Topological Data Analysis for COVID-19 Contact Tracing Data

Stefania Ebli, Jeanne Fernandez, Celia Hacker, Yann Mentha, Luca Nyckees

Abstract

Topological data analysis is a branch in the field of applied mathematics that has rapidly grown in the past few years. It offers toolboxes to study the shape of data and finds applications in various domains of machine learning [3]. Central concepts in topological data analysis are *persistent homology* [1] and *zigzag homology* [2], which both offer a way of analysing the behavior of topological features along a filtration or a diagram of simplicial complexes built from data. Contact tracing data has natural associated graph time-series representations from which one can generate diagrams of simplicial complexes. The way a pandemic spreads on a graph is a topic that has been widely studied with ideas coming from the framework of statistical analysis of network data (e.g. [4]). Here we use topological data analysis to investigate the nature of the changes in the topology of contact tracing graphs and provide a general pipeline to get higher-dimensional insights into the behavior of the COVID-19 spreading process.

Laboratory for Topology and Neuroscience
Spring semester of 2021

1 Introduction

A recent center of interest for the domain of discrete mathematics is the study of the COVID-19 pandemic evolution. The aim of this work is to use tools coming from topological data analysis (TDA) in order to understand contact tracing data collected in the Canton of Geneva. In particular, we model interactions among users using simplicial complexes. These are natural multi-dimensional extensions of graphs that encode not only pairwise relationships but also higher-order interactions between vertices, represented geometrically by triangles, tetrahedra, and higher dimensional simplices. Using these structures we can model n-fold interactions between users and attach a label to their context.

The main object of our study is a time-series of natural data-built graphs, and we look at how the topology of the graphs evolves through time. More precisely, we search for meaningful differences of topological invariants (e.g. the behavior of betti numbers) when comparing this to various kinds of random-equivalent time-series of graphs. We also provide possible socio-political factors explaining abrupt changes in some of the observed topological features. For example, we look at how *contamination waves* and *restriction policies* translate in terms of structural changes in our graphs.

In section 2, we introduce the theoretical basics of our method. In particular, we define the notions of *simplicial complex* and *betti numbers*. Section 3 consists of a complete description of the real contact-tracing data sets that our study is based on. In section 4, we describe the methods put in practice, namely the type of statistics that were used for the exploratory data analysis performance, the type of tools coming from TDA, and the general pipeline that puts it all together. This also involves a time window-based interactive method enabling the visualisation of fundamental graphical and topological features, thus allowing to isolate one-time critical events in a single time period.
The general pipeline is implemented in Python. The source code as well as illustrative notebooks are available at the following GitHub repository.

GitHub Repository

We also provide a way to interact with the pipeline and the data via an application packed with *Docker* (*cf.* the README.md file on GitHub).

2 Theoretical Background

In this section, we introduce the concepts necessary to the understanding of this work. In particular, we define the notions of *simplicial complex*, graph generalizations that help us model contact-tracing data in a way that is feasible with TDA, and *Betti* *numbers*, an important topological feature that encodes, in particular, the number of connected components, loops and voids in a graph.

2.1 Simplicial Complexes

A simplical complexes is, roughly speaking, a set of simplices (points, segments, triangles, tetrahedrons and so on) that are glued together in a specific hierarchic way. More precisely, we have the following.

Definition 2.1 (Simplex). *Let $k \in \mathbf{N}$. A* k-dimensional simplex *(or simply* k-simplex*) α is a convex hull built on a set of affinely independent points $V = \{v_0, ..., v_{k+1}\} \subset \mathbf{R}^k$. Symbolically, one has*

$$\alpha = \{\sum_{i=0}^{k} a_i v_i | (a_i)_{i=0}^{k} \in C_k\},$$

where we define the set $C_k = \{(a_i)_{i=0}^{k} | \sum_{i=0}^{k} a_i = 1\} \cap \mathbf{R}_+^{k+1}$.

It follows that simplices are a generalization of the notions of triangles and tetrahedrons, to any higher dimension. For example, a 2-simplex is a triangle, a 3-simplex is a tetrahedron and a 4-simplex is a 5-cell (homeomorphic to the disk D^5). *Regular* k-simplices are just k-simplices where, with notations as in the definition above, the points $v_1, ..., v_k$ are the standard unitary vectors of $\mathbf{R}^k$ and $v_0 = \mathbf{0}$. Simplices are most often studied up to homotopy, *i.e.* continuous deformation. Thus, one can always consider the case of regular simplices and imagine nice regular triangles and tetrahedrons. We define the *face* of a k-simplex to be a convex hull built on a subset of the points that constitute the k-simplex. Faces are simplices themselves. For two simplices α and β, we denote by $\alpha < \beta$ (or equivalently $\beta > \alpha$) the fact that α is a face of β.

Definition 2.2 (Simplicial Complex). *A* simplicial complex *X is a set of simplices such that both conditions (1) and (2) expressed below are satisfied.*

(1) $\beta < \alpha \in X \implies \beta \in X$

(2) $\alpha, \beta \in X$ *and* $\alpha \cap \beta \neq \emptyset \implies \alpha \cap \beta < \alpha$ *and* $\alpha \cap \beta < \beta$

2.2 Betti Numbers

Generally speaking, Betti numbers of a space allow us to get insights on its topological features. In fact, loosely speaking, k-Betti numbers (denoted b_k) represent the number of k-dimensional holes in the space. More precisely, for a given space X, b_0 is the number of connected components of X, b_1 is its number of loops and b_2 is its number of voids. For example, the torus is connected, generated by two loops and has one void and it follows that $(b_0, b_1, b_2) = (1, 2, 1)$. We introduce now those notions in a more formal way, with *simplicial homology.*

We consider a simplicial complex X. Let $C_k(X)$ be the free abelian group generated by the ordered k-simplices in X. We look at a k-simplex α as an ordered $k+1$-tuple of vertices $(\alpha_0, ..., \alpha_k)$. We have a sequence of group homomorphisms

$$\cdots \xrightarrow{\delta_3} C_2(X) \xrightarrow{\delta_2} C_1(X) \xrightarrow{\delta_1} C_0(X) \xrightarrow{\delta_0} \mathbf{0}$$

where we define, for each $k \in \mathbf{N}$, the so-called *boundary map*

$$\delta_{k+1} : C_{k+1} \longrightarrow C_k$$

by setting $\delta_{k+1}((\alpha_0, ..., \alpha_k)) = \sum_{i=0}^{k} (-1)^i (\alpha_0, ..., \alpha_{i-1}, \alpha_{i+1}, ..., \alpha_k)$. One can observe that compositions of consecutive boundary maps is zero, thus allowing us to formulate the following definition.

Definition 2.3. *With notations as above, we define the k-simplicial homology group of X, denoted $H_k(X)$, to be the abelian group* $\ker(\delta_k)/\mathrm{im}(\delta_{k+1})$.

Definition 2.4 (k-th Betti number)**.** *The k-th Betti number of a simplicial complex X, denoted $b_0(X)$, is defined as the rank of the homology group $H_k(X)$.*

2.3 Persistent Homology

We give the very basics of the theory of persistent homology. This is a way of getting insights on the evolution (through time, perhaps) of topological features such as connected components, cycles, voids and more generally n-dimensional holes. One of the main ideas to explore the behavior of such features is to define their birth and death coordinates, as described below.

