f)([z]) = H_n(g)([z])$ for all $z \in Z_n \subset C_n$. By definition of map induced in homology $H_n(f)[z] - H_n(g)[z] = [f(z)] - [g(z)] = [(f-g)(z)]$. Because f and g are homotopically equivalent through h, then $f - g = \varphi h + h\delta$ and $[(f-g)(z)] = [\varphi h(z) + h\delta(z)]$. But z is a cycle in C, thus $[h\delta(z)] = 0$ and $[(f-g)(z)] = [\varphi h(z)] = 0$ being $\varphi h(z)$ a boundary in D. □

3.2 Simplicial Complexes

An **abstract simplicial complex** is a non empty family X of finite subsets, called simplices, of a vertex set with the two constraints:

- a subset of a face in X is a face in X,

- the intersection of any two simplices in X is a face of both.

We assume that the vertex set V of the simplicial complex is finite and totally ordered. A simplex of $n+1$ vertices is called $n-$face and denoted by $[p_0, \ldots, p_n]$. The dimension of a simplicial complex is the highest dimension of the simplices in the complex.
Every abstract simplicial complex X has an associated topological space $|X|$ called the **geometric realization**.
If X is a finite simplicial complex, to define its geometric realization we have to choose an embedding of the vertex set V as an affinely indipendent set of points in some eucledian space $\mathbb{R}^M$ of sufficiently high dimension.

Definition 3.2.1. *A set of $n+1$ points $(p_0 \ldots p_n)$ in $\mathbb{R}^M$ is affinely independent if no hyperplane of dimension $n-1$ contains all the points.*

In this definition, we require two points to be distinguishable, three points not to be aligned, four points not to lie on a plane and so on. Every simplex in X can then be

identified with the convex hull of the corresponding embedded vertices, called geometric simplex.

Definition 3.2.2. *The convex hull of the affinely independent set of points* $(p_0 \dots p_n)$ *in* $\mathbb{R}^M$ *is the set of points* $x \in \mathbb{R}^M$ *such that there exist non negative numbers* $\lambda_0 \dots \lambda_n$ *with the property:*

$$x = \sum_{i=0}^{n} \lambda_i p_i \quad \textit{and} \quad \sum_{i=0}^{n} \lambda_i = 1$$

Figure 3.1: Low dimensional geometric simplices.

The geometric realization $|X|$ of the simplicial complex X is the union of all the geometric simplices associated to X. In particular a $0-$face is a vertex, a $1-$face is a segment, a $2-$face is a full triangle, a $3-$face is a full tetrahedron. We identify a simplicial complex with it's geometric realization.

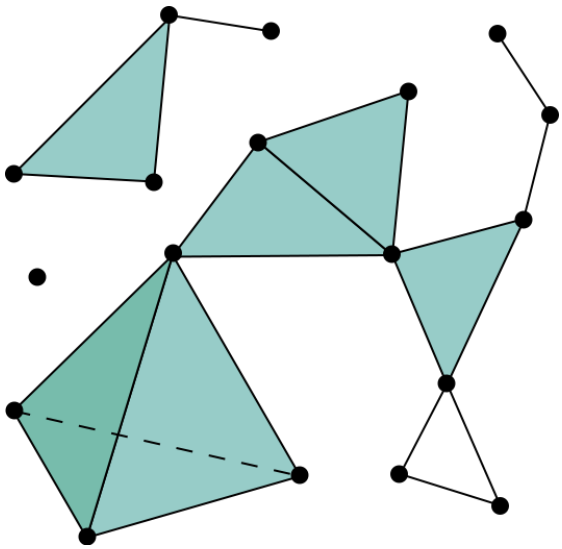


Figure 3.2: A three dimensional simplicial complex.

For the computation of simplicial homology, it is necessary to give an orientation to a simplicial complex. An oriented $n-$simplex is a simplex with an ordering on its vertices. Pair permutations of the order give the same orientation, odd permutations

instead reverse the order. A simplicial complex whose simplices are oriented is called an **oriented simplicial complex**.
A common method to orient a simplicial complex is to choose an orientation for all its vertices and consider the induces orientation on the simplices in the complex.

Remark 3.2.3. *Different orientations of a simplcial complex to not change the simplicial homology modules.*

Morphisms between simplicial complexes are called **simplicial maps** and share the property that the image of a vertex is a vertex and the image of a $n-$face is face of dimension $\leq n$.

