Detecting potential product segments using topological data analysis*

Avi Goldfarb, Jun Bum Kwon, and Trevor Snider
Rotman School of Management
University of Toronto

March 2016

Abstract

We introduce a method for identifying potentially related products using topological data analysis (TDA). From both simulated and real consumer purchase data, we show that "loopy segments" in TDA can connect regionally separated local products through national products, while standard clustering methods such as hierarchical clustering cannot. In addition to connections across locations, we also show that TDA can detect potentially co-purchased products between the salty snack and beer categories.

Keywords: Topological Data Analysis, Competitive analysis, Cluster analysis, Product segmentation

* Avi Goldfarb thanks SSHRC and AIMIA for research support. We thank Henry Adams, Donna Hoffman, and Daniel Ringel for helpful comments.

1. Introduction

Market structure analysis describes the relationships between brands and products in order to define the market (Elrod et al. 2002). Analysis of market structure is a key step in the design and development of new products, the repositioning of existing products, pricing, marketing communications, and marketing strategy (Srivastava, Alpert, and Shocker 1984; Urban, Johnson, and Hauser 1984; Kamakura and Rusell 1989; Urban and Hauser 1993; DeSarbo, Manrai, and Manrai 1993; Erdem and Keane 1996; Bergen and Peteraf 2002; Lattin, Carrol, and Green 2003; DeSarbo, Grewal, and Wind 2006). Until very recently, the bulk of published work focused on competitive market structure with a limited number of products (Erdem 1996; Cooper and Inoue 1996; DeSarbo and Grewal 2007; Kim, Albuquerque, and Bronnenberg 2011; Lee and Bradlow 2011).

In the last few years, new methods have arisen to identify and visualize market structure with many products. These new methods are a response to two developments. First, the variety of products in the marketplace has increased (Ailawadi and Keller 2004), increasing demand for such methods. Second, faster computers and increasing digital storage capacity have broadened the set of potential tools to make sense of this variety, enabling the supply side. This has created renewed interest among marketing scholars in market structure and segmentation (Netzer et al. 2012; France and Ghose 2016; Ringel and Skiera 2016).

In this paper, we apply a new data analysis technique, Topological Data Analysis (TDA hereafter, Carlsson 2009), to the problem of market structure segmentation with many products. Standard clustering methods work well for distinctly grouped data (Figure 1a). However, as the number of data points rises, the data set becomes more connected. One particular example of such connected data is a loopy segment (Figure 1b), where products locate closely together with their neighboring products but are indirectly connected to, and seemingly far apart from, some other products. TDA is particularly well-suited to identifying such segments.

Loopy segments can occur in analyzing national level market structure with customer level data on purchases. In many cases, not all products are available in each local market, and thus there can be no common customers among some related products. For example, suppose that a manufacturer launched

products W and M in Wisconsin and Massachusetts, respectively. Suppose that products W and M serve similar types of consumers in the different markets. No consumer can purchase both products due to local availability. Instead, some consumers purchase the local one. These consumers also purchase other products that are available in both markets. As a result, products W and M can be in the same segment, connected through the products that are available nationally, although no consumer purchases both W and M. This same framework can also connect products sold at different stores. The indirect connection between W and M will be identified in topological data analysis through a loopy segment.

Why is it useful to detect loopy segments? As described in the preceding paragraph, a loopy segment can include products that occupy the same product space in different markets, but that no consumer purchased together. If preferences are transitive, in the sense that if objects share a relationship to a common object, then they would be related if they were in the same domain, then loopy segments can help firms identify potentially competing or potentially co-purchased products that are not currently offered in the same market. This can help manufacturers who launch their products sequentially across regional markets. They can learn about (1) competitor products in one market and (2) indirectly connected products in the other market, and they can use that information to inform opportunities in both markets. This also helps retailers with limited shelf space to detect related products. For example, Costco and Walmart strategically keep a small number of products in each category. By identifying potentially related products, they can make better product assortment decisions.

Standard clustering methods such as hierarchical clustering are not good at identifying indirect connections such as those found in a loopy segment (Lesnick 2013). Recently developed community detection methods are particularly useful for segmenting many observations (Newman and Girvan 2004; Clauset, Newman, and Moore 2004; Pons and Latapy 2005; Raghavan, Albert, and Kumara 2007; Blondel et al. 2008). However, our simulations suggest that community detection methods are less effective than TDA at identifying product connections across markets because no community detection method assigns a product into multiple segments, unlike TDA. Thus, our results suggest that for the particular problem of identifying connected products in unconnected markets, TDA is a useful new tool.

Topology is a mathematical discipline that studies shape. TDA, developed by computational mathematician Gunnar Carlsson (2009), refers to the adaptation of this discipline to analyzing highly complex data (Ayasdi 2015). TDA assumes that all data has shape and shape has meaning, and thus tries to discover geometric relationships among data points. There are many applications across oncology, astronomy, neuroscience, image processing, and biophysics. Hoffman and Novak (2015) argue that TDA is useful in organizing potential applications of the 'internet of things'. There has been some commercialization efforts by analytics company Ayasdi (which counts Carlsson as one of its founders). For example, TDA analysis has helped identify new patient groups in breast cancer treatment, distinct playing styles of National Basketball Association players, and voting patterns of the members of the US House of Representatives (Lum et al. 2013). Ayasdi's website also discusses potential marketing applications in customer segmentation, personalized marketing, churn analysis, and network optimization (Ayasdi 2016).

We find that TDA is particularly well-suited to a specific marketing problem. We use simulated data and the IRI marketing academic data set (Bronnenberg, Kruger, and Mela 2008) to demonstrate that TDA can connect products in different markets through national products, while standard hierarchical clustering methods and community detection methods have difficulty. Our analysis of beer and salty snack buyers in Pittsfield Massachusetts and Eau Claire Wisconsin shows that different locally popular brands appear to occupy similar product space in the different markets. For example, in salty snacks, two national salty snacks (Rold Gold and Tostitos) connect local segments, which include two products that sell well in Wisconsin (Barrel O Fun and Jays) and a product that sells well in Massachusetts (UTZ). This suggests that the positioning of UTZ in Massachusetts is similar to the positioning of Barrel O Fun and Jays in Wisconsin. We also find potential co-purchase behavior between certain beer and salty snack products. While the two product categories are quite separated when using hierarchical clustering, many TDA segments include both beers and salty snacks.

We view the core contribution of this paper as introducing TDA methods to marketing by providing a clear marketing application. This adds a new clustering tool to the rapidly growing literature on market structure analysis using big data (France and Ghose 2016; Ringel and Skiera 2016). We view TDA as a

useful exploratory new tool and we highlight a specific strength of this tool. It should not be viewed as a replacement for other forms of product segmentation because it is unlikely to outperform those methods for standard product segmentation purposes.

Because TDA is a new method to marketing, section 2 uses several simple examples to provide an extensive discussion of the intuition behind TDA. Section 3 shows the usefulness of TDA for connecting similar products in separated markets using a simulation study, comparing TDA to other clustering tools. Section 4 applies the method to the IRI data to demonstrate its practical application in marketing. Section 5 concludes with a discussion of opportunities and limitations.

2. TDA methodology

Computing topology based on simplical complexes has been well understood decades (for more details, see Armstrong 1983, Edelsbrunner, Letscher, and Zomorodian 2002, Hatcher 2002, Zomorodian and Carlsson 2005, Edelsbrunner and Harer 2010, and especially Carlsson 2009). However, computing simplical complexes is resource intensive and so TDA had limited application until recently (Lum et al. 2013). Below, we provide a description of the TDA methodology.

2.1. Vietoris-Rips Complex

We use the most common and easily implemented TDA method, the Vietoris-Rips complex. Let $d(\cdot,\cdot)$ denotes the distance between two product points in customer purchase quantity space X. The complex $\mathrm{VR}(X, v)$ is defined as

- A set of vertices (data points or 0-simplices) is defined as X
- For vertices (data points) q and r, a line (edge or 1-simplex) $[qr]$ is included in $\mathrm{VR}(X, v)$ if $d(q, r) \leq v$
- A higher dimensional (k>1) simplex such as a triangular face (2-simplex) or a tetrahedron (3-simplex) is included in $\mathrm{VR}(X, v)$ if all of the lines (1-simplices) that make up the high dimensional simplex are in $\mathrm{VR}(X, v)$.

All the points within a simplex are directly connected each other. Given that our goal in this study is to find potentially related products, the simplex itself does not include indirectly connected products through other products. Next, because $\mathrm{VR}(X, v)$ includes a set of k-simplices $[x_0, x_1, \ldots x_k]$, where $x_i \in X$, at filtration value v, it is also called a filtered simplicial complex. Note that the complex $\mathrm{VR}(X, v)$ grows in filtration value v. In other words, data points are connected from their nearest neighbor to more distant ones. Lesnick (2013) label this the "thickening" process.

To enhance the formal description and explain how it works with marketing data, we provide several example cases on how TDA creates clusters of products based on purchases by two or three sample customers. Figure 2 presents the 9 different cases and Table 1 summarizes the key TDA output for each: filtration values, VR complexes, and Betti numbers which we define below.

2.2. Clustering distinctly grouped data (Cases 1 and 2)

Case 1 illustrates how TDA segments distinctly grouped data. The data, or vertex set, X consists of four products. Customer 1 purchased 0, 1, 5, and 5 units of products *a*, *b*, *c*, and *d* respectively. Customer 2 purchased 1, 1, 1, and 2 units. The points are plotted in the graph labeled *v*=0 (Data). At filtration value *v*=0, no product pair is connected yet, and thus $\mathrm{VR}(X, v = 0)$ includes only four data points $x_0 = \{a, b, c, d\}$. In the thickening process, we gradually increase filtration value by 0.01. The graph labeled *v*=1 (*Betti0*=2) shows that at v=1, we can now connect two groups of dots within the circles to generate two lines $x_1 = \{ab, cd\}$. These two groups are maintained until *v*=4, when all four dots become connected.

Case 2 provides a similar example. Product *d* is purchased more by both customers and has a distinct positioning. At *v*=1, products *a*, *b*, and *c* become one body, Then, at *v*=3.61, product *d* joins the others. It suggests that there are two product segments in Case 2. In Cases 1 and 2, the linking process of TDA is similar with that of standard hierarchical clustering.

2.3. Homology groups, Betti numbers, and loopy segments

Now, we show how TDA summarizes the shape of data. As described above in Cases 1 and 2, TDA generates a simplex (e.g. a line or a triangular face) by connecting data pairs which locate within filtration value v. The filtered simplex complex $\mathrm{VR}(X, v)$ can be summarized by what are labeled homology groups. The value $Betti_h$, where $h \in \mathbb{N}$, counts the number of h-th homology groups in the topological space, which is $\mathrm{VR}(X, v)$ here. "*Betti* numbers" was coined by Poincaré (1894) after Enrico Betti. The meaning of $Betti_h$,where $h \in \{0,1,2\}$, is as follows.

- $Betti_0$: the number of connected components
- $Betti_1$: the number of holes or loops
- $Betti_2$: the number of voids or cavities

It is possible to define higher $Betti_h$ numbers,where $h > 2$, but they are difficult to conceptualize and do not appear to matter in our empirical context. Thus we use up to $Betti_2$ in our study.

To provide examples of connected components and loops, the TDA literature often uses the shape of upper case letters. The letters that are qualified as $Betti_1$ with a single loop (and a single hole) are {A, R, D, O, P, Q}. In contrast, {B} is $Betti_1$ with two loops. All the other upper case letters have no loops, and can be thought of as a point if compressed.

A torus (or empty donut shape) is an example of $Betti_2$. A torus has a void inside of the donut as well as two loops: one with a hole in the center and the other with a hole inside the donut.

For each case in Figure 2, the bar graph shows the number of segments by *Betti* type for each filtration value v. For example, for Case 1, for $Betti_0$ (*Betti* dimension 0), we have four distinct groups until v=1. After v=1, there are two groups until v=4 when there is just one group as all the dots are joined together. Similarly, for Case 2, for $Betti_0$ we have four distinct groups until v=1, then two groups until v=3.61 and one group for $v \geq 3.61$. In this way, $Betti_0$ provides similar results to a standard clustering algorithm.

In contrast to $Betti_0$, both $Betti_1$ and $Betti_2$ count holes and voids, providing distinct insights into the data structure from a standard clustering algorithm. Following the literature, we label a hole or void as a

loopy segment in our paper. Given that we aim to detect related products that are not directly competing and that cannot be identified with standard clustering techniques, we focus on loopy segments ($Betti_1$ and $Betti_2$), where each product is indirectly connected with some others.

2.4. A loopy segment in a two dimensional plane (contrasting Cases 3 and 4)

In Cases 1 and 2, no hole exists. $Betti_1$ and $Betti_2$ are zero throughout. In particular, there is no empty space inside a simplex. Once all dots are connected in a triangle at a filtration value, the space within the triangle is covered. For example, in Case 2, when products *a* and *c* are connected at v=1.42, the inside of the triangle among products *a*, *b*, and *c* is shaded rather than blank because distance 1.42 covers all space within the triangle. This means that at least four products are necessary to form a hole in two-dimension space.

When can a loopy segment emerge in a two dimensional plane? The dots cannot be on a straight line (as in Case 1) and the diagonals should be longer than any of the four boundary lines. Case 2 does not form a loopy segment because products *a* and *c* are linked to each other before they link with product *d*.

Case 3, a square, does have a loopy segment. At v=0, there are four distinct dots and $Betti_0$=4. At v=2, lines can be drawn that connect the dots along the outside of the square and $Betti_0$=1. Importantly, the diagonals *{ac,bd}* are unconnected, meaning there is an unconnected simplex and so $Betti_1$=1. At v=2.83 the diagonals connect and there is no hole, and so for $v \geq 2.83$, $Betti_1$=0. The loopy segment suggests that the four products are indirectly related to their non-neighboring products because a grouping of size less than 2.83 shows no direct link between *b* and *d* or between *a* and *c*.

In contrast, Case 4 is a square with a dot in the middle: Product *e* is in the center of the other four products. Case 4 does not have a loopy segment. At v=1.42, four boundary products are connected with product *e* in the center, and so $Betti_0$=1 from this value. This linkage occurs before the boundaries are linked with each other, and so no hole is formed because product *e* made the diagonals shorter than the boundary lines. Case 4 shows that a hub structure, where one popular product competes with other products, is not likely to have a loopy segment.

2.5. Interval length of a loopy segment: Persistent homology (Cases 3 and 5)

When does the emerged hole disappear? Namely, how long does the hole persist? This is important to understand because more persistent holes suggest more robust connections that are distinct from standard clusters. Cases 3 and 5 provide a useful contrast for exploring persistent holes.

We now introduce a new concept, the *Betti interval*. The Betti interval describes how the homology of VR(X, v) changes with filtration value v. *Betti*$_1$ interval, with endpoints $[v_{start}, v_{end})$, corresponds to a hole that appears at v_{start}, remains open for $v_{start} \leq v < v_{end}$, and closes at v_{end}. The filtration range or the interval length, $v_{end} - v_{start}$, is the measure of *persistent homology*. Longer persistence suggests more robust features. In Case 3, at filtration value v=2.83, four simplex triangles {*abc, bcd, cda, dab*} arise when the additional two diagonals *ac* and *bd* fill in the square, and thus the hole disappears. In summary, the loopy segment is born at $v_{start} = 2$ and dies at $v_{end} = 2.83$, and thus its *Betti* interval has length (or filtration range) 0.83.

