Chapter 5
Clustering Consumers and Cluster-Specific Behavioural Models

Natalie Jane de Vries, Jamie Carlson, and Pablo Moscato

Abstract Social media has almost become ubiquitous in everyday communications and interactions between customers and brands. A novel clustering algorithm, that has shown high scalability in previous applications, is applied to analyse and segment an online consumer behaviour dataset. It is based on the computation of a Minimum-Spanning-Tree and a *k*-Nearest Neighbour graph (MST-*k*NN). Cluster-specific consumer behaviours relating to customer engagement are predicted using symbolic regression analysis which, in a commercial setting, would provide the basis for personalized marketing strategies. Five major clusters were found in the dataset of 371 respondents who answered questions from theoretical marketing constructs related to online consumer behaviours. They are labelled as follows: 'Brand Rationalists', 'Passive Socializers', 'Immersers', 'Hedonic Sharers' and 'Active Participators'. For each of these clusters, a linear model of customer engagement was predicted using symbolic regression analysis. These models inform possible personalized marketing strategies after proper segmentation of the customers based on their online consumer behaviour, rather than simple demographic characteristics.

Keywords Brand · Customer engagement · Engagement · Loyalty behaviour · Online customer engagement · Customer engagement prediction · Segmentation methodologies · Symbolic regression analysis

N. J. de Vries (✉) · P. Moscato
School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW, Australia
e-mail: natalie.devries@newcastle.edu.au; Pablo.Moscato@newcastle.edu.au

J. Carlson
Newcastle University Business School, The University of Newcastle, Callaghan, NSW, Australia
e-mail: Jamie.Carlson@newcastle.edu.au

P. Moscato, N. J. de Vries (eds.), *Business and Consumer Analytics: New Ideas*,
https://doi.org/10.1007/978-3-030-06222-4_5

5.1 Introduction

Online consumer behaviours with a brand focus have increased in terms of volume and types of behaviours displayed, across many different online communication platforms. Due to the increasing capabilities of online platforms such as social networking sites, consumers are now able to communicate, interact and engage with brands, and each other in real-time and in many new and different ways [9, 20, 36]. These high levels of interactions and new approaches to communication create the conditions for online customer engagement which produce high levels of heterogeneity amongst customers of a particular brand as every customer may wish to interact with their favourite brand in a personalized manner. Due to these recent advancements in communicating with their customers, organizations and brand managers in particular are able to gain detailed insight into the behaviour of their consumers. Consequently, the data rich social media environment enables brands to provide a more personalized experience for their customers which increases levels of customer engagement resulting in improved brand performance outcomes [11, 25].

Besides tracking online metrics of customers' online behaviours towards brands, insights into motivations for these behaviours are also of high importance when examining the consumer base in a market. Segmenting their customers and the identification of 'customer typologies' is becoming a strategic priority amongst many organizations and brands with operations in the online environment [15, 30]. Of particular relevance for organizations, technology empowered and highly interconnected customers do not form one homogenous group in terms of their online behaviours but rather, customer segments that dynamically change over time [1, 36]. In the specific context of the social media environment, brand managers need to be able to segment their online customers appropriately with relevant metrics and rigorous analytical approaches [36]. The development of such analytical frameworks would allow an organization with a brand page in the social media environment to profile customers, divide them in groups sharing common characteristics and apply tailored online experiences to optimize the overall customer experience to improve customer–brand relationships, and ultimate sales opportunities.

In addressing the issue for distinct measurement approaches to guide managerial decision making, researchers have since stated that it is preferable to have an analytical framework without *a priori* knowledge in the identification of segments to data processing and pre-classification since such an approach fails at capturing possible interactions between the salient variables in a particular market setting [29]. In response, this study provides a novel clustering methodological framework to tackle segmentation and the identification of customer typologies problems in an online marketing context of the social media environment, which has been underexposed in the literature thus far.

Furthermore, Aviad and Roy [3] state results of clustering exercises often require considerable effort by the end user to interpret and find use in the results. Therefore, we develop cluster-specific models predicting customer engagement which identify how the respondents in different clusters differ in terms of their motivations for

engaging with a brand through the social media platform. In addition to this, a new score, the CM1 score is used, which is an approach to identify the most salient features of particular clusters as introduced by Marsden et al. [31] in a study investigating Shakespearean era works.

From here on, this study provides a theoretical background to outline the relevant underlying literature in this area followed by a step-by-step description of our proposed methodology. After this, results of the computational experiments are presented and finally a discussion of results including future research directions is provided.

5.2 Theoretical Background

A theoretical background to the relevant areas in this study is provided. A brief overview of online consumer segmentation and existing clustering and Segmentation Methodologies is presented.

5.2.1 Online Consumer Segmentation

Customer segmentation has become a central concept in marketing and many organizations use segmentation to better serve and satisfy customer needs. As stated in Chap. 2, researchers have based the segmentation of markets on various factors, including cultural, geographic and socioeconomic variables as well as personality, life-style, user status and usage frequency. Customer segments based on these variables may be easy to understand and determine, but may not provide the best possible explanatory power [42]. As a consequence, marketing scholars highlight the need to account for heterogeneous customer perceptions and expectations in order to develop better marketing strategies and allocation of scarce organizational resources [13, 14]. Recent researchers have since argued that contemporary operative environments such as the Internet require sophisticated and sensitive segmentation methods and not 'blindly' follow the process of segmentation purely based on generic descriptors of consumers due to high levels of online consumer-to-consumer (C2C) interactions as well as more empowered consumers [30]. Given these underlying characteristics, researchers argue that segmentation should be seen as a tool in identifying factors with a causal relationship to future consumer behaviours towards a brand [4, 7]. Thus, consumers do not form one homogenous group when it comes to specific online consumer behaviours [1] and differences between these consumers need to be fully understood should marketers wish to target them.

As Bhatnagar and Ghose [4] argue, segmentation based on more than just demographic information has the capability to provide much more useful insights into consumers and understanding of specific consumer groups. Using only a

single base for segmentation, such as age or gender, limits understanding of consumer groups [7]. This is particularly true for consumers' behaviours on the Internet. In previous Internet-based studies such as the investigation of web buyers, demographics cannot discriminate between different types of web buyers as demographics alone do not provide sufficient diagnostic information about web users [4]. In the social media and marketing literature in general, it has become a management imperative for the investigation of consumer behaviours and customer engagement with brands in more detail in the online environment [11, 12, 25]. However, Foster et al. [15] state that only a few studies have aimed to differentiate (or segment) between brand page users to uncover specific behavioural profiles along a continuum of participation. Furthermore, they continue to explain that the research dedicated to segmentation as a marketing tool in the online, and social media context more specifically, falls far behind the volume of research investigating more 'traditional' segmentation strategies. On this basis, a lag exists between academic literature investigating online consumer segmentation based on behaviours towards brands and what is practised in reality by organizations in the online marketplace.

Dividing online consumers into distinct groups (segments) with regard to their different needs, attitudes and behaviours is already commonly used in market research and is used as a basis to target these specific segments with tailored marketing programs [6]. However, findings from McKinsey & Company's (2012) paper on the 'iConsumer' and 'The World Gone Digital' when it comes to understanding digital behaviours, simple categorizations are 'not possible or wise'. In this sense, clustering consumers in the online and social media space needs to account for behaviours and attitudes across a full range of variables in order to find the full levels of diversity existent within the consumer market.

As previously discussed, interpreting and finding use in the results of clustering exercises often requires considerable effort by the user of these results [3]. It is also important to have more than just descriptive information on clusters of consumers as Bhatnagar and Ghose [4] state, diagnostic information should help practitioners make better decisions about their marketing strategies. In other words, having a true understanding of the consumer groups (or clusters) within a market is more useful to brand managers than simple demographic descriptions of multiple groups of people. It is for these reasons that it is important to find a method for describing and uncovering 'hidden' details within clusters. If a segmentation of a group of consumers is done based on a spectrum of behaviours, it is illogical to describe the resulting segments based solely on demographic information. Furthermore, predicting actual online behaviours within clusters cannot be conducted using information based on consumers' age, gender or income. It is for these reasons the current study employs novel methodologies such as symbolic regression modelling and the calculation of the CM1 score (discussed in the next section) to build cluster-specific behavioural models as well as describe the clusters more in depth in terms of behaviours in the social media platform.

5.2.2 Clustering and Segmentation Methodologies

When interested in segmenting and targeting a set or group of consumers with similar behaviours, it is a common practice in marketing research to employ clustering methods in order to group consumers [9]. Clustering consumers is not new in marketing as Klastorin in 1983 [27] discussed that finding homogeneous groups of consumers is beneficial for marketing practitioners and strategists. As Jain [26] (and Chap. 3 of this book) explains, clustering methodologies can be split into two different types: supervised (classification) or unsupervised (clustering). The goal of unsupervised data clustering is to discover the natural groupings of a set of objects or patterns [26]. This means that in order to find the natural groupings (or segments) within a group of consumers, limited to no parameters should be set prior to the clustering exercise; otherwise, the exercise would be closer to classification. As explained by Aviad and Roy [3], clustering activities attempt to partition a dataset into groups so that the entities in one group (segment) are similar to each other and are as different as possible from the entities in other groups (or segments).

