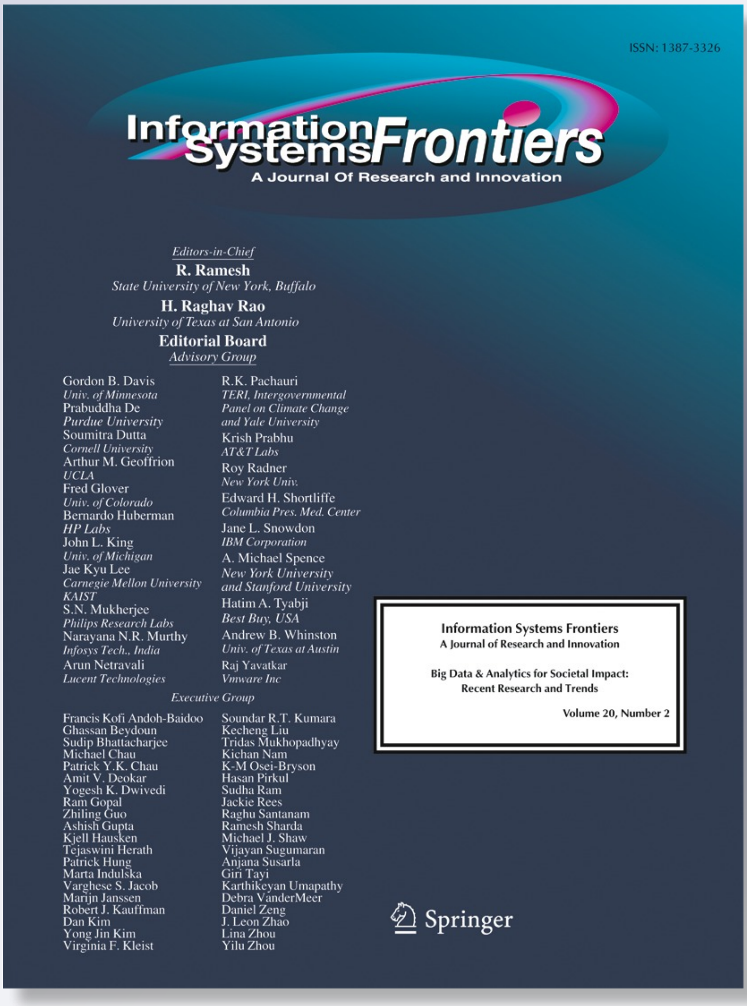
Keep the conversations going: engagement-based customer segmentation on online social service platforms

Nripesh Trivedi, Daniel Adomako Asamoah & Derek Doran

Information Systems Frontiers
A Journal of Research and Innovation

ISSN 1387-3326
Volume 20
Number 2

Inf Syst Front (2018) 20:239-257
DOI 10.1007/s10796-016-9719-x



Information Systems Frontiers
A Journal of Research and Innovation

Big Data & Analytics for Societal Impact:
Recent Research and Trends

Volume 20, Number 2

Springer

Springer

Inf Syst Front (2018) 20:239–257
DOI 10.1007/s10796-016-9719-x

Keep the conversations going: engagement-based customer segmentation on online social service platforms

Nripesh Trivedi[1] **· Daniel Adomako Asamoah**[2] **· Derek Doran**[3]

Published online: 12 November 2016

Abstract Most businesses and organizations develop online services as a value-added offering, which is a significant revenue stream from their existing user base. Such services may be enhanced with social elements to serve as value-added tools for user attraction and retention. Social elements may allow users to post content, share information and directly interact with each other. Investments in these social features are for naught if they do not encourage users to engage on the platform effectively. However, common ways to segment customers by their engagement is hindered by the statistical nature of behavioral data based on social elements. To address this important concern, this paper presents a methodological framework for engagement-based customer segmentation able to appropriately consider signals from social elements. It argues why the traditional approaches for user segmentation is ill-suited and advocates for the integration of kernel functions with clustering to segment, identify and understand user engagement profiles. The framework is demonstrated with real data from a large, very active OSS.

✉ Daniel Adomako Asamoah
daniel.asamoah@wright.edu

Nripesh Trivedi
nripesh.trivedi.apm11@iitbhu.ac.in

Derek Doran
derek.doran@wright.edu

[1] Department of Mathematical Sciences, Indian Institute of Technology, Banaras Hindu University, Varanasi, India

[2] Department of Information Systems and Supply Chain Management, Raj Soin College of Business, Wright State University, Dayton, OH, USA

[3] Department of Computer Science and Engineering, Kno.e.sis Research Center, Wright State University, Dayton, OH, USA

Keywords Analytics · Online social services · Engagement analysis · Kernel k-means · Clustering · Big data

1 Introduction

Online services created and managed by organizations offer novel, value-added services for its users. Such services have a direct, positive impact on a customer's loyalty to a brand (Ogonowski et al. 2014). Examples of online services include health portals by medical clinics that show patient lab results and other health related information (Gummerus et al. 2004) and peer-review and rating systems for e-business and auction websites. Some organizations are even entirely dependent on their online service offerings for revenue (Chuang et al. 2014). For example, organizations running web search engines, video streaming sites, online social networks, and news publishing sites rely on users to utilize their online service and view advertisements or register for subscriptions to generate revenue.

An emerging body of research suggests that the integration of *social elements* into online services yield a positive effect on purchase intention and brand loyalty, and that the extensive use of such features encourages users to continue to return to the service (Chen et al. 2013; Cyr et al. 2007; Huang and Benyoucef 2013). We define *any* kind of system that integrates social elements to be an *online social service* (OSS). Social elements manifest themselves on a platform through many functions, such as ones that let users define overt social connection (e.g. defining a concrete relationship between users, adding a "friend", or subscribing to a user's updates) or subtler relationships (e.g. leaving a comment on a user review in an e-commerce site or 'promoting' the content of another user on a company intranet). A canonical example of an OSS is a social media service or online social network.

Another example is Amazon.com, since viewers of a product can interact with each other by asking questions and leaving answers about a product. News sites that allow users to comment on an article, and respond to the comments of others, contain social interactions, also make them examples of an OSS. A hallmark of an OSS is its fostering of a virtual community. Studies of OSS's and their virtual communities consider peer-to-peer networks (Xia et al. 2012), networks of practice (Wasko and Faraj 2005), open source software developers (Shah 2006), Wikipedia editors (Iba et al. 2010; Nov 2007), and e-mails asking for technical advice in an organization (Constant et al. 1996).

Organizations that are able to study and quantify the ways in which users experience engagement in their online services, defined as a quality of the user experience that emphasizes the effects of positive aspects of interaction (Attfield et al. 2011), stand to gain a market advantage. For example, engagement analysis could separate users who choose to use only specific aspects of an online service, and who exhibit patterns similar to those who had left the service and never returned. Such analysis may be used to build models that predict user churn based on engagement features. Engagement analysis may further highlight parts of the service that are negatively associated with user engagement, or is seldom used by users who exhibit high engagement. Organizations could also use it to evaluate the total effectiveness of social features. For example, a social widget with which only a small percentage of users are engaged suggests the need to re-evaluate the widget's design and functionality. The degree to which a user experiences engagement with respect to social features thus requires careful consideration.

Engagement is an inherently qualitative concept, and most current approaches measure this using user surveys or other self-reporting mechanisms (Lalmas et al. 2013). Self-reported, qualitative measures of engagement are not desirable because the analysis is only limited to users who are willing to participate in a survey. Since different analysts may hold two different interpretations, that could lead to conflicting measurements and recommendations. An improved measure would be one that: (i) relies on user and behavioral features that are captured as users interact with the OSS; (ii) aggregates the readings of such features across all users into a big data set; and (iii) partitions users into groups according to these features in an automatic, unsupervised way with a minimum number of analyst defined hyper-parameters.

A framework that can meet these improvements to engagement analysis is necessary but challenging to develop. A simple approach may be to identify elements related to social engagement from a system, preprocess the data with common standardization and outlier removal procedures, and then partition users through an unsupervised clustering analysis (Chu et al. 2009). But this could fail when faced with data from an OSS because, based on measurements from a number of studies, features related to social elements exhibit heavy-tailed tendencies (Benevenuto et al. 2009). Heavy tails complicate engagement analysis because they can make linear decision boundaries in the data space, necessary for clustering algorithms such as *k*-means to find meaningful partitions, impossible to identify. While standardization and outlier removal can reduce variance in the data and make the emergence of linear boundaries more likely, they essentially mask the heavy-tailed nature of the data from the segmentation model. This act may bring unintended consequences, such as mapping users with social feature values that lie in the heavy-tail to be clustered alongside users whose feature values are very (potentially orders of magnitude) different (Lipsky 2009).

To address these challenges, this article proposes a framework for customer segmentation on an OSS that suitably separates users despite the heavy-tailed qualities of features related to social elements. The framework applies a non-linear transformation of the data that maps it to a high dimensional space with a Radius Basis Kernel Function (RBF), and then applies *k*-means to this projected data. This transformation aims to map the data to a space where linear decision boundaries may exist. The framework also offers guidance based on quantitative measures to select hyper-parameters that govern the number of engagement groups found and the way the data is projected to a higher dimensional space. This guidance enables the use of the framework in practice, even to those who are not familiar with kernel *k*-means. The contributions of this framework may be timely because social features are rapidly being integrated into existing online services built by businesses. Long standing and well known clustering-based methods to segment users with data standardization and the *k*-means algorithm (Chang et al. 2007) may not be valid with the heavy-tailed nature of social engagement features. The article includes a case study that applies the framework to a big data set from an active OSS.

2 Literature review

Studying the engagement of a person who participates in an activity is a research process that is based on the researcher's definition of what it means "to be engaged" and the context in which engagement is studied. A near universal trait, however, is to first define features that characterize the "engagement behaviors" of a person as it relates to the interactions he or she performs in the activity. One body of work considers physiological measures for this purpose (Ikehara and Crosby 2005; Konradt and Sulz 2001; Seah and Cairns 2008). The physiological data may come from sensors like eye trackers, mouse movements, blood pressure, heart pulses, and cameras (Attfield et al. 2011) and measure features such as eye movements, heart rate, and mouse clicks (Arapakis et al. 2014; Lagun et al. 2014). Such analysis reveals bodily reactions that are cognitively linked to mental engagement in a task. However, mental engagement is not easily relatable in the context of user analytics. For example, it is not feasible to measure the physiological state of large numbers of users as they interact with an OSS, and even if it were possible, it may be difficult to interpret into actionable insights.