Case 5 presents a rectangle. At v=2, two product groups are formed and then they are maintained until v=4, suggesting that there are two segments in this case. At v=4, a loopy segment consisting of all four products emerges with *Betti*$_1$ interval $[4, 4.48)$. Compared to Case 3, this loopy segment is born later (4 > 2) and is less persistent (0.48 < 0.83). The later birth suggests that, relative to Case 3, in Case 5 the *Betti*$_0$ segments are more distinct and that the indirect connections might provide insights into potentially related products that standard cluster methods might miss. The lower persistence suggests that the loopy (*Betti*$_1$) segment in Case 5 is, however, a less robust feature of the data.

2.6. Connecting loopy segments (Cases 6 and 7)

Using Cases 6 and 7, we explain how TDA connects segments and where it provides distinct insights from standard hierarchical clustering. In Case 6, there are two clearly separated product groups: one often purchased by two customers and the other not. At v=1, three segments are formed and so $Betti_0$ changes from 8 to 3. At v=2, the separate segments ab and dc are joined and so $Betti_0$ drops to 2. The rectangle and the square are separated until v=2.83 and $Betti_0$ becomes 1. This process is similar to the way hierarchical clustering methods group items.

In addition to identifying distinct segments, and unlike hierarchical clustering, TDA informs us whether each segment has a loopy structure or not. There are two loopy segments in Case 6. For the square ($efgh$), the loopy segment has interval length 0.42, starting at 1 and ending at 1.42. For the rectangle ($abcd$), the loopy segment has interval length 0.24, starting at 2 and ending at 2.24. This suggests a different kind of connection between the points in the rectangle and the points in the square, as in the above comparison between Cases 3 and 4. The loopy segment is more meaningful in the square because it recognizes that the four dots are more equally connected.

Case 7 shows two loopy segments that are connected through a common product, d. At v=2, TDA generates one whole segment ($Betti_0 = 1$) with two loopy segments ($Betti_1 = 2$). These segments persist until v=2.83. Product d in Case 7 serves as a gate product. TDA connects segments by assigning such gate products into multiple segments. This connection information helps to detect potentially related products across neighboring segments.

Products a and e, which are indirectly connected through the gate product d, do not appear to be direct competitors. Nevertheless, the common linkage with product d suggests that a and e are related. As we describe below, if a and e are primarily sold in different markets, the common gate product suggests that they may serve similar needs in the different markets.

This connecting ability enables TDA to yield distinct insights relative to other clustering methods such as hierarchical clustering, which forces full separation. In Case 7, most hierarchical clustering algorithms such as average and complete linkage or Ward's method, generate two segments: one with

products a, b, c, and d, and another with products e, f, and g. Moreover, single linkage algorithms, where, at each step, combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other, put all products into just one segment because all the products has same distance with their neighboring product. Thus, while single linkage algorithms closely resemble TDA in terms of $Betti_0$ groupings, the single linkage algorithm misses $Betti_k, (k \geq 1)$ groupings. In the simulation section below, we conduct a more comprehensive comparison across several clustering methods.

2.7. Voids in three dimensional space (Cases 8 and 9)

Next, we show when voids occur in three dimensional space using examples with three customers. Case 8 shows an example with four products. The simplex in three dimensional space is a tetrahedron, which also has four data points. Therefore, Case 8 cannot contain a void. At v=5.66, all four products are connected each other, resulting in a tetrahedron as well as four triangular faces. Because both a tetrahedron and a triangle are simplices, neither a void nor a hole occurs.

In Case 9, there are six product points that if joined together would form an octahedron. At v=2.83, each point is connected with four neighboring points, each of which is in the center of its neighboring square side, thus $Betti_1$ switches from 6 to 1. For example, product a is connected with products b, c, e, and f, but not product d on the opposite side. As a result, there are four triangles {abc, abf, ace, aef} that include product a. There are another four triangles that include d but not a {dbc, dbf, dec, def}. Only these 8 triangles, and no tetragons, are in each plane. Since a triangle is simplex, no hole is formed. As a result, $Betti_1$ remains at 0.

However, there is one void ($Betti_2 = 1$) inside the 8 triangles starting at v=2.83. First, intuitively, one can see that each point of the six points is connected with the other point in the opposite side indirectly through their neighboring products. Three product pairs ad, be, and cf have such an indirect connections. Second, to make sure that the inside is empty, we check whether any tetrahedrons with four data points occur. For example, product a is connected with products b, c, e, and f. However, product b is not connected e in its opposite side. There is also no link between products c and f yet. Therefore, no tetrahedron occurs.

At v=4, the three product pairs {*ad*, *be*, *cf*} on opposite sides connect. Now, the inside is occupied by twelve tetrahedrons {*abcd, abce, abcf, abdf, abef, acde, acef, adef, bcde, bcdf, bedf, cedf*}, leading to $Betti_2 = 0$. The length of the interval with this void is 1.17, and the interval is [2.83, 4).

The above cases outline how TDA identifies indirect connections between products. Before we analyze real world data, we provide simulation evidence that TDA generates a different type of insight than other commonly used methods.

3. Simulation study

Our goal is to demonstrate that TDA can identify potentially related products that have not been sold together in the same market. Our target application is to cluster products in two local markets which are regionally separated. In the IRI data analysis below, we examine sales across two cities, Eau Claire Wisconsin and Pittsfield Massachusetts. We cluster salty snacks and beers separately to see whether TDA can find products that occupy the same product space within a category in the two local markets. Then, we combine both product categories in the same analysis to see whether TDA can also connect products in different categories and different markets that could potentially be purchased by the same customers. In other words, for this simulation to be useful to marketers, we assume that preferences are transitive and examine whether TDA can unpack the relationships in the data.

Before analyzing real consumer purchase data from the two local markets, we do a simulation study to examine whether TDA can recover useful loopy segments in such a setting. We compare results from TDA with those from hierarchical clustering methods and community detection methods. The purpose of this section is not to demonstrate that TDA is always superior to other methods. Instead, the purpose is to highlight a particular case in which TDA does detect a pattern in the data when other methods do not.

3.1. Simulation study procedure

Our simulation study has the following 5 steps, as shown in Figure 3a. In the simulation, some products appear only in Wisconsin (W), some products appear only in Massachusetts (M), and some national products appear in both markets (N).

Step 1: True segments

We simulate two scenarios, shown in Figure 3b. In Scenario 1, there are two loopy segments, one in each local market. Each local segment includes one national product as well as its own local product. This shape is called a "wedge sum" in topology. The two local segments are connected through one national product. In other words, the national product is assigned into both segments. In Scenario 2, we add one local-only segment into each local market.

Step 2: Correlation matrix

We simulate a correlation structure among products only within the same local market because we assume that the two local markets are geographically separated and so no consumer can purchase both groups (M and W) of local products. To generate the loopy segment, we put higher correlation between neighboring products. Higher correlation means shorter distance. For example, we give correlation 0.5 and 0.6 between Wisconsin local product W1 and it neighboring local and national products (W2 and N4), while we assign a correlation of 0.2 between W1 and its non-neighboring product W3.

Step 3: Simulating consumer purchases

Using the above correlation structure and assuming a marginal Poisson distribution, we simulate 10,000 consumers' purchases in each market (20,000 consumers total). We chose the Poisson distribution to reflect the discrete nature of purchase quantity. The quantity purchased by each consumer of each product is therefore a draw based on correlated (across products) Poisson random variables. To generate correlated

Poisson random variables, we utilized an R implementation by Barbiero and Ferrari (2014). We ensuare that no consumer can buy both M and W products.[1]

Step 4: Distance or similarity matrix

With consumer purchases for each product, we calculate Euclidean distance among products across the 20,000 consumers. This distance matrix becomes input data for TDA and hierarchical clustering. We also construct similarity matrix for community detection methods that counts each product pair's joint purchase frequency as the number of consumers who purchase both products among the 20,000 consumers.

Step 5: Product clustering

We create segments from this data using TDA, four different hierarchical cluster algorithms, and five different community detection methods. Hierarchical clustering methods are perhaps the most commonly used tool for segmenting and positioning products and brands (Srivastava, Leone, and Shocker 1981; Punj and Stewart 1983; DeSarbo and DeSoete 1984; Zhai, et al., 2011). The first three hierarchical clustering algorithms we use are single, average, and complete linkage, which Johnson (1967) defines as the "standard" hierarchical clustering algorithms. The fourth is Ward's method (Ward 1963) which is known for working particularly well with marketing data (Punj and Stewart 1983). Among them, the closest algorithm to TDA is single linkage, where, at each step, combining two clusters that contain the closest pair of elements not yet belonging to the same cluster.

Recently, community detection methods have been proposed as segmentation tools in network analysis. The five community detection algorithms we use in this study are those developed by Newman and Girvan (2004), Clauset, Newman and Moore (2004), Pons and Latapy (2005), Raghavan, Albert and

[1] To ensure the existence of a hole structure, we assign a lower mean value for the national product than for the local products. Recall that a national product is sold in two markets, while local products are only sold in one market. A higher mean value of the national product results by construction in a longer distance between the national and local products. As a result, if a national product has too high of a mean value, it may not be part of a loopy segment.

Kumara (2007), and Blondel et al. (2008). They differ in terms of scalability and quality of detection.[2] Netzer et al. (2012) introduced the community detection method developed by Girvan and Newman (2002) for the first time in marketing, in order to segment discussion of 169 car models in an online forum. Newman and Girvan (2004) extended their previous paper by incorporating the weight of edge between vertices. Later, Clauset, Newman, and Moore (2004), Pons and Latapy (2005), Raghavan, Albert, and Kumara (2007), and Blondel et al. (2008) proposed new algorithms to process a large network quickly. Blondel et al.'s (2008) the Louvain method is known to show better performance in terms of speed and accuracy. Recently, Ringel and Skiera (2016) adapted the Louvain method as one component of their market structure map of more than 1,000 products from an online price and product comparison site.

To estimate TDA, we utilized a JavaPlex implementation by Adams, Tausz, and Vejdemo-Johansson (2014) through the MATLAB interface developed by Adams and Tausz (2015). For hierarchical clustering and community detection methods, we use the R "cluster" and "igraph" packages, respectively.

3.2. Simulation study results

Figure 4 shows the result of topological data analysis on the simulated data. In Scenario 1, there are two intervals under $Betti_1$ in the barcode chart, implying that TDA detects two loopy segments. TDA also generates segment members and their connection order. Each segment includes the appropriate local products and the national product N4, implying that the two local segments are connected through the national product. In Scenario 2, there are four intervals under $Betti_1$. As expected from the generated data, Segments 3 and 4 are connected through the national product N8, there is no overlapping product between Segments 1 and 2. In summary, TDA recovers the true segments well.

Next, we turn to the results from hierarchical clustering methods in Figure 5. For both Scenarios 1 and 2, the single linkage algorithm yields a different pattern than the others, putting the national brand, N8, into its own segment. Generally, in Scenario 1, the single linkage algorithm does successfully capture the

[2] Related to these methods, Henderson, Iacobucci, and Calder (1998) and John et al. (2006) used survey-based approaches to generate a brand-associative network.

different location groupings; however, in Scenario 2, the single linkage algorithm groups products together that should be completely separate (segments 1 and 2). The national product also connects to segments 1 and 2 when it should be disconnected. Although the single linkage algorithm is the most similar to TDA in terms of the intuition behind the algorithm, it performs poorly because it does not allow for loopy segments.

The other three hierarchical clustering methods perform better, in the sense that they do group the products into the appropriate two or four segments. Still, they do not capture the useful information that the national product connects the two groups of local products (segments 1 and 2 in Scenario 1 and segments 3 and 4 in Scenario 2) because the algorithms force each product into only one segment. While this feature of hierarchical clustering methods is often useful in marketing research analysis, it means that connections across products in different markets are better found using TDA.

We next examine a community detection method result from an R implementation of the Louvain method (Blondel et al. 2008). Because the other four community detection methods yielded the same results, the description that follows applies to all five methods. In Scenario 1, the community detection methods generate two segments: {W1, W2, W3} and {N4, M5, M6, M7}, failing to identify the gate product N4 because each product is assigned into only one segment, like the above hierarchical clustering. However, there is a potential way to detect the gate product using a node betweenness centrality measure (Freeman 1977) in network analysis with the assumption that a product (i.e. node) with high betweenness will connect local segments. We show this potential approach with the richer example in Scenario 2.

Scenario 2 results are shown in Table 2. Column 1 shows the "true" segments according to the simulation. Column 2 shows the TDA segments, and Column 3 shows the community detection method segments. Here the community detection methods split the sample into two groups, failing to capture the four distinct segments. This suggests that the community detection methods which we use segment products too broadly, perhaps because such network approaches use all the given connection information when they generate clusters. This problem may be solved by Ringel and Skiera (2016), who extend the Louvain method (Blondel et al. 2008) by (1) adding a "resolution" parameter and (2) combining a multilevel coarsening and refinement procedure (Rotta and Noack 2011). However, the new method by Ringel and

Skiera (2016) also does not identify indirect connection because it also does not allow for a product to be allocated into multiple segments (i.e. submarkets). Thus, we do not implement the extension of the community detection methods used by Ringel and Skiera (2016) here because detecting small segments is not the key aspect we emphasize in this paper as the key strength of TDA. Rather, we explore whether there is a potential way to identify indirect connections in the framework of network analysis as a benchmark model.

Next, while community detection methods do not directly identify gate products, it is possible to take the constructed network and identify products with high node betweenness. Column 4 shows that the national product has high betweenness centrality, suggesting that, by adding this step, the community detection methods can be used to help find the gate product. It is possible to then look at co-purchasing patterns with this gate product in Column 5 and identify indirectly connected local products identified through TDA. For example, W7 and M9 are especially likely to be purchased with N8, correctly suggesting a linkage between them. However, as we demonstrate in the empirical application below, this approach can be complicated if there are multiple potential gate products. For example, if Scenario 2 is adapted so that there is another national product N16, which is co-purchased often with W7 but rarely with M9 then it is not easy to decide whether W7 and M9 are potentially competing. The difficulty will increase as the number of national product increases.

In summary, TDA finds clear connections between the two local segments through the national product. In this small product network, more familiar clustering methods can also show such a link, but with additional effort required through manual checking of distances and values. As the number of products grows, however, such effort becomes impractical. Thus, we interpret the results of the simulation to suggest that TDA captures a potentially useful data pattern that is not captured by hierarchical clustering or community detection methods.

4. Marketing application

4.1. Data and computation time

The IRI Marketing data set (Bronnenberg, Kruger & Mela 2008) has individual-level consumer purchase data in two local cities: Eau Claire Wisconsin and Pittsfield Massachusetts. Consumers in these cities have distinct tastes and there are some differences in product availability. Therefore, this data set allows us to investigate whether TDA can detect potentially related products across local markets. We also look for potentially related products by looking across two categories, salty snacks and beers.