Definition 3.2.4. *A morphism $\phi : K1 \to K2$ between simplicial complexes is called a simplicial map if it sends the vertices of $K1$ to the vertices of $K2$, if given the simplex $[p_0 \dots p_n]$ in $K1$ the elements $\phi(p_i)$ are vertices of a simplex in $K2$, and on the corresponding simplices*

$$\phi(\sum t_i p_i) = \sum t_i(\phi(p_i)) \qquad t_i \in \mathbb{R}.$$

Simplicial complexes and simplicial maps determine a category that we will denote by SC.

We conclude with the fundamental definition of triangulable topological space.

Definition 3.2.5. *A triangulable topological space M is a simplicial complex X, homeomorphic to M, together with an explicit homeomorphism $h : X \to M$.*

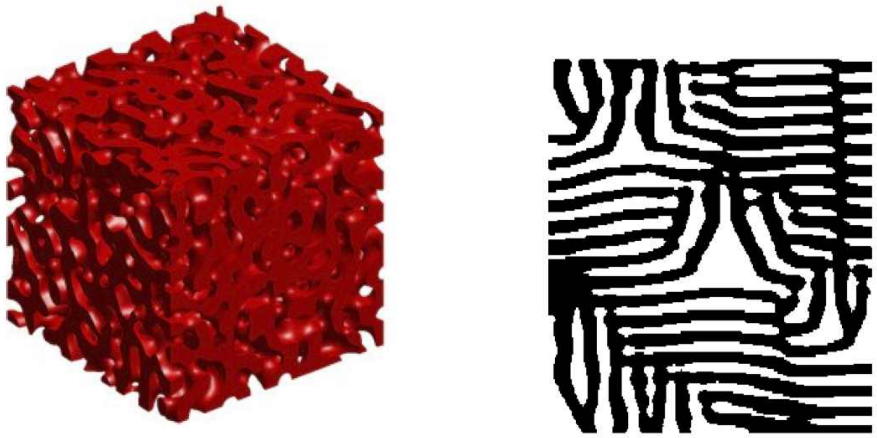
Examples of triangulable topological spaces are two dimensional surfaces, e.g. torus, projective plane, 2-sphere etc.. Also three dimensional topological varieties and differentiable varieties admit a triangulation. A necessary condition for a topological space to be triangulable is begin Hausdorff and metrizable. There are examples of varieties of dimension four that are not triangulable. As we will see, a triangulation is useful in determining the properties of a topological space. In particular we will be able to compute the homology of a triangulated space using simplicial homology.

We underline from now that the idea of persistent homology is to sample a set of points, possibly with rumor, from an underlying geometric space with the aim to reconstruct the topology of the space through a simplicial complex built from the points. The method is then not to triangulate a space but to approximate the space from one

or a sequence of triangulations.

3.2.1 Constructions from data sets and graphs

There are many ways to **construct simpicial complexes** from a graph or points in the Eucledian space $\mathbb{R}^M$, we will now go through some of them. The set of points in applications is a point cloud: a dataset, possibly sampled with error. Approximating the point cloud with a higher dimensional object, e.g a simplicial complex, it is possible to understand its shape distinguishing from the two settings in figure 3.4. Graphs are



(a) **(b)**

Figure 3.3: Point clouds with different shape.

instead the backbone of complex networks, see chapter 2. Associating a higher geometrical object to a graph, e.g a simplicial complex, proved to be highly informative about many body interactions within the network and other connectivity properties.

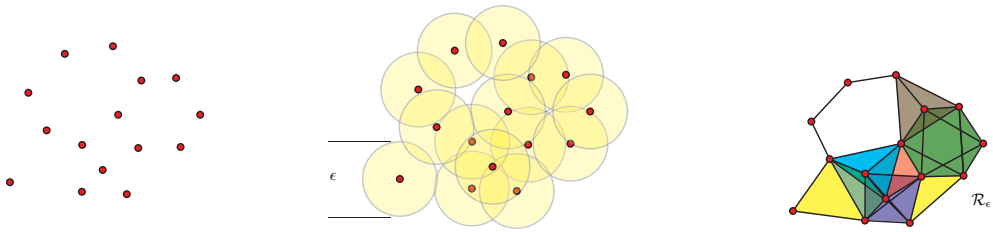
Clique complex

Fixed a graph G,we recall that a clique is a complete subgraph of G, see 2.2.4. The **clique complex** is a simplicial complex constructed from a graph. There is a $n-$face in the simplicial complex for every $(n+1)-$clique in the graph, the compatibility conditions between simplices are satisfied because subsets of cliques and intersection of cliques are cliques themselves.

