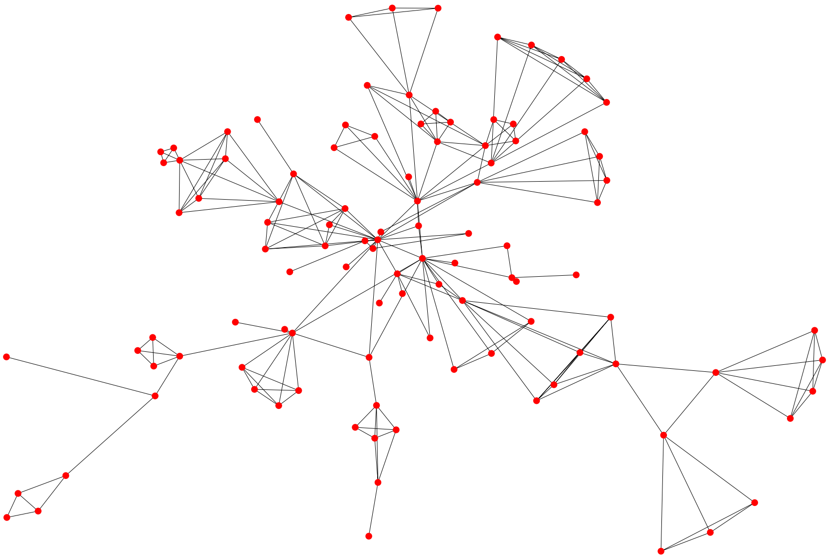
EPSRC
Engineering and Physical Sciences
Research Council

InFoMM
Industrially Focused
Mathematical Modelling

EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling



Understanding the Social Customer Using Topological Data Analysis

Ka Man (Ambrose) Yim

UNIVERSITY OF
OXFORD

Emirates

Contents

1 Introduction

A network describes pairwise relationships between objects, such as friendship between individuals.

Many systems in society and the physical world are organised into networks of interconnected elements, be it social networks such as Facebook, the electric grid, or the complex circuitary in our brain. In many of these systems, the network is vast, complicated, and messy. We would like to distill the information contained in a network into simple descriptors that are interpretable and informative. **Topological data analysis** (TDA) is one such means of distillation, describing the 'shape of data' using concepts from algebraic topology.

Our interest is in exploring the passenger social network in the Great Britain database of Emirates. The passenger social network is constructed using flight booking data. If passengers book flights together, they know each other. Thus we can draw edges between passengers on the same booking. This network is dynamical; as time goes by, new edges are formed between passengers, and a set of passengers may fly with Emirates multiple times. Hence given two customers a (for Alice) and b (for Bob), we can assign a weight $w_{ab}(t)$ that counts the number of bookings betwen Alice and Bob up to time t.

Our aim is to use TDA to analyse how a collection of passengers form connections with each other and create clusters. We are interested in clusters of frequent fliers since they represent brand loyalty on a *collective* and social level, which is more valuable than brand loyalty exhibited by isolated frequent fliers. We focus on nine networks in the Emirates data set, each consisting of about 100 passengers who flew with Emirates between 2012 and 2017 with no prior record with Emirates before 2012. We then use TDA to extract topological information from the network and use them to compare and contrast the nine communities of passengers.

Glossary of terms

- **Simplex (of passengers):** a collection of $(n+1)$ passengers form an n-simplex if each pair of passengers in this collection have flown together at least n times.
- **Simplicial Complex:** a collection of simplices. If a simplex σ is in the simplicial complex, then all sub-collections of passengers in σ are also in the simplicial complex.
- **Filtered Simplicial Complex:** an evolving simplicial complex. As time evolves, simplices are added to the simplicial complex. Simplices cannot be removed from a filtered simplicial complex during its evolution.
- **Boundary:** The boundary of an n-simplex is the collection of $(n-1)$-simplices, or subsets of n passengers, in the n-simplex. For example, the boundary of a 2-simplex of three passengers are the three unique pairs of passengers in the 2-simplex.
- **Cycle:** an n-cycle is a collection of n-simplices that has no 'end'. For example, the three unique pairs of passengers that can be made from three passengers form a graph that starts and ends in the same place; hence the three pairs form a 1-cycle.
- **Homology:** the k^{th} homology is the set of k-cycles in a simplicial complex, excluding cycles that are boundaries of some collection of $(k+1)$-simplices.
- **Persistent Homology:** A record of the birth and death of cycles in a filtered simplicial complex as time evolves.