Like most other clustering methods, TDA use the distance matrix among products as input data. We calculate Euclidean distance among products across all the consumers' purchase quantities during a particular year, 2003. There are 6,352 consumers who meet IRI's reporting criteria (Kruger and Pagni 2011 page 16) across the two cities in salty snacks and 3,101 who meet the reporting criteria in beer.

As we discussed above, TDA is computationally intensive, increasing exponentially as the number of products increases. To explore computational feasibility, we choose the top 10, 20, 30, 40, and 50 products in salty snacks and beer in each local market. Table 3 summarizes the results. For the top 10 case, there are 15 salty snacks and 17 beers across two local markets because national products are available in two regions. Some products which are sold in two regions have very low sales quantities in one local market. In this case, we classify it as a local product. We define a national product as a product that makes up more than 0.5% of category sales in each market. With 32 products, TDA took just 0.5 seconds. With 119 products (top 40 in each market, both categories), TDA took 30 minutes. Finally, with 146 products (top 50 in each market), our computer kept running without generating a TDA result. This demonstrates the computational limits of TDA without a high performance computer. The 119 products cover 94% salty snack sales and 89% of beer sales in these two markets. In most of what follows, we show results on the 32 products (row 1 of Table 3) because the smaller number of products allow for visual comparison of results with hierarchical clustering methods.

4.2. Potential competitors within a category

Table 4 reports the results. We focus on loopy segments ($Betti_1$ and $Betti_2$) in order to highlight the distinct results given by TDA. Table 4a shows the loopy segments for salty snacks. There are two loopy segments with hole ($Betti_1$) structures. Figure 6 visually summarizes the members of each segment. Two national products 'Rold Gold' and 'Tostitos', connect two neighboring segments. This connection information is useful in identifying products that serve the same role in different markets. The two national products are competing against (1) Wisconsin local products Barrel O Fun and Jays in segment 1 and (2) Massachusetts local product UTZ in segment 2.

From this indirect relationship, a marketing manager learns that those local products have similar positioning. In other words, if a marketing manager plans to launch the Midwestern (Wisconsin) local product Barrel O Fun or Jays in Massachusetts, she can predict that it will be likely to compete against East Coast (Massachusetts) local product UTZ, although those three local products do not compete in the same market in our data.

We next examine whether these relationships appear using hierarchical clustering and community detection methods. Figure 7 shows the results of hierarchical clustering the salty snacks products. Massachusetts local product UTZ does not seem to be related to Wisconsin local products Barrel O Fun and Jays. It is hard to see a connection between them in any of the four hierarchical clustering methods. These results are driven by the fact that no consumer purchases both Wisconsin and Massachusetts products. In summary, the standard hierarchical clustering cannot capture the pattern of indirect connection, unlike TDA. Table 5 shows the results of five different community detection methods. Again, no segment includes a mix of local brands from the two regions. As in the simulation, it is possible to use betweenness measures to try to identify connecting products. In this case, all the national products yield similar betweenness measures, meaning that all nine national product connect all the local products in the two regions. Then, to find potentially competing local products, one may need to check joint product purchases with each of the nine national products, as in Column 5 in Table 2. As we discussed in the simulation section, however, it is

hard to see which local product in one region is potentially competing against whom in the other region due to local product's different co-purchasing pattern with each of national products.

Next, we analyze a beer category. Table 4b shows that TDA generates 10 loopy segments with 8 holes ($Betti_1$)) and 2 voids ($Betti_2$)). Here national beer products connect products from the different local markets, even within a segment. For example, row 8 contains two national brands, a Massachusetts brand, and two Wisconsin brands (rows 1, 6, 9, and 10 have similar diversity). Figure 8a visualizes the segment in row 8 of Table 4b. Two national products Bud Light and Heineken connect the only Massachusetts brand in the 32 product data set (Michelob Light) with two Wisconsin brands (Miller Genuine Draft and Miller Genuine Draft Light). Furthermore, national products also connect local products across segments as described in the salty snacks category and in the simulation: Heineken, Smirnoff Twisted V, Corona Extra and other national brands appear in multiple segments. Table 4b also highlights a limitation of looking for loopy segments using TDA: There is some repetition of products across segments. This means that TDA is a useful starting point for identifying potentially interesting connections between products, but further analysis is needed to assess the strength and validity of those connections.

4.3. Potentially related products across categories

Next, we combine the salty snack and beer data together (32 total products) to see whether TDA can find products that might be purchased together, if they were available in the same market. The rightmost columns in Table 3 show that TDA generates many more segments from the combined data (beer + salty snack) than separate product data. For example, in the top 10 product case, there are 2 salty snack segments, 10 beer segments, and 29 combined (salty snack and beer) segments.

Table 4c shows the segment members from the combined data. Most segments (19 of 29) have both salty snacks and beer products, providing insight into why the combined data have more segments than the separate data. Given the underlying data, this makes sense: Even if a customer always buys the same beer brand and the same salty snacks brand, these brands are connected in the combined data and provides insight into which categories and products tend to be purchased by the same customers.

We also find potentially related products across categories in seven segments, as marked in the rightmost column in Table 4c. Figure 8b visualizes the segment in row 5. Massachusetts salty snack UTZ and Wisconsin beer Miller Genuine Draft are in the same segment. Once again, this is mainly due to their connection with national products. TDA provides the order of connection: (1) UTZ + Miller High Life, (2) Miller Genuine Draft + Heineken, (3) Miller High Life + Heineken, and (4) UTZ + Miller Genuine Draft. Each local product connects with a national product first and then the Massachusetts salty snack and Wisconsin beer get connected. This suggests that the purchase behavior of people who buy UTZ in Massachusetts is similar to the purchase behavior of people who buy Miller Genuine Draft in Wisconsin. This information could be used to inform product launches across markets. Alternatively, it might help generate advertising ideas, for example UTZ ads could borrow elements from a successful Miller Genuine Draft campaign.

4.4. Relationship between a segment's birth and its product diversity

In the above analysis, we focused on a relatively small number of products in order to facilitate comparison with hierarchical clustering and to ease the communication of the content of the various segments. When more products are included, TDA can generate more loopy segments. In this section, we explore how TDA measures of birth filtration value help identify the interesting segments. To do so, we now use the 119 total products (top 40 in each category in each market) that make up 94% of salty snacks sales and 89% of beer sales.

Birth filtration value is a useful metric because it measures how unusual a particular grouping is likely to be. TDA groups products that are close to each other first. Segments that emerge late are more likely to leverage the distinct insights that the topological approach offers. In particular, we are interested in detecting loopy segments that connect regionally distinct local products through national products. Because the connections are indirect, those loopy segments tend to form later.

We next correlate birth filtration value with the diversity of product members within a segment. We focus on diversity because, as argued above, a key use of TDA is to identify connections that other

methods would not. In this paper we have emphasized separate local markets. We measure diversity as follows. We first order all the products in the same local market by quantity. We assign each a rank based on this ordering, and take the difference of the rank across the two local markets. This difference is positive if the product is a Massachusetts product, negative if Wisconsin, and close to zero if national. Finally, we calculate the standard deviation of the rank gaps within a segment.

This gives a sense of the variation of the location of sales for the products in the segment: If the segment has a mix of strongly Wisconsin and strongly Massachusetts products, this diversity measure will be high. If the segment is mostly national products (or mostly from just one region), the diversity measure will be low. If the segment contains both national products and products from one region, the diversity measure will be in the middle.

Table 6 shows the relationship between a segment's birth filtration value and its product diversity. We run separate regressions of filtration value on diversity for $Betti_1$ and $Betti_2$ groupings. We also show results that drop short-lived segments, which may occur due to noise in data (Lesnick 2013). From visual inspection, we chose 3 as the cut-off value for eliminating segments. Thus, we show twelve regressions: three product cases, two *Betti* groupings, and with/without the short-lived segments. The coefficients are all positive and 10 of 12 are significant, implying that the segments that are formed late are more likely to have mixed local products across the two cities (i.e. product diversity), as expected. The two non-significant slopes are for salty snacks $Betti_1$, which has fewer segments than beer or the combined analysis, suggesting that this might be an issue with statistical power.

In summary, we show that TDA can detect high diversity segments that include local products in regionally distinct markets, particularly as the filtration value increases. If there are many high birth filtration value segments, a final step in identifying the potentially most interesting segments is to look for those with longer filtration range as longer intervals suggest more robust segments.

5. Conclusions

In this paper, we have applied Topological Data Analysis to a particular marketing application. We have shown that TDA is effective at identifying connections between products that are not purchased together but hold similar positioning in geographically distinct markets.

A key open question is whether the assumption of transitive preferences holds across settings. In particular, our framework assumes that two objects that have no direct relationship with each other, but are bought with a third object, are indirectly related. We have not directly tested this assumption because we do not have data on several cross-market product launches and data on pre-launch sales across locations. Furthermore, it is worth exploring whether assumption holds across market types. For example, it might hold in our setting for consumer non-durables, but it might not hold for durables or in business-to-business markets. Relatedly, while the IRI data are ideal in the sense that they have rich customer-level data in two distinct markets, the products do not have sufficiently rich attribute information to check that they serve similar roles by clustering products on attributes. In other words, we have shown the potential usefulness of TDA but leave a field test for future work.

Generally, TDA is a new data mining tool and we anticipate other marketing applications. We anticipate that those applications will be primarily identifying opportunities and a complement to other types of analysis. In this way, TDA should not be seen as a final step for segment analysis, but as a useful part of a more comprehensive analysis. It is exploratory and, as with all segmentation methods, it does not yield a legitimate causal interpretation. Nevertheless, we believe Topological Data Analysis should be seen as a useful tool in market structure and segmentation analysis.

References

Adams H, Tausz A (2015) JavaPlex tutorial. http://www.math.colostate.edu/~adams/research/javaplex_tutorial.pdf

Adams H, Tausz A, Vejdemo-Johansson M (2014) JavaPlex: A research software package for persistent (Co) homology. Proceedings of ICMS 2014, H. Hong and C. Yap (Eds.), *Springer-Verlag* Berlin Heidelberg 129–136.

Ailawadi KL, Keller KL (2004) Understanding Retail Branding: Conceptual Insights and Research Priorities. *Journal of Retailing* 80(4):331-342.

Armstrong MA (1983) *Basic topology*. Springer, New York, Berlin.

Ayasdi (2015) TDA and machine learning: Better together. http://www.ayasdi.com/resources/tda-and-machine-learning-better-together-via-intro-tda/.

Ayasdi (2016) website, http://www.ayasdi.com/industries/communications/personalized-marketing/, accessed on February 29, 2016.

Barbiero A and Ferrari PA (2014) Simulation of correlated Poisson variables, *Applied Stochastic Models in Business and Industry* 31(5):669–680

Bergen M, Peteraf MA (2002) Competitor identification and competitor analysis: A broad-based managerial approach. *Managerial and Decision Economics* 23(4-5):157-169.

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* (10):P10008.

Bronnenberg BJ, Kruger MW, Mela CF (2008) Database paper: The IRI marketing data Set. *Marketing Science* 27(4):745-748.

Carlsson G (2009) Topology and data. *Bulletin of the American Mathematical Society* 46(2):255–308.

Clauset A, Newman MEJ, Moore C (2004). Finding community structure in very large networks. http://www.arxiv.org/abs/cond-mat/0408187.

Cooper LG, Inoue A (1996) Building market structures from consumer preferences. *Journal of Marketing Research* 33(3):293–306.

DeSarbo WS, Grewal R (2007) An alternative efficient representation of demand-based competitive asymmetry. *Strategic Management Journal* 28(7):755-766.

DeSarbo WS, Grewal R, Wind J (2006) Who competes with whom? A demand-based perspective for identifying and representing asymmetric competition. *Strategic Management Journal* 27(2):101-129.

DeSarbo WS, Manrai AK, Manrai LA (1993) Non-spatial tree models for the assessment of comparative maket structure: An integrated review of the marketing and psychometric literature. Eliashberg J, Lilien G, eds. *Handbook in operations research and marketing science*, North Holland, Amsterdam, 193-257.

DeSarbo WS, Soete GD. 1984. On the Use of Hierarchical Clustering for the Analysis of Nonsymmetric Proximities. *Journal of Consumer Research* 11(1) 601-610.

Edelsbrunner H, Harer J (2010) *Computational topology: An introduction*. American Mathematical Society, Providence RI.

Edelsbrunner H, Letscher D, Zomorodian A. (2002) Topological persistence and simplication. *Discrete and Computational Geometry* 28:511-533.

Elrod T, Russell GJ, Shocker AD, Andrews RL, Bacon L, Bayus, Carroll JD, Johnson RM, Kamakura WRA, Lenk P, Mazanec JA, Rao VR, Shankar V. (2002) Inferring market structure from customer response to competing and complementary products. *Marketing Letters* 13(3): 221–32.

Erdem T (1996) A dynamic analysis of market structure based on panel data. *Marketing Science* 15(4):359-378.

Erdem T, Keane MP (1996) Decision-Making Under Uncertainty: Capturing Dynamic Choice Processes in Turbulent Consumer Good Markets *Marketing Science* 15(1): 1–20.

Freeman L (1977) A set of measures of centrality based on betweenness. *Sociometry* 40: 35–41.

France S, Ghose S (2016) An analysis and visualization methodology for identifying and testing market structure. *Marketing Science* 35(1): 182 – 197.

Girvan M, Newman ME (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12):7821-7826.

Gromov M (1987) Hyperbolic groups. Essays in group theory, *Mathematical Sciences Research Institute Publications* 8, Springer-Verlag, 75–263.

Hatcher A (2002) *Algebraic topology*. Cambridge University Press, Cambridge

Hausmann JC (1995) On the Vietoris–Rips complexes and a cohomology theory for metric spaces. Prospects in Topology: Proceedings of a conference in honour of William Browder, *Annals of Mathematics Studies* 138, Princeton Univ. Press, 175–188.

Henderson GR, Iacobucci D, Calder BJ (1998), Brand Diagnostics: Mapping Branding Effect Using Consumer Associative Networks. *European Journal of Operational Research*, 111 (December), 306–327.

Hoffman, D. Novak, T (2015) Emergent Experience and the Connected Consumer in the Smart Home Assemblage and the Internet of Things. Working paper, George Washington University.

John DR, Loken B, Kim K, Monga AB (2006) Brand concept maps: A methodology for identifying brand association networks. *Journal of Marketing Research* 43(4):549–563.

Johnson SC (1967) Hierarchical clustering schemes. *Psychometrika* 32(3):241-254.

Kamakura WA, Russell GJ (1989) A Probabilistic Choice Model for Market Segmentation and Elasticity Structure *Journal of Marketing Research*, 26 (November), 87–96.

Kim JB, Albuquerque P, Bronnenberg BJ (2011) Mapping online consumer search. *Journal of Marketing Research* 48(1):13-27.

Kruger MW, Pagni D (2011) IRI academic data set description. Information Resources, Inc. page 16

Lattin JM, Carrol DJ, Green PE (2003) *Analyzing multivariate data.* Duxbury Resource Center, Pacific Grove.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *Journal of Marketing Research* 48(5):881-894.

Lesnick M (2013) Studying the shape of data using topology. *The Institute Letter.* Institute for Advanced Study, Summer Issue, page 10-11.

Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J, Carlsson G (2013) Extracting insights from the shape of complex data using topology. *Scientific Reports* 3, 1236.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Science* 31(3):521-543.

Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. *Physical Review E* 69(2):026113.

Pons P, Latapy M (2005) Computing communities in large networks using random walks. http://arxiv.org/abs/physics/0512106

Punj G, Stewart DW (1983) Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research* 20(2):134-148.

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. *Physical Review* E 76, 036106.

Ringel DM, Skiera B (2016) Visualizing asymmetric competition among more than 1,000 products using big search data. Forthcoming at *Marketing Science*.

Rips E (1982) Subgroups of small cancellation groups. *Bulletin of the London Mathematical Society* 14 (1):45–47.

Rotta R, Noack A (2011) Multilevel local search algorithms for modularity clustering. Journal of Experimental Algorithmics 16:2-3.

Srivastava RK, Leone RP, Shocker AD (1981) Market Structure Analysis: Hierarchical Clustering of Products Based on Substitution-in-use. *Journal of Marketing* 45(3):38-48.

Srivastava RK, Alpert MI, Shocker AD (1984) A Customer-Oriented Approach for Determining Market Structures. *Journal of Marketing* 48 (1):32–45.

Urban GL, Johnson PL, Hauser JR. (1984) Testing competitive market structures. *Marketing Science* 3(2):83-112.

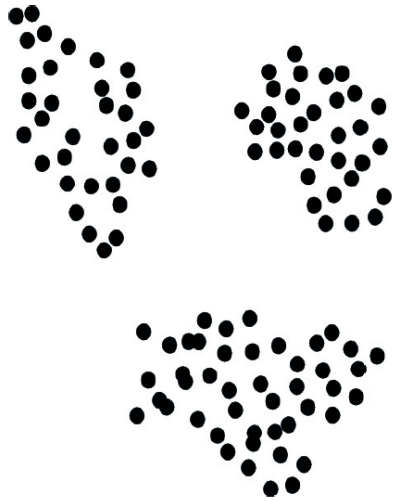
Vietoris L (1927) Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen* 97(1):454–472.

Ward JH (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58:236–244.

Zhai Z, Liu B, Xu H, Jia P (2011) Clustering Product Features for Opinion Mining. *In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. New York, NY. ACM, 347-354.

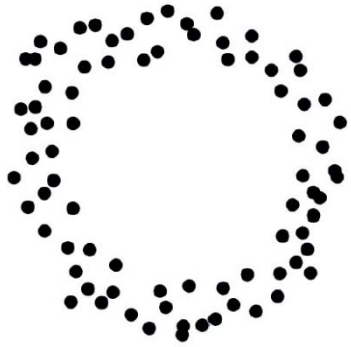
Zomorodian A, Carlsson G (2005) Computing persistent homology. *Discrete and Computational Geometry* 33:249-274.

Figure 1a: Distinctly grouped data



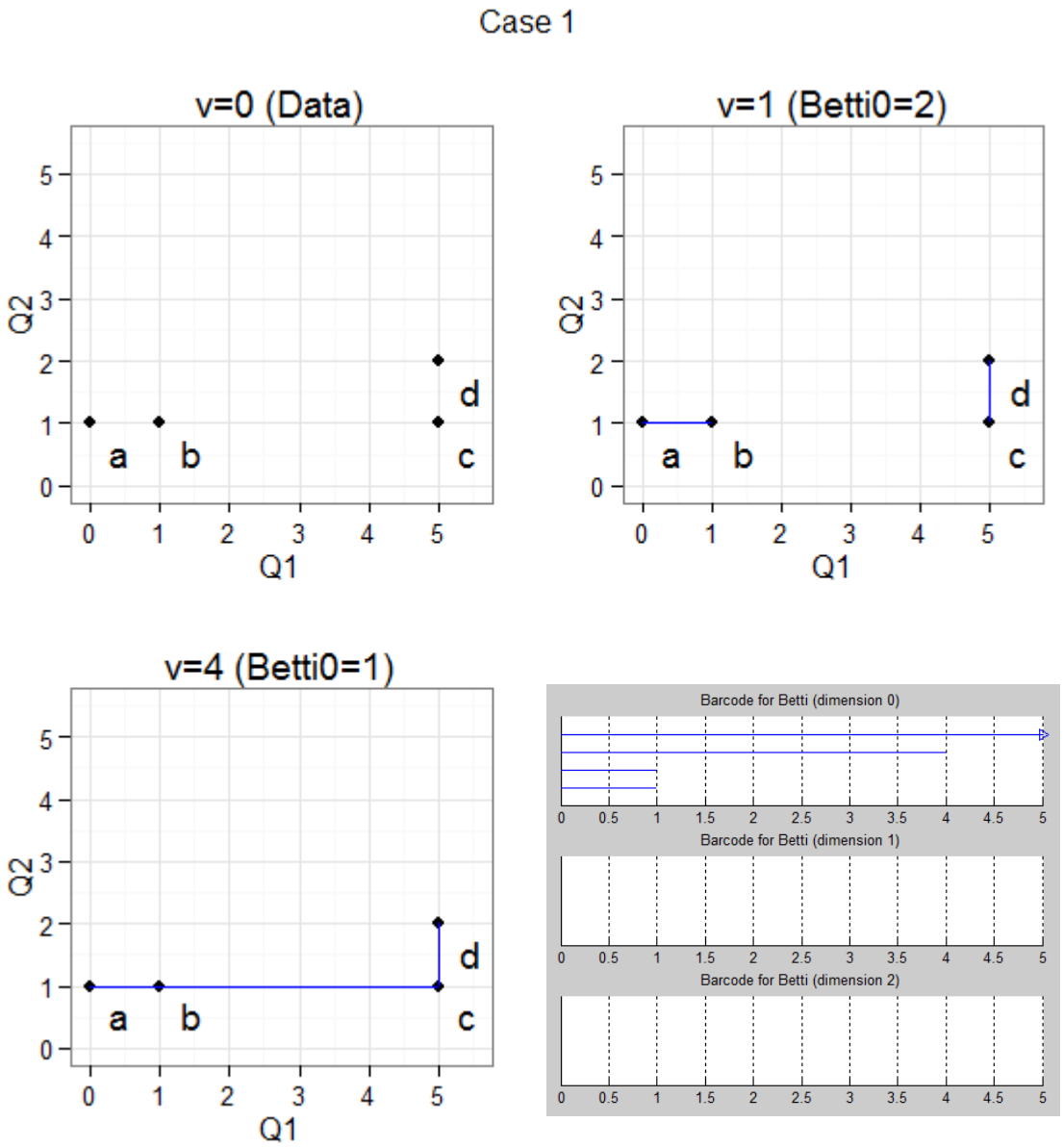
Source: Lesnick (2013)

Figure 1b: A loopy segment

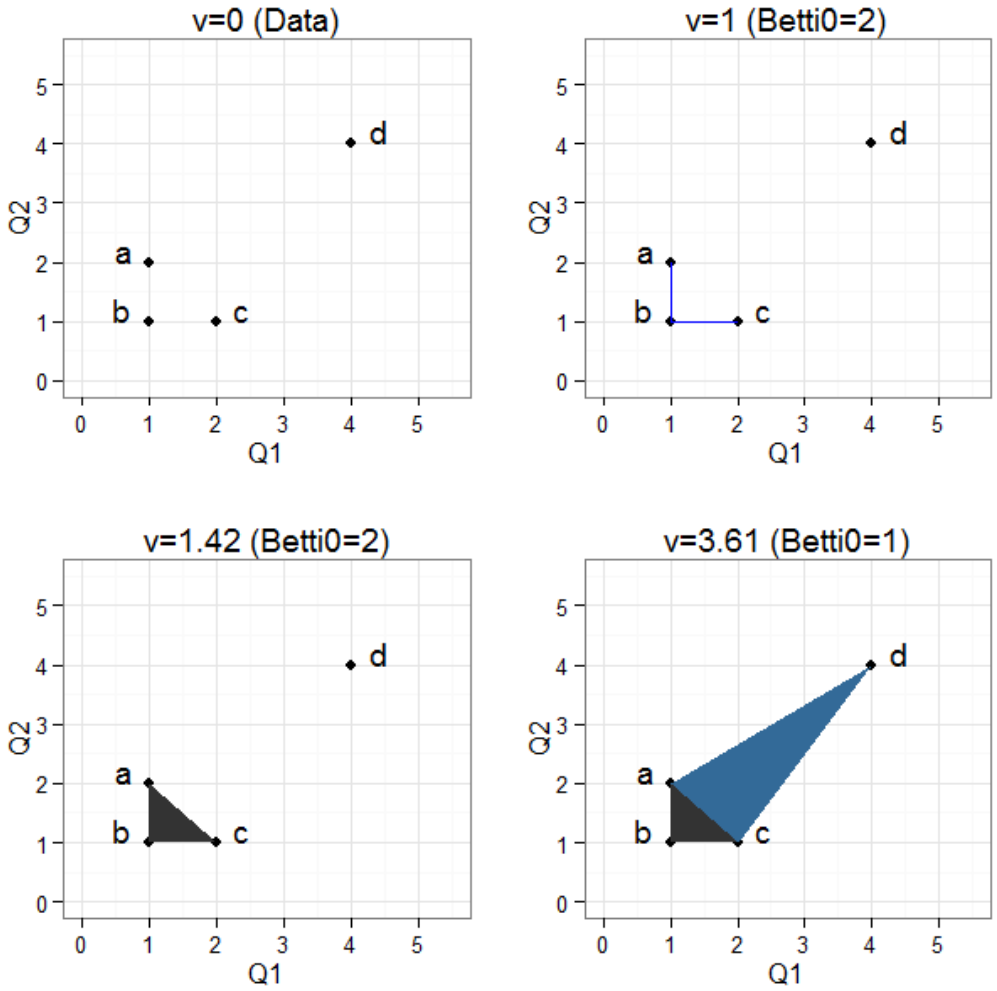


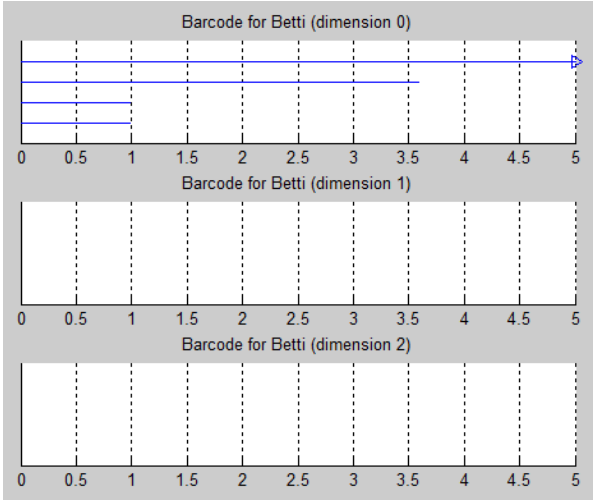
Source: Lesnick (2013)

Figure 2: TDA examples with two customers (Case 1-7) or three customers (Case 8 and 9)

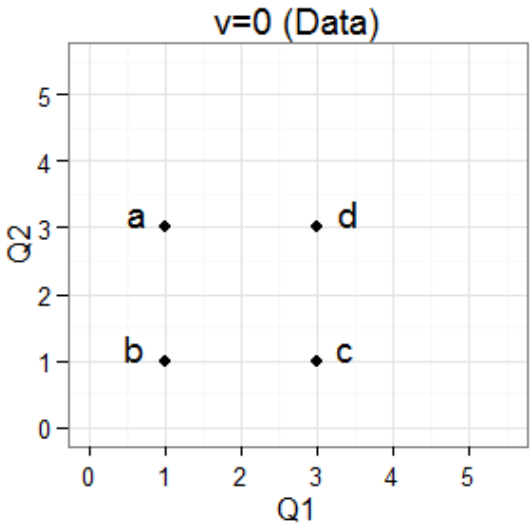


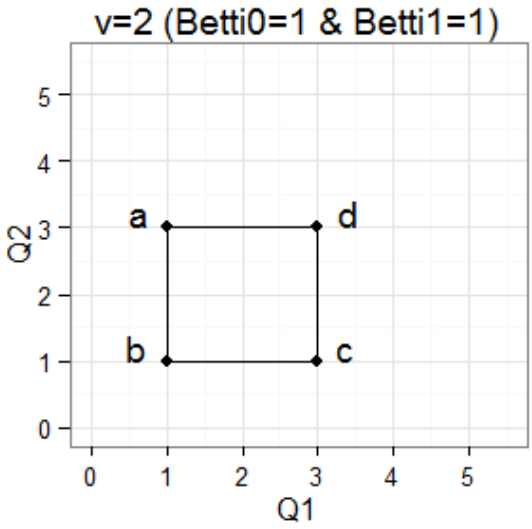
Case 2

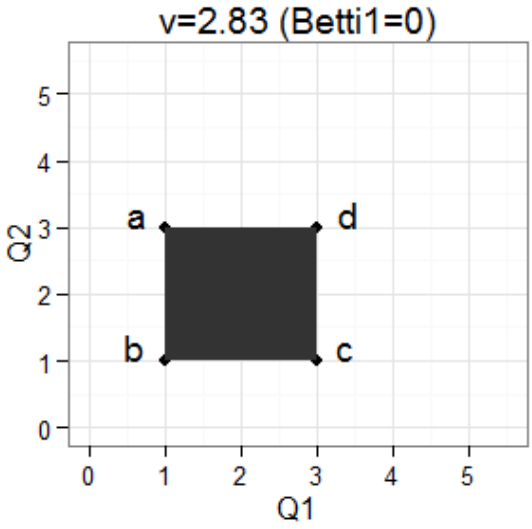

v=0 (Data)
v=1 (Betti0=2)
v=1.42 (Betti0=2)
v=3.61 (Betti0=1)
Q1
Q2
a
b
c
d

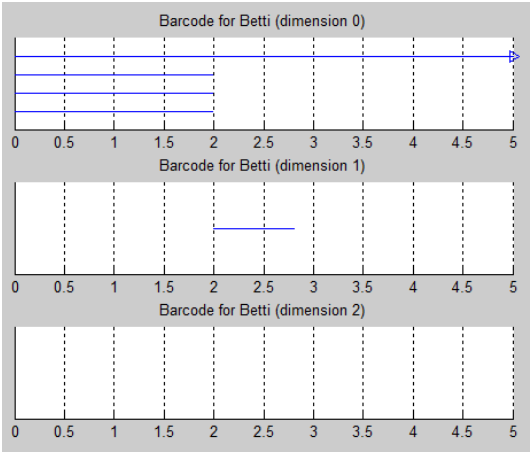

Barcode for Betti (dimension 0)
Barcode for Betti (dimension 1)
Barcode for Betti (dimension 2)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Case 3