Rips-Vietoris complex

The most popular simplicial complex for data analysis is the Rips-Vietoris complex, (29). The Rips-Vietoris complex is a simplicial complex associated to a set of points S in a metric space in the following way:

- every point $p \in S$ is the center of a radius ϵ ball $D(p, \epsilon)$
- $n+1$ points $\{p_0, \ldots, p_n\}$ determine a $n-$face if the corresponding radius ϵ balls intersect two by two, i.e $D(p_i, \epsilon) \cap D(p_j, \epsilon) \neq \emptyset$ for all $i \neq j \in \{0 \ldots n\}$.



(a) **(b)** **(c)**

Figure 3.4: Rips-Vietoris complex construction: point cloud in figure (a), ϵ balls centered in the points in figure (b), simplicial complex in figure (c).

Remark 3.2.6. *The Rips-Vietoris complex is the clique complex of a graph, namely the ϵ-neighborhood graph. Fixed the set of points S and the parameter ϵ, the associated neighborhood graph has S as vertex set; there is an edge between two vertices x and y if $d(x, y) \leq \epsilon$, being $d(x, y)$ the distance between x and y in $\mathbb{R}^M$.*

Cech complex

The Cech complex is a **nerve comple**x, we will thus give the definition of nerve complex before.
Let X be a topological space, and let $U = \{U_v\}_v$ be an open cover of X, this is a family of subsets U_v of X whose union is the whole X. The nerve of the covering U is a simplicial complex defined as follows:

- there is one vertex for each element U_v of the cover,
- there is one edge for each pair U_v and U_w in the cover such that $U_v \cap U_w \neq \emptyset$,

- in general, there is one n-simplex for each $n+1$-element subset $\{U_0, \dots U_n\}$ for which $U_0 \cap U_1 \dots U_n \neq$.

Geometrically, the nerve is essentially a dual complex for the covering.
Going back to our setting of a finite set of points $S \subset \mathbb{R}^M$ let's consider, as in the Rips-Vietoris complex, radius ϵ open balls around every point $p \in S$. The balls $D(\epsilon, p)$ for $p \in S$ are an open covering of S. The Cech complex $C_\epsilon(S)$ of the point set S with parameter ϵ, is the nerve of the cover of S given by ϵ-balls.
The construction is the following:

- the vertices of $C_\epsilon(S)$ are the points in S,
- the $n+1$ points $p_0, \dots, p_n$ generate a n-simplex if:

$$D(p_0, \epsilon) \cap D(p_1, \epsilon) \dots D(p_n, \epsilon) \neq \emptyset.$$

If the cover of the space is good, that is if the cover sets and all nonempty finite intersections of covers sets are contractible, then the Cech complex captures the topology of the cover from which it is defined, (47).

Theorem 3.2.7. *The Cech complex of a good cover has the homotopy type of the union of the cover sets.*

The Cech complex captures the topology of the cover from which it is defined because a union of intersection open balls is contractible just as a simplex. The computational complexity in calculating the intersection of open balls in the definition of Cech complex makes this complex not popular in applications.
The Rips-Vietoris complex is not homtopically equivalent to the Cech complex and thus to the open cover.

Example 3.2.8. *Let us consider the vertices P_0, P_1 and P_2 an equilateral triangle and ϵ greater than half of the side length but smaller than the circumcircle radius, then the three points generate a full triangle with the Rips-Vietrois complex and its boundary with the Cech complex.*

The Rips Vietoris complex is anyway easier to calculate. Using the following theorem, in (124) that relates the Rips-Vietoris complex to the Cech complex; the latter one is usually approximated with the Rips-Vietoris complex in computations.

Theorem 3.2.9. *Let S be a set of points in R^M and $C_\epsilon(X)$ the Cech complex of the cover of X by balls of radius $\epsilon/2$. Then there is chain of inclusions*

$$RV_{\epsilon'}(X) \subset C_\epsilon(X) \subset R_\epsilon(X) \text{ whenever } \frac{\epsilon}{\epsilon'} \geq \frac{2d}{d+1} \tag{3.2.1}$$

3.3 Simplicial Homology

We will now briefly introduce simplicial homology with coefficients in an arbitrary ring A, and the specify to the case $A = \mathbb{Z}$ and $A = k$; a general treatment of the subject and more algebraic topology can be found in (74), (88), (92), (93).