Definition 2.5 (Filtration.)**.** *A* filtration *on a simplicial complex K is an increasing chain of subcomplexes of K of the form $K_0 \subseteq K_1 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n = K$.*

Remark 2.6. *Such a chain can be given by a strictly monotone (i.e. such that $f(\alpha) < f(\beta)$ for any pair of simplices $\alpha < \beta$) function $f : K \to \mathbb{R}$ by considering as subcomplexes of K the corresponding level sets of f.* ♠

Definition 2.7 (p-th persistent Betti numbers.)**.** *Let $K_0 \subseteq K_1 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n = K$ be a filtration of a simplicial complex K, and consider the induced p-homology sequence $H_p(K_0) \to \cdots \to H_p(K_n)$ for a $p \in \mathbb{N}$. We define the p-th* persistent Betti numbers *as $\beta_p^{i,j} := \mathrm{rank}(H_p^{i,j})$, where $H_p^{i,j} := \mathrm{Im}(f_p^{i,j})$ and $f_p^{i,j} : H_p(K_i) \to H_p(K_j)$, for $0 \leq i \leq j \leq n$, are the natural group homomorphisms.*

Remark 2.8. *We have, for $0 \leq i, j \leq n$ and $p \in \mathbb{N}$,*

$$\begin{aligned}\beta_p^{i,j} &= \operatorname{rank}(H_p^{i,i}) \\ &= \operatorname{rank}(\operatorname{Im}(f_p^{i,i} : H_p(K_i) \to H_p(K_i)) \\ &= \operatorname{rank}(\operatorname{Im}(\operatorname{id}_{H_p(K_i)} : H_p(K_i) \to H_p(K_i)) \\ &= \operatorname{rank}(H_p(K_i)) \\ &= b_p(K_i),\end{aligned}$$

which is simply the p-th Betti number of the complex K_i. ♠

Remark 2.9. *One has $H_p^{i,j} \subseteq H_p^{j,j}$ for any $0 \leq i \leq j \leq n$.* ♠

Definition 2.10 (Birth and Death.). *A feature $\gamma \in H_p^{i,i} \setminus H_p^{i-1,i}$ is said to be* born at K_i. *Moreover, if the feature γ satisfies the set conditions $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ and $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$, we say γ* dies entering K_j.

Definition 2.11 (Persistence.). *Consider the feature γ of the definition above. In the case where the filtration is given by a map $f : K \to \mathbb{R}$, so that we consider subcomplexes $K_i := f^{-1}(a_i)$ for some $a_i \in \mathbb{R}$, we define the* persistence of γ *to be the real value $a_j - a_i$, interpreted as its lifetime.*

Definition 2.12 (Persistence Diagram.). *Consider the context of the previous definition. The* persistence diagram *of the filtration on K is the plot given by the identity diagonal of the real plane $\mathbb{R}^2$ along with the set of points (a_i, a_j) with multiplicities $\mu_p^{i,j}$, where $\mu_p^{i,j}$ is defined as the number of classes that are born at K_i and die entering K_j.*

2.4 Zigzag Persistent Homology

Zigzag persistent homology aims at generalizing the setting of persistent homology, by considering a class of diagrams that encapsulate the case of persistence modules.

Definition 2.13 (Zigzag module). *Consider a diagram of vector spaces of the form*

$$V_0 \longleftrightarrow V_1 \longleftrightarrow V_2 \longleftrightarrow \cdots \longleftrightarrow V_n,$$

where each double-sided arrow $\longleftrightarrow$ is either a left arrow $\longleftarrow$ or a right arrow $\longrightarrow$, that is a linear map of vector spaces. We call such a diagram of vector spaces a zigzag module. The type τ *of a zigzag module represents the orientation of its arrows and is denoted by a sequence of f and g, respectively denoting forward and backward maps.*

Example 2.14 (Type of a module). *The zigzag module below has type $\tau = gfg$.*

$$V_1 \longleftarrow V_2 \longrightarrow V_3 \longleftarrow V_4$$

Here, we study a specific type of zigzag modules formed by applying a homology functor $H_p(\cdot)$ to a diagram of embeddings between simplicial complexes of the form

$$K_0 \longrightarrow K_1 \longleftarrow K_2 \longrightarrow \cdots \longleftarrow K_n.$$

This yields, for a fixed $p \in \mathbb{N}$, the following zigzag module.

$$H_p(K_0) \longrightarrow H_p(K_1) \longleftarrow H_p(K_2) \longrightarrow \cdots \longleftarrow H_p(K_n)$$

Given a collection of simplicial complexes, there is a natural way to obtain such a zigzag module, as desribed in the next definition.

Definition 2.15 (Natural p-zigzag module). *Let $\mathcal{X} = \{X_i\}_{i=0}^n$ be a collection of simplicial complexes. For a given $p \in \mathbb{N}$, we define the* natural p-zigzag module associated to $\mathcal{X}$ *to be the module formed by applying the functor $H_p(\cdot)$ to the following diagram.*

$$\begin{tikzcd}
& X_0 \cup X_1 & & X_1 \cup X_2 & & \ldots & \\
X_0 \arrow[ur] & & X_1 \arrow[ul] \arrow[ur] & & X_2 \arrow[ul] \arrow[ur] & & X_n \arrow[ul]
\end{tikzcd}$$

Definition 2.16 (Remak decomposition). *A* Remak decomposition *of a zigzag module $\mathbb{V}$ is a decomposition of the form $\mathbb{V} = \mathbb{W}_1 \oplus \cdots \oplus \mathbb{W}_n$, where all $\mathbb{W}_i$ are indecomposable.*

Remark 2.17. *Any zigzag module admits a Remak decomposition.* ♠

Definition 2.18 (Interval module). *One can define the* type-τ interval module $\mathbb{I}_\tau(b, d)$ *as the only τ-module formed by the spaces*

$$I_i^\tau(b,d) = \begin{cases} \mathbb{F} & \text{if } b \leq i \leq d \\ 0 & \text{otherwise,} \end{cases}$$

and equipped with identity maps between two adjacent copies of and zero maps elsewhere.

Theorem 2.19. *Any zigzag module $\mathbb{V}$ admits a Remak decomposition of the form*

$$\mathbb{V} = \mathbb{I}_\tau(b_1, d_1) \oplus \cdots \oplus \mathbb{I}_\tau(b_N, d_N).$$

Remark 2.20. *By the Krull-Schmidt principle (see Proposition 2.2 in [2]), the Remak decomposition into interval modules of a given zigzag module $\mathbb{V}$ is an isomorphism invariant of $\mathbb{V}$. This allows us to formulate the next definition.* ♠

Definition 2.21 (Zigzag persistence). *The persistence of a zigzag module $\mathbb{V}$ is defined as the multiset* $\mathrm{Pers}(\mathbb{V}) = \{[b_i, d_i] | i = 1, ..., N\}$ *induced from a Remak decomposition*

$$\mathbb{V} = \mathbb{I}_\tau(b_1, d_1) \oplus \cdots \oplus \mathbb{I}_\tau(b_N, d_N).$$

3 Datasets

3.1 Structure of Data

The data was collected in the Canton of Geneva with either fill-in forms handed in to positive cases or the Swiss *social pass* app. The main data set on which our study is based encodes a little more than 70000 entries, each corresponding to the reporting of a contact by an individual tested positive. There are around 25000 positive cases and thus each one of them reports a bit less than 3 contacts in average. The way we build our graphs focuses on one entry attribute in particular, namely the date on which a reported person was in contact with the corresponding positive case for the last time. We also have access to a label attributing a social context to a given reported contact, e.g. *work-related*, *living together*, or *school-related*. This enables us to manually add connections according to this context and interpret the changes in the resulting statistics. We also report results obtained with a smaller secondary dataset encoding a little more than 6000 entries. This is the dataset on which we first explored our main pipeline features.

3.2 A Word on Noise and Interpretation

We use the secondary dataset as an illustrative example only. Indeed, we consider it, due to is small size, too prone to over interpretation mistakes. This is motivated by the fact that the data can be noisy and the main reason for this is that contact-tracing data is extremely hard to collect properly, as there are numerous ways in which information could be missing, since it is a human-based process. For example, people may forget about precise dates, contacts or even willingly omit to mention them. Another factor of noise is the relatively large proportion of false negative and false positive results when testing for the COVID-19. For that reason, we only base our interpretation on the results obtained with the main dataset, considered big enough to properly represent the spread of the pandemic. Also, it covers a time interval where the data-collection policies don't change, which is a crucial aspect for our study.

4 Method and General Pipeline

In this section, we describe the general objects we deal with, how we extract them from the data sets, and what kind of results we expect from the tools we presented in Section 2. Moreover, we introduce a general pipeline that puts together various sub-methods, in a way that can be reproducible for any other data set based on the dated reporting of contacts by positive individuals.