Analysis of engagement with respect to user interactions are more practical to perform because it does not depend on physiological readings. Furthermore, since user interactions are based on the functionality and interface of a platform, the engagement analysis may lead to actionable insights for the platform owner. Previous research has proposed a variety of measures for quantifying or measuring user interactions for this purpose. The most popular approach involves self-reporting, where questionnaires, interviews, reports, and product reaction cards are given to users in the hope that they respond honestly. For example, Webster and Ho (1997) developed a seven item questionnaire with items that include the degree of attention, challenge, intrinsic interest, and variety in the context of presentation software. Jacques (1996) developed a 13-item survey to evaluate user engagement in online e-commerce environments. His survey included items about user's attention, perceived time, motivation, needs, control, and attitudes. Brooke (1996) addresses the functionality of a system and users' satisfaction of those systems. Previous studies have used this technology independent test to assess user engagement in hardware, software and web systems. Self-reported measures of engagement carry few technological hurdles to adopt and are relatively cheap to run (Gagné and Godin 2005; Hawkshead and Krousel-Wood 2007). Lalmas et al. (2013) characterize such a user engagement measure as subjective and optimized for only small-scale applications.

Rather than asking users to self-report, emerging approaches seek to measure user behaviors directly from a software system. Depending on the software, analysts may define various key behavioral indicators (KBIs) and their relationship to user engagement. Media websites, for example, may use average page views, bounce rate (people arriving at a website and leaving immediately), user satisfaction, and top internal search phrases as different types of KBIs. An e-commerce website may use a different set of KBIs, such as conversion rates (e.g. ratio of users who buy an item), average order value, and user loyalty metrics (Booth and Jansen 2008). By themselves, KBIs reveal the engagement characteristics of an individual user. Recent work has considered how the collection of all KBIs across all users may be used to build models that classify users into broad segments, each corresponding to a different mode or type of engagement profile. For example, Lehmann et al. (2012) considered user interactions across a variety of different websites, from social media to e-commerce, to discover a small number of Web user engagement profiles. Their study considered users grouped a priori into different segments based on the number of days in a month a user visited the website. van Dam and van de Velden (2014) consider different features of Facebook users to classify them into different types.

We hypothesize that measuring user behaviors directly from a software system is a promising approach for engagement analysis. This is because user behaviors are often tracked automatically, enabling the collection of big datasets that capture enough activity to build complex engagement models. The framework proposed in this article enables this data-driven approach to engagement analysis in practice for any kind of OSS. The proposed framework is specified in a generic fashion, and offers guidance for its application in any applied setting.

3 Engagement analysis framework

In this section, we introduce our framework for segmenting customers based on how they engage on an OSS. We first consider how to summarize the notion of engagement along three orthogonal dimensions, and discuss the types of user behaviors that embody each dimension. We then present the heavy-tailed characteristics of the distributions of these features and discuss why traditional *k*-means clustering is unsuitable. As part of a case study, we present a framework in which we introduce kernel *k*-means clustering as a method better suited to handle heavy-tailed online social features.

3.1 Dimensions of engagement

Instead of basing engagement analysis on specific platform interactions, we instead propose a breakdown of the concept into multiple generic dimensions that KBIs germane to any OSS may be associated with. These dimensions are:

(i) *Initiation*: This describes how often a user chooses to enter the service, and once he or she is on, how frequently the user initiates a social action. Once a user becomes part of an OSS, O'Brien and Toms (2008) notes that his ability to initiate and sustain social actions is necessary for sustaining engagement. For example, the frequency with which the user logs into a site or establishes a "connection" with another user represents an initiation. This is especially vital for an OSS that is still in its infancy. It captures the notion that engaged users often use the site and are interested in expanding their 'footprint' in a virtual community. They thus exhibit a profound initiative in using the site and its services.

(ii) *Interaction*: This describes the volume of social interactions and activities a user performs, emphasizing the utilization of the platform's social elements (Lehmann et al. 2012). For example, users who often upload content or share their thoughts, frequently respond and react to content posted by others, or establish virtual ties with a large number of others may be highly engaged.

(iii) *Loyalty*: Loyalty indicates the extent to which a user finds a website to be effective and his likelihood to re-engage with it another time (Webster and Ahuja 2006). Previous literature generally refers to this as "intention to return" (O'Brien and Toms 2010). Loyalty also reflects the length of time and frequency with which a user has actively participated in a service. The extent to which some services are utilized and the total length of time a user is active on the site are examples of loyalty. Whereas initiation features capture how often a service is used and the user's penchant to increase their social footprint on the OSS, loyalty corresponds to the length of time and breadth of services the user taps into.

The KBIs (hereafter referred to as user or behavioral features) best reflecting each of these dimensions depend on the type of online service, the kinds of social elements available on it, and the type of recorded user behavior data.

3.2 Segmenting users by engagement

After *n* behavioral features across all of the above dimensions are identified for a platform, the next task is to segment each user into an engagement vector *x*. A natural way of segmenting users based on their engagement activities is to utilize a clustering algorithm that separates the set of vectors $\{\boldsymbol{x}\}$, into non-overlapping groups based on a measure of vector distance or similarity. For this purpose, studies turn to *k*-means clustering, a simple algorithm often used for user behavior modeling (Farajian and Mohammadi 2010). It identifies a collection of *k* (specified a priori) centroid vectors C_k in the feature space, and assigns vectors to classes by the centroid vector it is closest to. The optimal set of *k* centroid positions C_k are those that minimize the sum of all distances from the engagement vectors to their assigned centroids:

$$C_k = min_{\{e_m \in C_k\}} \sum_{m=1}^{k} \sum_{i=1}^{n} \|x_i - e_m\| 2 \tag{1}$$

where k is specified a priori and $\sum_{m=1}^{k}\sum_{i=1}^{n}\|x_i = e_m\|^2$ is the squared Euclidian distance between an engagement vector x_i and the centroid e_m of its assigned cluster m. Note that the resulting assignment of data points to clusters define linear boundaries between points that fall into different clusters in the feature space (Jain 2010). An iterative algorithm is typically used to find the set of centroids C_k minimizing Eq. (1):

1. Randomly choose *k* points in the space of all data vectors to serve as centroids.
2. Assign each vector *x* to the group that has the closest centroid.
3. When all objects have been assigned, re-calculate the *k* centroid positions.
4. Repeat steps 2 and 3 until the change in centroids between iterations fall below a threshold, Δ.

The set of engagement vectors across all users $\{\boldsymbol{x}\}$ needs to satisfy specific requirements for *k*-means to be applicable. An important one is that all behavioral features represented in $\boldsymbol{x}$ must exhibit similarly small variances so that outlier points do not heavily influence the position of a centroid. This is because *k*-means clustering can only identify linear separations in the data that may be impossible to find when one dimension of the data exhibits significant variance or large numbers of outliers (Shukla and Naganna 2014). Such a variance requirement may not be satisfied for the social behavioral features of an OSS; countless past studies have shown how features related to social behaviors on online services follow heavy-tailed distributions. A heavy-tailed distribution is one whose complementary cumulative distribution function, *R*(*X*)=Pr(*X*>*x*), drops at a rate slower than exponential, i.e. when:

$$\lim_{X\to\infty} e^{sX} R(X) = \infty \tag{2}$$

This slow decay in the right tail of the distribution causes a significant variation in *X*. A particular type of heavy-tailed distribution is one that has a *power-tail*, where *R*(*X*) satisfies the above limit and has the functional form:

$$R(X) = cX^{-\alpha} \tag{3}$$

where *c* is some constant and α is a scaling parameter. Power-tailed distributions exhibit even more variance than heavy-tailed ones, so much so that that all of its moments greater than $\lceil \alpha \rceil$ is infinite (Lipsky 2009). For example, the sampling distribution of a power-tailed random variable with α < 2 has infinite variance, and at α < 1 has an infinite mean.

Table 1 illustrates how a wide variety of user behaviors on OSSs, as reported in previous research, exhibit heavy and power tails. The table lists the particular type of heavy-tailed distribution fitted to the data.

In order to mitigate the negative effect of heavy-tailed data on a practical, even distribution of data into separate segments, some studies standardize feature values so that they all fall within the same range. This transforms the features to exhibit smaller variations and hence make them become more amenable to *k*-means clustering (Chu et al. 2009). However, while standardization may be important in the presence of outliers that skew sample statistics and model fits, data in heavy tails are meaningful observations representing a characteristic trend in the data. Standardization would essentially eliminate the fact that this data in the tail lies through many orders of

Table 1 Heavy-tailed distributions in the social behaviors of different OSS services

Behavioral feature	Online social service and best distribution fit
Number of social connections	Facebook; heavy tailed (Ugander et al. 2011)
	Political blogs; log-normal (Adamic and Glance 2005)
	Flickr; power law (Mislove et al. 2008)
	LiveJournal; power law (Viswanath et al. 2009)
	Orkut; power law (Viswanath et al. 2009)
	YouTube; power law (Viswanath et al. 2009)
	Google+; power law (Gong et al. 2012)
	Pinterest; power law (Zhong et al. 2014)
	GitHub; power law (Lima et al. 2014)
	Twitter: power law (Kwak et al. 2010)
Number of responses to a user contribution	Flickr; power law (Mislove et al. 2008)
	Pinterest; power law (Zhong et al. 2014)
	GitHub; power law (Lima et al. 2014)
	Twitter; power law (Kwak et al. 2010)
	Online games; power law (Szell and Thurner 2010)
Session duration	Orkut; heavy-tailed (Benevenuto et al. 2009)
	MySpace; heavy-tailed (Benevenuto et al. 2009)
	Orkut; heavy-tailed (Benevenuto et al. 2009)
	Hi5; heavy-tailed (Benevenuto et al. 2009)
Number of other users a user sends messages to	Slashdot; power law (Kunegis et al. 2009)
	StackOverflow; power law (Anderson et al. 2012)
	Online games; power law (Szell and Thurner 2010)
Time between logins	Orkut; heavy-tailed (Benevenuto et al. 2009)
	MySpace; heavy-tailed (Benevenuto et al. 2009)
	Orkut; heavy-tailed (Benevenuto et al. 2009)
	Hi5; heavy-tailed (Benevenuto et al. 2009)
Frequency of logins	MySpace; right-skewed (Torkjazi et al. 2009)

magnitude, which is an important aspect. For example, empirical observations of heavy-tailed data features a large number of points whose value is orders of magnitude larger than the mean (Lipsky 2009), and a standardization will map these many points to values that are of the same or similar magnitude as the standardized mean.

3.3 Kernel *k*-means clustering

Recognizing that standardization of heavy-tailed data to enable *k*-means clustering is not an appropriate transformation, we instead consider the incorporation of a kernel function that projects data vectors into a higher dimensional space where the data is clustered. Intuitively, data that lives in a higher dimensional space has additional directions in which it varies, thus increasing the likelihood of finding a linear separation in the data. A simple illustration of projecting generic data into higher dimension is shown in Fig. 1.

In the left panel, points are shaped and colored by their true class (cluster) label, with no way to define a linear boundary between the two classes that *k*-means attempts to find. If we were to project that same data to a higher three-dimensional space using the transformation $(x, y) \rightarrow (x, y, x^2 + y^2)$, as shown in the right panel of Fig. 1, however, we now see the linear separation that *k*-means can identify to correctly cluster the groups.