Important topological information about simplicial complexes such as number of connected components and holes is given by simplicial homology. Fixed an oriented simplicial complex X of dimension d, the sets X_n of oriented n-simplices in X are linked by the set maps:

$$\begin{array}{rcll} d_i : X_n & \longrightarrow & X_{n-1} & 0 \leq i \leq n \\ [p_0, \ldots, p_n] & \to & [p_0, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n]. & \end{array}$$

The free A-modules C_n on the sets X_n are called n-chain modules. The set maps d_i yield module maps $C_n \to C_{n-1}$ which we also call d_i.
The morphism $\partial_i := \sum_{i=0}^{n} (-1)^i d_i$ defined sending an $n-$face to the alternate sum of it's $(n-1)$-faces and then extended by linearity, shares the property

$$\partial_{n-1} \circ \partial_n = 0.$$

This implies that the modules C_n and the differential operator ∂_n define the simplicial chain complex.

Definition 3.3.1. *The simplicial chain complex of X with coefficients in A is the chain complex*

$$C_X : \qquad 0 \to C_d \overset{\partial_d}{\to} C_{d-1} \overset{\partial_{d-1}}{\to} \cdots \to C_1 \overset{\partial_1}{\to} C_0 \to 0$$

The assignment $\mathcal{X} \to C_{\mathcal{X}}$ induces a functor from SC to Ch_A.
The simplicial chain complex C_X, in turn, determines the simplicial homology modules of the simplicial complex X.

Definition 3.3.2. *The n-th simplicial homology of X with coefficients in A, is the n-th homology module of the chain complex C_X.*

The rank of the module $H_n(X)$ is called n-th Betti number of the simplicial complex. The first Betti numbers of X have an easy intuitive meaning: the 0-th Betti number is the number of connected components of X, the first Betti number is the number of two dimensional (poligonal) holes, the third Betti number is the number of

three dimensional holes (convex polyhedron).

In terms of categories, simplicial homology is the restriction of the functor H_n to the category simplicial chain complexes. It is fundamental to note that by composing the two functors $SC \stackrel{C}{\to} Ch_A$ and $Ch_A \stackrel{H_n}{\to} mod(A)$ we can view simplicial homology as a functor from SC to $mod(A)$ implying the following fundamental proposition.

Proposition 3.3.3. *Let X and Y be two simplicial complexes, a simplicial map $f : X \to Y$ determines a homomorphism map between the homology modules $H_n(f) : H_n(X) \to H_n(Y)$ for all n.*

Usually homology modules are computed with coefficients in the ring of integers $\mathbb{Z}$. A $\mathbb{Z}$-module is a group, and this is the reason for which it is common to refer to homology modules as homology groups. In applications instead homology is usually computed with coefficients in a field k, for computational reasons. It is possible to pass from one ring of coefficients to another using tensor products.

Definition 3.3.4. *Let A be a ring, M a right A-module, and N a left A-module. The tensor product of $M \otimes_A N$ of the two modules over A can be defined as the A-module, generated by $a \otimes b$ with relations:*

$$\begin{aligned} (a_1 + a_2) \otimes b &= a_1 \otimes b + a_2 \otimes b \\ a \otimes (b_1 + b_2) &= a \otimes b_1 + a \otimes b_2. \\ ar \otimes n &= m \otimes rn \end{aligned}$$

for $a \in M$, $b \in N$ and $r \in A$

If we are considering the simplicial homology of a simplicial complex, with coefficients A, and there is a homomorphism $\phi : A \to B$, it is possible to calculate simplicial homology of that complex using coefficients in B using tensor products.
The ring B has in fact a A module structure given by $b \cdot a := \phi(b) \cdot a$ for all $a \in A$ and $b \in B$. If (C, ∂) is a simplicial chain complex with coefficients in A, the chain complex $(C \otimes_A B, \partial_B)$ gives simplicial homology with coefficients in B. The modules in the chain complex $C \otimes_A B$ are $(C \otimes_A B)_i := C_i \otimes B$; the differentials are defined as $\partial_B(c_n \otimes b) := \partial(c_n) \otimes b$.

3.4 Persistent Homology

The starting point in persistent homology is a **filtration**. A filtration is an increasing sequence of simplicial complexes that give a multi scale representation of an underlying space, this can be a point cloud or network. Persistent homology consists in computing the simplicial homology of the spaces in a filtration and comparing them with the maps induced in homology by the inclusions. We can claim that persistent homology is the homology of a filtration. The idea is to identify features (homology generators) that are in the simplicial complexes for many values of the parameter. Such features are called persistent features and identified with connectivity properties for the point cloud under study, (66). For networks, the interpretation of persistent features is different and in fact strongly depends on the filtration under study. For example we use a filtration for weighted networks to distinguish between scale properties of real world networks and features persisting at all scales.
We will now go through one filtration for a point cloud based on the Rips-Vietoris complex and three filtrations on networks based on the clique complex that are exploited in this thesis.