2 Topological Data Analysis

Topological data analysis (TDA) is the process of computing persistent homology on a set of time-dependent combinatorial data, such as a time-evolving network. We summarise the TDA pipeline in Figure 1.

A simplicial complex encodes relationships between any number of individuals, generalising the notion of a network.

The Filtered Simplicial Complex

The first step in TDA is to assemble the data into a **filtered simplicial complex**. An abstract simplicial complex, or simplicial complex, is a generalised version of a network.

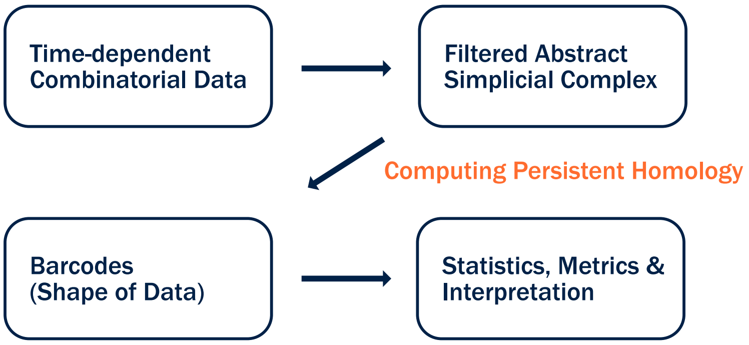

Time-dependent Combinatorial Data
Filtered Abstract Simplicial Complex
Computing Persistent Homology
Barcodes (Shape of Data)
Statistics, Metrics & Interpretation

Figure 1 – The TDA pipeline. Firstly, we organise the data into a filtered simplicial complex. We then compute the persistent homology of the filtered simplicial complex and produce a barcode that we can interpret and analyse.

A network only describes relationships between pairs of nodes, whereas a simplicial complex is a collection of relationships that can occur between any number of nodes. We call a relationship a simplex. For example, take three passengers, Alice (a), Bob (b) and Charlie (c). We can construct different relationships between them:

- A 0-simplex, describing one individual, is a node, e.g. $\langle a \rangle$, $\langle b \rangle$, $\langle c \rangle$;
- A 1-simplex, describing pairwise relationships, is an edge, e.g. $\langle ab \rangle$, $\langle ac \rangle$, $\langle bc \rangle$;
- A 2-simplex describing triadic relationships, e.g. $\langle abc \rangle$.

If an n-simplex between $n+1$ individuals exists in the simplicial complex, then all combinations of relationships between the $n+1$ people in the simplex must also exist in the simplicial complex. For example, if there is a relationship between Alice, Bob and Charlie $\langle abc \rangle$, then Alice and Bob $\langle ab \rangle$, Alice and Charlie $\langle ac \rangle$ and Bob and Charlie $\langle bc \rangle$ must also have a relationship with one another respectively and thus exist in the simplicial complex. The people involved in the relationship $\langle a \rangle$, $\langle b \rangle$ and $\langle c \rangle$ must also exist. In this case, the simplicial complex is

$$\{\langle abc \rangle, \langle ab \rangle, \langle ac \rangle, \langle bc \rangle, \langle a \rangle, \langle b \rangle, \langle c \rangle\}.$$

If Bob does not have a relationship with Charlie, then we take the relationship $\langle bc \rangle$ out of the complex. Since Alice, Bob and Charlie cannot have a collective triadic relationship if Bob and Charlie does not have a relationship with each other, we also need to remove the $\langle abc \rangle$ from the complex. Thus the simplicial complex in this case is simply

$$\{\langle ab \rangle, \langle ac \rangle, \langle a \rangle, \langle b \rangle, \langle c \rangle\}.$$

which reduces to a network with only pairwise relationships.