The kernel *k*-means algorithm applies *k*-means clustering to vectors that have been mapped to a higher dimensional space. It seeks a set C_k of k centroid positions satisfying:

$$C_k = argmin_{\{e_m \in C_k\}} \sum_{m=1}^{k} \frac{1}{n|e_m|} \sum\nolimits_{x_i \in E_m, x_j \in E_m} k\left(x_i, x_j\right) \qquad (4)$$

where $k(x_i,c_j)$ is a kernel function that measures the distance (i.e. similarity) between two engagement vectors x_i and x_j assigned to the same cluster E_m with centroid e_m in some higher dimensional space. Different choices for the kernel function k encode different ways to perform the data

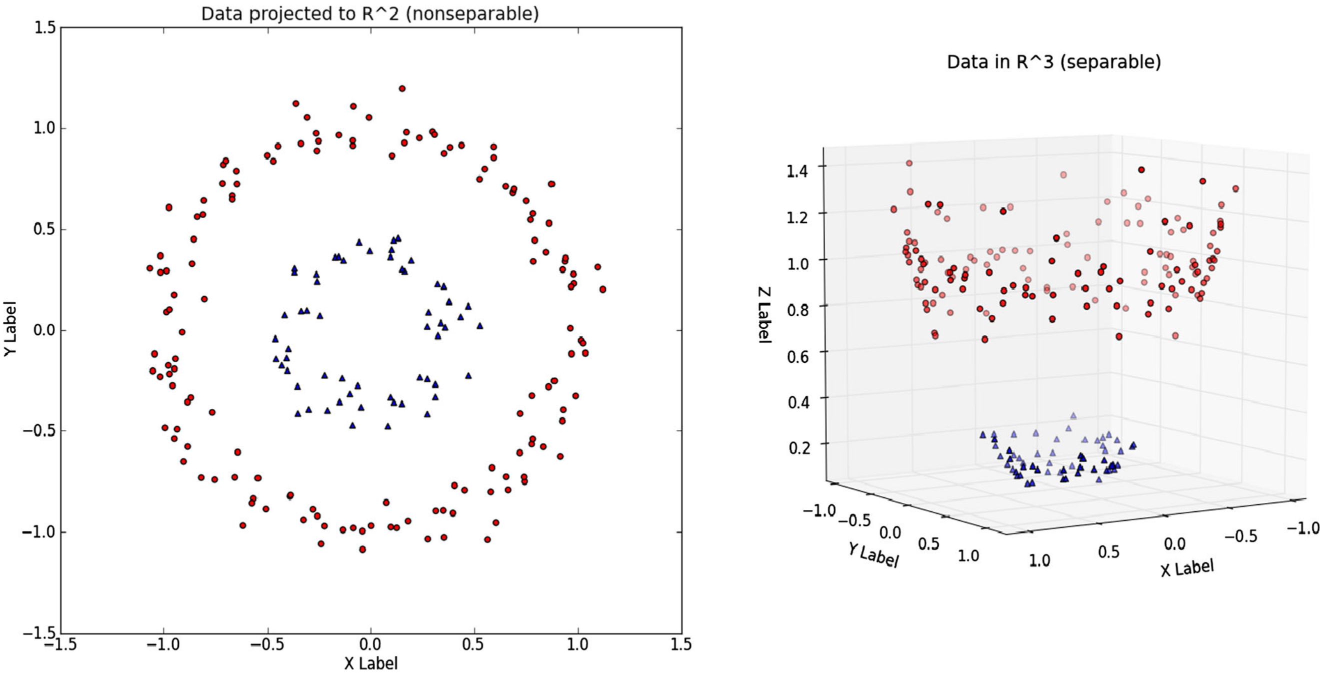


Fig. 1 Projection from a lower dimension (*left*) to a higher (*right*) dimensional space where a linear separation of the data exists. reprinted with permission from (Kim 2013)

projection. Many choices of kernel functions are possible, and the proposed framework considers the RBF kernel:

$$\kappa\left(x_i, x_j\right) = \exp\left(-\frac{\left(x_i - x_j\right)^2}{2\sigma^2}\right) \tag{5}$$

where σ is a scaling hyper-parameter. The RBF kernel is an appropriate choice for cases in which there is little or no information available about how to project the data in a way that yields a linear separation (Romero et al. 2014). It has been shown to perform well in a variety of contexts (De la Torre and Vinyals 2007; Romero et al. 2014).

3.4 Hyper-parameter fitting

Both σ and k play a pivotal role in the performance of the kernel function. When σ is large, the exponential in the RBF kernel will take on a linear shape, and hence, will yield results similar to traditional k-means. On the contrary, if underestimated, the function may over fit the data set and hence yield a clustering result that is not generalizable to future users of the OSS. An analyst therefore needs to tune both k (the number of clusters to be used) and $\boldsymbol{\sigma}$ intuitively, using a measure of clustering quality to identify the best pair of parameter settings (Taniar 2008).

Choosing the "best" clustering among a set of candidate results is an inherently subjective process. The process is equivalent to deciding on an optimal choice of hyper-parameters (e.g. the number of clusters in a solution), which is a long standing problem in machine learning (Bergstra and Bengio 2012). To pinpoint a suitable value of k and σ, the framework considers combinations of values within a grid of feasible values, generates a clustering result, and quantitatively evaluates the result with respect to the following criteria:

- *Clustering Quality*: In practice, an ideal set of parameter values should strongly discriminate points among clusters, that is, have points in the same cluster "close" to each other and points in different clusters "far" from each other. It should also be able to assign an equitable distribution of points into clusters. Strong discrimination ensures that each cluster represents a unique kind of engagement. We use the silhouette (SI) coefficient, described below, to measure quality.
- *Clustering Equity*: It could be that the highest quality clustering result is one that skews the distribution of users in clusters. Skewed cluster size distributions may not be "actionable" results for engagement analysis because it lumps most users into a single cluster whereby the diversity of user interactions is masked. This lack of diversity limits the extent to which business decisions, differentiated treatments, engagement campaigns, and other actions can be applied to users based on their engagement group. Thus, we also consider how equitable the distribution of users into clusters are by its entropy.

Recognizing the subjective nature of choosing a clustering solution, the analysis framework does not offer a quantitative weighting to the importance of quality and equity in the hyper-parameter fitting process. This is because, depending on the intention of the analyst and his or her subjective view of the importance of quality and equity, different weightings may be suitable. To guide an analyst in choosing a clustering result, the framework offers two guidelines:

1. *The measure of quality should always be emphasized over the measure of equity.* It should be clear that the quality of any clustering model needs to be high because a low quality measure implies that no clustering structure was found in the data. For example, one can imagine a degenerate clustering where equally sized random samples of users are assigned to each case. This clustering maximizes equity since the same number of users fall in each class, but the random, scattered positions of users in the same class are not indicative of any kind of clustering.
2. *When comparing models*, *one of comparable quality but significantly better equity is more desirable*. Since higher equity is indicative of a result that is more actionable for an organization, one may choose a model of comparable quality in exchange for significantly increased diversification of users into engagement classes.

3.4.1 Measuring cluster quality

The framework measures how well a solution discriminates points among clusters using the silhouette (SI) coefficient, a typically used validation measure (Pang-Ning et al. 2006). It is defined by computing a score s for the j^th^ data point assigned to the i^th^ cluster x_{ij}, taking the average of these scores across all points in a cluster, and then taking the average of these averages over all clusters (Kodinariya and Makwana 2013):

$$S(k;\sigma) = \frac{1}{k}\sum_{i=l}^{k} \frac{1}{m}\sum_{j=1}^{m} s\left(x_{ij};\sigma\right) \tag{6}$$

Where σ is the hyper-parameter needed for the RBF kernel. s(X_{ij};σ) is defined by: (Camps-Valls et al. 2007):

$$s\left(x_{ij};\sigma\right) = \frac{w^{\varnothing}\left(x_j\right) - v^{\varnothing}\left(x_j\right)}{\max\left(w^{\varnothing}\left(x_j\right), v^{\varnothing}\left(x_j\right)\right)} \tag{7}$$

Where:

$$v^{\varnothing}\left(x_j\right) = \frac{1}{\left(|C_i|-1\right)}\kappa\left(x_j, x_{j,}\right) - \frac{2}{\left(|C_i|-1\right)}\sum_{x_j \in C_{i,x_l \neq x_j}} \kappa\left(x_j, x_l\right) + \frac{1}{\left(|C_i|-1\right)}\sum_{x_j \in C_{i,x_l \neq x_j}} \kappa\left(x_l, x_l\right) \tag{8}$$

is the average distance from the projection of x_{ij} in a higher dimensional space to every other engagement vector assigned to the same cluster as x_j in the projected space, and:

$$w^{\varnothing}\left(x_j\right) = min_{h=1,\ldots,i-1,i+1\ldots.k}\left(\frac{1}{|C_h|}\kappa\left(x_j, x_{j,}\right) - \frac{2}{|C_h|}\sum_{x_l \in C_h}\kappa\left(x_j, x_l\right) + \frac{1}{|C_h|}\sum_{x_l \in C_h}\kappa\left(x_l, x_l\right)\right) for\, x_j \in C_i \tag{9}$$

defines the average distance from x_j to every other engagement vector x_i that is not in the cluster to which x_j is assigned. The values of $s(x_{ij};\sigma)$ will range between −1 and 1. Values close to 1 indicate that x_{ij} lie very close to vectors within its cluster, and lie far away from vectors in different clusters. Values close to −1 indicate the opposite. We thus seek a k and σ that maximizes $S(k;\sigma)$ since it indicates that, on average, vectors lie close to others within their cluster and far away from vectors in other clusters.

3.4.2 Measuring cluster equity

To measure the equity of the distribution of users that fall into clusters, the framework uses the information theoretic concept of *entropy*. Entropy is a quantification of the "randomness" of a probability distribution. Intuitively, a skewed probability distribution $p(x)$ is "less random" in the sense that samples taken from the distribution will have a value from a highly likely small subset of the set of possible values. On the other hand, a distribution where any value could be draw with similar probability, is a "more random" distribution. The entropy $H[X]$ of a distribution quantifies this concept: the higher the entropy of a distribution, the higher the diversity of samples taken from it. The entropy of a distribution is defined by:

$$H[X] = -\sum\nolimits_{x} p(x) log_2 p(x) \tag{10}$$

where $p(x)$ is the probability that the random variable X takes the value x. In the proposed framework, x corresponds to the cluster membership of a user and $p(x)$ is the probability that a randomly sampled user is a member of cluster x. Observe how entropy increases as the number of possible values x could take increase and as the probabilities all take on similarly small values.