A frequently used method in market segmentation is latent class analysis. It is one of the earliest methods adopted by social science and marketing researchers to separate a group of people into latent 'classes' or segments [17]. In more recent years, latent class analysis has been used to segment online consumers. Bhatnagar and Ghose [4] segment e-shoppers based on consumer perceptions and behaviour with respect to online commerce using a latent class analysis. Furthermore, Campbell et al. [7] segment 883 consumers based on their reactions to social network marketing using latent-class analysis and find a total of five segments with distinct characteristics. Other recent methods used in a market segmentation of consumers include k-Means cluster analysis, [4, 41], finite-mixture models in Structural Equation Modelling using Partial Least Squares [30] and a cluster analysis using a set number of variables using Ward's method for hierarchical clustering [19, 35]. However, what these methods all have in common is the number of algorithmic parameters that must be specified by the user. In other words, these methods become less 'unsupervised' when more parameters are introduced a priori. As Jain [26] explains about k-Means clustering methodology, for example, it is that the number of clusters (k) needs to be determined by the user as well as the cluster initialization and he goes on to explain that automatically determining the number of clusters has been one of the most difficult problems in data clustering. Aviad and Roy [3] also explain that in real-life data mining problems, there is no a priori classification making the division process into groups very difficult to define and construct. For these reasons, it is important to consider the number of parameters that are set prior to the clustering algorithm and attempt at limiting the number of user-defined parameters.

One advantage of data clustering consumers into homogeneous segments is that of the identification of central points of segments, which can be treated as ideal points to reflect the customer requirements of the consumers inside the segment [8]. In classification (supervised) problems, for example, a common objective is feature selection where the goal is to determine the best or most parsimonious set of variables for the algorithm. In clustering exercises, it is also important to find distinctive characteristics of the resulting groups but the goal is different to classification as the aim is to determine those features (or variables) that lead to a maximum differentiation of each cluster (segment) based on its members and their characteristics [3]. In the context of consumer segmentation in marketing, statistical analyses for significance such as an ANOVA test between clusters are conducted, e.g. [19, 41]; however, these methods are not capable of describing the clusters or finding those features that truly distinguish between clusters of consumers. In other scientific fields, various methods have been introduced to address this problem. One of these is the computation of the CM1 score which finds those features that are the most clearly identifiable characteristics of each cluster. This score has previously been applied in the identification of panels of biomarkers for breast cancer subtypes. In this instance it was used to identify the transcriptional state of genes that are consistently 'over-expressed' or 'under-expressed' in each subtype [32]. Analogously, it was used in a computational stylistic study that aimed to identify words which were used differently by Shakespeare and his contemporary peers [31]. In these studies, the CM1 score helped to analyse and describe the various clusters (or subtypes) in further detail and identify salient features of each group. We refer the reader to the publicly available online publication where the score was introduced [31] for its definition and its use in a different study. Methodological explanations of the use of this score in this context are provided in the following section.

5.3 Materials and Methods

This chapter uses the small online customer engagement dataset which is outlined in Sect. 26.2.4 of Chap. 26. In this dataset, 371 respondents answer questions about their online engagement behaviour with brands through Facebook. As stated in Sect. 5.1, insights are needed on top of online metrics in order to truly understand and capitalize on consumers' motivations for engaging with brands and companies in the online environment. Here, the method of this study is outlined including the design, distance measure used for the clustering methodology, the CM1 score calculation and symbolic regression predictive modelling process.

5.3.1 Method Design

There are several stages to the methodology used in this study. Firstly, a brief outline will be given of the questionnaire tool construction and explanation of the dataset's basic characteristics. Then, the methodology for our clustering method is also conducted by going through several stages. Firstly, the Spearman's rank correlation coefficient is computed for all items for all respondents which results in the generation of a distance/similarity matrix. These distances will provide the basis for a graph which is a combination of a Minimum Spanning Tree (MST) and a k-Nearest Neighbourhood (kNN) Graph. The MST-kNN agglomerative algorithm and graph will provide the resulting clusters found by this study. Whilst generating these graphs, they will be visualized and analysed in order to describe and outline the segments found through our clustering method. In this, concepts such as 'node betweenness centrality', see Chap. 8 for the definition of this one and other measures. Basic layout algorithms are also used for visualization purposes. A brief inspection of the demographic and technology usage information will be done followed by the use of the CM1 score [31]. After this description of the clusters, each cluster will be analysed using symbolic regression analysis. The purpose of doing so is to identify mathematical models for average customer engagement levels to suit each cluster. This will further identify the characteristics that set each cluster apart, as it may be expected that different variables affect the level of customer engagement in each cluster.

5.3.2 Distance Measure

Before we continue with the clustering analysis in this chapter, a distance or similarity measure needs to be calculated for each of the data points. In this study, the Spearman's rank correlation coefficient was preferred. The Spearman's rank correlation coefficient is a non-parametric measure of statistical dependence between two variables which is appropriate for discrete variables, including ordinal ones. A perfect Spearman correlation between two variables may indicate that they are related via a monotonic function, whilst, in contrast, the Pearson correlation will only attain the maximum value when the two variables are related by a linear function. Herlocker et al. [21] have been critical about the assumption of linearity and since the Spearman's rank correlation does not rely on the assumption of linearity or other assumptions, it is preferred to be used in this case. Particularly due to the reason that we have a range of variables on a Likert scale we have preferred it as a measure of similarity [5].

5.3.3 Background of the MST-kNN Clustering Algorithm

Firstly, a brief introduction to Graph Theory is presented here in order to provide context. A simple undirected graph is denoted as $G(V, E)$ in which V is a non-empty set of vertices (also called nodes) and E is a set of unordered pairs of distinct elements of V called edges. An edge weighted graph is denoted as $G(V, E, W)$ in which V and E are defined as before but each edge now has associated a weight and W is a set of weights. We refer to the sets E and V as $E(G)$ and $V(G)$, to indicate that they are the set of edges of G and analogously, the set of nodes of G, respectively [18].

A path in a graph $G(V, E)$ is a sequence of edges which connects a sequence of vertices. In an undirected graph $G(V, E)$, we say that two nodes a and b are connected if the set of edges E contains a subset of them that form a path between nodes a and b. A graph is said to be connected if every pair of vertices in the graph is connected. A connected component of a graph is a maximal connected subgraph of G; in this case, each node and each edge belong to exactly one connected component. A simple undirected graph is a tree if in it any two vertices are connected by exactly one simple path. A graph is a forest if it is a disjoint union of graphs that are all trees, which means that in a forest all the connected components are trees. Given a connected, simple undirected graph $G(V, E)$, a spanning tree of that graph ($MST(G)$) is a subgraph that is a tree and connects all the vertices together. Given a graph G we can enumerate all its spanning trees and order them according to the total sum of weights of all edges of the tree. Accordingly, a tree is a Minimum Spanning Tree of G, denoted as $MST(G)$, or the Minimum Weight Spanning Tree if it is a spanning tree of G with the total sum of weights of its edges (its weight) being less than or equal to the weight of every other spanning tree of G.

Inostroza-Ponta et al. [22] have proposed a clustering algorithm known as MST-kNN; an agglomerative method combining the outputs given by a Minimum Spanning Tree and the k-Nearest Neighbour (kNN) algorithms. This method has been further explained in Chap. 3. The MST-kNN method has been tested on comprehensive studies on large-scale biological weighted networks and it has been successfully applied in various areas, see, for instance [2].

MST-kNN performs better than some other known classical clustering algorithms (e.g. k-Means and SOMs) in terms of homogeneity and separation [22, 24] in spite of not using an explicitly defined objective function, but using a clear stopping criterion instead. Due to its characteristics, it performs well even if the dataset has clusters of different mixed types (i.e. MST-kNN is not biased to 'prefer' convex clusters or when the data has clusters that are embedded in subspaces of different dimensionalities). Most importantly, the MST-kNN algorithm scales very well, allowing the possibility that the methods described in this paper can be extended to the analysis of very large datasets including questionnaires and other marketing datasets involving millions of consumers, online behaviours, brand pages or even brands on dedicated hardware.

The MST-kNN can be classified as a constructive heuristic that is not biased for the choice of a particular objective function, yet it provides a strong guarantee of optimality of a property of the final solution. We explain this property after we explain the algorithm. First, the algorithm's input can be either a distance matrix between all pairs of nodes or a weighted graph. In this study, a dissimilarity matrix that it is computed from the Spearman's rank correlation matrix is the input for the algorithm. Formally, if $r(a, b)$ is the Spearman's rank correlation between two respondents a and b over the set of questions, then the corresponding distance matrix $D = [d(a, b)]$ with each coefficient is calculated as $d(a, b) = 1 - r(a, b)$. Given this input matrix D, the output of the MST-kNN algorithm is a forest. This means that the MST-kNN generates a partition of a set of nodes given as an input using the information of similarities/dissimilarities between each pair. It not only gives a partition of the nodes but some of them are connected by edges.

We mentioned that the algorithm returns a forest that satisfies a property. The set of nodes are the ones that are part of the input. In the forest given as output, any edge of the forest that connects two nodes does so if the edge is one of the edges of the minimum spanning tree ($MST(G)$) and, at the same time, it is also an edge present in the set of edges of the k-nearest neighbour graph ($kNN(G)$). The k-NN graph is the graph that has one node per object and that has an edge between each pair of nodes, for example, a and b, if either a is one of the k nearest neighbours of b or if b is one of the k nearest neighbours of a, or both. We note that edges of the minimum spanning tree are not bound to have this property regarding 'k-neighbourness' and, the addition of this extra constraint has the effect of disconnecting the MST, thus creating a multi-tree forest and consequently leading to a natural partition of the set of nodes.

There are several variations of this scheme. In one of them, the value of k is set up to a relatively large value which is linked to the total number of nodes, and then, when the MST is fragmented in different components, a different value is selected for the different connected components using the same formula but now having for each of the connected components the number of nodes in each of them as input, thus leading to different values of k for each component. Another approach is the one we have used in this work, in which a value of k is fixed. In this paper we studied the cases of $k = 1$ and with the automatic selection of k in this study (i.e. $ln[n] = ln[371]$, since we have 371 samples in this dataset). Inostroza-Ponta et al. [22–24] outline the details of these methods and their applications to other real-world problems.