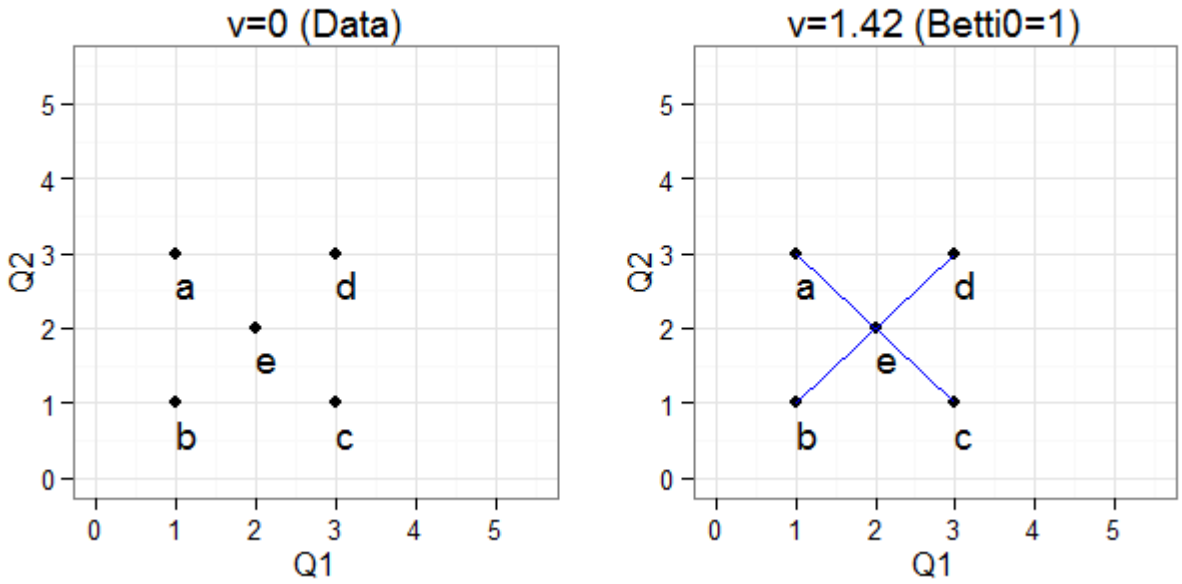

v=0 (Data)
a
b
c
d
Q1
Q2

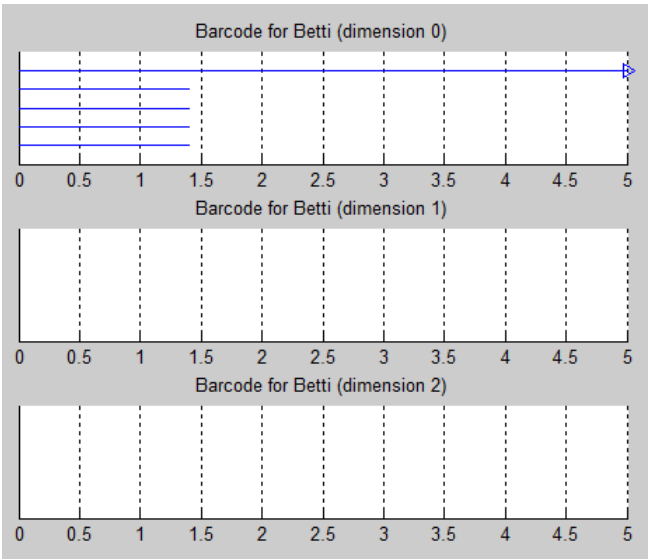

v=2 (Betti0=1 & Betti1=1)
a
b
c
d
Q1
Q2


v=2.83 (Betti1=0)
a
b
c
d
Q1
Q2

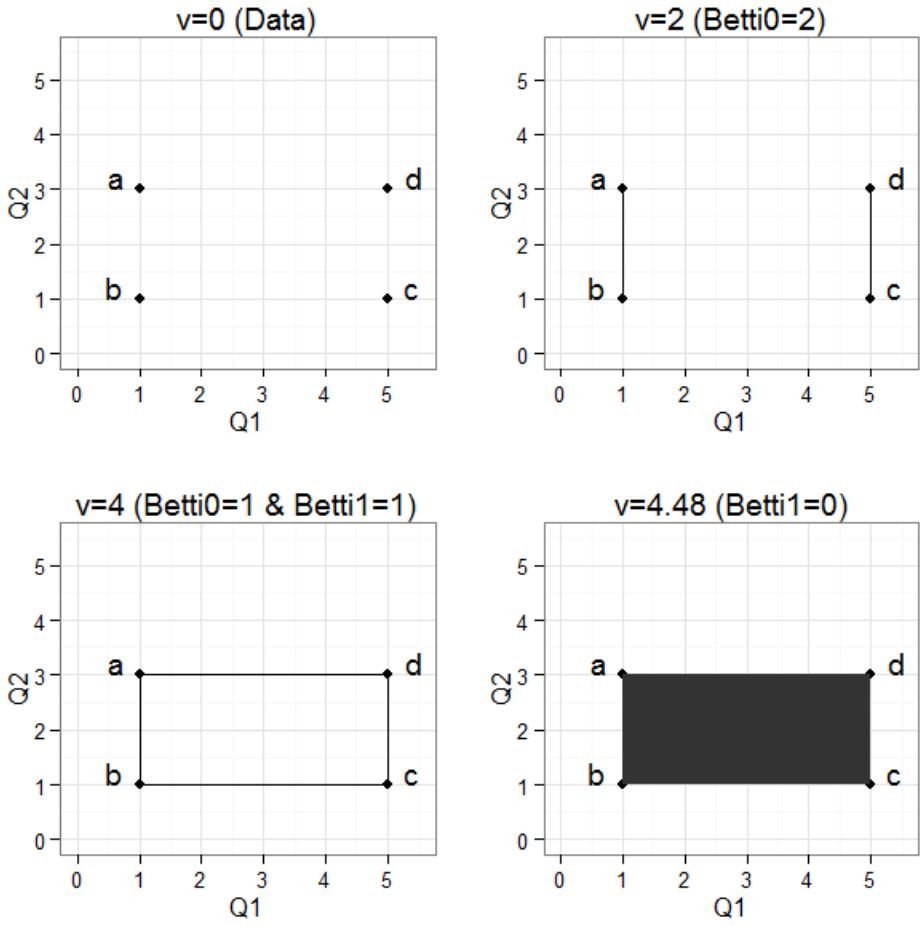

Barcode for Betti (dimension 0)
Barcode for Betti (dimension 1)
Barcode for Betti (dimension 2)

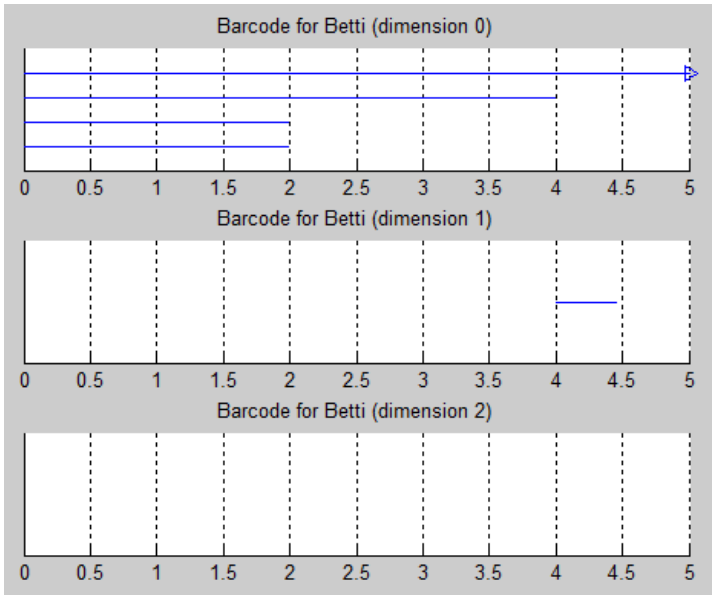
Case 4


v=0 (Data)
v=1.42 (Betti0=1)
Q1
Q2
a
b
c
d
e

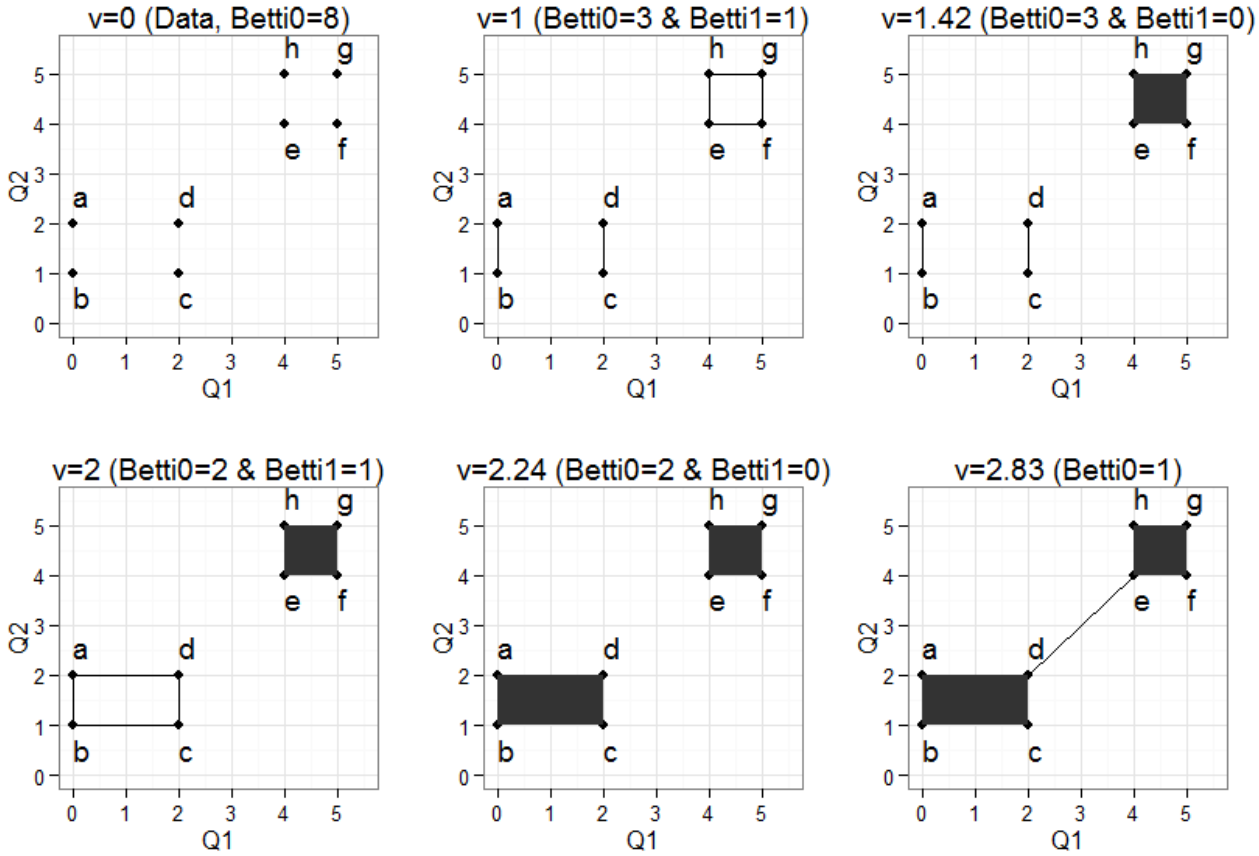

Barcode for Betti (dimension 0)
Barcode for Betti (dimension 1)
Barcode for Betti (dimension 2)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

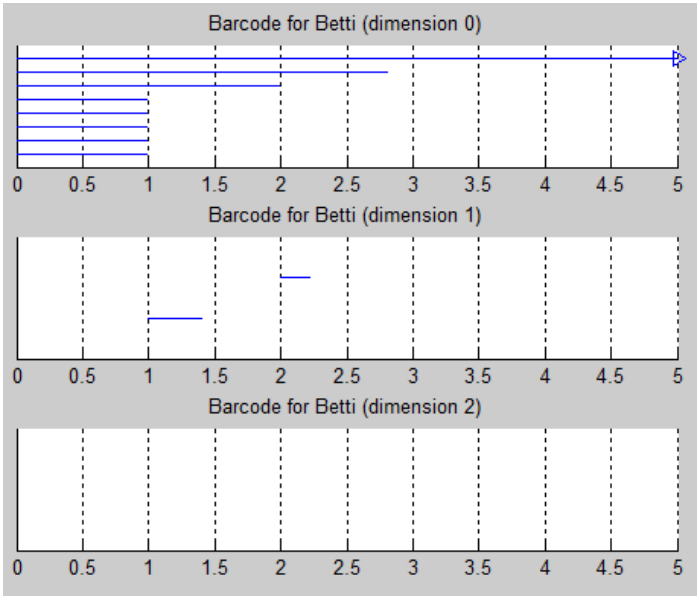
Case 5


v=0 (Data)
v=2 (Betti0=2)
v=4 (Betti0=1 & Betti1=1)
v=4.48 (Betti1=0)
a
b
c
d
Q1
Q2

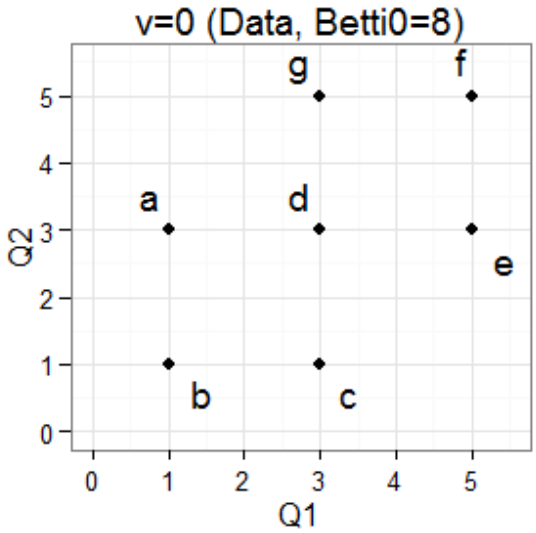

Barcode for Betti (dimension 0)
Barcode for Betti (dimension 1)
Barcode for Betti (dimension 2)

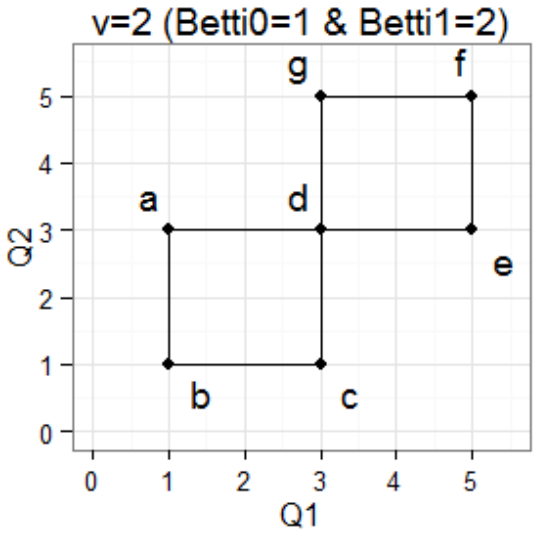
Case 6

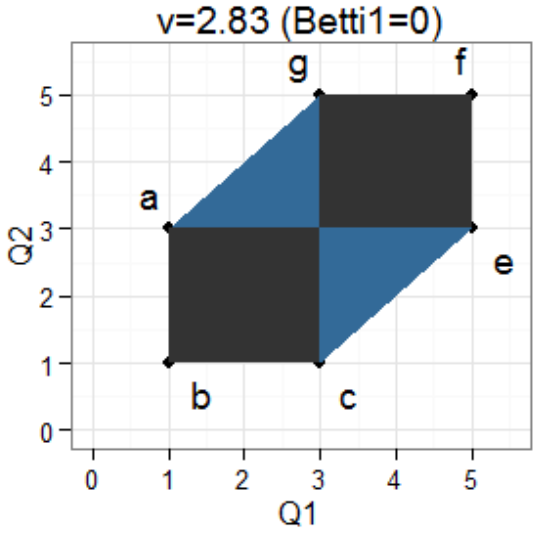

v=0 (Data, Betti0=8)
v=1 (Betti0=3 & Betti1=1)
v=1.42 (Betti0=3 & Betti1=0)
v=2 (Betti0=2 & Betti1=1)
v=2.24 (Betti0=2 & Betti1=0)
v=2.83 (Betti0=1)
Q1
Q2
a
b
c
d
e
f
g
h