4 A case study from a large active OSS

We next present a case study to demonstrate our framework. It substantiates the utility of the kernel approach to segment users in a vibrant and active OSS service. The case study also gives a concrete example of evaluating hyper-parameter settings based on quality and equity, and presents an engagement analysis along the three dimensions' discussion earlier. The case study also contrasts the results of kernel *k*-means method against those obtained from feature correlation analysis, traditional *k*-means clustering and a recently proposed method for high dimensional data clustering (SUBCLU).

4.1 Platform and data description

The subject platform for our case study is an emerging OSS called *7 Cups of Tea* (*7cot*). 7cot is an online service offering private, anonymous, quick, and live emotional support. This support is usually delivered through a one-on-one *conversation*, which is defined as a chat session between two people. With rich gamification mechanisms (e.g. public points and badges for talking to those needing support), the platform fosters an active community or crowd of "listeners" who are trained to help "members" that are facing a wide range of emotional problems. Besides one-on-one conversations, members and listeners communicate on a vibrant forum, or in group chat rooms. In less than 2 years, 7cot has attracted a community of over 130,000 users who collectively held over 1.2 million one-on-one conversations.

With direct communications from one user to another, 7cot clearly fits the description of an OSS. The social elements of the platform include connections between members and the listeners they communicate with, group chats defining relationships among participants, and forums that connect members across the user base by the threads in which they participate. Interactions among members and listeners imply the existence of a bipartite social network structure (Doran et al. 2015) that is not explicitly represented in the site. We thus do not classify 7cot into an online social network or social network service. It also is not a form of social media because users do not use 7cot as a medium to broadly share information with others.

A database capturing the attributes of all users, interactions, and activities performed since the inception of 7cot on December 5th, 2013 through November 18th, 2014 is considered in a case study. The database includes metadata about every user except for those attributes related to the user's true identity, contact information, and transcripts of the one-on-one conversations. In fact, the attributes of each conversation record were limited to participant identifiers, the date the conversation commenced, the number of messages exchanged by the participants, whether the conversation was for a teenager or adult member, if the conversation was terminated by the member or listener, and the timestamp of the last message sent. User behaviors on the site were captured between May 7th and November 18th. For privacy reasons, the number of messages sent, requests made, forum posts made, logins, forum views, help guide views, and page views through the mobile app or Web browser per user per day are the only attributes captured. Table 2 summarizes the participations and actions of the user base. The participation statistics underscore the size and volume of activities on 7cot, while the action statistics reveal the average activity of a user during days when they have logged in at least once. For example, the table shows how members connect to an average of 1.83 listeners on the days they access 7cot. With a deeper characterization of this data and an evaluation of user interactions available in our previous studies (Calzarossa et al. 2016; Doran et al. 2015), Table 2 demonstrates that the platform's levels of participations and daily actions make it a suitable platform for a case study.

4.2 Kernel *k*-means clustering

In this section, we apply kernel *k*-means clustering over a set of nine attributes. Table 3 presents a list of various types of user behaviors representative of each of the engagement dimensions discussed earlier in the paper.

In Fig. 2, we explore the distributions of all the features in our data set to confirm if 7cot is a reasonable candidate for the application of a kernel *k*-means method. The figure demonstrates that all the attributes are skewed, and exhibit a heavy-tailed shape. In each of the nine, the x axis represents the magnitude of a quantity while the y-axis represents the probability of the corresponding values of x is greater. We label the smallest value of x after which a tail follows a power law distribution as x_{min}. The common approach to estimate this value is by visual inspection of the distribution on log-log scale. However, such an approach is highly subjective and prone to error. Clauset et al. (2009) provide a recommendation for estimating this lower threshold using the Kolmogorov-Smirnov test. We implement this approach to test for power law distributions in the attributes. The parameters obtained using this approach are shown in Table 4. Since the parameter α lies between 2 and 3 for all but one feature, the mean exists but higher order moments for all but one feature are infinite.

To build the clustering model using the kernel *k*-means method, we chose to vary the range of σ between 0.001 and 0.01. This is because, for values of σ above and below 0.01 and 0.001 respectively, solutions for any value of k undesirably placed the majority of points in just one or two clusters. Such a solution does not provide any actionable feedback, and hence is undesirable. We also limited the parameter k to be less than 7 so that users are not divided into a large number of small clusters that only represent a very specific behavior pattern.

Table 2 Summary of 7cot user participations and actions

Participation		Actions (average per user per active day)	
Number of Conversations	1,145,797	Logins	2.41
Distinct Forums	53	Conversational Messages	34.48
Number of Messages	39,509,790	Conversational Requests	1.83
Forum Posts	82,223	Forum Posts	2.93
Number of Members	87,232	Forum Post Views	6.38
Number of Listeners	33,601	Page Views	15.98
Number of Hybrid	12,038	Help Guide Views	4.12

We also chose this number based on the practical number of business strategies and feedback initiatives that the business could undertake to encourage user engagement. Figure 3 shows a trace of SI values for different clustering solutions as k and σ vary. No matter the value of k, we note that smaller values of σ have a tendency to yield results with smaller SI values. This may be due to the fact that, as the value of σ is lowered, the algorithm fails to compute reasonable linear separators with dropping values of σ. Setting $k = 5$ or 6 reveals the best separation of clusters for $\sigma > 0.001$.

Table 5 presents the entropy of the cluster size distributions for all clustering solutions whose SI values are among the top five in Fig. 3. We chose the top five heuristically noting that the SI values begins to steeply decrease beyond the fifth value. We find that the parameter settings $k = 6$ and $\sigma = 0.009$ yields a distribution of users into clusters with the highest entropy, and hence, is a balanced solution.

Table 3 Dimensions and attributes used for clustering

Dimension	Attribute	Description
Initiation	Number of account logins (L)	Total number of times a user has logged into the website since they registered. It captures the idea that, the more frequently a user accesses the OSS, the more interested she may be in taking advantage of the site's services.
	Number of conversation requests (R)	Total number of times a user has requested a conversation from another user on the site. Conversation requests are a very strong form of interaction, with a user explicitly reaching out to connect to another person on the social service.
Interaction	Number of messages exchanged (M)	Total number of one-on-one, directed messages sent between users. It offers a direct insight into the extent to which a user is involved with other website users.
	Number of forum posts (P)	Total number of times a user has posted in the forum of the website since they registered. Intuitively, larger numbers of forum posts suggest that a user is experiencing strong site engagement.
	Number of forum views (V)	Total number of times a user has viewed the websites forum. This subtle feature captures the degree to which users are interested in the content shared on the service. The more the views, the more captured a user's interest is.
Loyalty	Number of help guide views (H)	Total number of times a user has viewed the help guide from the website for different problems. This depicts how much a user trusts the website's help resources.
	Number of page views from web (PW)	Total number of pages viewed by user surfing the website through web. This directly measures involvement of users who surf the web version.
	Number of page views from app (PA)	Total number of pages viewed by a user surfing the website through app. This directly measures involvement of users who surf the app version.
	Number of active days (A)	Total number of days that a user has logged into the website. This measures the user's loyalty. The greater the active days, the more loyal a user is.

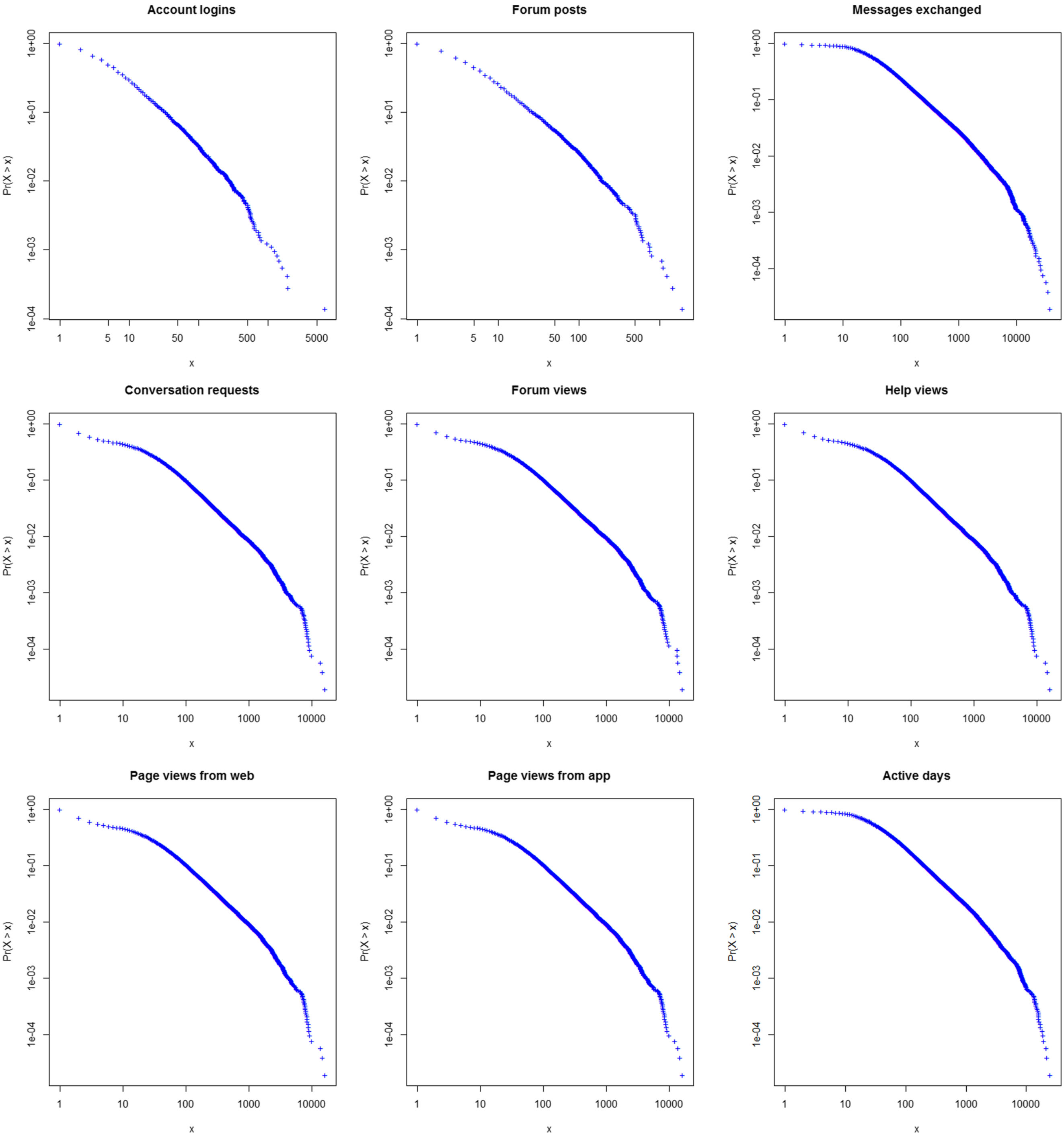


Fig. 2 Plots for heavy tailed distribution for the nine features as in Table 3

4.3 Comparative analysis

In this section, we make a case for the need to use kernel *k*-means for analyzing user engagement on a large OSS. First, we run a simple correlation analysis on the attributes to determine if we could be obtain similar insights. Next, we compare the clusters that emerge from our kernel *k*-means framework against clusters generated by a traditional *k*-means clustering algorithm and a leading high dimensional data clustering method (SUBCLU).