4.1 Data-built Graph Time-series

With the data set coming from contact tracing policies, this is the way we build our graph time-series. The idea is to built a graph G from the data, representing the spreading of the pandemic and to generate graph time-series from a natural time-filtration on the nodes and edges for further analysis. Here, the map $c(\cdot)$ assigns to a graph its corresponding clique complex. First, we define a few mathematical objects that can be created from the data.

1. The graph G denotes the undirected weighted graph built created as follows. Nodes represent individuals (either tested positive or reported as recent contacts by one or more positive persons), and edges represent basic interactions between nodes. More precisely, for an individual x tested positive, we draw edges between him and its reported contacts and label those edges in the following way. Suppose x reported a contact c with last contact date d. Then we draw an edge (x, c) with weight $f(d)$, where we define $f(d)$ as the number of days that went by from the 1rst of March (the first date in the dataset) to date d.

2. The increasing chain of graphs $\{G_i\}_{i=0}^n$ is obtained as follows. We can take the weights on the edges of G to obtain a filtration based on the natural timeline. We get an increasing chain of graphs where we choose the convention that nodes are added at the same time as they generate an edge, so that the 0-skeleton evolves through time. We obtain a filtration of the form
$$G_0 \subset G_1 \subset \cdots \subset G_n = G.$$

3. The increasing chain of simplicial complexes $\{K_i\}_{i=0}^n$ is obtained by applying the map $c(\cdot)$ to the chain above, leading to a chain
$$K_0 \subset K_1 \subset \cdots \subset K_n = K,$$
where $K_i := c(G_i)$.

4. As contact tracing inherently shows a non-negligible proportion of missing values, we choose to perform some feature imputation on our graph G, by manually adding missing edges according to some well-defined criteria. For example, we can choose to complete our graph by adding an edge between two reported contacts whenever their context is *living together*. This process of *edge imputation* is motivated by the idea that some social relations are positively correlated with the contaminating actions, and that some unobserved edges are mistakenly missing. This process creates a new graph that we denote by $\tilde{G}$. The induced graph and complex chains are denoted by $\{\tilde{G}_i\}_{i=0}^n$ and $\{\tilde{K}_i\}_{i=0}^n$.

5. We consider for each time window (t_1, t_2) the graph $G(t_1, t_2)$ consisting of the interactions between a positive individual and its reported contacts whose last contact date x is situated in the window (t_1, t_2). This way, we create a chain of possibly overlapping graphs $\{G_i(\Delta, w)\}_{i=0}^{m(\Delta,w)}$, where $w = t_{i+1} - t_i$ is the time length of every window, and Δ is the step-size. This means we have $G_0(\Delta, w) = G(0, w)$ and $G_1(\Delta, w) = G(\Delta, w + \Delta)$. More generally, one has

$$G_i(\Delta, w) = G(i \cdot \Delta, w + i \cdot \Delta).$$

Now, the main tools we use for the data analysis are the following.

1. **Global exploratory data analysis.** We also consider G as it is, for the basic graph exploratory data analysis part.
2. **Time-window method.** We consider the graph chain built with the time-window process. This enables the time-window-based visualisation method, aiming to isolate abrupt changes in various graphical and topological features in the evolution of the contamination process.
3. **Persistent Homology.** We generate and study the persistence diagrams obtained from the chains $\{K_i\}_{i=0}^{n}$ and $\{\tilde{K}_i\}_{i=0}^{n}$.
4. **Zigzag Homology.** We fix a choice of time-window parameters (Δ, w), and then generate and study the zigzag persistence diagrams obtained from the time-window chain $\{G_i(\Delta, w)\}_{i=0}^{m(\Delta,w)}$.

4.2 Random Equivalent Graph Time-series

One important direction of this work is based on the idea that the evolution of a pandemic, when translated to a graph time-series, has a specific structure that should differ from most random equivalent graph time-series. For that reason, we look at various random constructions and compare the results through all three methods mentioned above.

Example 4.1 (Erdõs-Renyi equivalents.)**.** *Given the time-series $(G(t_1, t_2))_{t_1,t_2}$, we look at the time-series $(H(t_1, t_2))_{t_1,t_2}$, where $H(t_1, t_2)$ is an Erdõs-Renyi equivalent of $G(t)$. Moreover, given the whole weighted graph G, we also look at its random weighted equivalent H. In order to perform a filtration of the natural timeline on these graphs as well, the number of days $f(x)$ of each day is randomly sampled from a kernel density estimation of these dates on the graph $G(t_1, t_2)$*

4.3 Time-window-based Visualisation

Multiple statistics can simultaneously be displayed for different graphs in a given sliding time-windows (Figure 1). In this case, we investigate the statisctics of three different graphs : the global graph, the edge-tuned on place of residence criteria graph as described in Section 4.1, and a random ER equivalent. A slider allows to control which time-window is displayed. As an illustrative example, we report the results obtained with the secondary dataset. Here, we present the statistics from top-left to bottom-right.

- **Simplex Count**: we plot the number of simplices counts with respect to the simplex dimension. As expected, the three graphs exhibit the same number of 0 dimensional simplices (vertices), while the edge-tuned graph counts several more higher-order simplices, as implied by the induced fully connected subgraphs.
- **Betti Number counts**: the natural and edge-tuned graphs display a similar number of connected components (0-betti numbers). Moreover, we observe that the Erdos-Reyni equivalent is the only graph to exhibit cycles (1-betti numbers).
- **Component size distribution**: as expected, the natural and edge-tuned graphs share the same distribution as no inter-component edge is added when adding edges. The difference is more significant when comparing with the Erdõs-Renyi graph, whose distribution reveals the presence of a giant component with several smaller satellites of the same magnitude of the natural/edge-tuned graphs.
- **Degree Distribution**: distribution of the degree of the nodes. As seen on Figure 1, a majority of the nodes degrees range from 0 to 5.
- **Log Degree Distribution**: Same plot using a log plot: we see that the number of nodes tends to decrease exponentially with the degree. Notice the super-connected outsider in the natural and edge-tuned graphs with degree > 35 that is not present for the Erdõs-Renyi equivalent.
- **Number of nodes**: We plot the number of nodes of the graphs for the current time window with respect to time (as day), starting from March 1. The greatest amount of data ranges in the August period.

We are able to get a closer look at a graph in a given time-window, as shown in Figure 2.

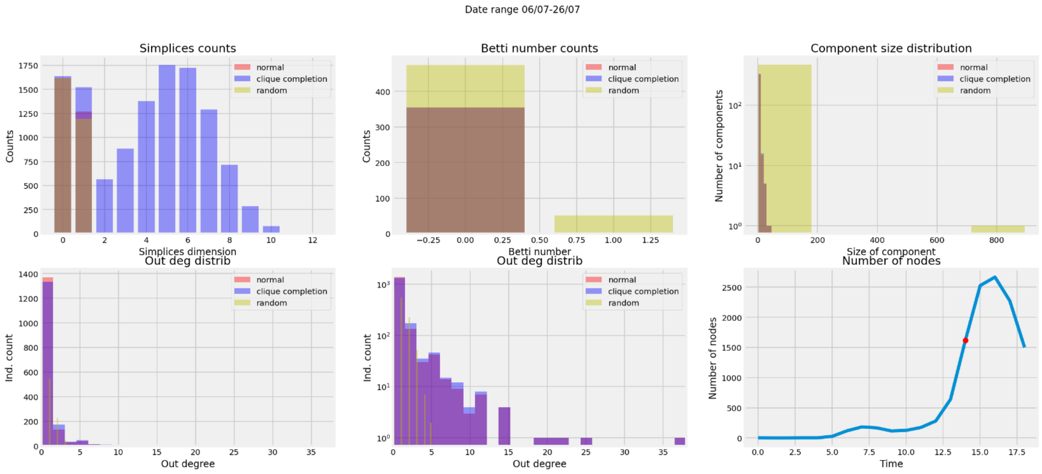


Figure 1: Statistics displayed for a given time-window (secondary data).