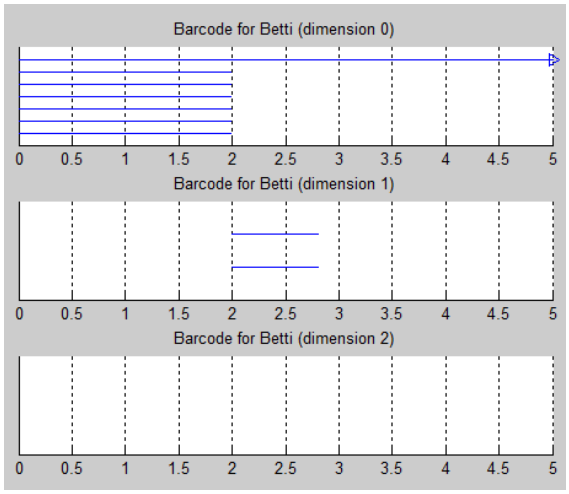

Barcode for Betti (dimension 0)
Barcode for Betti (dimension 1)
Barcode for Betti (dimension 2)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Case 7


v=0 (Data, Betti0=8)
g
f
a
d
e
b
c
Q2
Q1
0
1
2
3
4
5

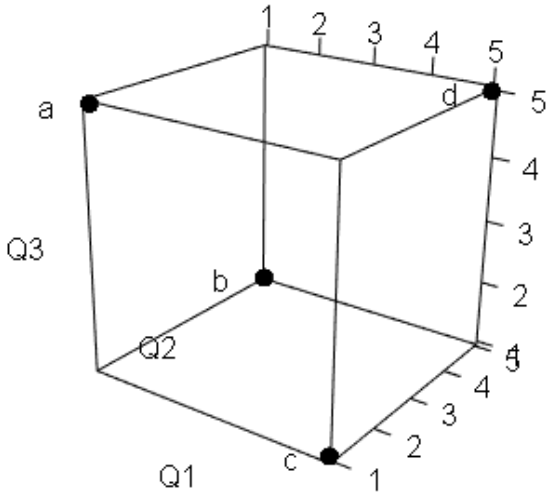

v=2 (Betti0=1 & Betti1=2)
g
f
a
d
e
b
c
Q2
Q1
0
1
2
3
4
5


v=2.83 (Betti1=0)
g
f
a
e
b
c
Q2
Q1
0
1
2
3
4
5

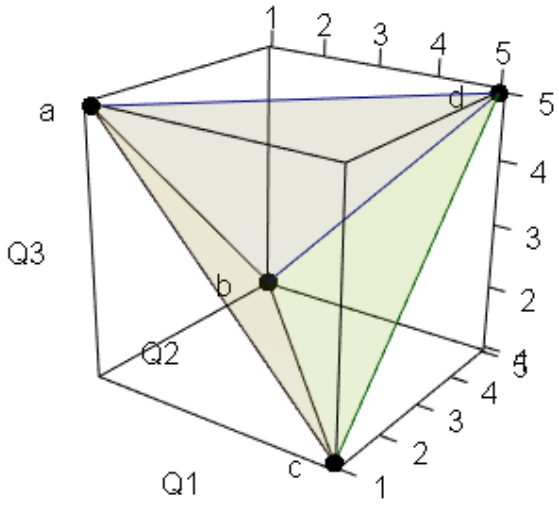

Barcode for Betti (dimension 0)
Barcode for Betti (dimension 1)
Barcode for Betti (dimension 2)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

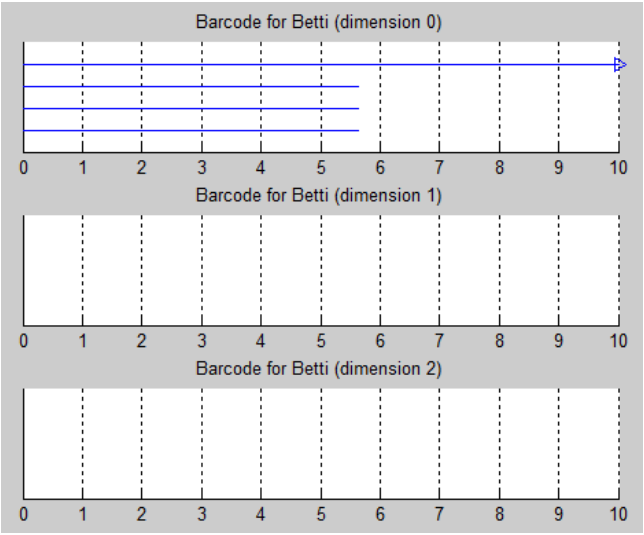
Case 8

V=0 (Data, Betti0=4)



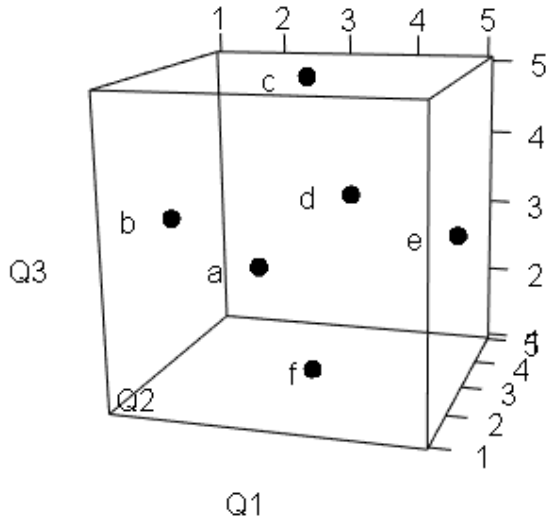
V=5.66 (Betti0=1)



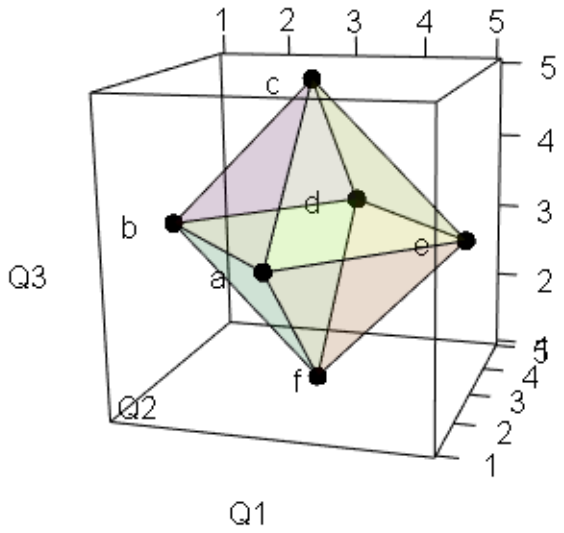


Case 9

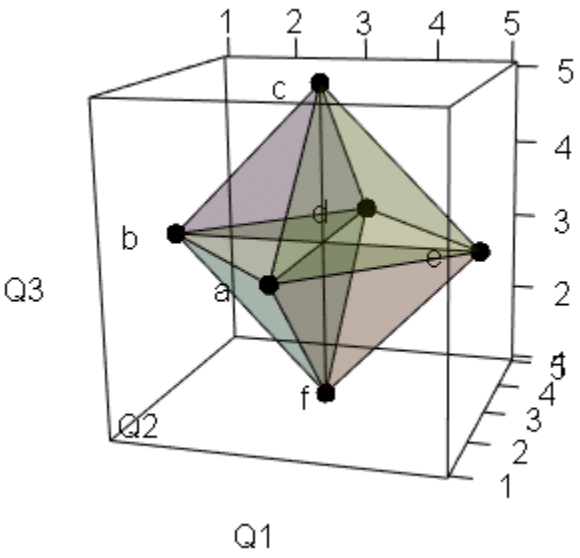
V=0 (Data, Betti0=6)



V=2.83 (Betti0=1, Betti2=1)



V=4 (Betti2=0)



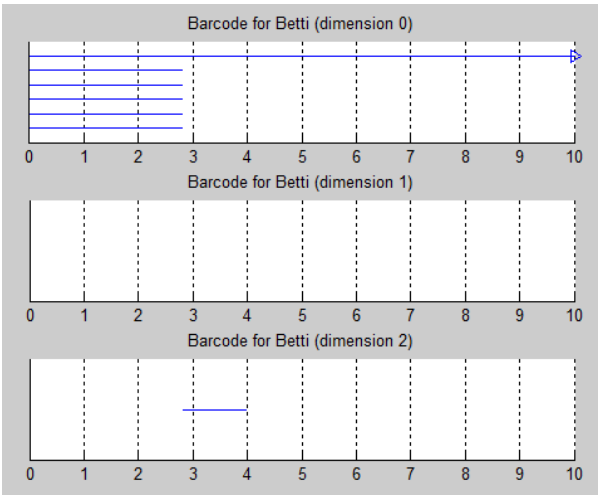


Figure 3a: 5 steps for simulation study

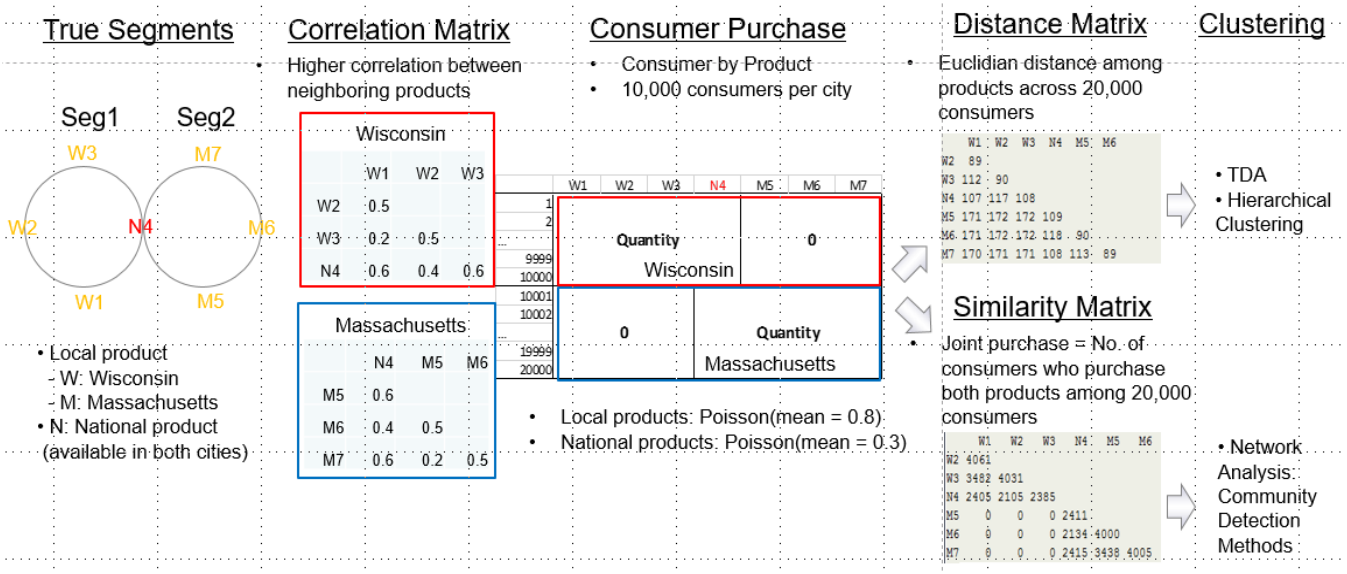
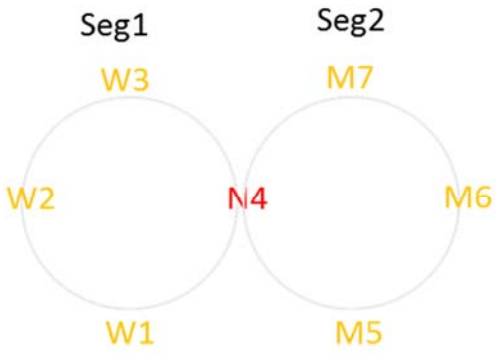
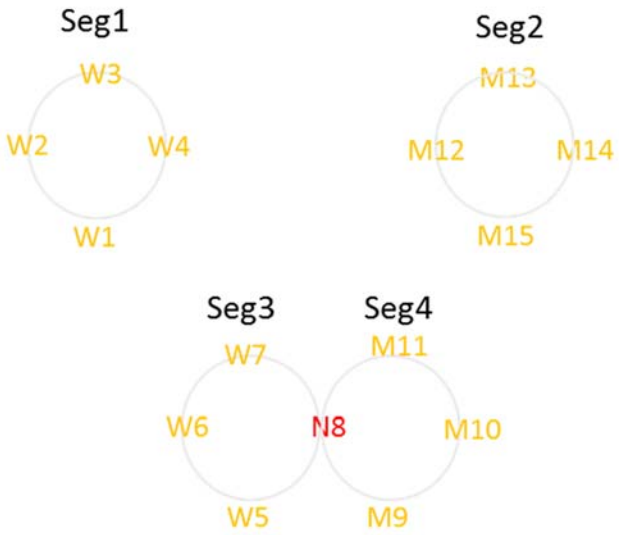


Figure 3b: True segments in simulation study

Scenario 1



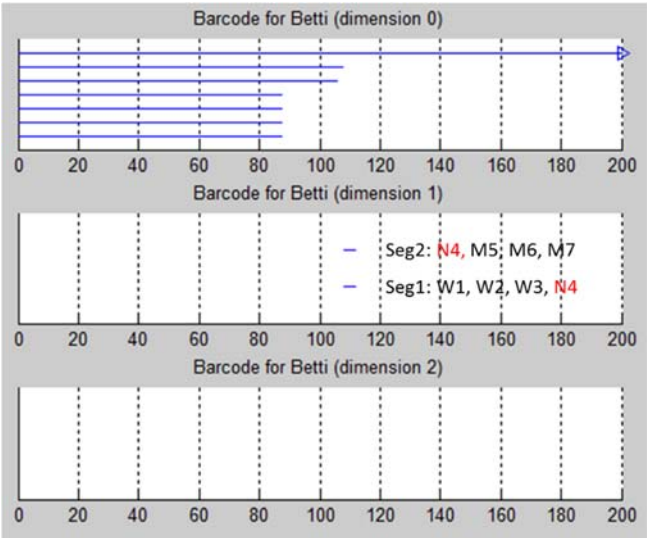
Scenario 2



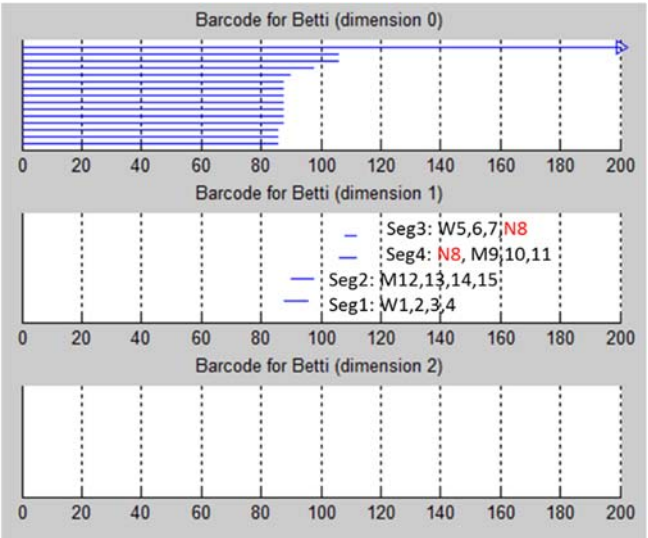
Product type W means Wisconsin product, M means Massachusetts product, and N means national product.