4.3.1 Correlation analysis

We run a correlation analysis on the attributes using Pearson's correlation coefficient. We illustrate the results with a heat map in Fig. 4. As evident from the figure, most

Table 4 Parameters of power law fits for each feature (estimated by Clauset et al.'s (2009) test)

Feature	Parameter (α)	x_{min}
Account logins (L)	2.095	100
Forum posts (P)	2.417	95
Messages exchanged (M)	1.950	87
Conversation requests (R)	2.112	158
Forum views (V)	2.065	101
Help views (H)	2.101	101
Page views from web (PW)	2.077	101
Page views from app (PA)	2.076	98
Active days (A)	2.037	109

attributes show no significant correlation among themselves except for three cases. The first case exhibits a correlation between the attributes, Account login (L) and Active days (A). This is intuitive in the sense that users who log in often are expected to use the website over a large number of days. The second case is the correlation between Forum Posts (P) and Forum views (V). This observation is consistent with the fact that people who often post on the forum would also view the forum often to track responses on the post and other posts. The third observation is that the Messaging attribute (M) exhibits very little correlation with any other attribute. This can be explained by the fact that the total number of messages sent (39,509,790) greatly exceeds the values of all other attributes as evidenced by the total number of conversations sent in Table 2. Moreover, Table 4 identifies M as having the smallest value of α, indicating that its distribution exhibits the greatest amount of variation. In fact, since α <2, the variance of the sampling distribution of the number of messages sent per user would have increasing variance as the size of the sample increases. Besides these simple insights, which match intuitive reasoning, correlation analysis fails to give out any significant insight on the engagement experienced by the members of the platform.

Table 5 Entropy of the top five SI values

Hyper-parameter (σ)	*k*	SI value	Entropy
0.01	5	0.193	1.243
0.007	6	0.196	1.673
0.009	6	0.190	1.682
0.01	6	0.199	1.626
0.007	5	0.184	1.250

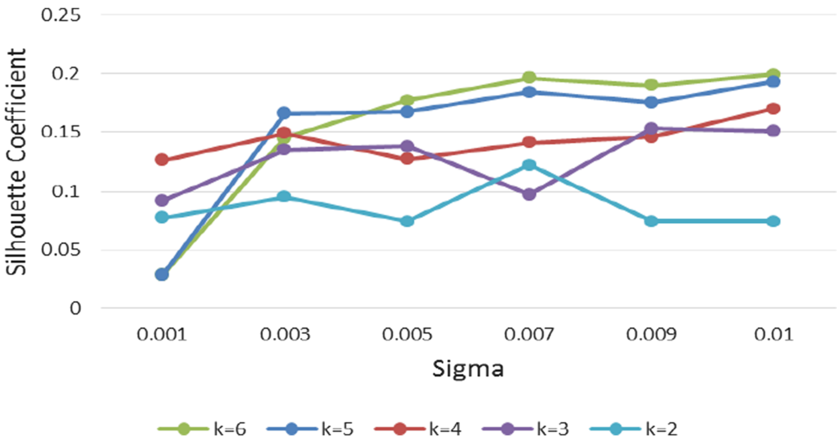
4.3.2 Traditional *k*-means

We next compare our results with the results generated by the traditional *k*-means algorithm. Table 6 presents the SI value and the entropy of the clustering results for varying *k*. While a very strong SI value of 0.922 is achieved for any *k*, we note that the entropy values are very small compared to the kernel *k*-means approach. The effect of these low entropy clustering become apparent when examining the distribution of cluster sizes. For example, Fig. 5 gives this distribution for *k* = 6, where over 96 % of all users fall into one cluster, with the scant few remaining distributed among the others. Such a model, which lumps all users into a similar engagement profile, cannot support practical business decision making.

For a parsimonious comparative analysis that would utilize only consistent parameters, we further consider a range of cluster counts between 3 and 6 for both the traditional and kernel *k*-means methods. Bearing the discussion in the hyper-parameter searching section for our choice of *k*, we deemed these cluster sizes as the most practicable and actionable for our target OSS in terms of the business strategy aligned with improving user engagement. Also, for the kernel *k*-means method, we chose a sigma value of 0.009 as was determined during the hyper-parameter search.

As shown in Fig. 6, the kernel *k*-means algorithm performs better than the traditional *k*-means algorithm across this range of *k* based on entropy values. Specifically, the average entropy value of the kernel *k*-means algorithm (1.339) is an order of magnitude greater than that of the traditional *k*-means

Fig. 3 Chart of SI values for different assignments of *k* and *σ*



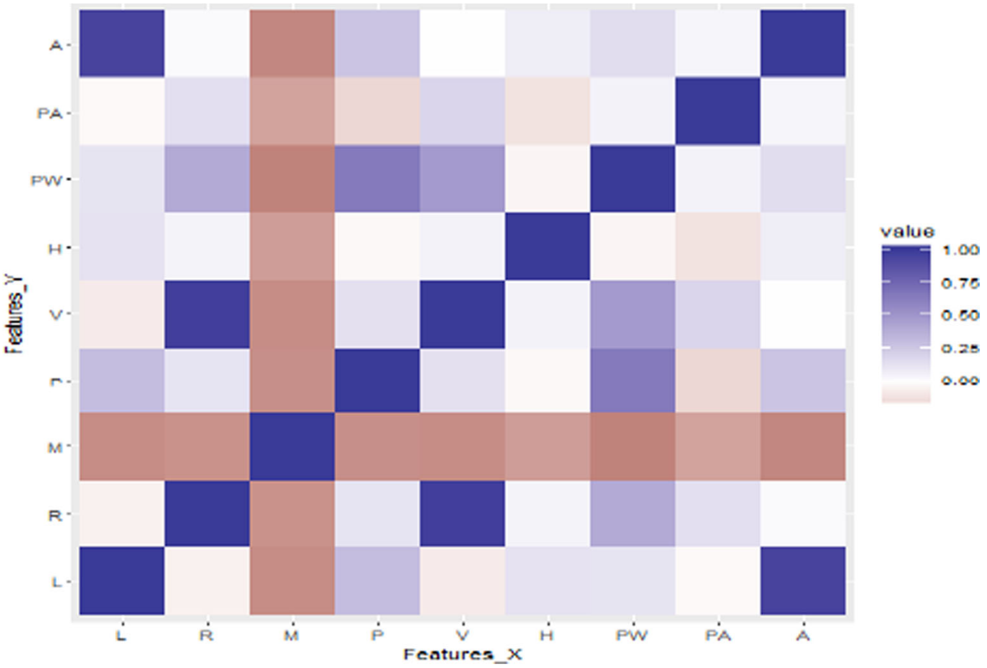


Fig. 4 Correlation between attributes (sequence of attributes based on Table 4)

algorithm (0.123). Whereas the traditional *k*-means cluster distribution (for $k = 6$) as shown in Fig. 5 indicates that at least 96 % of the users were grouped into just one cluster, the kernel *k*-means cluster distribution in Fig. 7 for $k = 6$, $\sigma = 0.009$ shows a relatively even distribution with the largest cluster containing approximately 30 % of the users.

4.3.3 High dimensional clustering (SUBCLU)

We also compare our framework against a method for clustering high dimensional data, namely, the SUBCLU algorithm (Kailing et al. 2004). SUBCLU overcomes the limitations of the CLIQUE (Agrawal et al. 1998), a pioneering approach in subspace clustering (Kriegel et al. 2005), which essentially identifies projections of the input data into a space where a subset of the attributes in the data set is represented with regions that have high density of points. Whereas CLIQUE prunes parts of the data space whose point density falls below a certain threshold (Agrawal et al. 1998), thus possibly missing out on the clustering information present in the data set, SUBCLU does not take this pruning approach. Past studies confirm that SUBCLU's approach yields superior clustering results (Kailing et al. 2004).

Table 6 SI and entropy values for traditional *k*-means clustering

SI value	Entropy	*k*
0.983	0.024	2
0.973	0.027	3
0.971	0.054	4
0.945	0.124	5
0.922	0.186	6

Following the guidance of Kailing et al. (2004), we ran the SUBCLU algorithm with parameter settings MinPts = 8 and $\varepsilon = 2.0$ over the 7cot dataset. SUBCLU was able to identify four clusters whose size distribution is shown in Fig. 8. While this distribution appears to be more equitable than traditional *k*-means, it still does not offer an even distribution over a larger number of clusters as compared to the kernel *k*-means. Figure 9 quantifies this difference. The entropy of the SUBCLU solution (0.872) is much better as compared to traditional *k*-means (0.091 for $k = 4$), but less than the proposed kernel *k*-means framework (1.68 for $k = 4$, σ=0.009).

5 Engagement analysis

In this section, we discuss findings that we generated based on the kernel *k*-means algorithm and explain the dynamics of user segmentation in terms of their engagement activities on the OSS used in the case study.

5.1 Kernel *k*-means

In Fig. 10, we show the normalized centroid coordinates with respect to a particular feature on log scale. Consistent with previous cluster analysis studies by Davis et al. (1988); Fredline and Faulkner (2000) and Madrigal (1995), we identified and created social representations based on common themes among the clusters. We identified three groups based

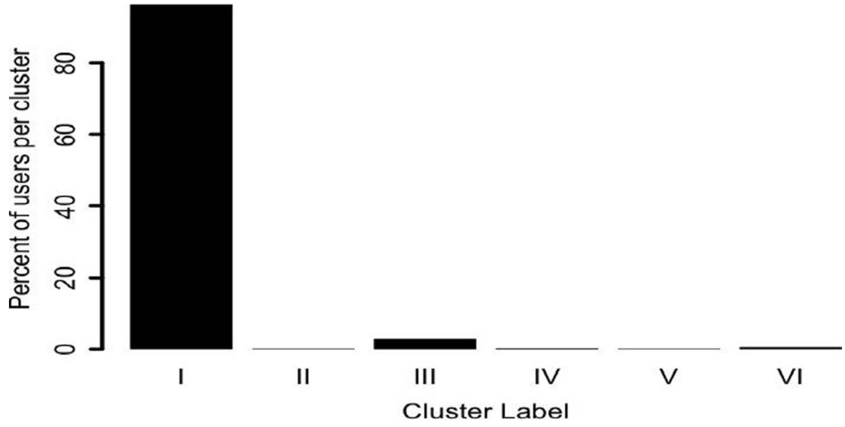


Fig. 5 Distribution of users per cluster using traditional *k*-means

on common characteristics and cluster patterns in the community of users.