We are free to define what we mean by a relationship in the context of the data. In our application, we define an n-simplex to be a collection of $n+1$ passengers, where each pair of passengers in the simplex have shared at least n bookings between them. In other words, the edge weight between the individuals of the network is at least n. For example, Alice, Bob and Charlie is a 2-simplex if $w_{ab} \geq 2$, $w_{ac} \geq 2$ and $w_{bc} \geq 2$. An n-simplex is thus a cluster of $n+1$ individuals who have demonstrated a collective loyalty to Emirates.

As bookings accumulate over time, the simplicial complex evolves. Initially, we assume that all passengers are disconnected. Edges are then formed between passengers, and as the same set of passengers fly repeatedly with Emirates, they emerge as clusters. The series of simplicial complexes evolving over time is called a **Filtered Simplicial Complex**, or simply a filtration. In our data, time is a natural parameter over which we evolve the filtration. We show an example in Figure 2. At $t=2$, a booking between the three passengers highlighted in red increments the weights of the edges between them by one and adds an extra edge to the simplicial complex compared with $t = 1$. At $t = 3$, a booking between the two passengers highlighted in red increments the edge between them by 1. As a result, this introduces a 2-simplex as we have three passengers who have travelled with each other at least twice as pairs.

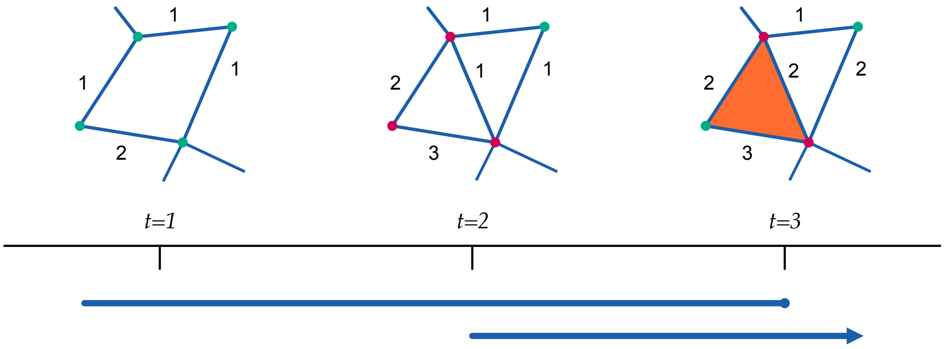


Figure 2 – Top: a filtered simplicial complex of passengers. The weight associated to the edge between a pair of passengers is the number of times the pair have booked together. Edge weights do not decrease with time. A two-simplex between three collectively loyal customers, such as the orange triangle in the diagram, closes a hole, or 1-cycle, in the simplicial complex. Below: we can parametrise the birth and death of such holes with a barcode. Each bar represents an independent 1-cycle. The edge added at $t = 2$ splits an existing cycle into two, creating an extra bar in the barcode. The death of a cycle at $t = 3$ terminates one of the bars. The remaining *persistent* cycle is represented by a bar with an arrow tip.

Persistent Homology

In simple terms, the **homology** groups of a simplicial complex represent the collection of connected components, loops, cavities or higher dimensional 'holes' in a simplicial complex. Consider the simplicial complex shown in Figure 3. The simplicial complex has two components that are not connected with each other, a one dimensional hole enclosed by 1-simplices, and a two dimensional cavity enclosed by 2-simplices in the hollow tetrahedron on the left hand side. In the language of homology, we say the 0th homology group H_0 (describing the connected components) has two elements, the 1st homology group H_1 (describing the one-dimensional loops) has one element, and the 2nd homology group H_2 (describing the two dimensiona cavities) also has one element.