4.4 Persistent Homology

Using the weights of the natural timeline as described in Subsection 4.1, we are able to build a persistent diagram for the natural graph and its Erdõs-Renyi equivalent. This time, no time-window is involved: instead, nodes are gradually appended to the graph so as to get a proper filtration. Once again, we display the results obtained with the secondary dataset as an illustrative example. We plot the persistence diagrams for 0 and 1-betti numbers of the natural data-based graph (Figure 4) and its Erdõs-Renyi version (Figure 5).

Interpretation. The 0-dimension (resp. 1-dimension) plot displays the life span of the connected components (resp. cycles) via birth-death coordinates. For the 0-dimension plot, points situated on the upper horizontal dashed line correspond to connected components that never merge with an older component, and dots on the diagonal correspond to components whose birth date correspond to the death date. The latter dots represent edges that connect a vertex belonging to a pre-existing component to a new vertex. We see that most of the connected components appear in August, along with the amount of data.

Example 4.2. *The Erdõs-Renyi equivalent exhibits less ever lasting connected components than the natural graph does. Instead, a majority of components ends up joining older components, confirming the idea of a giant component as observed in the* component size distribution *of Figure 1.*

4.5 General Pipeline in a Nutshell

We illustrate the general pipeline in Figure 3.

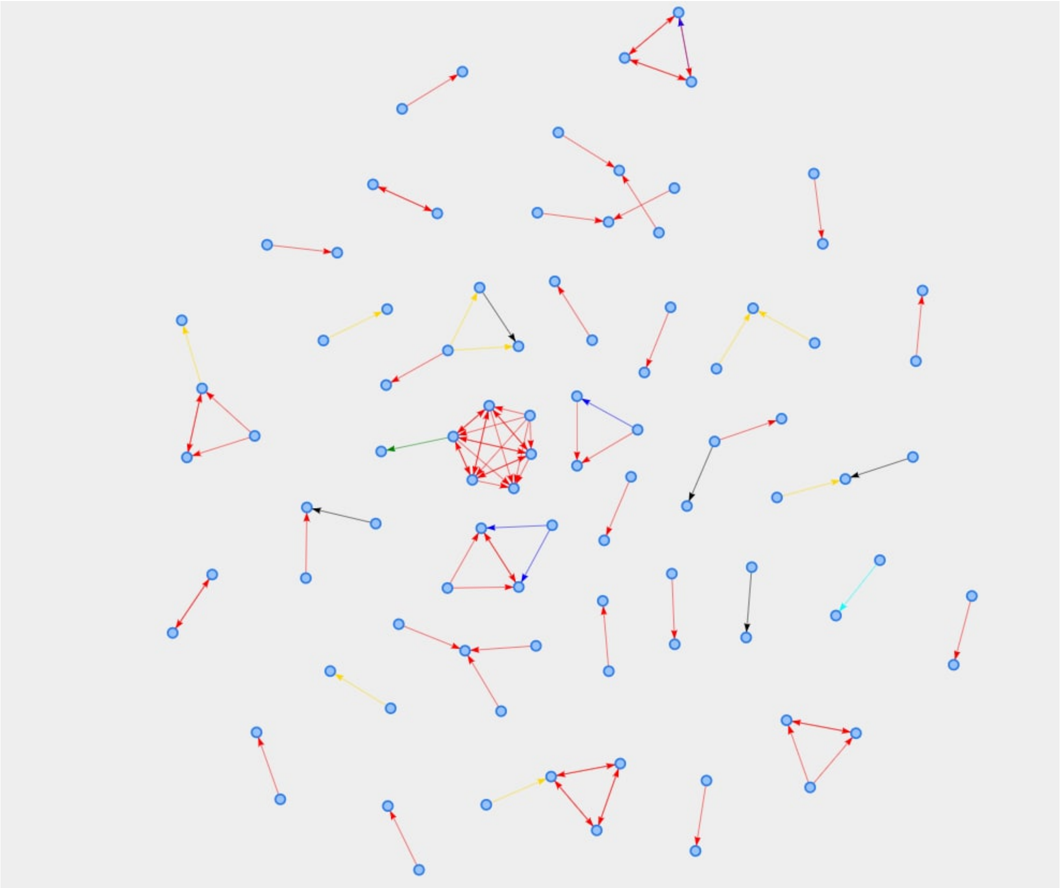


Figure 2: Graph in a given time-window visualized with *Pyviz*. Directed edges represent a positive subject reporting a contact and the nature of their interaction is indicated with a coloring scheme convention.

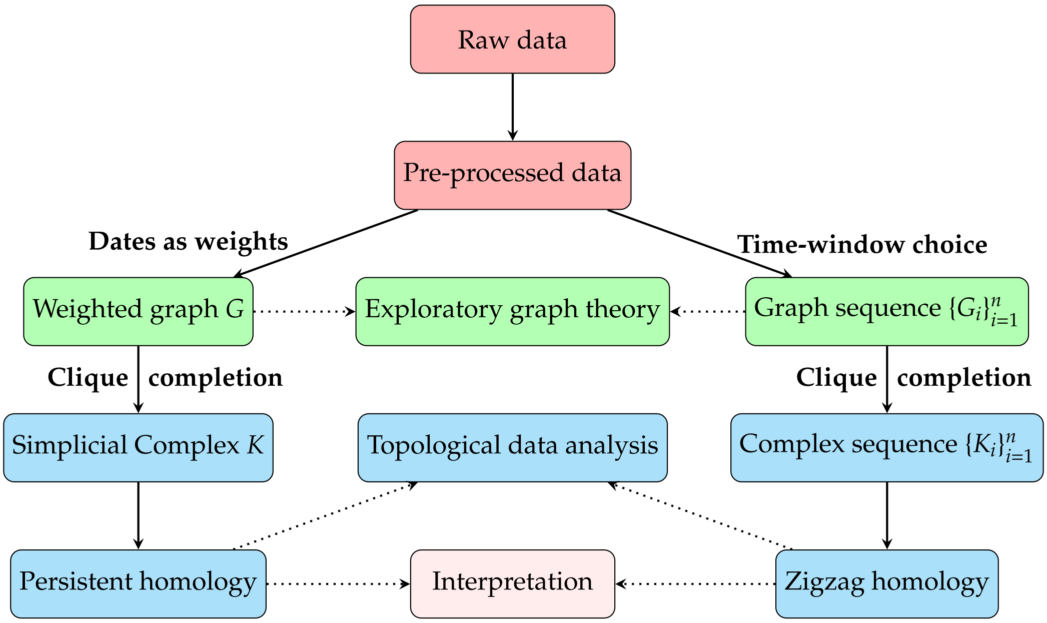


Figure 3: General pipeline illustration. We use tools from both graph theory and topological data analysis to study the topology of graphs naturally formed based upon timed spreading process data (contact tracing data).

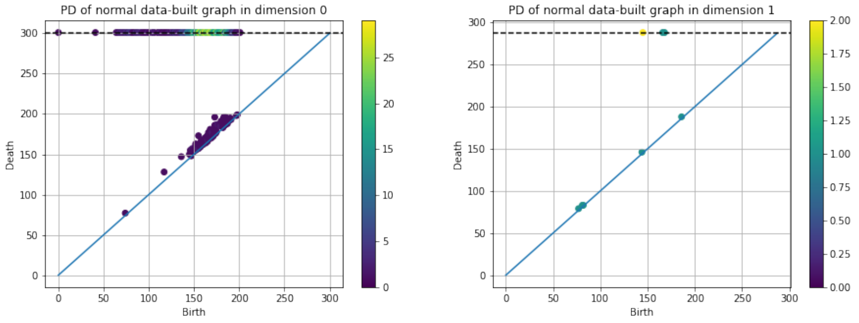


Figure 4: **Persistence diagrams for the small data-based graph.** On the left, the dimension 0 plot. On the right, the dimension 1 plot.

5 Results

We now display and interpret the results obtained with the main dataset.

5.1 Time-window-based Visualisation

We start by summarizing the statistics explored in the time-window process (we display the results for the time window 21.10.20 – 10.11.20 in Figure 6).