Group One: This is the largest group of users and is represented by cluster IV. Figure 10 shows that these users rank high in all of the three key dimensions, initiation, interaction and loyalty. They are more engaged in the sense that, they log in and surf the website often, either using a web or a mobile version of the service. They also post messages, respond to other posted messages and generally engage in conversations with other users. Group one essentially forms the participant base of the platform who maintains and popularizes the functions of the website.

Group Two: The next group of users is represented by cluster III. In Fig. 10, it is evident that the users in this cluster have high values in terms of initiation and moderate values in terms of interaction. They also have extremely low values in terms of loyalty. These users tend to visit the website frequently and log in often. However, their participation is limited in terms of exchange of messages with other participants on the platform. Even though they have a lot of interest in the website (which their frequent logins demonstrate), their full potential as loyal participants is not realized yet. That is, they have interest in the website yet, they probably do not interact extensively because there are not enough suitable functions on the platform that serve their need. They are at the borderline of fully engaging in or disengaging from the platform if the OSS does not meet their interests in a timely manner. In a situation where the business wants to expand its participant-base, this group will be the most important group to target.

Group Three: The third group of users belong to the remainder of the clusters: clusters I, II, V, and VI. Based on Fig. 10, we see that these users have no loyalty; neither do they initiate visits to the platform often. Their prime engagement with the platform is via the interaction dimension. We also realize that these users send a lot of conversation request that results in a relatively high number of messages exchanged with other users. In reference to an earlier description of the platform, these users are "members" who utilize the free social services of the 7cot platform by seeking support from "listeners' about their condition. This group of users are not necessarily the dominating participant base of the platform. However, they are very purposeful in terms of exact services they need from the platform. They have a niche interest, and hence not all the numerous functions on the platform would appeal to them. As a result, the business will have to identify the specific needs and interest of such a group of users.

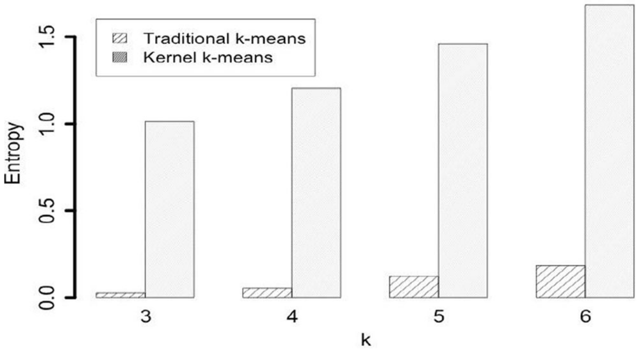


Fig. 6 Entropy values for kernel and traditional *k*-means for different values of *k*

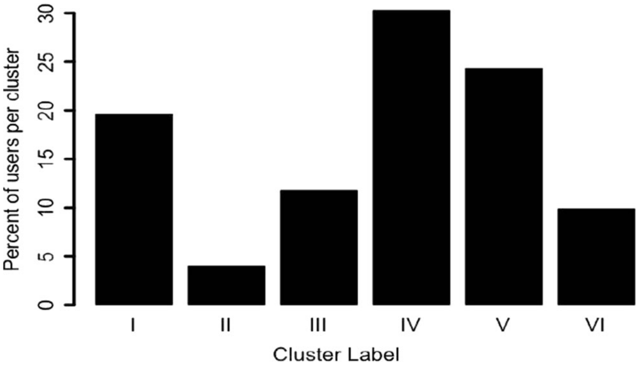


Fig. 7 Distribution of users per cluster using kernel *k*-means

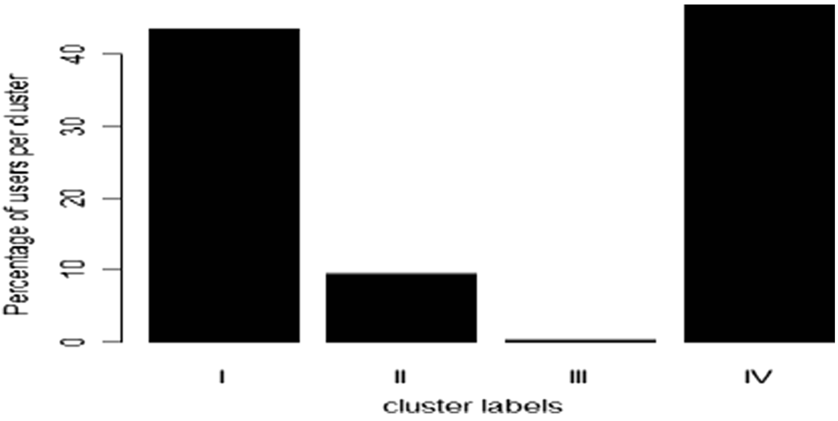


Fig. 8 Distribution of users per cluster using SUBCLU

5.2 Comparison with traditional *k*-means

While Section 4.3.2 established that kernel *k*-means yields a more equitable and actionable clustering, we next see if traditional *k*-means could still have delivered meaningful insights. Figure 11, shows the normalized centroid coordinates with respect to a particular feature on log scale with traditional *k*-means. Unfortunately, analysis reveals a number of confusing or difficult to interpret observations. First, cluster VI contains users that sent the most number of messages as shown in the interaction dimension, but this is the only kind of activity they perform on the platform. This is inconsistent with one's expectations about platform use since they do not send any conversation requests to listeners, or use other actions of the site. Users in cluster III rate the highest along attributes of the loyalty dimension even though they do not send many messages. Again, this is a paradox since the set of users that are most significant along the loyalty dimension hardly use OSS' most important feature (interacting with listeners) and also do not bother to register on the website. Moreover, cluster II rates among the second highest set of users for sending messages to listeners. However, they have second lowest values along loyalty and initiation dimensions. Given the growing community of members, this may be contradictory since users with high messaging rates should at least have some significant engagement along other dimensions. These hard to interpret observations and the results presented in Section 4.3.2, underscore the unsuitability of *k*-means for engagement analysis on a large OSS data.

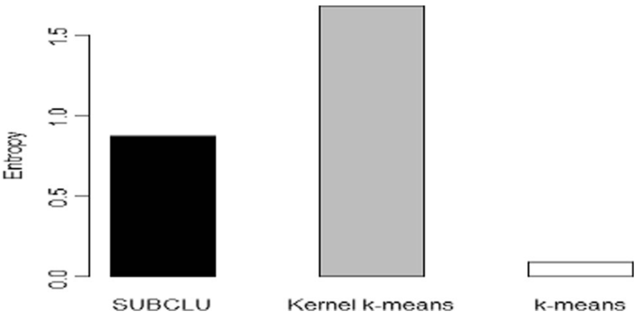


Fig. 9 Comparison of entropy values between SUBCLU, kernel k-means and k-means

6 Discussion

Our study, along with previous studies such as Fredline and Faulkner (2000) and Jain (2010) shows that organizing data into sensible groups such that practical and informed actions can be taken is a useful requirement of any unsupervised learning methodology. This is evident in the use of our framework, which allows one to identify a coherent group of data elements with a common set of characteristics distinct from other groups of elements. The role of kernel k-means to generate clusters is often times a preliminary step to a more targeted approach to understanding patterns of user profiles on an OSS. Based on our analysis, we have demonstrated that kernel *k*-means algorithms are capable of capturing non-linear relationships between data points in a large data set, and hence highly useful for real world applications such as studying user engagement on OSS platforms (Chitta et al. 2014). For instance, in our case study, we are able to identify different clusters of individuals (group one, group two and group three). To a practitioner, a closer examination of the common characteristics of these groups would provide vital insights for enhancing engagement by catering and responding to users' needs.

By comparing with leading clustering methods, traditional *k*-means and a recently devised high clustering algorithm, SUBCLU, we show that the kernel *k*-means algorithm produces relatively higher entropy values on average. This indicates that the kernel algorithm generates clusters that have relatively even uniformity in terms of cluster sizes as shown in Fig. 7. For businesses and organizations who have online presence via OSS platforms, this allows for the deployment of practical and actionable analysis and decisions on how to improve user engagement.

Being able to determine the number of clusters needed for any application using the kernel *k*-means approach could be a

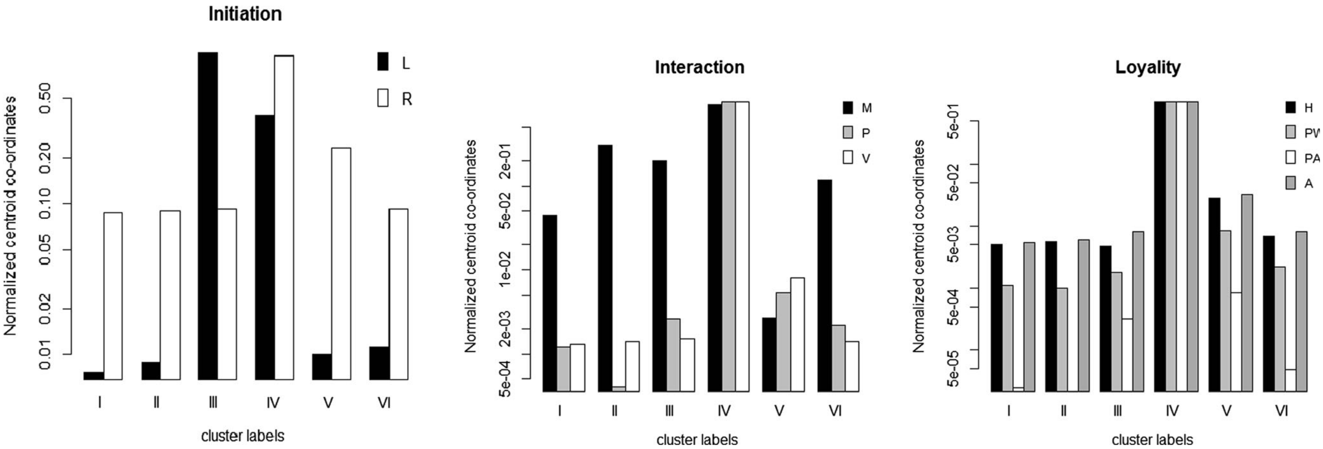


Fig. 10 Plot of normalized centroid co-ordinates obtained using kernel k-means for all features grouped according to their respective dimensions (Graph legend is explained in Table 3)

difficult experience that may require both science and art. Based on our case study, we identify a four-step approach for determining the best cluster solution for a kernel *k*-means clustering analysis that is applicable to other business segmentation problems. First, we consider a practicable value for *k* based on the business application for such segmentation. Secondly, we consider the hyper-parameter, σ, which is most important factor that forms the basis of the kernel function utility. We liberally chose the value for σ such that at least 80 percent of the data points are grouped into at least two clusters. We deemed this practical for the user engagement application we explore in the case study. In our dataset, we found this value to be between 0.01 and 0.001. Next, we choose an SI value by picking the top few value after which there is a significant drop in the SI values. Lastly, we choose the most practical clustering solution by choosing a clustering solution with the highest entropy.