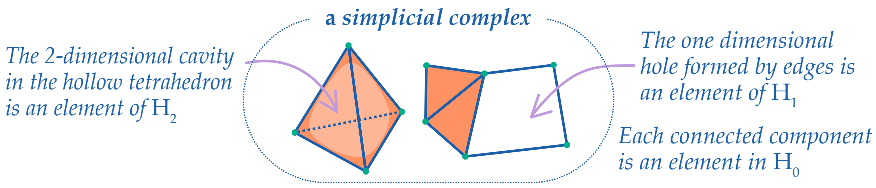

a simplicial complex
The 2-dimensional cavity in the hollow tetrahedron is an element of H_2
The one dimensional hole formed by edges is an element of H_1
Each connected component is an element in H_0

Figure 3 – A simplicial complex with two H_0 elements, one H_1 element and one H_2 element. The shaded orange triangles represent 2-simplices between three passengers.

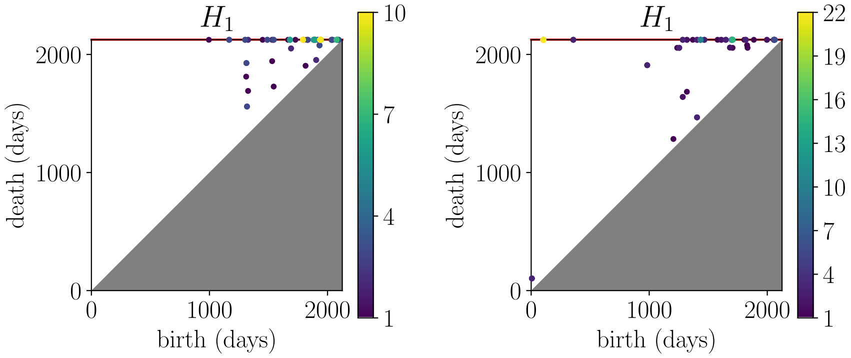
Persistent homology describes how these homologies appear and disappear over time as the simplicial complex evolves in the filtration. We can represent the persistent homology of a filtered simplicial complex as a **barcode**. We show the H_1 barcode of a filtered simplicial complex at the bottom of Figure 2. At $t = 2$, the additional edge splits an existing 1-cycle into two, creating two independent 1-cycles. At $t = 3$, the additional two-simplex, represented by the orange triangle in the figure, kills one of the 1-cycles in the simplicial complex. This series of events is recorded in the H_1 barcode below the filtered simplicial complex. Each bar in the H_1 barcode represents an independent 1-cycle. A new

Persistent homology summarises changes in the 'shape' of data as time evolves.

Persistence diagrams provide a succinct summary of the changes in the topology of the simplicial complex in the filtration.

bar emerges at $t = 2$ when a new 1-cycle is created and a bar is terminated at $t = 3$ when a 1-cycle dies. The surviving 1-cycle, denoted by a bar with an arrow tip, is said to *persist*.

Using persistent homology, we parametrise the evolution of combinatorial data, the simplicial complex, as a collection of points (b, d) representing the birth and death of bars in the barcode; therein lies the real power of persistent homology. Bars that persist are assigned a death time $d = \infty$. In figures 4a and 4b, we plot the birth and death of cycles in networks 3 and 4 respectively. These plots are called persistence diagrams.



(a) Network 3

(b) Network 4

Figure 4 – **H_1 persistence diagrams for networks 3 and 4 respectively. Points lying on the top red line at 'infinity' represent cycles that do not close. Points close to the diagonal have a short lifetime. Each site on the diagram may be occupied by multiple points, the multiplicity being indicated by the colour of the point.**