5.3.4 CM1 Score Calculation

After the clustering analysis, the clusters were investigated based on demographic and technology usage information. However, as stated, we also found that simple demographics do not provide intricate detail in the true underlying behavioural characteristics of each cluster. Therefore, in order to further describe and investigate

the clusters in terms of online consumer behaviours, customer engagement and attitudes, we use the CM1 score to rank the answers of respondents that belong to two clusters. The score was recently introduced in a comprehensive study of the identification of word usage that would discriminate between authors. It was applied to produce models of authorship of a large group of plays from the Shakespearean era in [31]. Like the t-test, in order to calculate the CM1 score of the participants' responses to a question we first need to compute the difference between means of samples in the two groups of participants X and Y. Typically, X is the set of participants in a cluster and Y is the set of participants which are not in it. The distinctive characteristic of this score in comparison with the t-test is that it is moderated by the range of values of the set that has the largest set of samples, rather than the combined standard deviation of X and Y. We refer to Marsden et al. for details of this score [31].

5.3.5 Symbolic Regression Analysis

In this study the aim is to identify clusters (segments) of consumers that appear naturally based on their online behaviours, rather than examining descriptive information and without setting a priori parameters. To continue this data-driven approach, symbolic regression is used in order to find models for customer engagement for each of the clusters, which is a concept that has received increasing interest in recent years in the area of online consumer behaviour [25, 39]. Considering we are segmenting consumers based on their online behaviours, it is of considerable interest to generate cluster-specific models to predict online customer engagement with brands. As we have argued, interpreting results of clustering and segmentation exercises could include considerable effort by those marketing managers who are the end users of the clustering results. Therefore, we argue that investigating cluster-specific behavioural models assists marketing and brand managers in interpreting clustering and segmentation results.

Unlike numerical regression methods, in which model hypotheses are generated and fit to available data, symbolic regression discovers not only the coefficients within a structure, it also involves the discovery of the structure in the data and, consequently searches for the structure of the resulting models. Symbolic regression is defined as 'finding a mathematical expression, in symbolic form, that provides a good, best or even perfect fit between a given finite sampling of values of the independent variables and the associated values of the dependent variable(s)'. Stated more simply, the process of symbolic regression involves finding a model that best fits a given set of data. The main advantage of this method is that the researcher does not have to specify the structure of the regression model in advance [16].

In order to keep the method proposed in this study easily adoptable for future research by marketers and researchers, we use an open access software package named Eureqa [37]. Eureqa provides the user with a clear user interface, is free for academic use and the output is a Pareto optimal curve that trades model fitting for its complexity which aids the researcher in making the decision of this trade-off which is a significant problem as Smits and Kotanchek [38] point out. In this contribution, we have used the Eureqa Desktop package for the identification of models of average customer engagement in different clusters. We have restricted the search to only models that employ the basic 'building blocks' of 'addition', 'subtraction', 'multiplication', 'introduction of a constant value', 'the introduction of integer and real values' and 'the introduction of new input variables' in the expressions it generates. The two objectives of the Pareto Optimality Curve are the fitness of an error metric which represents the expression accuracy (which is user defined) and the 'complexity' of the model. Eureqa uses an ad hoc approach to define what the 'complexity' of a model is, which is the sum of the complexities values attributed to the use of each of the individual operations used to generate the formula. Our selected basic 'building blocks' have the minimum individual complexity. The 'error metric' that we have used was the 'Correlation coefficient', this means that during the evolutionary computation procedure Eureqa iteratively tries to find models that maximize the normalized covariance. Putting these things together, Eureqa helps to try to find a scale and offset invariant model that has the 'shape' of the data whilst at the same time uses as few input variables and mathematical operations as possible, giving a good trade-off of input selection and trend behaviour without risking over-fitting the data.

Through employing symbolic regression analysis, we aim to find a function for levels of online Customer Engagement within each cluster. In doing so, we iterate the same process for each of the clusters found by the MST-kNN clustering algorithm and subsequently select the best-fitting linear solution as found by Eureqa. This means that at the end of this process we have a function for Average Engagement (ENG) for each of the clusters that shows which of the input variables are relevant to that cluster. The results of this process are presented in the following section.

5.4 Results

As outlined in the materials and methods section, there are several stages to this study. Therefore, the results will be presented according to the order of these stages.

5.4.1 Distance Measure

As explained in the previous section, the basis for the clustering algorithm in this study is a Spearman-based distance matrix. As the sample for this study contains 371 respondents, this is a 371 × 371 matrix containing all the values for the Spearman correlation. In order to investigate this information, we identified those respondents who are the 'closest' in terms of Spearman correlation and those who are the 'furthest'. In so doing, we also aim to emphasize the level of heterogeneity amongst customers in terms of their online behaviours with a brand focus.

Figure 5.1 shows two participants who have similar answer profiles of which one has selected an online clothing retailer (BlackMilk Clothing) and one a photography business (see Bliss Photography). These brands are two different types of businesses, an online fashion retailer and a photography service. However, these two respondents are the most similar in terms of their Spearman correlation in the whole sample, showing that those customers of the same brand do not necessarily have to be similar in their behaviours towards the brand. This highlights the focus of this study of finding homogenous groups of people in terms of their behaviours rather than their basic or demographic information. This also means that companies may need differentiated online strategies for their various customers across varying brand levels or sub-brands.

Figure 5.2 shows two participants who have selected a travel webpage (HIS Travel) and a clothing brand retailer (Portmans). These two figures illustrate that there exists wide heterogeneity in the responses of the participants regarding their online consumer behaviours and were therefore found to have the 'furthest distance' from each other based on Spearman correlation. As shown in Fig. 5.2, for the first half of the questions in the survey the respondent who had selected Trip H.I.S. answered significantly higher than the respondent who answered Portmans. Amongst these questions were questions about the respondent's Usage Intensity of the branded social media page, their brand involvement with that brand, their level of self-brand congruency with that brand's image as well as the levels of functional value and hedonic value of the brand's Facebook page. It is in the questions regarding 'flow' and the level of perceived social value of the brand's Facebook page where the respondents are closer in their responses. However, for the remainder of the questions, the respondent with the brand Portmans answers higher than the respondent with Trip H.I.S. These questions relate more to the actual experience on the brand's Facebook page and their intention to interact and engage with the brand online in the future. We can speculate one possible reason for these differences in responses. The person who selected Trip H.I.S. may be highly involved with this brand but may not have such a good experience online, whereas the opposite may be true for the Portmans respondent.

As shown in both Figs. 5.1 and 5.2, heterogeneity exists amongst online consumers in terms of their behaviours and perceptions of the online experience. Therefore, we continue our analysis of this sample. After the Spearman-based distance matrix is computed, a minimum spanning tree is created. In this minimum spanning tree, the k-Nearest Neighbour algorithm is subtracted to reduce the number of edges which results in clusters (or, a subset of trees) that contain nodes that are only connected when they are nearest neighbours as explained in the previous section.

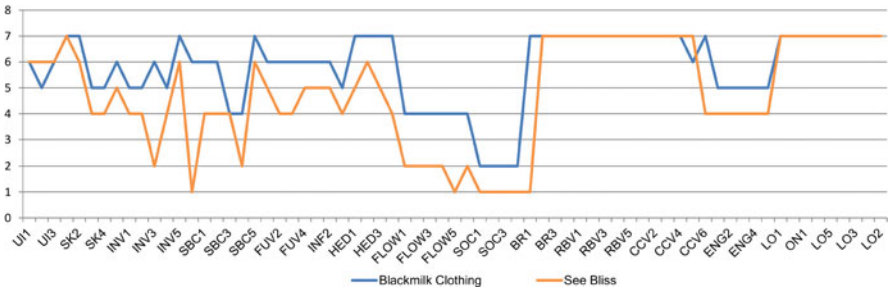

Fig. 5.1 Homogeneity between respondents—the pair of respondents that have shown the highest degree of homogeneity (closest in Spearman's Rank correlation) and their selection of two different brand choices

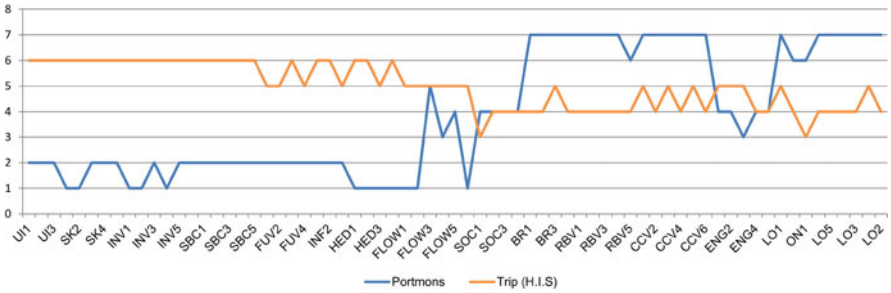

Fig. 5.2 Heterogeneity between respondents—the pair of respondents that have shown the highest level of heterogeneity in their responses ('furthest' in Spearman's Rank correlation) and their choices of brands

5.4.2 MST-kNN Clustering Results

In this section the results of the MST-kNN clustering method are presented that have been applied as described in our methodology. Firstly, we present the clustering method with the value of $k = 1$, followed by the automated value of k using the natural logarithm of n. Figures of these outcomes are displayed.

5.4.2.1 MST-*k*NN Algorithm with $k = 1$ Results

As stated, we have used the similarities produced by the Spearman's rank correlation method to create a distance matrix amongst all the respondents of the questionnaire. If we denote with $r(a, b)$ the Spearman correlation between two respondents a and b, then we have computed their distance as $d(a, b) = 1 - r(a, b)$. Using this distance matrix between respondents, we then applied the MST-*k*NN algorithm to find a clustering of the respondents in highly similar groups.