- **Simplices Counts.** Now, the main data set is approximately ten times bigger as the other one, and this resonates with the simplex counts. We had a total of around 1600 graph vertices in a 50-days time window for the old dataset, whereas the new one yields, for example, a total of around 4000 graph vertices in a 20-days time window. Even after proceeding to clique completion, we obtain a very low proportion of higher-dimensional data, as opposed to the results of the small dataset. Here, we formed around 2400 edges and 200 triangles. We note that the Erdõs-Renyi equivalent produces a bit more edges but contains no triangles at all (*i.e.* no 2-dimensional simplices).

- **Betti Number counts** The 0-th Betti numbers, *i.e.* the connected components, are reported below. The 1-Betti, *i.e.* the cycles, are not present in the data-based graphs, although we can find about 250 of them in the random equivalent. This is already a determining characteristic of our graphs : in a time window of 20 days, we don't expect any cycles to be formed, as they are interpreted as extremely improbable events, such as the same person being tested positive twice in 20 days, or at least reported after one of its own reported contacts has been tested positive.

- **Component size distribution** The significant change from going to the small dataset to the main dataset is reflected here, as we now look at a graph that is way more connected than before. Indeed, we have only 9 connected components in the graph of the given time window.

- **Number of nodes.** The curve of the number of nodes in function of time forms a spike centered around the month of October 2020. Hence our choice of time window for displaying and interpreting results.

5.2 Persistent Homology

We now try to get insights on the longevity and global behavior of basic topological features in our data by examining a set of persistence diagrams, as described in Section 4.

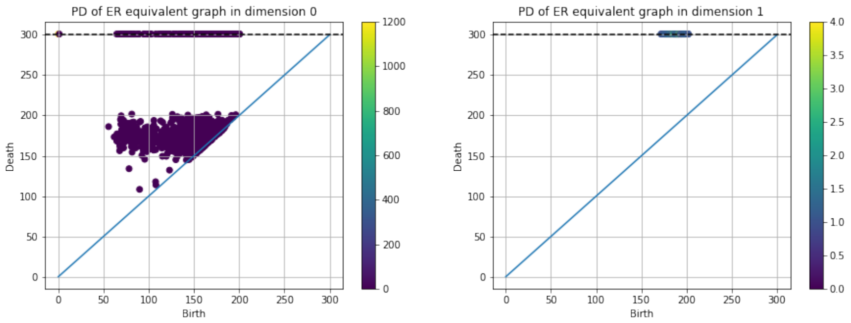


Figure 5: **Persistence plot for the Erdõs-Renyi equivalent (small data).** On the left, the dimension 0 plot, on the right, the dimension 1 plot.

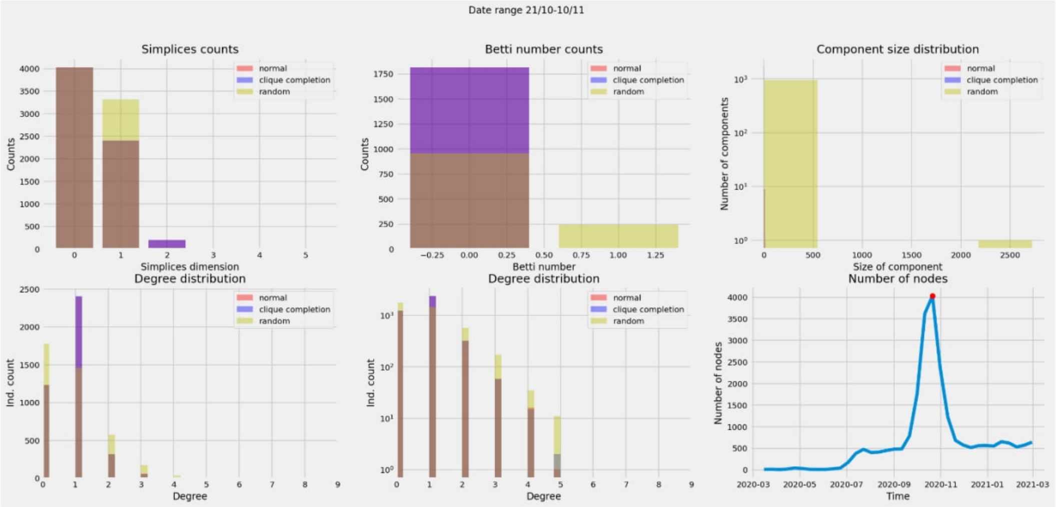


Figure 6: **Statistics displayed for a given time-window (main data).**

- **Connected components.** In comparison to the small dataset results, we observe a significant increase in life expectancy of connected components (for those that do die eventually). Indeed, we look at a distribution of coordinates that is rather far apart from the main diagonal, with features living up to 150 days (see Figure 7). As before, there is a large proportion of components that have infinite persistence, *i.e.* that never die. Those are the components that never merge with an older connected component. The color bar indicates the multiplicity of points on the diagram. We distinguish three main types of coordinates within the dimension-1 persistence diagram.

    1. Approximately 350 connected components have life coordinates $(b, d) = (0, \infty)$. Those represent the initial setting of the tracing. They form a set of disconnected patches that never merge between them (else the elder rule would randomly choose one of two merging components and create a death). This phenomenon accounts for the nature of the tracing : the incomplete reporting procedure leads to highly disconnected data.
    2. A few connected components have a very short life span (between 0 and 5 days). Those account for two specific case. First, when an individual a is tested positive and reports contact b that was itself tested positive without reporting a in the first place. Second, when an individual a reports a contact b that had previously been reported by a positive individual c.
    3. Lastly, some connected components do have a finite and non negligible persistence. Those represent non-trivial patches (consisting of more than one individual) that merge together. We note that this type of merging becomes active towards August 2020 (corresponding to days 150 to 180), which is the time-window where we observe an explosion of the pandemic spreading.

- **Cycles (1-holes).** We observe an increase in the number of cycles within the main data set, with respect to the smaller one. That said, the same distribution of coordinates is observed for both data sets, across the persistence diagram. First, we note that there is a very small quantity of cycles produced - 56. This goes in accordance with the intuition that the spreading pattern mainly forms tree-like structures producing very few closed paths. Moreover, closed paths are most often triangles which do not account for cycles. Secondly, in the main data diagram, we observe a few cycles - 21 - having finite persistence. Those are rare events in the spreading of the pandemic, corresponding to two individuals living in the same social bubble that get tested positive with significant time difference. Finally, we note that the Erdõs-Renyi model exclusively contains infinite persistence cycles.

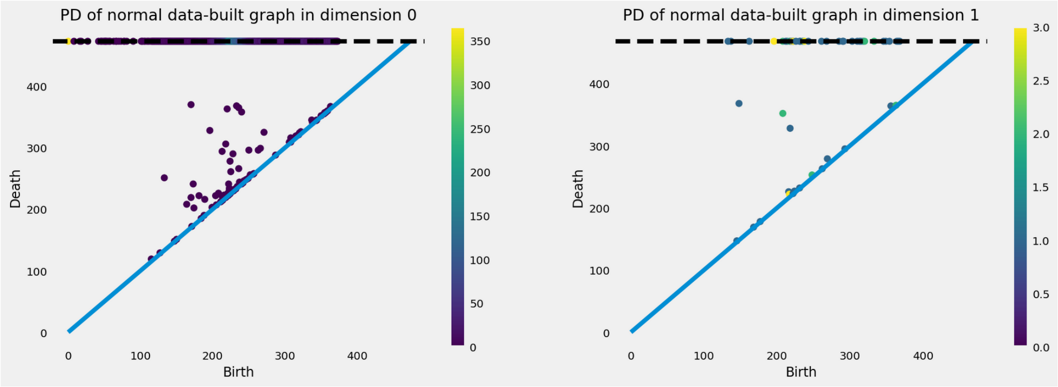


Figure 7: **Persistence diagrams for the main data-based graph.** On the left, the dimension 0 plot. On the right, the dimension 1 plot. The color bar indicates the multiplicity of a given coordinate.