Figure 4: TDA barcode chart for simulation study

Scenario 1



Scenario 2



Product type W means Wisconsin product, M means Massachusetts product, and N means national product.

Figure 5a: Hierarchical clustering for Scenario 1 in simulation study

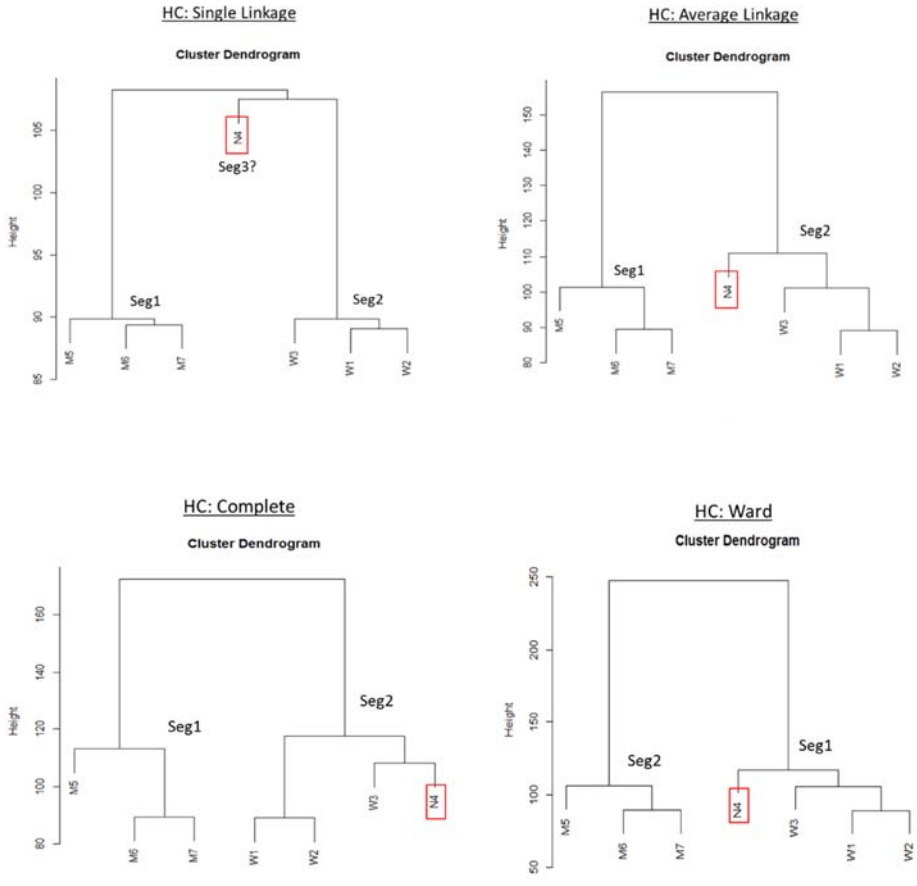
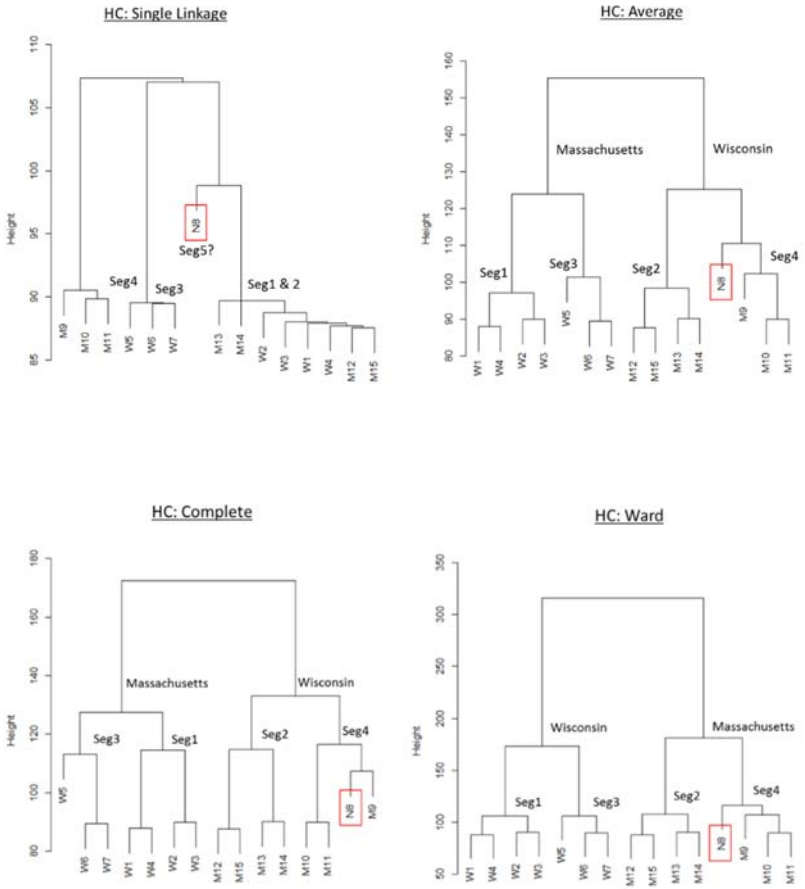


Figure 5b: Hierarchical clustering for Scenario 2 in simulation study



Product type W means Wisconsin product, M means Massachusetts product, and N means national product.

Figure 6: Potentially competing products across segments using IRI data

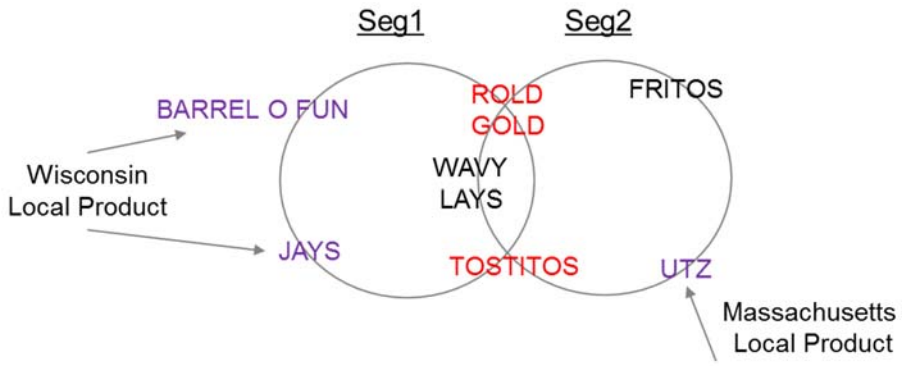


Figure 7: Hierarchical clustering for salty snacks using IRI data

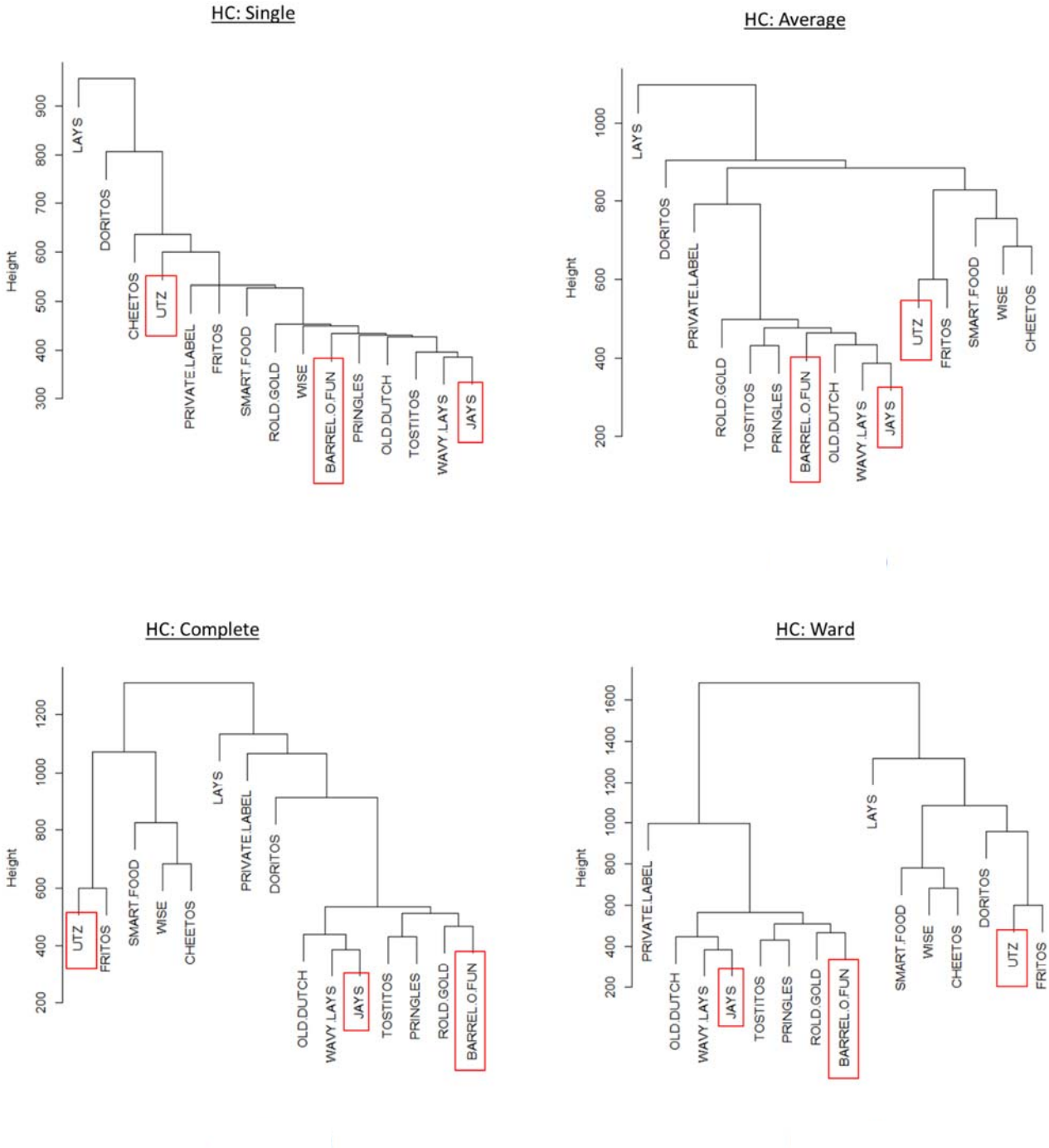


Figure 8a: Potentially competing products within a segment using IRI data with order of connection

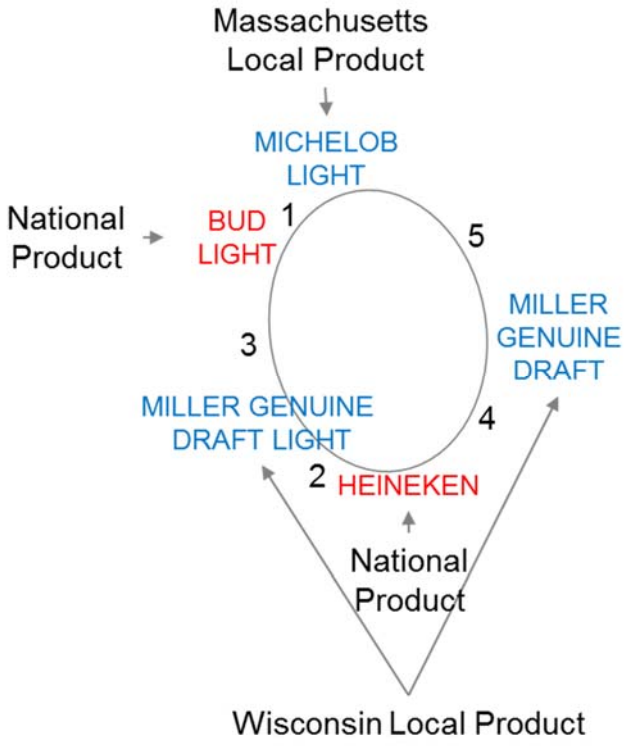


Figure 8b: Potentially related products across segments using IRI data with order of connection

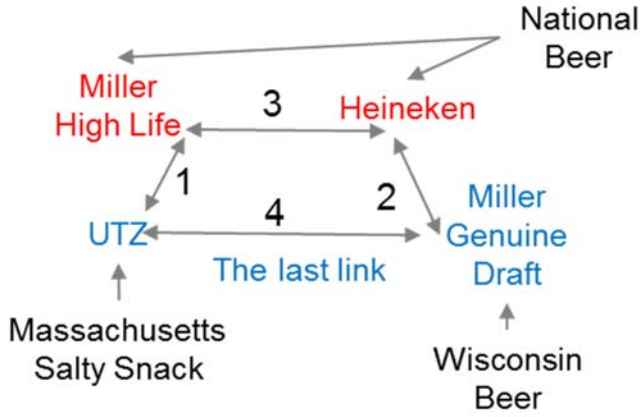


Table 1: TDA cases

Case	Description	Filtration Value (v)	*Filtered Simplicial Complex = VR(X, v)				Betti Numbers		
			Points	Lines	Triangles	Tetrahedron	B0	B1	B2
1	Two segments	0	{a, b, c, d}				4	0	0
		1		{ab, cd}			2	0	0
		4		{bc}	{bcd}		1	0	0
2	Tetragon	0	{a, b, c, d}				4	0	0
		1		{ab, bc}			2	0	0
		1.42		{ac}	{abc}		2	0	0
		3.61		{ad, cd}	{acd}		1	0	0
3	Square loopy segment	0	{a, b, c, d}				4	0	0
		2		{ab, bc, cd, ad}			1	1	0
		2.83		{ac, bd}	{abc, bcd, cda, dab}		1	0	0
4	Center point within square	0	{a, b, c, d, e}				4	0	0
		1.42		{ae, be, ce, de}			1	0	0
5	Rectangle loopy segment	0	{a, b, c, d}				4	0	0
		2		{ab, cd}			2	0	0
		4		{ad, ac}			1	1	0
		4.48		{ac, bd}	{abc, bcd, cda, dab}		1	0	0
6	Distant two loopy segments	0	{a, b, c, d, e, f, g}				8	0	0
		1		{ab, cd, ef, fg, gh, hd}			3	1	0
		1.42			{efg, fgh, ghe, hef}		3	0	0
		2		{bc, da}			2	1	0
		2.24			{abc, bcd, cda, dab}		2	0	0
		2.83		{de}			1	0	0
7	Neighboring two loopy segments with one connection	0	{a, b, c, d, e, f, g}				8	0	0
		2		{ab, bc, cd, ad, de, ef, fg, gd}			1	2	0
		2.83		{ag, ce, ac, bd, df, eg}	{abc, bcd, cda, dab, def, efg, fgd, gde, adg, cde}		1	0	0
8	Tetrahedron	0	{a, b, c, d}				4	0	0
		5.66		{ab, ac, ad, bc, bd, cd}	{abc, abd, acd, bdd}	{abcd}	1	0	0
9	Octahedron with void	0	{a, b, c, d, e, f}				6	0	0
		2.83		{ab, ac, ae, af, bc, bd, bf, cd, ce, de, df,ef}	{abc, abf, ace, aef, dbc, dbf, dec, def}		1	0	1
		4		{ad, be, cf}	{abe, acf, adb, adc, ade, adf, bcf, bec, bed, bef, cfd, cfe}	{abcd, abce, abcf, abdf, abef, acde, acef, adef, bcde, bcdf, bedf, cedf}	1	0	0

*Filtered simplicial complex, VR(X, v), is cumulative as v increases: The table shows the additional simplices for each filtration value v.