The engagement analysis framework provides a three-fold insight into an OSS platform such as 7cot. Just as is demonstrated in the case study, first, it helps explore, understand and group features that influence the level of user engagement into three deferent dimensions. Secondly, with the aid of kernel *k*-means, it is able to demonstrate how to segment the data set into different actionable segments. Finally, it can help generate in-depth insights about how users engage and use social networks on an individual basis. This in effect could help businesses identify and formulate strategies to increase customer retention, increase sales and roll out marketing campaigns. For instance, on the 7cot platform, if a social feature introduced shows a comparatively higher engagement among a group of users with respect to a particular feature, one may imply that the feature is very useful for the group and use it as a basis to promote products to other groups who might have similar group characteristics.

6.1 Utilization of insights to promote engagement

The insights that we obtained are in harmony with the way platform is structured. The unique service that the platform offers to its users to interact with people who are willing to listen to their grievances was the most important social feature

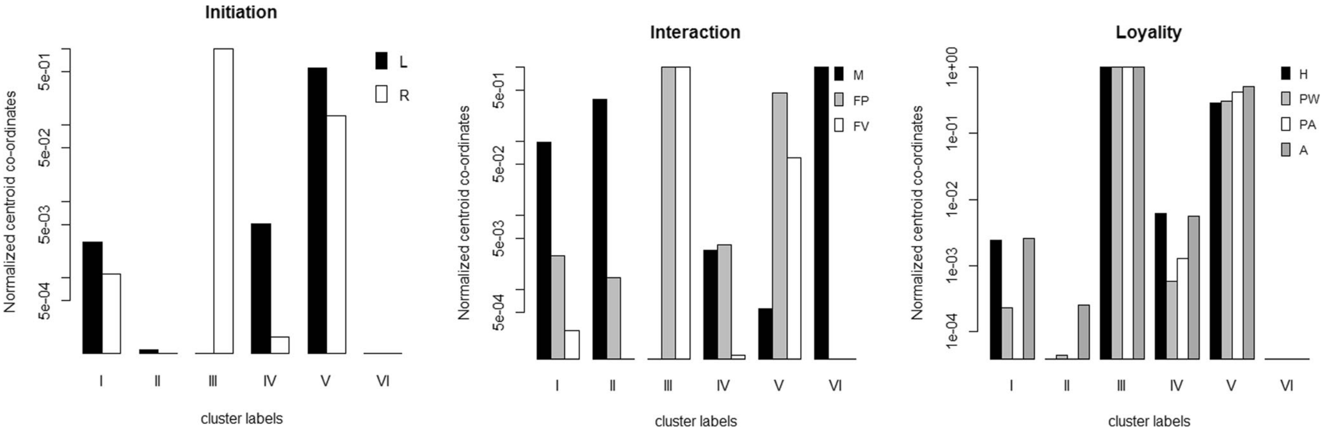


Fig. 11 Plot of normalized centroid co-ordinates obtained using traditional k-means for all features grouped according to their respective dimensions (Graph legend is explained in Table 3)

existing on the website at the time of this study. The way users do this is by sending conversation requests to listeners.

As is evident from Fig. 10, users in cluster IV are among the most engaged users along all three dimensions. They not only account for most of the interactions occurring on the website, but they also account for most of the conversation requests sent on the website. This is consistent with the structure of the website. They are also the most loyal set of users existing on the website. These users comprise the largest segment of the total users (Fig. 7). This confirms that the facility offered by the website to interact with listeners is actually engaging to the users visiting the platform of 7cot.

We have communicated our results to 7cot. Preliminary discussions suggest that as part of their overall development plan, the conclusions drawn from this study may help shape their business strategy, strengthen, and expand their user base. The three specific outcomes we identified for 7cots' implementation are listed below:

1. 7cot's mission is to mitigate different concerns of the users that arrive on its online platform. Quoting from the website of 7cot, "this website is for anyone who wants to live in a world free of stigma and stereotype. A world where all 7 billion of us can grow and feel we truly belong." It does so by offering users an opportunity to interact with an individual who is willing to listen to their grievances. For any company, the prime concern is usually how relevant is their mission to the existing population and whether the mission carries enough significance and substance to attract a strong user base. We observed that 7cot has a core set of users in Group one who would immediately relate to its mission. This set of users are the same set of users that account for most of the interactions happening on the website with the listeners. Thus, to maintain and increase their user base, the company should focus on making its services better and if possible make it more easily accessible and robust.
2. Group two users visit the website most often but the website fails to capture their interest as they are among the least loyal set of users. The most prominent reason (going by the website' mission and since they visit so often) that surfaces is loneliness or apathy. Maybe they just need to be prompted to kindle prolonged interest. To tap into this section of users, the website should focus on arranging prompters as soon as a visitor arrives on their website. It only needs to do this for users who actually have a high login rate since in this case these users have most number of logins.
3. Lastly, we observed that Group three users initiate activities on the website and also perform other interactions but the website fails to capture their interests. This is reflected in them being not loyal to the OSS as depicted by the loyalty dimension. Since they perform a fair amount of interaction on the website, these users have something in common with the mission of the website. Thus, these users could be tapped into to further strengthen the loyal user base of the website. The website offers a service where it sends periodic updates according to the preference of the user. This period ranges from immediately to weekly. The company may need to target these users by specifically catering to their specific needs. For instance, this could involve informing them of something happening on the website that may hold their interest. Holding online webinars for users struggling with a particular ailment is one such example.

A surprising observation we made in this study was that there was a set of users (Cluster III) that had surprisingly high number of logins (Fig. 10), yet mostly choose not to interact with listeners despite this being the most important social feature on 7cot's OSS. This set of users not only have surprising high number of logins in comparison to the interactions they perform on the website, but also, they have the most number of logins among all sections of users. These users might be in the phase of familiarizing themselves with the website. Another reason could be the presence of loneliness or apathy. Presence of such users on this platform should be a common phenomenon but they distinguished themselves on how strongly they emerge in our study by accounting for the most logins on the website.

7 Conclusion and future work

This study provides a framework for studying user engagement, presenting a set of three dimensions for exploring user engagement on online social service (OSS) platforms. Using the specifics of a case study, we further demonstrate how the kernel *k*-means method surpasses commonly used techniques in terms of performance, when analyzing high-dimensional big data sets on OSS platforms.

Our study mainly seeks to present a framework that supports the study of user engagement on an OSS platform. By nature, cluster analysis tends to function as a method that teases out underlying models and structure in data based on a set of inherent characteristic. Certainly, understanding of the data, signals and user segments is vital in this process. We utilized these steps in our case study as we generated the three groups of users among the clusters. Understanding the data and user segments informed how we utilized the concept of social representation (Fredline and Faulkner 2000) to create the three groups (or themes) of user.

We utilized real data from an online social service platform, 7cot for our analysis. The study can be extended to develop frameworks to study other customer relations issues in other domain areas such as retail. We also intend to extend our study by delving deeper into other features and services that support user engagement on an OSS. The obvious challenge would be to maintain the generalizability of such an extension while maintaining its usability for specific purposes. The present work took into account the kernel *k*-means clustering method. Incorporating and devising different clustering methods and algorithms that work for different case scenarios would be another avenue for extending this study. The most general extension would be where the framework can itself decide on the kind of algorithm that would work, given the nature of data collected from the users. For example, the present dataset had heavy tails along all the nine features. A general framework would not only predict the kind of algorithm applicable based on the kind of data-set but would also offer insights on the profitability of the specific services offered by the website.

References

Adamic, L.A., & Glance, N. (2005). The Political Blogosphere and the 2004 U.S. Election. In *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05* (pp. 36–43). http://dl.acm.org/citation.cfm?id=1134271.1134277\nhttp://portal.acm.org/citation.cfm?doid=1134271.1134277.

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). *Automatic subspace clustering of high dimensional data for data mining applications* (vol. 27). ACM.

Anderson, A., Huttenlocher, D., & Kleinberg, J. (2012). Discovering Value from community activity on focused question answering sites : a case study of stack overflow. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 850–858).

Arapakis, I., Lalmas, M., Cambazoglu, B. B., Marcos, M., & Jose, J. M. (2014). User engagement in online news: under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology, 65*(10), 1988–2005.

Attfield, S., Kazai, G., Lalmas, M., & Piwowarski, B. (2011). Towards a science of user engagement (position paper). *WSDM Workshop on User Modelling for Web Applications*. http://www.dcs.gla.ac.uk/~mounia/Papers/engagement.pdf.

Benevenuto, F., Rodrigues, T., Cha, M., & Almeida, V. (2009). Characterizing user behavior in online social networks. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, (17313374869111495992related: ON2QmLZ2RfAJ) (pp. 49–62).

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.

Booth, D., & Jansen, B. (2008). A review of methodologies for analyzing websites. *Handbook of Research on Web Log Analysis*, 141–162. http://faculty.ist.psu.edu/jjansen/academic/jansen_website_analysis.pdf.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry* (pp. 4–7).

Calzarossa, M. C., Massari, L., Doran, D., Yelne, S., Trivedi, N., & Moriarty, G. (2016). Measuring the users and conversations of a vibrant online emotional support system. In *IEEE Symposium on Computers and Communications*, Messina, Italy.

Camps-Valls, G., Rojo-Alvarez, J. L., & Martinez-Ramon, M. (2007). *Kernel methods in bioengineering, signal and image processing*.

Chang, H.-J., Hung, L.-P., & Ho, C.-L. (2007). An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis. *Expert Systems with Applications*, *32*(3), 753–764. http://linkinghub.elsevier.com/retrieve/pii/S0957417406000327.

Chen, Y.Y., Lai, F.W., Goh, K. N., & Daud, S.C. (2013). The effect of integrating social plugins into e-commerce website : a study on online consumer behaviour. In *International Conference on Ubiquitous Information Management and Communication. ACM* (pp. 0–5). doi:10.1145/2448556.2448612.

Chitta, R., Jin, R., Havens, T. C., & Jain, A. K. (2014). Scalable Kernel clustering : approximate Kernel k-means. *arXiv*.

Chu, C.-W., Holliday, J.D., & Willett, P. (2009). Effect of data standardization on chemical clustering and similarity searching. *Journal of Chemical Information and Modeling*, *49*(2), 155–61. http://www.ncbi.nlm.nih.gov/pubmed/19434820.

Chuang, H. H. C., Lu, G., Peng, D. X., & Heim, G. R. (2014). Impact of value-added service features in e-retailing processes: an econometric analysis of web site functions. *Decision Sciences Journal of Innovative Education, 45*(6), 1159–1186.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review, 51*(4), 661. doi:10.1137/070710111.