The results for the MST-*k*NN clustering method with $k = 1$ are presented in Fig. 5.3. We can see that all the respondents are grouped in 45 different clusters.

Fig. 5.3 Results of MST-*k*NN with $k = 1$. A result of 45 small clusters is found, which means that this result is likely to be the 'upper bound' of the heterogeneity in this dataset

Edges connect the respondents and each of these edges belong to a minimum spanning tree computed from the distance submatrix obtained by only taking into consideration the participants in that cluster.

An important property to highlight to the reader, when examining the figure, is that when the value of k is set to 1, in the shown solution of the MST-*k*NN, each edge that connects two respondents a and b indicates that either a is the most similar respondent to b in the entire dataset, or b is the most similar respondent to a in the entire dataset, or both. This means that this partition and visualization also offers an interesting alternative for data exploration and provides some insights on cluster structure. Results with $k = 1$ are a natural bound for this method, indicating that 45 is likely the upper limit of the heterogeneity present in the natural clusters in this dataset when this approach is used.

5.4.2.2 MST-*k*NN Algorithm with Automatic Setting of the Value of *k*

The alternative approach for computing a high-level description of the dataset, and a partition in clusters having a larger number of respondents is by direct application of the MST-*k*NN algorithm now with the automatic selection of k. The automatic selection of the value of k takes the natural logarithm of n, in this case $ln(371) = 5.92$ which is rounded up to $k = 6$. The algorithm will produce both a partition of the set of respondents in clusters, but also, like before, produce a set of edges connecting those respondents. An edge between two respondents a and b will indicate, in this case, that either person a is the closest (or up to the sixth closest) person of person b, or that b is the closest (or up to the sixth closest) person of person a, or both. The final result is found in Fig. 5.4 indicating five clusters (segments) of respondents.

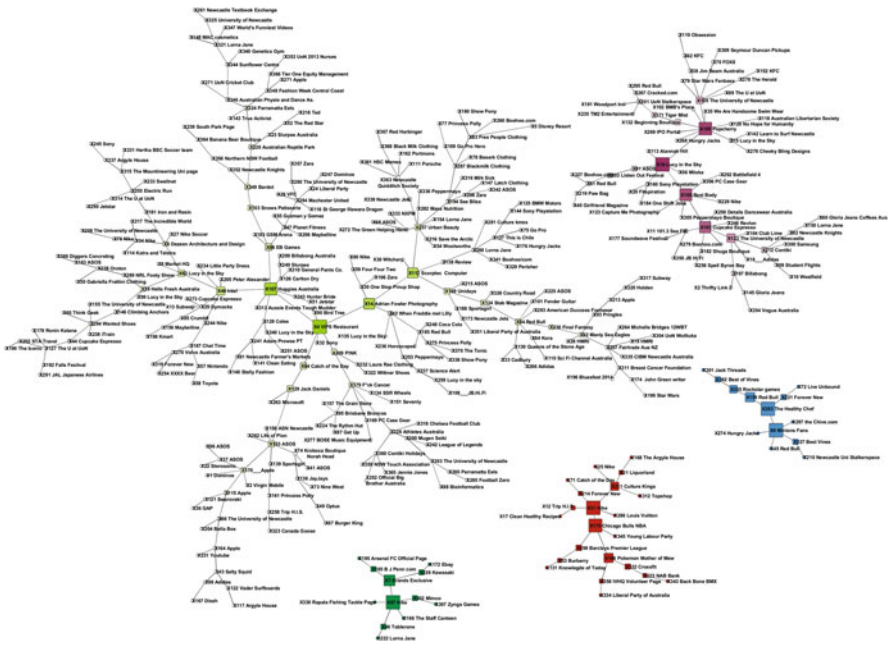

Fig. 5.4 Results of MST-*k*NN with the default setting for k. Five clusters are found when the value of k is automatically determined (and rounded up so $k = 6$ in this case). The size of the nodes indicates 'node betweenness centrality'. Different colours represent which cluster the nodes are part of and the shade of the node colours also represents 'node betweenness centrality'

The clustering with the automatic selection of k gives us five clusters of consumers in the data. One of these is clearly much larger than the other clusters; however, we are still able to examine these clusters in order to describe them. The largest cluster identified, Cluster One ($n = 250$) is shown in light green in Fig. 5.4 and has several large nodes based on computing the 'node betweenness' centrality measure.

The large nodes at the very centre of Cluster One represent respondents who selected a restaurant, a nappy brand, a photographer's brand page and a consumer electronics brand page in the questionnaire survey. The second largest cluster identified is Cluster Two ($n = 68$) and is shown in purple in Fig. 5.4. The largest nodes in this cluster represent respondents who selected a cafe, two online shopping retailers and a fitness page. Cluster Three ($n = 12$) is shown in dark green in which the two central largest nodes stand for a respondent who answered in reference a sports apparel brand and another respondent who answered in reference to an online shopping retailer.

Cluster Four ($n = 13$) is shown in blue with the central nodes in this cluster representing respondents who selected an energy drink brand, a healthy food brand page and a movie brand fan page. Finally, Cluster Five ($n = 22$) shown in red and labelled has several large nodes including respondents who answered a 'street wear' clothing brand, a sports apparel brand, an American basketball team page and a commercial television series brand page. Again, the central nodes in these clusters show that people who are homogenous in terms of their online behaviours and ways in which they communicate and interact with their 'favourite' brands online could come from a very heterogeneous group in terms of actual brand preferences and other descriptive information.

5.4.3 Describing the Cluster Results

In order to describe and analyse the clusters found by the MST-*k*NN algorithm we will conduct several steps. Firstly, each of the five clusters will be described in terms of their demographic information and their technology usage profile. Secondly, we will compute the CM1 score to identify those behaviours (variables) that are of particular importance in identifying each cluster. Finally, the symbolic regression analysis is conducted as described in our proposed methodology. Following this, we finish with a discussion of our results and a guide for future research suggestions.

5.4.3.1 Basic Description of Clusters

As shown in Fig. 5.4, Cluster One is the largest cluster comprising 250 respondents. Within this cluster, the age ranged from 17 to 49 years old and consisted of 41.8% males and 58.2% females. Furthermore, 73.8% of respondents in this cluster chose a service brand in the survey questions and 26.2% a product brand. In the

technology usage profile of this cluster, we can find that the majority are experienced Facebook users having had an account for over 3 years (82.0%) who have 'liked' the page they selected for 6 months or more (62.5%). What is also interesting about these experienced social media users is that the majority of respondents access the Facebook platform from a mobile telephone or device. Consumers accessing Facebook from a laptop, mobile tablet or mobile telephone account for 86% of the sample.

Figure 5.4 also shows that Cluster Two is the second largest cluster found in the data. In this cluster, the age ranges from 18 to 45 years old with an average age of 21.38 years old. This cluster comprises of 33.8% male respondents and 66.2% female respondents. Furthermore, in cluster two, 83.8% of respondents selected a service brand and only 16.2% selected a product brand. In examining the technology usage profile of cluster two, it is clear that this cluster has similarities with cluster one. The majority of respondents in this cluster have had a Facebook account for longer than 3 years (80.9%) and have used and interacted with the Facebook brand page they selected for more than 6 months (55.9%). This shows that clusters may not necessarily be identified solely based on basic information. Furthermore, the majority of the cluster accesses the Facebook brand page from a mobile phone device (52.9%) and state that they are signed in all the time (45.6%). On this basis, we can deduce that this cluster forms an experienced group of Facebook users who are 'tech-savvy' accessing the social media platform mostly from a mobile device. This may have implications in the way that they interact with brands through social media due to the different functions and displays of mobile devices to desktops.

An examination of Cluster Three's descriptive information indicates similarity to the whole sample and other clusters, where the average age in Cluster Three is 21.25 years old ranging from 19 to 30 years. Furthermore, there was an exact 50% split between male and female respondents in this cluster and 53.8% of those respondents selected a service brand, whilst 41.7% of respondents selected a product brand. The technology usage profile of cluster three shows that although overall this cluster is still similar in terms of average age, duration of having a Facebook account and so forth, there are a few differences in this cluster. For example, 25% of this cluster accesses the Facebook brand page from a home desktop PC which is higher than the other clusters. This may impact on the way in which these consumers interact online and the way in which they like to use the social media site. Moreover, only 33.3% of this cluster indicate they are 'signed in all the time' through a mobile device which is also lower than the other clusters.

An examination of Cluster Four indicates an average age of 21 years old with the age ranging from 18 to 30 years. 61.5% of the respondents in this cluster are male and 38.5% are female. Similar to the other clusters, more respondents selected a service brand (76.9%) than a product brand (23.1%).

Analysing Cluster Four's technology usage shows that these respondents are experienced Facebook users with 84.6% having had a Facebook account for 3 years or more and 61.5% of the respondents indicate being 'signed in all the time'. This figure is almost twice in size as that of Cluster Three where only 33.3% of respondents indicate being signed in all the time. Furthermore, consistent with the respondents of Cluster Four being signed in all the time, 53.8% of these respondents access Facebook from a mobile device.

An examination of Cluster Five ($n = 25$) indicates an average age of 22 years old with the age ranging from 18 to 41 years which is slightly older than Cluster Three and 4. In this cluster, 50% are male and 50% female. Again, more respondents chose a service brand in their responses (72.2%) than those who chose a product brand (27.3%). Cluster Five again shows a group of respondents who are experienced Facebook users with 81.8% of respondents having had an account for 3 years or more and 59.1% indicating that they are signed in all the time. Exactly 50% of this sample answer that they access the brand's Facebook page from a mobile device and 40.9% from a laptop. This may indicate that these respondents engage in these online activities on the go which may impact on their actual online behaviours towards brands.