3.4.1 Filtrations

Most of the constructions, explained in subsection 3.2.1, depend on a parameter. As the parameter grows, simplices are added to the simplicial complex; determining an increasing sequence of simplicial complexes, a filtration.

Definition 3.4.1. *A simplicial complex X is filtered if we are given a family of subspaces $\{X_v\}$ parametrized by $\mathbb{N}$, such that $X_v \subseteq X_w$ whenever $v \leq w$. The family $\{X_v\}$ is called a filtration.*

Rips-Vietoris Filtration

The Rips-Vietoris complex depends on the radius ϵ of the balls associated to points in the point cloud. If $\epsilon_1 < \epsilon_2$ the complex with ϵ_1 radius balls is contained in the complex with ϵ_2 radius balls. To the growth of ϵ we obtain an increasing sequence of simplicial complexes, a filtration, the Rips-Vietoris filtration. In this context persistent topological features of the filtration are considered as features of the point cloud.

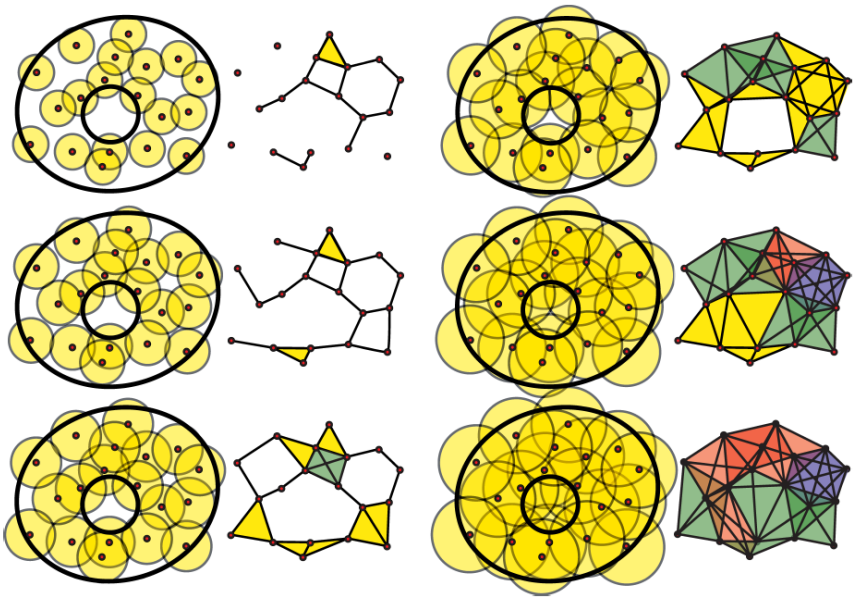


Figure 3.5: Rips Vietoris filtration.

Clique filtration

For unweighted networks, the clique filtration is used in (77) to analyze the difference between the connectivity properties of random networks, networks with exponential connectivity distribution and scale-free networks.

Definition 3.4.2. *The $k-$skeleton X_k of a simplicial complex X is the subcomplex of X containing all the faces of dimension smaller or equal to k.*

Consider a complex network and the corresponding clique complex X, the clique filtration is obtained by filtering the clique complex according to the dimension of the skeleton:

$$X_0 \subseteq X_1 \subseteq X_2 \subseteq \ldots \subseteq X.$$

Note that persistent features of the clique filtration are generators of the homology groups of the clique complex. These generators can be directly calculated from the clique complex of the graph, thus the filtration gives no extra information. This is not the case for the following neighbor filtration and weight clique rank filtration that we have introduced for weighted networks in which persistent features cannot be determined from a single simplicial complex in the family. In the case of the weight clique rank filtration we also reveal the intricate multiscale relation between weights and links

in a weighted indirect network.

Neighbor filtration

The **neighbor filtration** was introduced by us and used in chapter 4 for the study of social networks. The Neighbor filtration is a sequences of graphs, specifically open Ego networks of increasing radius, see chapter 2. Chosen a node v in a network, the first step $N^1(v)$ of the filtration is the graph of v's direct neighbors and the links between them. The second step is the graph induced by v's direct neighbors and nodes at path distance 2 and so on, to the growth of the path distance we determine the sequence:

$$N^1(v) \subseteq N^2(v) \subseteq \ldots \subseteq N^t(v). \tag{3.4.1}$$

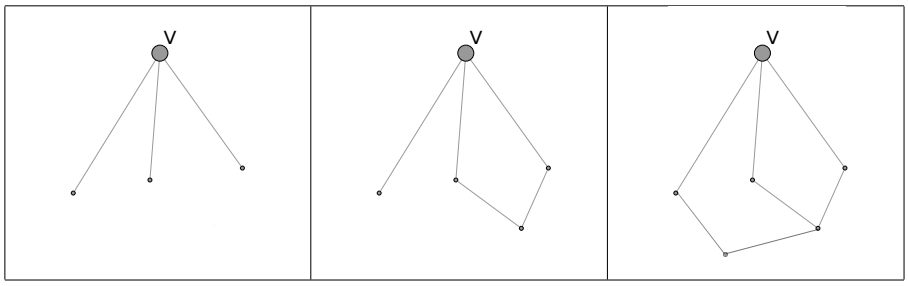


Table 3.1: Network neighbor filtration of v's Ego network.