Constant, D., Sproull, L., & Kiesler, S. (1996). The kindness of strangers: the usefulness of electronic weak ties for technical advice. *Organization Science, 7*(2), 119–135.

Cyr, D., Hassanein, K., Head, M., & Ivanov, A. (2007). The role of social presence in establishing loyalty in e-service environments. *Interacting wih Computers, 19*(1), 43–56. doi:10.1016/j.intcom.2006.07.010.

Davis, D., Allen, J., & Cosenza, R. M. (1988). Segmenting local residents by their attitudes, interests, and opinions toward tourism. *Journal of Travel Research, 27*(2), 2–8.

De la Torre, F., & Vinyals, O. (2007). Learning Kernel expansions for image classification. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on* (pp. 1–7).

Doran, D., Yelne, S., Massari, L., Calzarossa, M.-C., Jackson, L., & Moriarty, G. (2015). Stay awhile and listen: user interactions in a crowdsourced platform offering emotional support. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 667–674). Paris, France: ACM.

Farajian, M., & Mohammadi, S. (2010). Mining the banking customer behavior using clustering and association rules methods. *International Journal of Industrial Engineering*.

Fredline, E., & Faulkner, B. (2000). Host community reactions: a cluster analysis. *Annals of Tourism Research, 27*(3), 763–784.

Gagné, C., & Godin, G. (2005). Improving self-report measures of non-adherence to HIV medications. *Psychology & Health, 20*(6), 803–816. doi:10.1080/14768320500386441.

Gong, N. Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., & Song, D. (2012). Evolution of social-attribute networks: measurements, modeling, and implications using Google+. *Internet Measurement Conference*, 131–144. doi:10.1145/2398776.2398792.

Gummerus, J., Liljander, V., Pura, M., & Riel, A. V. (2004). Customer loyalty to content-based web sites: the case of an online health-care service. *Journal of Services Marketing, 18*(3), 175–186.

Hawkshead, J., & Krousel-Wood, M. A. (2007). Techniques for measuring medication adherence in hypertensive patients in outpatient settings: advantages and limitations. *Disease Management and Health Outcomes*, *15*(2), 109–118. http://ideas.repec.org/a/wkh/dmhout/v15y2007i2p109-118.html.

Huang, Z., & Benyoucef, M. (2013). From e-commerce to social commerce: a close look at design features. *Electronic Commerce Research and Applications, 12*(4), 246–259. doi:10.1016/j.elerap.2012.12.003.

Iba, T., Nemoto, K., Peters, B., & Gloor, P. A. (2010). Analyzing the creative editing behavior of wikipedia editors: through dynamic social network analysis. *Procedia - Social and Behavioral Sciences, 2*(4), 6441–6456.

Ikehara, C., & Crosby, M. (2005). Assessing cognitive load with physiological sensors. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, *00*(C), (pp. 1–9).

Jacques, R. D. (1996). *The nature of engagement and its role in hypermedia evaluation and design*. Retrieved from http://ethos.bl.uk/OrderDetails.do?uin = uk.bl.ethos.336369.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. doi:10.1016/j.patrec.2009.09.011.

Kailing, K., Kriegel, H.-P., & Kröger, P. (2004). Density-connected subspace clustering for high-dimensional data. In *Proc. SDM* (vol. 4). SIAM.

Kim, E. (2013). Everything you wanted to know about the Kernel trick (but were too afraid to ask). http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html. Accessed 1 Aug 2014.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of cluster in K-means clustering. *International Journal of Advance Research in Computer Science and Management Studies, 1*(6), 90–95.

Konradt, U., & Sulz, K. (2001). The experience of flow in interacting with a hypermedia learning environment. *Journal of Educational Multimedia and Hypermedia*, *10*(1), 69–84. http://www.editlib.org/index.cfm?fuseaction = Reader.ViewAbstract&paper_id = 7992&from = NEWDL.

Kriegel, H.-P., Kroger, P., Renz, M., & Wurst, S. (2005). A generic framework for efficient subspace clustering of high-dimensional data. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE.

Kunegis, J., Lommatzsch, A., & Bauckhage, C. (2009). The Slashdot zoo: mining a social network with negative edges. *Proceedings of the 18th international conference on World wide web* (pp. 741–750). doi:10.1145/1526709.1526809.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, s social network or a news media? *The International World Wide Web Conference Committee (IW3C2)* (pp. 1–10).

Lagun, D., Hsieh, C.-H., Webster, D., & Navalpakkam, V. (2014). Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 113–122). ACM.

Lalmas, M., O'Brien, H. L., & Yom-Tov, E. (2013). Measuring user engagement. In *Tutorial s of the 22nd International World Wide Web Conference*.

Lehmann, J., Lalmas, M., Yom-Tov, E., & Dupret, G. (2012). Models of user engagement. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7379 LNCS* (pp. 164–175). doi:10.1007/978-3-642-31454-4_14

Lima, A., Rossi, L., & Musolesi, M. (2014). Coding together at scale: GitHub as a collaborative social network. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (pp. 295–304). http://www.scopus.com/inward/record.url?eid=2-s2.0-84909956490&partnerID=tZOtx3y1.

Lipsky, L. (2009). *Queueing theory : a linear algebraic approach*. doi: 10.1007/ 978-0-387-49706-8.

Madrigal, R. (1995). Residents' perceptions and the role of government. *Annals of Tourism Research, 22*(1), 86–102.

Mislove, A., Koppula, H.S., Gummadi, K.P., Druschel, P., & Bhattacharjee, B. (2008). Growth of the flickr social network. In *Proceedings of the first workshop on Online social networks - WOSP '08* (pp. 25–30). doi:10.1145/1397735.1397742.

Nov, O. (2007). What motivates Wikipedians? *Communications of the ACM, 50*(11), 60–64.

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology, 59*(6), 938–955.

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology, 61*(1), 50–69.

Ogonowski, A., Montandon, A., Botha, E., & Reyneke, M. (2014). Should new online stores invest in social presence elements? The effect of social presence on initial trust formation. *Journal of Retailing and Consumer Services, 21*(4), 482–491. doi:10.1016/j.jretconser.2014.03.004.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining. Library of Congress*. doi:10.1016/0022-4405(81)90007-8.

Romero, R., Iglesias, E., & Borrajo, L. (2014). A linear-RBF multikernel SVM to classify big text corpora. *BioMed Research International*. http://www.hindawi.com/journals/bmri/aa/878291/.

Seah, M., & Cairns, P. (2008). From immersion to addiction in videogames. In *Proceedings of BCS HCI* (pp. 55–63). http://eprints.whiterose.ac.uk/68181/.

Shah, S. K. (2006). Motivation, governance, and the viability of hybrid forms in open source software development. *Management Science, 52*(7), 1000–1014.

Shukla, S., & Naganna, S. (2014). A review on K-means data clustering approach. *International Journal of Information and Computation Technology, 4*(17), 1847–1860.

Szell, M., & Thurner, S. (2010). Measuring social dynamics in a massive multiplayer online game. *Social Networks, 32*(4), 313–329. doi:10.1016/j.socnet.2010.06.001.

Taniar, D. (2008). *Data mining and knowledge discovery technologies*. IGI Global.

Torkjazi, M., Rejaie, R., & Willinger, W. (2009). Hot today, gone tomorrow : on the migration of myspace users. *... of the 2nd ACM workshop on Online ...* (pp. 43–48). doi:10.1145/1592665.1592676.

Ugander, J., Karrer, B., Backstrom, L., Marlow, C., & Alto, P. (2011). The anatomy of the facebook social graph. *Arxiv preprint arXiv, abs/1111.4*(July) (pp. 1–17). doi:10.1.1.31.1768.

van Dam, J.-W., & van de Velden, M. (2014). Online profiling and clustering of Facebook users. *Decision Support Systems*, *70*, 60–72. http://www.sciencedirect.com/science/article/pii/S0167923614002796.

Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. P. (2009). On the evolution of user interaction in Facebook. *Proceedings of the 2nd ACM workshop on Online social networks - WOSN '09*, 37. doi:.10.1145/1592665.1592675.

Wasko, M. M., & Faraj, S. (2005). Why should i share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly, 29*(1), 35–57.

Webster, J., & Ahuja, J.S. (2006). Enhancing the Design of web navigation systems: the influence of user disorientation on engagement and performance. *Mis Quarterly*, 661–678.

Webster, J., & Ho, H. (1997). Audience engagement in multimedia presentations. *ACM SIGMIS Database, 28*(2), 63–77.

Xia, M., Huang, Y., Duan, W., & Whinston, A. B. (2012). Research note-to continue sharing or not to continue sharing? An empirical analysis of user decision in peer-to-peer sharing networks. *Information Systems Research, 23*(1), 247–259.

Zhong, C., Salehi, M., Shah, S., Cobzarenco, M., Sastry, N., & Cha, M. (2014). Social bootstrapping : how pinterest and last. fm Social Communities Benefit by Borrowing Links from Facebook. *Www2014* (pp. 305–314). doi:10.1145/2566486.2568031.

Nripesh Trivedi received an Integrated Master's Degree in Mathematics and Computing from the Indian Institute of Technology (Banaras Hindu University), Varanasi in October, 2016. His past research experience include working as a Research Associate and Research Assistant at the LSIR Lab at EPFL in 2014 and 2016, as a Research Assistant at Kno.e.sis (at Wright State University) in 2015 and as a Software Engineering Intern at Polotsk State University in 2013. He is also a summer alumnus of the Indian Academy of Sciences. His research interests are in data mining, machine learning and data management systems.

Daniel Adomako Asamoah, Ph.D. is an Assistant Professor in Management Information Systems in the Raj Soin School of Business at the Wright State University. He received his Ph.D. in Management Information Systems from the Spears School of Business at the Oklahoma State University. His research focuses on business analytics and intelligence, big data applications, decision support systems in health care and electronic social media. His research has been published in multiple journals, including Communications of the Association of Information Systems, Decision Support Systems, Journal of Computing and Information Technology and Health Care Management Science. Major conferences such as the annual meetings of the Americas Conference on Information Systems (AMCIS), the Decision Sciences Institute (DSI), the Institute for Operations Research and the Management Sciences Healthcare (INFORMS Healthcare) and the Pre-ICIS Business Analytics Congress have accepted his papers. He has won multiple awards including the prestigious Phoenix Doctoral Award and the Philips Dissertation Fellowship Award. As a principal investigator, his research has been funded by the Wright State University Research Council. He is also an editorial board member of the International Journal of Experimental Algorithms (IJEA).

Derek Doran, Ph.D. is an Assistant Professor in the Department of Computer Science & Engineering at Wright State University. He directs the Web and Complex Systems Laboratory, part of the Kno.e.sis Research Center. His research interests are in network and relational data analysis as well as web, social and geospatial informatics. Dr. Doran has authored more than 50 publications and is an inventor on multiple patents. He received his Ph.D. in Computer Science & Engineering from the University of Connecticut.