Based on the above analysis, the demographic information is not providing novel insights into understanding each cluster and its characteristics. This being the case, we propose the use of the CM1 score in the context of online consumer behaviour and segmentation studies as a better approach to investigate and describe clusters (or segments) of consumers.

5.4.3.2 CM1 Score Results for Clusters

The characterization of the major differences between the clusters is presented using the CM1 score to describe and label the clusters in further detail and provide further insights. Figure 5.5 shows the curves of the CM1 scores ordered from smallest to largest values which show the 'lowest' to the 'highest' scoring identifiable features for each cluster. At the very bottom and very top of each of the bar charts, several features protrude further than other bars which we call the 'shoulders' of each of the CM1 score curves.

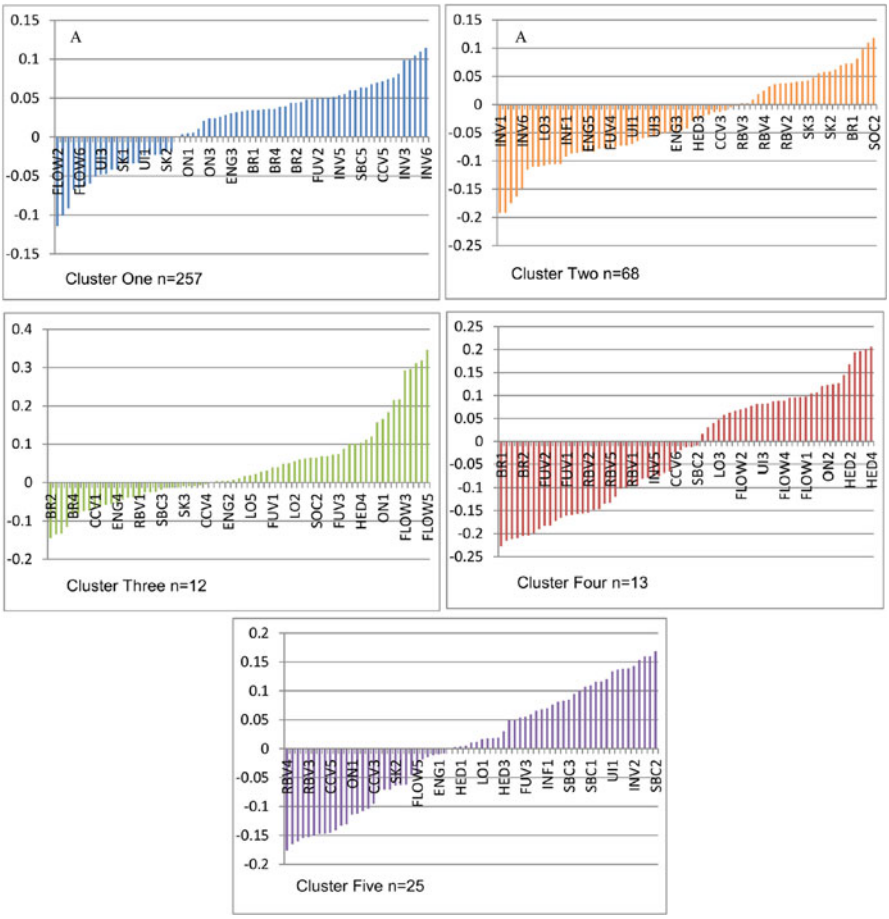

Fig. 5.5 CM1 Score computations for each cluster—the CM1 score has been computed for each cluster. The values were sorted from lowest to highest showing the 'bottom' to 'top' CM1 scores for each cluster. They are presented in the five graphs with data from Clusters 1 to 5, respectively. For most clusters bottom and top 'shoulders' can be seen in which there are several features that are higher or lower than the general trend in each curve

These features are also shown in Table 5.1 where each of the bottom and top five features are shown for all clusters. Those features that have formed a 'shoulder' in the curves of Fig. 5.5 are shown in bold in Table 5.1. Some clusters did not have a 'shoulder' on one or both ends and for these clusters the bottom five or the top five features are used for analysis.

Starting with Cluster One, we see that all bottom five features include the features relating to the Flow construct. When a person experiences 'flow' when using a web page, social media page or is engaged in an activity that person is in a perceived state of effortless action, loss of time and sense that the experience stands out as being

Table 5.1 Five bottom and top CM1 Scores for each cluster—highlighted in bold are those features that represented a 'shoulder' in the CM1 score graph for that cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Five 'lowest' CM1 scores				
FLOW2	**INV1**	**BR2**	BR1	RBV4
FLOW5	**INV4**	**CCV5**	SOC1	RBV2
FLOW1	**INV2**	**BR1**	INF2	RBV5
FLOW4	**INV3**	**SK2**	INV6	BR2
FLOW6	**INV6**	BR4	BR2	RBV3
Five 'highest' CM1 scores				
INV3	BR1	**FLOW3**	**HED2**	INV2
INV2	SK5	**FLOW4**	**HED1**	**UI3**
INV1	**SOC4**	**FLOW2**	**FLOW5**	**UI2**
INV4	**SOC1**	**FLOW1**	**ON1**	**INV4**
INV6	**SOC2**	**FLOW5**	**HED4**	**SBC2**

exceptional when compared to 'normal' activities [34]. For example, the question relating to FLOW2 (the very most bottom CM1 score) stated: 'When I am visiting (using/operating) my favourite Facebook brand page: I lose track of time' (rated on a scale 1–7). For this cluster this means that these consumers, on average, have lower scores for questions relating to flow than all other clusters. When inspecting the technology usage profile of this cluster, it may provide a reason why this cluster scores low on flow. Nearly 60% (59.4%) of this cluster indicated that they engage with the Facebook brand page through a mobile telephone or a mobile tablet. For a consumer to experience flow, the environment needs to be appropriate to 'lose time and sense of self' which means that this may not be possible on small mobile device screens to deliver a media rich experience for flow to occur. Furthermore, the top features of this cluster show only variables relating to the 'Brand Involvement' construct which aimed at gathering an understanding of the respondents' previous (offline) involvement with the brand. For example, the very most top feature, INV6 stated: 'I am involved in/with this brand'. Considering that the top CM1 features only included brand involvement variables, we can understand that the consumers in this cluster engage with this brand online as they feel as though they are 'involved' in/with this brand, that the brand is significant and important to them or whether the brand means a lot to them and not just to entertain themselves and 'lose track of time' online. This being the case, we label this cluster as 'brand rationalists' since members of this group concentrate their time in consuming brand pages via a mobile device for convenience and given they are more highly involved fans of the brand than other respondents, content containing cognitive aspects of a utilitarian nature can be argued to be of greater importance to this segment than other respondents.

As can be seen in Table 5.1, Cluster Two shows an interesting contrast where all bottom CM1 scores are exactly those variables that are in the top of Cluster One's CM1 scores. That is, five of the six items relating to the '**Brand Involvement**' construct. This means that the respondents in this cluster have, on average, a lower 'involvement' with the brand they engage with online than all other clusters. The

three most salient features for this cluster, and those that represent a 'shoulder' in the CM1 curve all relate to the 'Social Value' construct. The strongest variable, SOC2 stated 'I can meet new people like me on this Facebook brand page' (rated from 1 to 7). This indicates that these respondents perhaps use the Facebook brand page in a different way; to socialize and interact, rather than become truly involved in the brand. This being the case, we label this cluster as 'passive socializers' since members of this group concentrate their time interacting with other consumers on the brand pages via a laptop or mobile device at least once a week. Given these consumers are lowly involved fans of the brand, content of a hedonic, utilitarian and co-creative nature is of greater importance to this segment than other respondents.

Cluster Three also provides an interesting case to compare to Cluster One as all the items relating to 'Flow' that are found in Cluster One's bottom features appear in Cluster Three's top CM1 score features. Oppositely to Cluster One, this means that these respondents, on average, have higher scores for questions relating to 'flow' than all other respondents. Interestingly, 50% of this cluster responded that they access the Facebook brand page from either a home desktop PC or a laptop computer, as shown in this cluster's technology usage profile. As stated, for 'flow' to occur, the environment has to be suitable and this could be a reason why this cluster is more likely to experience 'flow' when using the Facebook brand page.

An examination of Cluster Three's bottom CM1 scores indicates a small variety of additional features. For instance, BR2, the lowest CM1 score belongs to the 'Brand Interaction Value' construct and specifically the item 'I can communicate with the brand on this Facebook brand page'. The second lowest CM1 score, CCV5 comes from the 'Co-creation Value' construct and the item: 'The Facebook brand page allows my involvement in providing services to me to get the experience that I want'. This suggests that these respondents are perhaps less likely to seek direct interaction and dialogue with the brand and more likely to 'browse' and 'lose track of time' whilst using and consuming content on the Facebook brand page. These stark contrasts between clusters provide great insights to the motivations and reasons for consumers interacting and engaging with the Facebook brand pages. This being the case, we label this cluster as 'immersers' since members of this group concentrate their time interacting with the brand page from home desktop PC, laptop and mobile device at least once a week and given they are lowly involved fans of the brand, content of a utilitarian and hedonic nature to induce flow experiences is of greater importance to this segment than other respondents.