Weight Rank clique filtration

The **weight Rank Clique filtration** defined by us on a weightenetwork Ω combines the clique complex construction with a thresholding on weights. The results of applying persistent homology to this filtration are illustrated in chapter 5.
To construct the weight rank clique filtration we follow these steps :

- rank the weights of links from ω_{max} to ω_{min}: the discrete parameter ϵ_t scans the sequence,

- at each step t of the decreasing edge ranking, consider the thresholded graph $G(\omega_{ij}, \epsilon_t)$.This is the subgraph of Ω with links of weight larger than ϵ_t.

- For each graph $G(\omega_{ij}, \epsilon_t)$, build the clique complex $K(G, \epsilon_t)$.

The clique complexes are nested to the growth of t and determine the weight rank clique filtration. Persistent one dimensional cycles represent weighted loops with much weaker internal links.

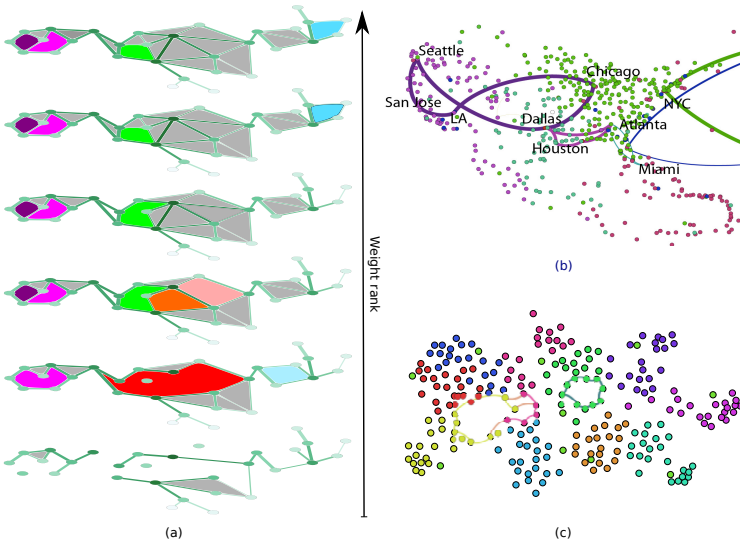


Figure 3.6: Weight rank clique filtration.

Remark 3.4.3. *Note that there is a conceptual difference in interpreting H_1 persistent homology of data with the Rips-Vietoris filtration, see chapter 3 and H_1 persistent homology of weighted networks with the weight rank clique filtration. While in the first case persistent generators are relevant and considered features of the data, short cycles are more interesting for networks. This is because random networks, or randomizations of real networks, display one dimensional persistent generators at all scales, while short lived generators testify the presence of local organization properties on different scales.*

3.4.2 Persistence homology modules

Given a filtration, the next step is to compute the homology modules of the simplicial complexes in the filtration and compare them using the maps induced in homology by inclusions.

Definition 3.4.4. *The persistence homology module of a filtration $\{X_v\}_v$ is given by the groups $H_n(X_v)$, that are the homologies of the simplicial complexes and the maps $H_n(i_{v,w}) : H_n(X_v) \to H_n(X_w)$ induced in homology by the inclusions $X_v \hookrightarrow X_w$ for all $v \leq w$.*

Persistent homology is often computed with coefficients in a field k, the groups $H_n(X_v)$ are vector spaces and the maps $H_n(i_{v,w}) : H_n(X_v) \to H_n(X_w)$ are linear maps. Every ordered set is a category, the objects of the **order category** associated to $\mathbb{N}$ are natural numbers and there is a morphism between two numbers v and w if and only if $v \leq w$. A sequence of vector spaces indexed by $\mathbb{N}$ can be seen as a functor $F : \mathbb{N} \to SC$ associating to every index v the simplicial complex X_v and to every morphism $v \leq w$ the linear inclusion $X_v \hookrightarrow X_w$.