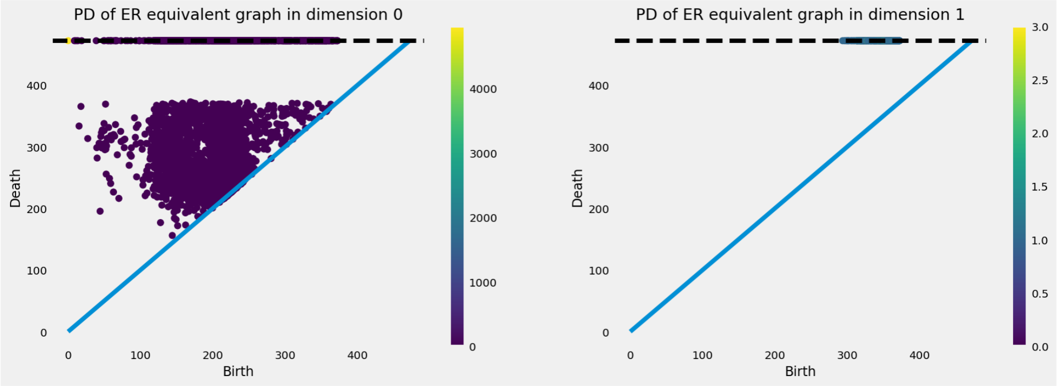


Figure 8: **Persistence diagrams for the ER equivalent (main data).** On the left, the dimension 0 plot. On the right, the dimension 1 plot. The color bar indicates the multiplicity of a given coordinate.

5.3 Calibration with Erdõs-Renyi equivalents

We investigate fundamental graph features of the obtained graph time-series by means of *graph calibration*. To this end, we fix parameters (w, Δ) where w is the time-window size and Δ is the time-step used to build the graph time-series. For a given graph feature $A(\cdot)$, we compute the ratio over the i-th time-window as

$$r_A^i := \frac{A((G_i)_{\mathrm{eq}})}{A(G_i)},$$

where G_i is the graph living in the i-th time-window and $(G_i)_{\mathrm{eq}}$ is an Erdõs-Renyi equivalent of G_i. Concretely, we take the graph characteristic $A(\cdot)$ to be

1. the global network efficiency,
2. the clustering coefficient,
3. the average degree,
4. the global efficiency of the largest component.

Calibration computations are shown in Figure **??**. In terms of average degree $\bar{d}(\cdot)$ of the network, the ratio $r_{\bar{d}}^i$ oscillates around the baseline $y = 1$ with amplitude $A \leq 0.2$. Hence, from this point of view we see that the data time-window graphs have an Erdõs-Renyi-like behavior. Now, we observe an interesting phenomenon that characterizes the way the pandemic spreads over time. Let $E(\cdot)$ and $E^C(\cdot)$ respectively denote the global efficiency of the network and the global efficiency of the largest connected component in the network. Then, the ratio r_E^i explodes around index $i = 30$, where $|G_i| = 8399$, whereas the ratio $r_{E^C}^i$ is significantly lower than one for all indices i. This altogether indicates that a peak of COVID-19 spreading (around time-window $i = 30$) leads to the generating of many connected components with very high within-global efficiency, whereas as a whole the global efficiency is a lot lower than for an Erdõs-Renyi model.

Finally, we observe an interesting phenomenon with the clustering coefficient C. This quantity represents the probability, for a given graph, that two nodes having a common neighbor are adjacent. It gives a measure of how much a network is modularisable. For small time-window graphs, e.g. for early time-windows, we observe an Erdõs-Renyi-like behavior in the sens that the ratio of clustering coefficients do not present extreme values. Instead, around the peak at time-window $i = 30$, the inverse ratio $\frac{1}{r_C^i}$ explodes as

$$\frac{1}{r_C^{30}} = \frac{C(G_{30})}{C((G_{30})_{\mathrm{eq}})} = 780.7.$$

Roughly speaking, this means when the spreading of the pandemic is maximal, the data time-window graph organizes itself in clusters with an extremely high modularity.

5.4 Calibration with Fixed Degree Membership Models

Section 5.3 shows evidence of a powerful clustering choice of the data graphs, for which each cluster has a high information-traveling speed. This observation motivates us to try fitting our data graphs with graphs that represent a more heterogeneous degree distribution than Erdõs-Renyi models and we choose the class of fixed degree membership models (FDM models). Furthermore, we evaluate the result of FDM fitting by plotting the time-evolution of graph calibration ratios, as in Section 5.3.

Regarding the clustering coefficient of time-window graphs, the fixed degree membership modeling shows a much higher accuracy than the Erdõs-Renyi models. That said, the data-built time-window graphs still present a much higher clustering capacity, attaining a maximal inverse ratio at time-window $i = 30$ (Figure 9, first plot):

$$\frac{1}{r_C^{30}} = \frac{C(G_{30})}{C((G_{30})_{\text{eq}})} \simeq 340.$$

Now, we further observe that the fixed-degree membership models allow for a much higher performance fitting with respect to the global efficiency coefficient of data time-window graphs. Indeed, we obtain a bounding condition (Figure 9, second plot):

$$1 \leq r_E^i \leq 4.2 \; \forall i = 1, ..., 50.$$

The ratios of the global efficiency of the largest component coefficients present a behavior that is similar to the context of Erdõs-Renyi modeling. Indeed, we observe an oscillating ratio with three major local minima at time-windows $i = 5, 19, 30$ and bounding condition (Figure 9, third plot):

$$0.1 \leq r_{EC}^i \leq 1.2 \; \forall i = 1, ..., 50.$$

5.5 Zigzag Persistent Homology

We present some results about zigzag homology in Figures 12, 13 and 14. More precisely, we plot the zigzag persistence diagrams in dimensions $d = 1, 2$ for three different parameter choices $(w, \Delta) \in \{(7, 4), (14, 7), (28, 14)\}$. For each fixed choice of parameters, connected components as well as cycles both present a very regular pattern, in the sense that their coordinates are distributed along the diagonal, with an almost constant persistence. For any choice of parameters, connected component representatives (resp. cycles) persist for approximately 1 to 6 periods (resp. 1 to 3 periods). Here, a period corresponds to Δ days.

6 Conclusion

Any data analysis of contact tracing data is to be handled carefully, as the reporting process on which the data collecting is based presents a large variety of incomplete recordings. We process the data and tackle the problem of inaccurate reporting by inferring a hypothesis on the spreading of the pandemic. This leads to a method we call *edge imputation*. The spreading of the pandemic can be modeled via graphs whose edges correspond to a potential contamination. We generate various mathematical objects induced by this graph representation, such as time-series of overlapping simplicial complexes. Now, this allows for different ways of analysing the data. In particular, we make use of exploratory data analysis, basic graph theory techniques, as well as topological data analysis tools such as persistent homology and zigzag persistent homology.

We observe tree-like spreading within the global data-graph G, which presents a high clustering coefficient. More precisely, we have almost disconnected clusters - accounting for the incomplete nature of the reporting process - with very high intra-cluster efficiency coefficient. The tree-like spreading leads to the formation of only very few closed paths, which happen to most often be triangles.

An important part of the analysis is drawn from comparing the data-produced graphs with random equivalent models. We conclude from the graph calibration on Erdõs-Renyi and fixed-degree membership models that the data graphs present complex non-random structure. In particular, we observe a high modularity in the global data graph G as it follows a division into clusters being mutually almost disconnected and having a very significant intra-cluster information travelling speed. Graph calibration underlines the possibility of fitting the global data graph G with a stochastic block model. Now, a part of the comparison analysis consists of observing the links and differences between the persistence diagrams of both the global graph's natural filtration and an Erdõs-Renyi equivalent filtration. [...]