Cluster Four and Five show different and unique cases. Cluster Four specifically does not have a conspicuous bottom 'shoulder' of CM1 scores; however, when analysing the bottom five features, two 'BR' and two 'INF' items appear. As stated, BR related to 'Brand Interaction Value', whereas INF relates to 'Informational Value'. Similarly to Cluster Three, perhaps these respondents are less concerned about having direct interaction with the brand through the Facebook page and less likely than other respondents to seek valuable information through the Facebook

platform. The top CM1 scores for Cluster Four include three items from the 'Hedonic Value' construct, one 'Flow' item and one 'Online Loyalty Behaviour' (ON) item. Hedonic value relates to how entertaining and 'fun' that respondent perceived the Facebook brand page to be. For example, the very top CM1 score, HED4 stated: 'The content of the Facebook brand page is entertaining'. One could argue that when a person has fun, or is entertained by a certain online activity, it is more likely for that person to enter a 'state of flow' which could explain the reason for these items appearing in this cluster's top CM1 scores together. The item relating to online 'loyalty' behaviours, ON1 specifically stated: 'I will share this brand's Facebook page content in the future'. Logically, if a person thinks that specific content is 'fun' and 'entertaining' they are more likely to want to share this with their friends. This being the case, we label this cluster as 'hedonic sharers' since members of this group concentrate their time interacting with the brand page for the enjoyment, fun and entertainment of the social media experience on the brand page whilst sharing content with others to support the need for enjoyment and hedonic gratification. Therefore, content predominately hedonic in nature combined with social and utilitarian content is of greater importance to this segment than other respondents.

Finally, Cluster Five also does not possess a specific bottom 'shoulder' of CM1 scores, which is why we investigate the bottom five features. Four of these features relate to the 'Relationship Building Value' construct (RBV) and one 'Brand Interaction Value' construct. These two constructs include statements about interaction, relationships and direct communication between the respondent and the brand on Facebook. For example, the most bottom feature (RBV4) stated 'The Facebook brand page is committed to delivering add-in values (e.g. special offers, member programs) to keep me loyal to the brand'. This shows that these respondents in Cluster Five score, on average, lower in these questions than all other respondents. On the other hand, the top CM1 scores of this cluster include four features that form a top 'shoulder' and include SBC2 which is part of the 'Self-Brand Congruency' construct, INV4 (Brand Involvement) and UI2 and UI3 which are part of the 'Usage Intensity' construct. Usage intensity questions aimed at gaining an idea how often consumers 'used' the Facebook brand page. For example, UI3 stated 'I regularly use the Facebook brand page' (rated from 1 to 7). This indicates that although these respondents may value interactions with the brand through Facebook less than other clusters, they may engage with those brands they feel that are congruent to their own personality and they so at a higher usage intensity than other clusters. This being the case, we label this cluster as 'active participators' since members of this group identify themselves and their self-concept with the brand and regularly visit and use the brand page, which is reflected in the high result for how often the brand uploads new content on its Facebook page (i.e. to facilitate regular interaction and revisits to the brand page), how these consumers visit once a week or more and the high percentage of consumers who are signed in all of the time into Facebook. Therefore, regular posting of content is of greater importance to this segment to the sampled population to than other respondents.

From this CM1 score analysis, we see varying clusters of consumers forming in terms of their online behaviours and specific characteristics. Whilst Cluster One [brand functionalists] respondents are more 'involved' with the brand where they choose to interact with through social media, Cluster Two [passive socializers] respondents are the exact opposite and place more value on the social experience online. Furthermore, Cluster One is the least likely out of the whole sample to experience a state of 'flow' owing to the need for convenience and access to utilitarian oriented content, whilst for Cluster Three [immersers], flow was the highest scoring CM1 score.

We now take one step further in investigating the clusters found in this study by mathematically modelling cluster-specific customer engagement models. That is, using a data-driven symbolic regression analysis, we aim to find predictive models for online Customer Engagement for each cluster separately. In doing so, we provide a basis for future targeted marketing strategies with the objectives of driving higher levels of customer engagement with the brand. Using all other variables in the study, models to predict Customer Engagement are found that are specific to each cluster which go above and beyond the description and explanation of clusters using basic information.

5.4.4 Symbolic Regression Analysis: Cluster-Specific Behavioural Model Building

In this section, we provide an identification of models for Customer Engagement for each of the clusters. As stated, using and interpreting the findings of a clustering or segmentation analysis remains a challenge in practical settings. As we have highlighted, online customer engagement has been of considerable interest to marketing scholars and practitioners alike in recent years. This provides the motivation for cluster-specific model building in order to predict customer engagement in the social media environment. After segmenting consumers into distinct segments, the next logical step is to prepare targeted marketing strategies. However, in order to achieve this, the marketer needs to thoroughly understand their consumers. Whereas the objective of using the CM1 score was to describe the clusters, here we aim to find predictive models of online customer engagement with brands to provide greater insight into each of the clusters. These predictive models could subsequently be used for guiding targeted marketing strategies.

As the theoretical construct of Customer Engagement contained five items, the average of these items for each respondent has been taken to create one variable 'AVENG' to be modelled by our method. As explained, Eureqa builds a Pareto optimality curve to plot the trade-off between complexity and the error value of each model it finds in the data. This curve assists the user in selecting the best model found, considering the appropriate level of complexity as well as accuracy

depending on the selected error metric. Furthermore, as we explained about Eureqa's 'building blocks', we restricted the search to only use models that involve the mathematical operations of addition, subtraction, multiplication of variables and the introduction of an integer constant. These integer constants can then appear as a coefficient multiplying a variable or as an additive term somewhere in a formula. The guiding function used for the search is the Correlation Coefficient, this means that the program's task is to find a model of the Average Engagement (labelled AVENG) that highly correlates with the observed values, rather than, for example, minimizing actual error. Here we report the best simple linear models that do not include or integer constants found by Eureqa and provide some initial insights. The meaning of results for each cluster is explained further in the next section.

For all clusters, the best (in terms of error metric) simple linear models that do not include any integers constants are shown in Table 5.2. As can be seen, varying qualities of models are found across the five clusters all with considerably low complexity levels. Included in the table are the correlation coefficients for each of the models as provided by Eureqa.

Table 5.2 Best 'simple' linear models for AVENG—for each cluster, the Pearson correlation coefficient (*Corr. Coef.*) and complexity value (*Compl.* as defined by Eureqa default values) are shown as well as the best model found by Eureqa

	Corr. Coef.	*Compl.*	Symbolic regression model
Cluster 1: Brand rationalists	0.59	3	AVENG = UI3 + SOC2
Cluster 2: Passive socializers	0.64	5	AVENG = UI1 + CCV5 + LO1
Cluster 3: Immersers	0.99	7	AVENG = UI3 + BR1 + RBV2 − FUV1
Cluster 4: Hedonic sharers	0.98	5	AVENG = SK2 + SK5 + FUV3
Cluster 5: Active participators	0.92	5	AVENG = CCV5 + LO1 + ON3

Specifically, from these simple models we can gather information about what type of variables are important in each cluster to predict online customer engagement. Starting with Cluster One—brand rationalists, although this model does not have a perfect fit and not a very high correlation coefficient (0.59), it provides us with two variables that Eureqa has found to be in a model which is correlated to AVENG. UI3 is part of the Usage Intensity construct and SOC2 as part of the Social Value construct. This shows us that these two variables correlate with customer engagement in our largest cluster, Cluster One. Having the relationship of addition between these two variables also indicates that these two activities need to be happening together to predict AVENG and drive higher values of customer engagement. Therefore, as the consumer uses the Facebook page more intensely, they are likely to derive higher social value from the experience which combinedly explain the level of customer engagement with the brand in this cluster of 'brand rationalists'.

For Cluster Two—passive socializers, a model was found with a slightly higher correlation coefficient (0.64) and complexity level 5. The variables in this model

are all found to be positively correlated with AVENG and they are as follows: UI1, CCV5 and LO1. Like Cluster One, Usage Intensity is again part of a linear model which is correlated to AVENG. This could be expected as it is likely that those customers who use the brand page more intensely are more likely to be engaged with the brand online. The other two variables belong to the theoretical constructs of 'Co-Creation Value' and 'Loyalty', respectively. As with Cluster One that means these three variables collectively contribute to AVENG in this cluster.

Cluster Three, immersers, has a model for AVENG with the highest correlation coefficient (0.99). There are four variables, three of which Eureqa found to positively contribute to a linear model that correlates with AVENG; however, one variable, FUV1 (Functional Value) is found to be contributing in an inverse sense to the model of customer engagement. A possible interpretation of this is that consumers who are part of this cluster perhaps do not like 'simple', useful and functional information on the Facebook page and do not use the brand page to search this kind of information. Again, Usage Intensity is found in this cluster together two variables relating to interaction and relationship with the brand: BR1 (Brand Relationship Value) and RBV2 (Relationship Building Value). This means that, rather than visiting and using the Facebook brand page for functional reasons, these customers prefer to have interactions with the brand and feel like they are creating a brand relationship.

The AVENG model for Cluster Four, Hedonic sharers, also has a high correlation coefficient (0.98) and is also quite simple. Three variables are found to positively correlate with customer engagement: SK2 and SK5 which belong to the Subjective Knowledge construct and FUV3 (Functional Value). This is the only cluster for which subjective knowledge is found to predict AVENG by Eureqa. These variables are about how 'knowledgeable' and confident the respondent feels about their skill and expertise in using Facebook and the brand page. It is logical to expect that when a consumer feels more confident about how to use a certain type of technology to engage with a brand, the more they will display behaviours towards that brand. However, considering these variables were not included in any other cluster, it shows that perhaps, it matters more for these respondents than others. Furthermore, conversely to Cluster Three, a variable of the Functional Value construct is found to be positively correlated with AVENG for Cluster Four. That means that these respondents do like to find useful and functional information on the Facebook brand page and that they are more likely to engage with the brand online if this is available to them.