Using a simple construction, that gives an equivalence of categories between functors from $\mathbb{N}$ to SC and graded modules over the polynomial ring in one variable $k[t]$, a persistence homology module is mapped to a $k[t]$-module.

Construction 3.4.5. *To a family of vector spaces $\{X_v\}_{v\in\mathbb{N}}$ and linear maps $\varphi_{v,w} : X_v \to X_w$ for all $v \preceq w$, such that $\varphi_{v,w} = \varphi_{z,w} \cdot \varphi_{v,z}$, for all $v \preceq z \preceq w$, we can associate the vector space $X = \bigoplus_v X_v$ with $k[t]$-module action*

$$\begin{aligned} t : X_v &\longrightarrow X_{v+1} \\ m &\longrightarrow \varphi_{v,v+1}(m) \end{aligned}$$

A **persistent homology generator** is a generator of $H_n := \oplus_{v\in\mathbb{N}} H_n(X_v)$ according to the $k[t]$- module structure, i.e an element $g \in H_n(X_v)$ such that there is no $h \in H_n(X_w)$ for $w < v$ with the property that $t^{v-w}h = g$.

The **structure theorem** of modules over a Principal Ideal Domain completely classifies modules over the polynomial ring $k[t]$. This theorem is a generalization of the fundamental theorem of finitely generated abelian groups and roughly can be stated saying that finitely generated modules over a PID can be decomposed in prime factors just as the integer numbers. The decomposition divides the module in a free submodule and a torsion submodule.

Definition 3.4.6. *A principal ideal domain A is a commutative ring that:*

- *A has no zero divisors, i.e if $ab = 0$ for $a, b \in A$ then $a = 0$ or $b = 0$*

- *every ideal is principal, i.e it can be generated by single element.*

A primary ideal is an ideal $I \subset A$ such that if $xy \in I$ then $x \in I$ or $y^n \in I$ for some integer n.

Theorem 3.4.7. *Let M be a finitely generated module over a principal ideal domain A, then the following decomposition holds:*

$$M = \oplus_{i=1}^{m} A \oplus_j A/(q_i)$$

where (q_i) are primary ideals in A.

The polynomial ring in one variable $k[t]$, is a principal ideal domain because it is an Eucledian Domain (there is the Eucledian division algorithm). In this case, all the primary ideals are of the form t^v for $v \in \mathbb{N}$. The decomposition theorem for modules over the polynomial ring $k[t]$ is then:

$$M = \oplus_i k[t](-a_i) \oplus_j k[t](-c_j)/t^{d_j}.$$

In this expression we have denoted with $k[t](-a_i)$ and $k[t](-c_j)$ the shift of the polynomial ring by degree a_i and c_j respectively, considering k_t as a graded ring, with grading given by the exponents of the monomials. This means that $(k[t](-a_i))_s = k[t]_{s-i}$ and the same for c_j. This notation highlights the degree of the module generators.

Remark 3.4.8. *The degrees of the generators $(a_i,\ c_j)$ and the torsion degrees d_j completely determine a $k[t]$-module.*

The degree of each generator g is called its birth of the generator and denoted with β_g, and the degree in which the generator is annihilated by the module action, called death of the generator, will be denoted byδ_g. The persistence (lifetime) of a generator is measured by $p_g := \delta_g - \beta_g$.
Generators that are in the free part of the module or have a long lifetime are called persistent generators.

Persistent homology generators can be computed using the libraries **javaPlex** (Java) or **Dionysus** (C++), which are both available from the Stanford's CompTop group website (`http://comptop.stanford.edu/`), and presented using the barcode or the persistence diagram. We developed a Python module to wrap the javaPlex library, consisting of a number of scripts able to preprocess complex networks and store the resulting homological information in a manageable form. The module is available online at `https://github.com/lordgrilo/Alchemy`.

3.4.3 Barcode

The complete classification of persistence modules in theorem 3.4.7 is given by the persistence intervals of the generators in H_n. It is possible to visualize this information through a handy complete invariant, the barcode, (29) .