It is worth wile mentioning that the main limit of this type of contact tracing analysis stems from the uncertainty lying within the data acquisition procedure. Indeed, on the one hand, contact tracing reports cannot be taken as reliable and on the other hand, tests can also be unreliable, being false-negatives or false-positives. Furthermore, it is not possible to evaluate the accuracy of most results, as there is no ground truth data to compare with. Finally, we note that the use of topological data analysis is not the best analytical fit when facing the contact tracing data set. Indeed, the patterns produced by the spreading of the pandemic form almost disconnected patches of tree-like structures with very few 2-cells and cycles, and this means the data contains very little topological information.

References

[1] Gunnar Carlsson. *Persistent Homology and Applied Homotopy Theory*. 2020. arXiv: 2004.00738 [math.AT].

[2] Gunnar Carlsson and Vin de Silva. *Zigzag Persistence*. 2008. arXiv: 0812.0197 [cs.CG].

[3] Frédéric Chazal and Bertrand Michel. *An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists*. 2021. arXiv: 1710.04019 [math.ST].

[4] Sam Spencer and Lav R. Varshney. *Social Bubbles and Superspreaders: Source Identification for Contagion Processes on Hypertrees*. 2020. arXiv: 2010.11350 [eess.SP].

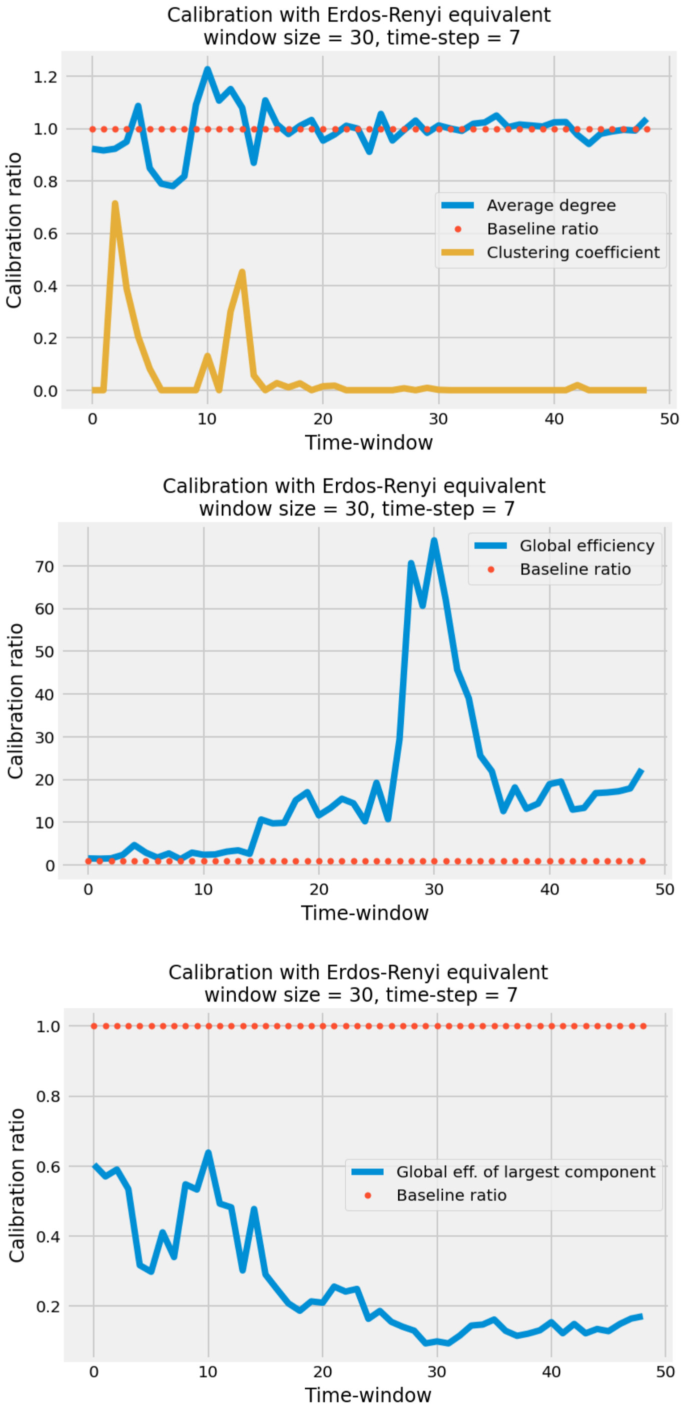


Figure 9: Graph calibration with an Erdos-Renyi equivalent time-series.

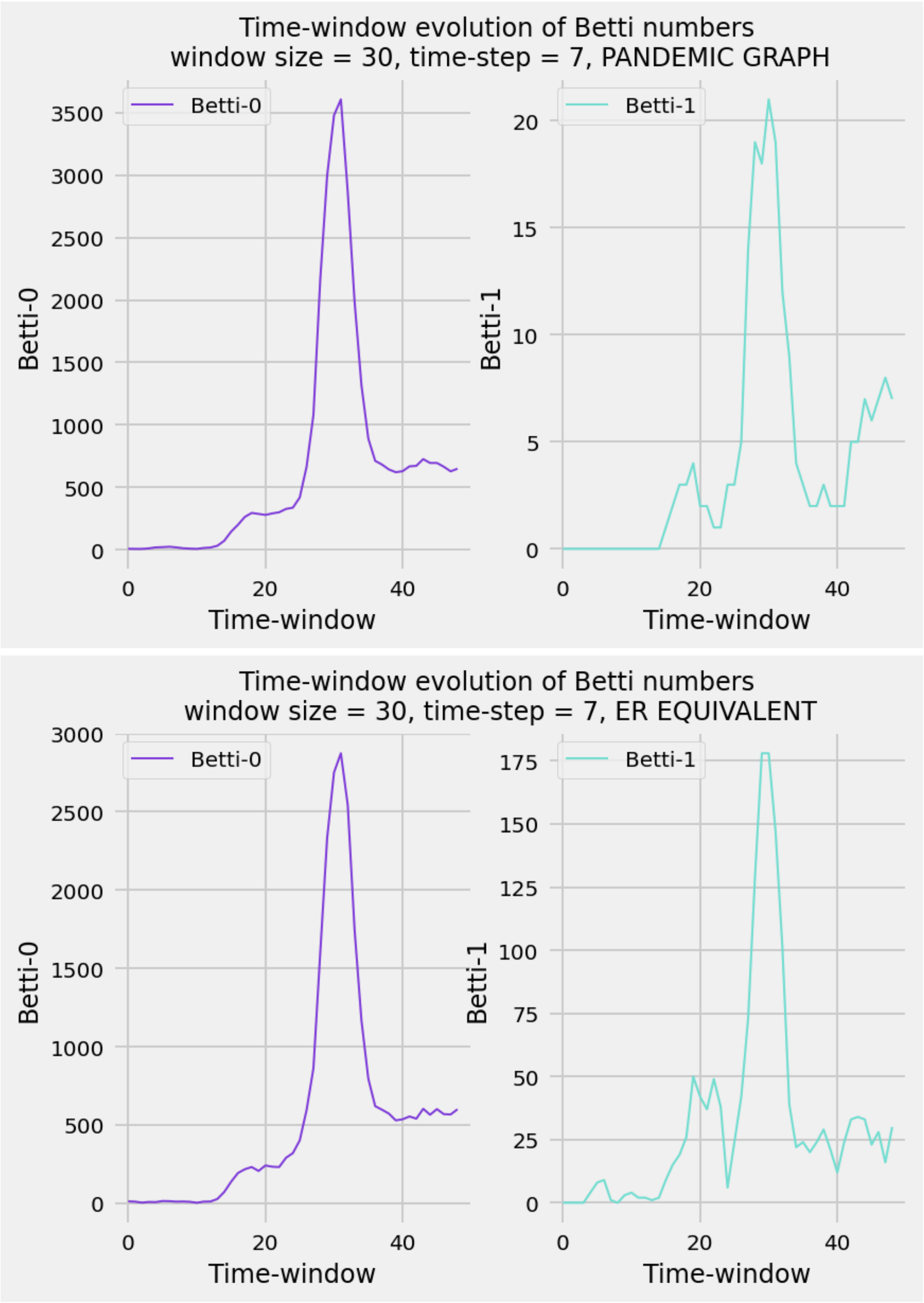
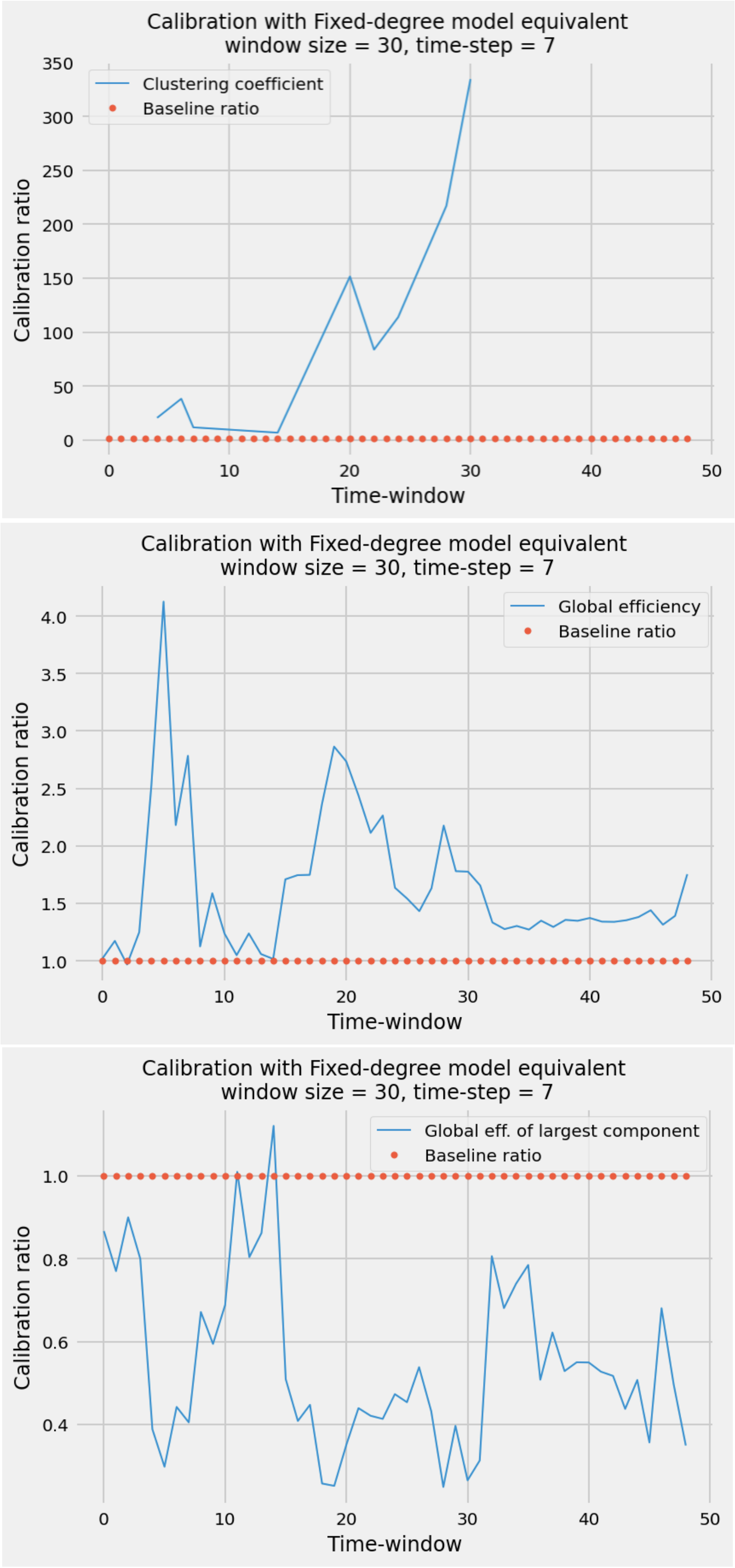