Finally, a simple, positive linear model for Cluster Five, active participators, was found with a correlation coefficient of 0.92. The three variables are CCV5 (Co-Creation Value), LO1 (Loyalty) and ON3 (Online Loyalty Behaviours). Considering that two variables relating to loyalty behaviours are found in a predictive model for customer engagement means that certain customers may need to be loyal customers prior to them engaging with the brand through social media. That is, once consumers have developed a brand relationship in the offline environment, they then continue to pursue this relationship in the online environment by interacting with the brand via social media. Furthermore, the co-creation value items all asked respondents

how they feel the brand tries to create value for them in using the online social media platform. Naturally, if the brand provides an effective process in enabling co-creative interactions, a customer is more likely to engage with that brand online. The implications of these findings for marketing practitioners will be discussed in the following section.

5.5 Discussion

In this paper, we apply a novel and innovative methodology to the objective of clustering consumers in terms of online behaviours towards brands. More specifically, we examine a seemingly homogeneous sample obtained through survey research of customer's behaviours towards brands through a social media platform. We identify the existence of heterogeneity in the data amongst customers and the need to cluster customers in terms of their behaviours and attitudes rather than their demographic descriptive information or by the 'type' of brand they like and follow online. Through this study, we advance several contributions, both to literature and practice. In this section, we outline these contributions as well as recognize the limitations of this study, provide a guide for future research and present final conclusions.

The principal contribution in this paper is to propose a viable alternative to existing market segmentation and clustering methodologies. In doing so, we provide an ability to better describe and understand resulting clusters for the purposes of informing strategy formulation for facilitating customer engagement with specific market segments which enable insights for guiding target marketing strategies. Specifically, in this study, we have taken a data-driven approach with its roots in the natural sciences and applied it in this social scientific context. We show that the novel MST-*k*NN method is useful for finding clusters of consumers that are similar in their behavioural profile, rather than their demographic information. Here, we elaborate on our findings further and in particular, articulate what the findings for each cluster means for marketing practitioners.

Our findings entice marketing and brand managers to pursue effective segmentation and targeting strategies by using their behavioural and psychographic profiles rather than only relying on their more basic or demographic behaviours. Furthermore, the consumer behavioural models found by our data-driven approach go a step further in describing and interpreting the clustering results and provide a guideline for practitioners targeted marketing strategies. Starting with Cluster One (i.e. the *'brand rationalists'*) we have observed consumers are already involved with the brand prior to online interactions. This information would not have been available if the segmentation process had used demographic variables only. Furthermore, from the symbolic regression analysis, we have seen that social interaction is important for predicting online engagement with the brand for these consumers. What this means for marketers is that they need to provide a social online atmosphere by, for example, allowing their customers to comment, like and

share with the brand, and more importantly for this cluster, with each other on the social media brand page. Eliciting conversations between different users on the brand page and the encouragement of active participation in conversations will generate higher levels of customer engagement for Cluster One consumers, the 'brand rationalists'. Considering the characteristics of this cluster, targeting those consumers who are already involved with the brand would yield more successful outcomes of their online customer engagement strategies. This can be done by, for example, targeting those users who have followed the brand online for a longer period of time or those consumers in existing databases of the brand and then providing supporting processes on the social media brand page to enable socialization.

Moving to Cluster Two (i.e. the 'passive socializers'), conversely to the 'brand rationalists' (in Cluster One), consumers from this group are more likely to derive 'Social Value' in engaging with a branded social media page as these variables were part of the highest CM1 scores. Furthermore, they are also less involved with the brand than other clusters. The results of the symbolic regression modelling give us information to guide targeted strategies for these consumers. For instance, what drives customer engagement with brands in the social media environment in this cluster is when those customers feel like they are allowed to be involved in the provision of services to them and that they can customize the online experience they want (as the Co-Creation construct appears in the average Customer Engagement model). This insight, together with the finding that when customers feel like recommending the brand to others as well as higher usage rates, drives higher levels of engagement from these consumers. For marketing and brand managers, this means that they need to provide customers an opportunity and supporting processes to co-create with the brand page to receive the consumption experience they want. For example, supporting processes include responding to questions about the product or service via the social media page, requesting feedback or input from customers, and allowing mechanisms for customization of the social media experience will all lead to higher levels of customer engagement with brands from this cluster.

Cluster Three (i.e. the 'Immersers') was found to be the cluster that is more likely to experience a state of flow when using a branded social media page than other clusters. There could be various reasons for this such as that they use the social media platform from a larger screen such as a computer or tablet (50% of them report that they access Facebook from a desktop or laptop computer). Other reasons could be that they spend more time on the branded social media page or that they enjoy it more than others. Together with the outcomes from the symbolic regression analysis, these findings indicate that the online atmosphere delivered via the social media platform needs to facilitate this type of online experience. In doing so, these customers are likely to 'lose track of time' during their usage when their skill and challenge presented to them are in balance, and the more they experience a flow state, the more likely high levels of engagement will occur. For the brand to receive higher levels of engagement on the social media platform from these consumers, online material needs to be existent for customers to enable a flow state to occur

whilst customers are using the Facebook page. Furthermore, 'Relationship Building Value' and 'Brand Interaction Value' are also found to be predictive factors of online customer engagement in this cluster, which are important objectives marketers can take into consideration when targeting this cluster of consumers. For instance, this means actively building rapport and a relationship (e.g. special offers, useful value adding content, member programs and benefits) with consumers through the online social media page is likely to lead to higher levels of online customer engagement with the brand for this cluster as well as actively interacting with brand-related communication (e.g. video showcasing the brand) and encouraging conversations and dialogue specifically relating to the brand via the social media page.

Next is Cluster Four: the 'hedonic sharers'. This cluster is the only cluster in which their knowledge of in using social media technology, and their confidence on it, affects their customer engagement levels with brands. This, together with functional value (how functional the content of the Facebook page is) predicts and leads to higher levels of customer engagement in this cluster. What this means for the brand is that perhaps they need to educate those consumers in this cluster on the usage of new technologies. This may not only apply to the Facebook platform, but other new technological advances in customer brand interactions. A great deal of research has been conducted in this field using, for example, 'Technology Acceptance Models' to examine consumers' 'readiness' and knowledge to use new technology, see, for instance [28, 40]. As such, this type of research can aid those marketers who are seeking to better serve consumers who are not yet knowledgeable or confident in using new technology. Mechanisms that allow these consumers to become more confident, which might be needed in the case of Cluster Four, will lead to higher levels of online customer engagement with the brand for them. Furthermore, the fact that all four items of Hedonic Value in Cluster Four were in the top CM1 scores means that these respondents derive, on average, greater hedonic (fun, entertaining) value from using a branded Facebook page. If brands continue to provide these respondents with fun and entertaining content, together with a useful experience, combined with education mechanism for those consumers who need it on using new technologies, they can expect higher levels of customer engagement with brands from cluster four consumers.

Finally, Cluster Five (i.e. 'active participators') consisted of respondents already involved with the brand who responded, on average, higher to a question relating to their Self-Brand Congruency. This means that these customers are more likely to feel that their personality, and the 'personality' that the brand they like online portrays is congruent. From the symbolic regression model we find that, as with Cluster Two, Co-Creation Value is a driver of customer engagement for this cluster. Also predicting customer engagement for this cluster are a loyalty item and an online loyalty behaviour item. This means that for the consumers in this cluster to engage with a brand online, they may need to be loyal customers first. Commonly in

customer engagement research, customer engagement with a brand is modelled as a driver of loyalty, rather than loyalty to engagement [11, 25]. However, for this small cluster this relationship may actually occur from loyalty to engagement which is an issue that should be taken into consideration by marketers and could be explored in future research studies.

5.5.1 Limitations and Future Research

To clarify our perspective on the findings and as a possible guide for future research in this field we start by outlining some of the known limitations of our study. First, the relatively small student sample for this study poses some limitations. Even though it has been contended that a student sample is deemed appropriate for online and social media studies [10], future research should consider using a larger and more varied sample in order to improve generalizability. Second, this data was collected using a paper-and-pen survey in an offline setting. To further advance research in this field, 'real-time' behavioural data could be used as well as online surveys presented to a consumer as they are engaging in behaviours towards, and interacting with, brands online.

Third, in the analysis of the symbolic regression results, we have presented only those linear models that were both present in the Pareto frontier and have the maximum number of variables (thus they are the most fitting linear models). This is a self-imposed limitation; however, these models were competing with much more complex non-linear models that perhaps tried to 'over-fit' the data rather than predict the trend. An advantage of selecting these linear models stems from the ease of interpretation as they could easily be converted into real advice to marketing practitioners. In large-scale studies however, some more complex models could be used to analyse each of the clusters as they would provide the researcher with further information and understanding of the drivers for customer engagement in that cluster.

Fourth, we have based our study using consumer-reported data with brands only through one online platform, Facebook. Whilst this one is indeed an omnipresent social media platform, we could also argue that we may have not gathered all possible online interactions consumers could have with brands as other social media platforms allow for close interactions between brands and consumers. Consumer's brand interactions on other platforms such as Twitter, Foursquare or Pinterest would need to be investigated in the future to gather consumer insights on a broader spectrum as they may lead to other personalization strategies. This would also approximate what marketing and brand managers attempt to do in real-life more closely as practitioners need to manage their relationships with consumers across all online and offline communication platforms.

Other areas of future research include the use of community detection algorithms to address segmentation problems such as the one in this study. We have recently shown the application of a community detection algorithm in an analysis of

consumer behaviour in the charitable and non-for-profit sector where we compared it to clustering methods [33]. Amongst the currently existing community detection algorithms, some could provide partitioning results of the set of consumers in groups ('communities') that in turn could be matched against and compared to the five clusters described in this study. Such a study on 'ensemble segmentation' (the use of different partitioning algorithms in one study and followed by an analysis of observed commonalities) could be useful for further characterization of 'core' subgroups of people with even better defined saliency features. In addition, other types of algorithms for community detection allow the possibility of having 'overlaps', allowing fuzzy membership of consumers to more than one community [43] or allowing a 'percentage membership' to clusters in studies of fuzzy clustering algorithms [8]. The use and comparison of such 'fuzzy' methods would in turn give new insights for marketers wishing to target specific clusters or groups. Together with the use of hierarchical clustering methodologies, this would allow companies (which may operate under budget constraints) to identify similarities between clusters and design a relatively smaller set of marketing strategies to target their consumer base.