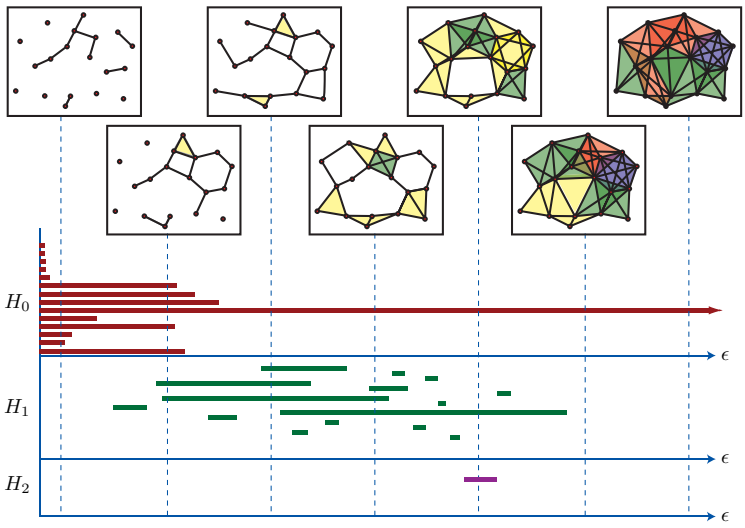


Figure 3.7: Barcode of a point cloud. The associated filtration is the Rips-Vietoris filtration.

Definition 3.4.9. *The* **barcode** *of a filtration is the set of intervals* $[\beta_g; \delta_g]$ *for all generators* $g \in H_n$.

By persistent topological features we intend generators of H_n such that the interval $[\beta_g; \delta_g]$ is large with respect to the filtration length.
These are considered as the features of a point cloud when using the Rips-Vietoris filtration.
When using network filtrations instead, homology recognizes motifs and patterns in the graph. For weighted networks, a persistent homology generator represents a set of nodes linked by strong interactions two by two but not cohesively connected in general.

In figure 3.8 *A)* The two 3-clique components shown are equivalent expect for the right one being closed in a ring-like configuration. The Betti numbers distinguish between them. Both have one connected component (H_0), but only the right one creates a H_1 (1d) cycle, thus $\beta_1 = 0, \beta_1' = 1$.

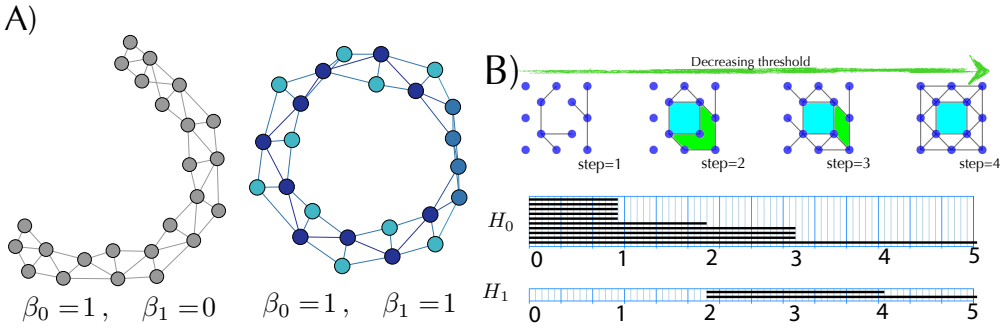


Figure 3.8: Topology of networks and weight rank clique filtration.

In figure 3.8 *B)* weighted clique complex of a weighted network. H_1 persistence generators, in the figure are represented by the light blue and green hulls; the former persists to the end of the filtration, while the latter has a shorter timespan and is quickly filled. Note that all triangles are not H_1 cycles because they form a 3-clique, which is mapped to a 2-simplex and thus invisible to homology.

Remark 3.4.10. *Note that a barcode is only possible for the description of modules over the polynomial ring or over a principal ideal domain in general. Given a module over the multivariate polynomial ring for example, that represents multiperistent homology modules as we will see in chapter 6.*

An alternative way to represent persistent homology modules is the **persistence diagram** (55), (113). A persistence diagram is a set of points in the plane counted with multiplicity, it can be recovered from the barcode considering the points $(\beta_g, \delta_g) \in \mathbb{R}^2$ with multiplicity given by the number of generators with the same persistence interval.

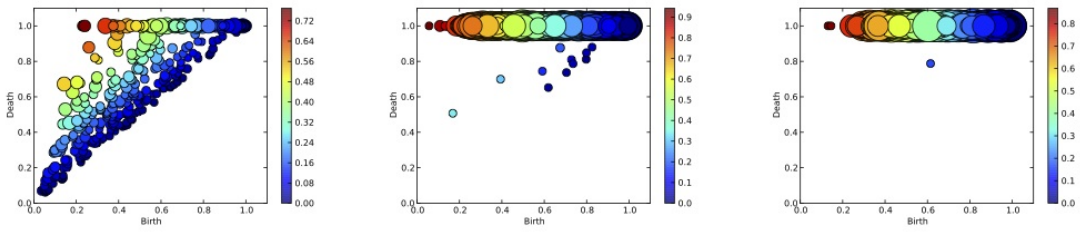


Figure 3.9: Persistence diagram of three weighted networks. The associated filtration is the weight clique rank filtration.