Statistics on Persistance Diagrams

Given the persistence diagrams of a network, we propose two measures that extract specific information about the evolution of the network's structure and passenger relationships. We note that adding an edge either joins two disjoint connected components, or connects nodes within the same component. The first measure tracks the tendency of a network to form cycles or grow tree-like branches as time goes by. We would like to compare these two mechanisms. If the first mechanism dominates, then the main component of the simplicial complex is tree-like, reaching out to connect more vertices as time goes by. However, if the second mechanism dominates, the complex folds on itself and reduces the distance between connected passengers. In the first case, the complex becomes more extensive; in the second case, it becomes more compact. We can use data from the barcodes to help us quantify these competing behaviours. Edges added between disjoint connected components cause H_0 bars to terminate and those added between passengers in the same component case create H_1 cycles. We can measure, at each time step t,

$$\eta(t) = (\#\, H_1 \text{ births before } t \;-\; \#\, H_0 \text{ deaths before } t) \quad \text{per node.}$$

Networks with lower η values are considered to be more tree like than those with higher values.

Another measure we can extract from the persistence diagrams is the number of 2-simplices created at each time step. We recall that a 2-simplex is a collection of three people with a high collective loyalty to Emirates. 2-simplices that are created early in the filtration are the trios of passengers who book more frequently with each other and thus accumulate weights quicker to reach the threshold of two bookings between each pair than trios corresponding to 2-simplices that are created later in the filtration.

Networks with more 2-simplices show a greater collective loyalty. We can assign to each network a score for collective loyalty which we call z, which is the number of 2-simplices, weighted such that the 2-simplices formed later in the filtration contributes a smaller value to the score. Since each 2-simplex birth must lead to a cycle death or a void birth, we can simply perform a weighted count of the number H_1 deaths and H_2 births.

3 Results on Emirates data

We compute $\eta(t)$ and z for each of the networks in the Emirates data set. In Figure 5, we plot $\eta(t)$ for the nine networks in our dataset. While all networks show positive η values at the last time step, networks 1, 3, 4, 5 and 8 clearly have a larger value of η; moreover, they show a positive rate of increase in η, sustained over approximately 800 days, whereas the other networks show little sign of increase. We conclude that η is useful in distinguishing between the dynamics of different networks.

In Figure 6, we plot a box and whiskers diagram of the distribution of 2-simplex birth times for different networks and their z scores. We see that network 4 and 3 have the highest z scores since they have the most 2-simplices. In particular, the distribution of 2-simplex birth times of network 4 is biased towards earlier times, which indicates that bookings between passengers in a 2-simplex tend to accumulate at a faster rate compared to those in network 3.

We also observe that η and z contain different information about the dynamics of the network. While networks 1, 5 and 8 score highly in η (see Figure 5) after 2000 days, they are ranked some of the lowest in z, having no 2-simplices, suggesting while passengers in these networks are separated by small distances on the social network, none of them fly frequently together as a collective. On the other hand, networks 3 and 4 are ranked high in the z score and have a high η value and rate of increase. This suggests that there are some collectives of individuals in networks 3 and 4 with a special affinity for Emirates, with plenty of mutual acquantainces who have also flown with Emirates, albeit less frequently.

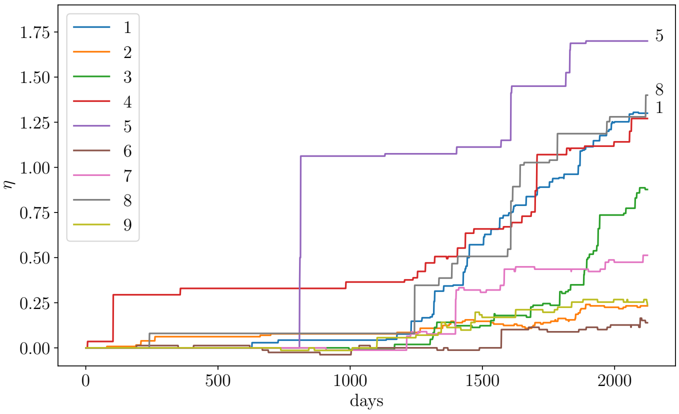
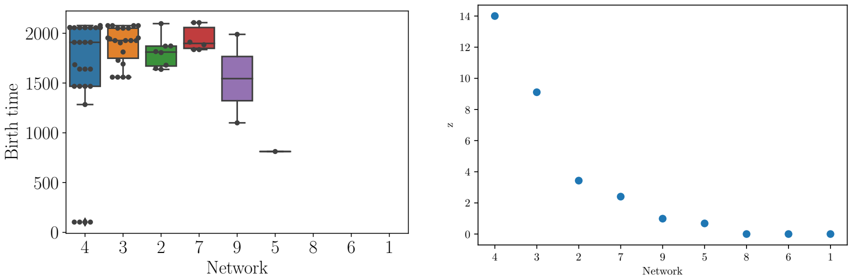