5.5.2 Final Conclusions

In summary, we have presented a comprehensive process for segmenting, analysing and guiding personalized marketing strategies to target consumers. It is important to remember that truly understanding consumer segments, based on their actual characteristics is what will guide marketers' targeting strategies and what will determine their successfulness in terms of positive outcomes for the brand. In this work, we have shown that segmenting consumers based on their behavioural profile, specifically in this case, their online behaviours towards a brand, and subsequently analysing these segments using more than just demographic information gives more insights and knowledge about consumers than demographics alone. As we have stated, the method used in this study is scalable to very large datasets making it feasible to scale this study to millions of consumers to segment and target brands' customers based on their real-time online behaviours. We have also shown the benefits of analysing clusters using symbolic regression modelling which is a method that is transferrable to many other instances and is extremely flexible with its options.

Acknowledgements We would like to thank Dr. Ahmed Shamsul Arefin for his help in providing the clustering result used in this study. Pablo Moscato acknowledges previous support from the Australian Research Council Future Fellowship FT120100060 and Australian Research Council Discovery Projects DP120102576 and DP140104183.

References

1. M Aljukhadar and S Senecal. Segmenting the online consumer market. *Marketing Intelligence & Planning*, 29(4):421–435, 2011.
2. AhmedShamsul Arefin, Mario Inostroza-Ponta, Luke Mathieson, Regina Berretta, and Pablo Moscato. *Clustering Nodes in Large-Scale Biological Networks Using External Memory Algorithms*, volume 7017 of *Lecture Notes in Computer Science*, book section 36, pages 375–386. Springer Berlin Heidelberg, 2011.
3. B Aviad and G Roy. A decision support method, based on bounded rationality concepts, to reveal feature saliency in clustering problems. *Decision Support Systems*, 54(1):292–303, 2012.
4. Amit Bhatnagar and Sanjoy Ghose. A latent class segmentation analysis of e-shoppers. *Journal of Business Research*, 57(7):758–767, 2004.
5. C Blattberg, Robert, Byung-Do Kim, and A Neslin, Scott. Database management: Analyzing and managing customers, 2008.
6. Petter Bae Brandtzæg, Jan Heim, and Amela Karahasanović. Understanding the new digital divide—a typology of internet users in Europe. *International Journal of Human-Computer Studies*, 69(3):123–138, 2011.
7. Colin Campbell, Carla Ferraro, and Sean Sands. Segmenting consumer reactions to social network marketing. *European Journal of Marketing*, 48(3/4):432–452, 2014.
8. Kit Yan Chan, C. K. Kwong, and B. Q. Hu. Market segmentation and ideal point identification for new product design using fuzzy data compression and fuzzy clustering methods. *Applied Soft Computing*, 12(4):1371–1378, 2012.
9. Anil Chaturvedi, J. Douglas Carroll, Paul E. Green, and John A. Rotondo. A feature-based approach to market segmentation via overlapping k-centroids clustering. *Journal of Marketing Research (JMR)*, 34(3):370–377, 1997.
10. S. C Chu and Y Kim. Determinants of consumer engagement in electronic word of mouth (eWOM) in social networking sites. *International Journal of Advertising*, 30(1):47–75, 2011.
11. Natalie Jane de Vries and Jamie Carlson. Examining the drivers and brand performance implications of customer engagement with brands in the social media environment. *J Brand Manag*, 21(6):495–515, 2014.
12. Natalie Jane de Vries, Jamie Carlson, and Pablo Moscato. A data-driven approach to reverse engineering customer engagement models: Towards functional constructs. *PLoS ONE*, 9(7):e102768, 2014.
13. Wayne S. DeSarbo, Kamel Jedidi, and Indrajit Sinha. Customer value analysis in a heterogeneous market. *Strategic Management Journal*, 22(9):845–857, 2001.
14. Arne Floh, Alexander Zauner, Monika Koller, and Thomas Rusch. Customer segmentation using unobserved heterogeneity in the perceived-value–loyalty–intentions link. *Journal of Business Research*, 67(5):974–982, 2014.
15. Mary Foster, Bettina West, and Anthony Francescucci. Exploring social media user segmentation and online brand profiles. *Journal of Brand Management*, 19(1):4–17, 2011.
16. O Giustolisi and D. A Savic. A symbolic data-driven technique based on evolutionary polynomial regression. *Journal of Hydroinformatics*, 8(3):207–222, 2006.
17. Paul E. Green, Frank J. Carmone, and David P. Wachspress. Consumer segmentation via latent class analysis. *Journal of Consumer Research*, 3(3):170–174, 1976.
18. J. L Gross and J Yellen. *Handbook of Graph Theory*. Discrete Mathematics and its Applications. CRC Press LLC, Boca Raton, Florida, 2004.
19. Chris Hand, Francesca Dall'Olmo Riley, Patricia Harris, Jaywant Singh, and Ruth Rettie. Online grocery shopping: the influence of situational factors. *European Journal of Marketing*, 43(9/10):1205–1219, 2009.

20. Thorsten Hennig-Thurau, Edward C. Malthouse, Christian Friege, Sonja Gensler, Lara Lobschat, Arvind Rangaswamy, and Bernd Skiera. The impact of new media on customer relationships. *Journal of Service Research*, 13(3):311–330, 2010.
21. J. L Herlocker, J. A Konstan, A Borchers, and J Riedly. An algorithmic framework for performing collaborative filtering. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 230–237. ACM New York, 1999.
22. Mario Inostroza-Ponta, Regina Berretta, Alexandre Mendes, and Pablo Moscato. *An automatic graph layout procedure to visualize correlated data*, pages 179–188. Springer, 2006.
23. Mario Inostroza-Ponta, Regina Berretta, and Pablo Moscato. QAPgrid: A two level QAP-based approach for large-scale data analysis and visualization. *PLOS One*, 6(1):e14468, 2011.
24. Mario Inostroza-Ponta, Alexandre Mendes, Regina Berretta, and Pablo Moscato. *An integrated QAP-based approach to visualize patterns of gene expression similarity*, pages 156–167. Springer, 2007.
25. Benedikt Jahn and Werner Kunz. How to transform consumers into fans of your brand. *Journal of Service Management*, 23(2):344–361, 2012.
26. Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.
27. T. D Klastorin. Assessing cluster analysis results. *Journal of Marketing Research (JMR)*, 20(1):92–98, 1983.
28. Paul Legris, John Ingham, and Pierre Collerette. Why do people use information technology? a critical review of the technology acceptance model. *Information & Management*, 40(3):191–204, 2003.
29. E. C. Malthouse, M. Haenlein, B. Skiera, E. Wege, and M. Zhang. Managing customer relationships in the social media era: Introducing the social CRM house. *Journal of Interactive Marketing*, 27(4):270–280, 2013.
30. R Mancha, M. T Leung, J Clark, and M Sun. Finite mixture partial least squares for segmentation and behavioral characterization of auction bidders. *Decision Support Systems*, 57(1):200–211, 2014.
31. J Marsden, D Budden, H Craig, and P Moscato. Language individuation and marker words: Shakespeare and his Maxwell's demon. *PLOS One*, 8(6):1–12, 2013.
32. Heloisa Helena Milioli, Renato Vimieiro, Carlos Riveros, Inna Tishchenko, Regina Berretta, and Pablo Moscato. The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the METABRIC data set. *PLOS ONE*, 10:e0129711, 2015.
33. Leila M. Naeni, Natalie Jane de Vries, Rodrigo Reis, Ahmed Shamsul Arefin, Regina Berretta, and Pablo Moscato. Identifying communities of trust and confidence in the charity and not-for-profit sector: a memetic algorithm approach. In *The 7th IEEE International Conference on Social Computing and Networking (Socialcom)*. IEEE, 2014.
34. A O'Cass and J Carlson. Examining the effects of website-induced flow in professional sporting team websites. *Internet Research*, 20(2):115–134, 2010.
35. H Ouwersloot and G Odekerken-Schröder. Who's who in brand communities – and why? *European Journal of Marketing*, 42(5/6):571–585, 2008.
36. K. Y Peters, A. M Chen, B. O Kaplan, and K Pauwels. Social media metrics—a framework and guidelines for managing social media. *Journal of Interactive Marketing*, 27(4):281–298, 2013.
37. M Schmidt and H Lipson. Eureqa, 2013.
38. G. F Smits and M Kotanchek. *Pareto-front exploitation in symbolic regression*, book section 17, pages 283–299. Springer, US, 2005.
39. Jenny van Doorn, Katherine N. Lemon, Vikas Mittal, Stephan Nass, Doreén Pick, Peter Pirner, and Peter C. Verhoef. Customer engagement behavior: Theoretical foundations and research directions. *Journal of Service Research*, 13(3):253–266, 2010.

40. Viswanath Venkatesh and Fred D. Davis. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2):186–204, 2000.
41. Elaine Wallace, Isabel Buil, Leslie de Chernatony, and Michael Hogan. Who "likes" you ... and why? a typology of Facebook fans. *Journal of Advertising Research*, 54(1):92–109, 2014.
42. Michel Wedel and Wagner A. Kamakura. *Market Segmentation: Conceptual and Methodological Foundations*. International Series in Quantitative Marketing. Kluwer Academic Publishers, Norwell, Massachusetts, USA, 2 edition, 2000.
43. Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Comput. Surv.*, 45(4):1–35, 2013.