Table 2: Community detection methods in Scenario 2 in the simulation study

Product Type	(1) True	(2) TDA	(3) Community detection	(4) Betweenness centrality	(5) Joint product purchase with national product N8 among 20,000 simulated customers
W1	1	1	1	0	1463
W2	1	1	1	0	1433
W3	1	1	1	0	1445
W4	1	1	1	0	727
W5	3	3	1	0	**2399**
W6	3	3	1	0	**2123**
W7	3	3	1	0	**2413**
N8	3,4	3,4	2	49	N/A
M9	4	4	2	0	**2538**
M10	4	4	2	0	**2203**
M11	4	4	2	0	**2473**
M12	2	2	2	0	1491
M13	2	2	2	0	1493
M14	2	2	2	0	1499
M15	2	2	2	0	675

Column 1 to 3 each show a different method. The numbers in the column represent the assigned segment according to that method. Therefore the numbers are not related across columns. Only the national product has nonzero betweenness centrality.

Table 3: TDA results by the top N products in each market

Top N Products in Each Market	No of Products across Two Markets			Market Coverage(%)*		Elapased Time (seconds)			No of Segment		
	S	B	S+B	S	B	S	B	S+B	S	B	S+B
10	15	17	32	69	61	0.4	0.4	0.5	2	10	29
20	29	33	62	85	76	1	0.7	11	12	66	163
30	41	49	90	90	84	4.8	2.9	97.9	35	189	486
40	56	63	119	94	89	15.5	5.4	1857.2	106	289	1035
50	68	78	146	96	92			Keep running			
*Market coverage is based on sales unit.											

Table 4a: TDA for salty snacks

Birth	Death	Interval Length	Salty Snacks	Product Type
Betti 1				
466.8	469.8	3	**ROLD GOLD**, BARREL O FUN, JAYS, **TOSTITOS**	N, W, W, N
649.6	694.9	45.3	**TOSTITOS**, UTZ, **ROLD GOLD**, FRITOS, WAVY LAYS	N, M, N, N, N

Product type W means Wisconsin product, M means Massachusetts product, and N means national product.

Table 4b: TDA for beers

No	Birth	Death	Interval Length	Beers	Product Type	Both Locals
	Betti 1					
1	148.6	162.6	14	SMIRNOFF TWISTED V, HEINEKEN, LEINENKUGEL, MICHELOB GOLDEN DRAFT LIGHT, MICHELOB LIGHT	N, N, W, W, M	Y
2	129.2	175.5	46.3	SMIRNOFF TWISTED V, HEINEKEN, MILLER GENUINE DRAFT, CORONA EXTRA, LEINENKUGEL	N, N, W, N, W	N
3	169.5	175.5	6	SMIRNOFF TWISTED V, CORONA EXTRA, HEINEKEN, MILLER GENUINE DRAFT LIGHT	N, N, N, W	N
4	126.3	195.4	69.1	SMIRNOFF TWISTED V, HEINEKEN, LEINENKUGEL, COORS LIGHT, MILLER GENUINE DRAFT	N, N, W, N, W	N
5	191.5	201.4	9.9	MILLER GENUINE DRAFT, CORONA EXTRA, MICHELOB ULTRA, SMIRNOFF ICE	W, N, N, N	N
6	144.4	225.7	81.3	MILLER GENUINE DRAFT, HEINEKEN, MICHELOB LIGHT, MICHELOB GOLDEN DRAFT LIGHT	W, N, M, W	Y
7	213.6	246.6	33	SMIRNOFF TWISTED V, HEINEKEN, MILLER LITE, MICHELOB LIGHT	N, N, N, M	N
8	181.7	286.1	104.4	BUD LIGHT, MICHELOB LIGHT, HEINEKEN, MILLER GENUINE DRAFT LIGHT, MILLER GENUINE DRAFT	N, M, N, W, W	Y
	Betti 2					
9	309.8	373.6	63.8	MILLER GENUINE DRAFT, MICHELOB LIGHT, MILLER GENUINE DRAFT LIGHT, MICHELOB ULTRA, HEINEKEN, CORONA EXTRA, SMIRNOFF TWISTED V, BUD LIGHT	W, M, W, N, N, N, N, N	Y
10	375.5	380.4	4.9	BUDWEISER, MILLER GENUINE DRAFT, OLD MILWAUKEE, HEINEKEN, MICHELOB LIGHT, MICHELOB GOLDEN DRAFT LIGHT	N, W, W, N, M, W	Y

Product type W means Wisconsin product, M means Massachusetts product, and N means national product.

Table 4c: TDA for the combined data

No	Birth	Death	Interval Length	Salty Snack & Beers	Product Type	Both Products	Potentially Complem-entary
	Betti 1						
1	237.4	259.8	22.4	bSMIRNOFF TWISTED V, bHEINEKEN, bLEINENKUGEL, bMICHELOB GOLDEN DRAFT LIGHT, bMICHELOB LIGHT	bN, bN, bW, bW, bM	N	N
2	206.4	280.4	74	bSMIRNOFF TWISTED V, bHEINEKEN, bMILLER GENUINE DRAFT, bCORONA EXTRA, bLEINENKUGEL	bN, bN, bW, bN, bW	N	N
3	270.7	280.4	9.7	bSMIRNOFF TWISTED V, bCORONA EXTRA, bHEINEKEN, bMILLER GENUINE DRAFT LIGHT	bN, bN, bN, bW	N	N
4	201.7	312.2	110.5	bSMIRNOFF TWISTED V, bHEINEKEN, bLEINENKUGEL, bCOORS LIGHT, bMILLER GENUINE DRAFT	bN, bN, bW, bN, bW	N	N
5	199.6	315.2	115.6	bMILLER HIGH LIFE, sUTZ, bMILLER GENUINE DRAFT, bHEINEKEN	bN, sM, bW, bN	Y	Y
6	250.2	315.2	65	bMILLER GENUINE DRAFT, sSMART FOOD, bHEINEKEN, bMILLER HIGH LIFE	bW, sM, bN, bN	Y	Y
7	305.8	321.6	15.8	bMILLER GENUINE DRAFT, bCORONA EXTRA, bMICHELOB ULTRA, bSMIRNOFF ICE	bW, bN, bN, bN	N	N
8	230.7	360.5	129.8	bMILLER GENUINE DRAFT, bHEINEKEN, bMICHELOB LIGHT, bMICHELOB GOLDEN DRAFT LIGHT	bW, bN, bM, bW	N	N
9	356.8	365.8	9	bMICHELOB LIGHT, sPRINGLES, bSMIRNOFF TWISTED V, bHEINEKEN	bM, sN, bN, bN	Y	N
10	341.2	393.9	52.7	bSMIRNOFF TWISTED V, bHEINEKEN, bMILLER LITE, bMICHELOB LIGHT	bN, bN, bN, bM	N	N
11	306.7	428.9	122.2	bSMIRNOFF TWISTED V, bCORONA EXTRA, bMICHELOB LIGHT, sBARREL O FUN, bSMIRNOFF ICE	bN, bN, bM, sW, bN	Y	Y
12	425.1	434.9	9.8	bMICHELOB ULTRA, bMILLER GENUINE DRAFT, bSMIRNOFF ICE, sROLD GOLD	bN, bW, bN, sN	Y	N
13	423.2	440.8	17.6	bSMIRNOFF ICE, sOLD DUTCH, bMILLER GENUINE DRAFT, sJAYS, sWAVY LAYS	bN, sW, bW, sW, sN	Y	N
14	422.6	442.9	20.3	bSMIRNOFF TWISTED V, sJAYS, sWAVY LAYS, bSMIRNOFF ICE	bN, sW, sN, bN	Y	N
15	290.2	456.9	166.7	bBUD LIGHT, bMICHELOB LIGHT, bHEINEKEN, bMILLER GENUINE DRAFT LIGHT, bMILLER GENUINE DRAFT	bN, bM, bN, bW, bW	N	N
16	423.7	477.7	54	bHEINEKEN, sFRITOS, bMILLER GENUINE DRAFT, bCOORS LIGHT	bN, sN, bW, bN	Y	N
17	430.9	511.9	81	bSMIRNOFF TWISTED V, bCORONA EXTRA, bHEINEKEN, sUTZ, bMILLER LITE	bN, bN, bN, sM, bN	Y	N
18	407.6	544.4	136.8	bCORONA EXTRA, sCHEETOS, bMICHELOB LIGHT, bHEINEKEN	bN, sN, bM, bN	Y	N
	Betti 2						
19	470.5	474.8	4.3	bSMIRNOFF TWISTED V, bSMIRNOFF ICE, sTOSTITOS, sWAVY LAYS, sOLD DUTCH, sBARREL O FUN, bMILLER GENUINE DRAFT, sJAYS	bN, bN, sN, sN, sW, sW, bW, sW	Y	N
20	440.4	477.2	36.8	bSMIRNOFF TWISTED V, bCORONA EXTRA, sBARREL O FUN, bMILLER GENUINE DRAFT, sTOSTITOS, bMICHELOB LIGHT, bHEINEKEN	bN, bN, sW, bW, sN, bM, bN	Y	Y
21	468.2	477.2	9	bSMIRNOFF ICE, sROLD GOLD, sBARREL O FUN, bSMIRNOFF TWISTED V, sTOSTITOS, bHEINEKEN, bCORONA EXTRA, bMILLER GENUINE DRAFT	bN, sN, sW, bN, sN, bN, bN, bW	Y	N
22	455.4	495.9	40.5	bSMIRNOFF TWISTED V, bLEINENKUGEL, sTOSTITOS, bMICHELOB ULTRA, bSMIRNOFF ICE, sROLD GOLD, bHEINEKEN, bMILLER GENUINE DRAFT, bCORONA EXTRA	bN, bW, sN, bN, bN, sN, bN, bW, bN	Y	N
23	466.8	511.9	45.1	bMILLER GENUINE DRAFT, bHEINEKEN, bMILLER HIGH LIFE, bSMIRNOFF TWISTED V, sUTZ, bCORONA EXTRA, bMICHELOB ULTRA	bW, bN, bN, bN, sM, bN, bN	Y	Y
24	413.1	534.1	121	bMICHELOB ULTRA, bSMIRNOFF ICE, sFRITOS, bSMIRNOFF TWISTED V, bHEINEKEN, bCORONA EXTRA	bN, bN, sN, bN, bN, bN	Y	N
25	494.8	596.8	102	bMILLER GENUINE DRAFT, bMICHELOB LIGHT, bMILLER GENUINE DRAFT LIGHT, bMICHELOB ULTRA, bHEINEKEN, bCORONA EXTRA, bSMIRNOFF TWISTED V, bBUD LIGHT	bW, bM, bW, bN, bN, bN, bN, bN	N	N
26	556.6	597.1	40.5	bMILLER GENUINE DRAFT, bMICHELOB LIGHT, bMILLER HIGH LIFE, bMICHELOB ULTRA, bHEINEKEN, bCORONA EXTRA, bSMIRNOFF TWISTED V, sSMART FOOD	bW, bM, bN, bN, bN, bN, bN, sM	Y	Y
27	599.7	607.7	8	bBUDWEISER, bMILLER GENUINE DRAFT, bOLD MILWAUKEE, bHEINEKEN, bMICHELOB LIGHT, bMICHELOB GOLDEN DRAFT LIGHT	bN, bW, bW, bN, bM, bW	N	N
28	613	634.8	21.8	bMILLER GENUINE DRAFT, sROLD GOLD, sOLD DUTCH, bSMIRNOFF TWISTED V, bOLD MILWAUKEE, bMICHELOB ULTRA, bHEINEKEN	bW, sN, sW, bN, bW, bN, bN	Y	N
29	652	654.2	2.2	bHEINEKEN, bCORONA EXTRA, sPRINGLES, sWAVY LAYS, bMILLER GENUINE DRAFT, sSMART FOOD	bN, bN, sN, sN, bW, sM	Y	Y

Prefix b and s mean beer and salty snack, respectively. Product type W, M, and N means Wisconsin, Massachusetts, and national product, respectively.

Table 5: Community detection methods for salty snacks using IRI data

Salty Snacks Brand	Product type	Blondel et al. (2008)	Raghavan, Albert, and Kumara (2007)	Pons and Latapy (2005)	Clauset, Newman, and Moore (2004)	Newman and Girvan (2004)
OLD DUTCH	W	1	1	1	1	1
BARREL O FUN	W	1	1	1	1	1
JAYS	W	1	1	1	1	1
LAYS	N	2	1	1	2	1
PRIVATE LABEL	N	2	1	1	2	1
DORITOS	N	2	1	1	2	1
WAVY LAYS	N	2	1	1	2	1
CHEETOS	N	2	1	1	2	1
PRINGLES	N	2	1	1	1	1
ROLD GOLD	N	2	1	1	2	1
TOSTITOS	N	2	1	1	2	1
FRITOS	N	2	1	1	2	1
SMART FOOD	M	2	1	2	2	1
UTZ	M	2	1	2	2	1
WISE	M	2	1	2	2	1

Each column shows a different method. The numbers in the column represent the assigned segment according to that method. Therefore the numbers are not related across columns. Product type W means Wisconsin product, M means Massachusetts product, and N means national product.

Table 6: The relationship between a segment's "birth" filtration value and its product diversity

		Betti 1		Betti 2	
		Effect	No of Segment	Effect	No of Segment
Salty Snack	All	0.039 (0.027)	56	0.065** (0.020)	50
	Cut at 3	0.039 (0.027)	55	0.060*** (0.019)	41
Beer	All	0.12*** (0.043)	127	0.022** (0.01)	162
	Cut at 3	0.12*** (0.042)	124	0.022** (0.01)	153
Combined	All	0.065*** (0.019)	370	0.068*** (0.009)	665
	Cut at 3	0.061*** (0.018)	364	0.063*** (0.009)	631

Each number is the coefficient on product diversity from a regression of birth on product diversity. The number of observations is the number of segments. ***p < 0.01; ** p< 0.05; *p<0.10