Figure 5 – η vs time of the filtration, which compares the number of edges responsible for linking previously disjoint connected components and the edges that connect passengers who are already connected in the same component. The coloured lines represent each network.

4 Discussion, conclusions, & recommendations

We have applied TDA to analyse nine passenger social networks in the Great Britain database of Emirates. Firstly, we reorganised the social networks into simplicial complexes, encoding cliques of passengers that show collective loyalty to Emirates as a simplex. We then computed the persistent homology of the evolving simplicial complexes, summarising the birth and death of cycles in the simplicial complexes as persistence diagrams. Analysing the persistence diagrams, we observe that networks 3 and 4 in the data merit special interest. In networks 3 and 4, new bookings tend to take place between passengers with mutual acquantainces and numerous passengers in those networks show collective loyalty to Emirates.

Insights from TDA point to a set of actions that could improve a customer's loyalty. Passengers adjacent to the collectives of loyal passengers in networks 3 and 4 may respond well to marketing strategies as they may be influenced by acquantainces who are deeply



(a) Each point marks the birth time of a 2-simplex in the given network. The box and whiskers plot for each network shows the median, quartiles, range and outliers of the distribution.

(b) The z score for each network. The z score is a measure of a component's collective loyalty. A higher z score means that it has a higher collective loyalty.

Figure 6 – Graphs showing the number of 2-simplices in a network and the time taken for 2-simplices to be created to generate a score z for the collective loyalty of passengers in a network. Networks 1, 6 and 8 have no 2-simplices and have score $z = 0$.

loyal to Emirates. Relationships with individuals who show collective loyalty should be cultivated. Given their high z values, networks 3 and 4 might represent organisations which have special incentives to fly with Emirates, and further data might be gathered to learn more about them as an organisation of people.

There is scope to extend our current body of work. We can transform persistence diagrams into 'vector representations' such as persistence landscapes and persistence images. These vector representations are amenable to analysis using traditional data analytics and machine learning algorithms. For example, if we are given a large collection of social networks, we can cluster them into different classes using vector representations of their persistence diagrams as inputs in clustering algorithms. This may drive investigations into developing different sales strategies for passengers in different classes of social networks. We can also compute the persistent local homology of a passenger, a variant of persistent homology which characterises the evolution of a passenger's connections with its neighbours in the social network. This would allow us to classify passengers based on the shape of their local neighbourhood, and target individuals who are influencers in their social network.

5 Potential impact

The field of topological data analysis is barely a decade old, and applications of TDA to network data is an even more recent development that has only started in earnest in the past few years. As a proof of concept, we have demonstrated that TDA has important contributions to furthering our understanding of time evolving network, due to its ability to summarise abstract network features as numbers that are interpretable and amenable to further analysis. Moreover, we have shown that TDA outputs can be translated into actionable business insights. This lends confidence to further research into applying TDA to network data in other aspects of business.

Dr Andrew Mellor, Oxford-Emirates Data Science Lab, said: "*Ambrose has done a deep-dive into topological data analysis, covering far more than just the traditional techniques. This has been extremely informative for the Lab as we have seen both where TDA gives interesting insight, and where it fails. Overall the project has gone very well, and we have learned new things about the social customer which can potentially be operationalised.* "