Figure 10: Time-window evolution of Betti numbers. We compare the pandemic graph time-series (with $(w, \Delta) = (30, 7)$) and the Erdõs-Renyi equivalent time-series.



Calibration with Fixed-degree model equivalent
window size = 30, time-step = 7

Calibration with Fixed-degree model equivalent
window size = 30, time-step = 7

Calibration with Fixed-degree model equivalent
window size = 30, time-step = 7

Figure 11: Graph calibration with a fixed-degree equivalent time-series.

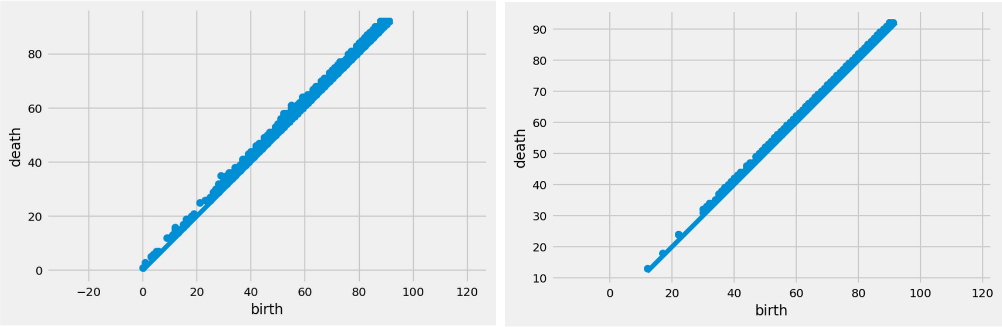


Figure 12: Persistence diagrams of zigzag homology on graph time-series with window size $w = 7$ and step size $\Delta = 4$. Left : dimension-0, right : dimension-1.

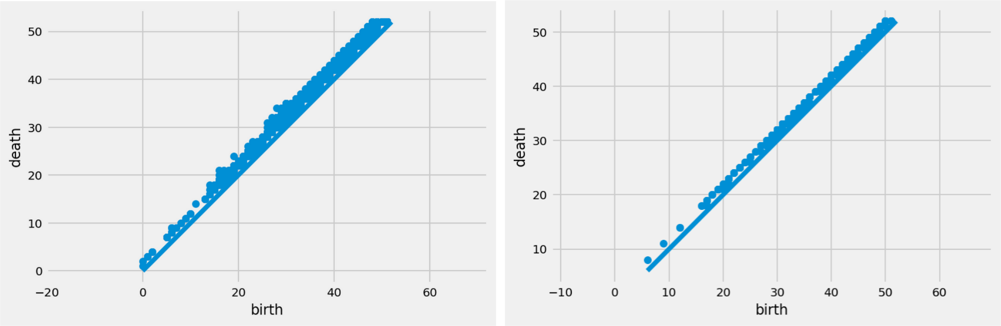


Figure 13: Persistence diagrams of zigzag homology on graph time-series with window size $w = 14$ and step size $\Delta = 7$. Left : dimension-0, right : dimension-1.

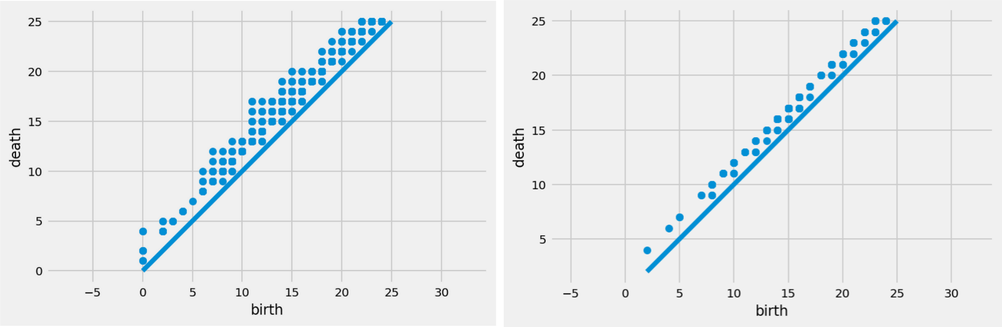


Figure 14: Persistence diagrams of zigzag homology on graph time-series with window size $w = 28$ and step size $\Delta = 14$. Left : dimension-0, right : dimension-1.