



## Identification and characterisation of Facebook user profiles considering interaction aspects

Pedro Henrique B. Ruas, Alan D. Machado, Michelle C. Silva, Magali R. G. Meireles, Ana Maria P. Cardoso, Luis E. Zárate & Cristiane N. Nobre

To cite this article: Pedro Henrique B. Ruas, Alan D. Machado, Michelle C. Silva, Magali R. G. Meireles, Ana Maria P. Cardoso, Luis E. Zárate & Cristiane N. Nobre (2019): Identification and characterisation of Facebook user profiles considering interaction aspects, Behaviour & Information Technology, DOI: [10.1080/0144929X.2019.1566498](https://doi.org/10.1080/0144929X.2019.1566498)

To link to this article: <https://doi.org/10.1080/0144929X.2019.1566498>



Published online: 22 Jan 2019.



Submit your article to this journal [↗](#)




Article views: 49



View Crossmark data [↗](#)



# Identification and characterisation of Facebook user profiles considering interaction aspects

Pedro Henrique B. Ruas <sup>a</sup>, Alan D. Machado<sup>a</sup>, Michelle C. Silva<sup>a</sup>, Magali R. G. Meireles<sup>a</sup>, Ana Maria P. Cardoso<sup>b</sup>, Luis E. Zárate<sup>a</sup> and Cristiane N. Nobre<sup>a</sup>

<sup>a</sup>Institute of Exact Sciences and Informatics, Pontifical Catholic University of Minas Gerais, Belo Horizonte, Brazil; <sup>b</sup>Postgraduate Program in Information Systems and Knowledge Management, FUMEC University, Belo Horizonte, Brazil

## ABSTRACT

The great number of social network users and the expansion of this kind of tool in the last years demand the storage of a great volume of information regarding user behaviour. In this article, we utilise interaction records from Facebook users and metrics from complex networks study, to identify different user behaviours using clustering techniques. We found three different user profiles regarding interactions performed in the social network: *viewer*, *participant* and *content producer*. Moreover, the groups we found were characterised by the C4.5 decision-tree algorithm. The 'viewer' mainly observes what happens in the network. The 'participant' interacts more often with the content, getting a higher value of closeness centrality. Therefore, users with a participant profile are responsible, for example, for the faster transmission of information in the virtual environment, a crucial function for the Facebook social network. We noted too that 'content producer' users had a greater quantity of publications in their pages, leading to a superior degree of input interactions than the other two profiles. Finally, we also verify that the profiles are not mutually exclusive, that is, the user of a profile can at determined moment perform the behaviour of another profile.

## ARTICLE HISTORY

Received 21 May 2018  
Accepted 7 December 2018

## KEYWORDS

Online social network;  
decision tree; data mining;  
user profile; Facebook;  
behaviour modelling

## 1. Introduction

In its origin, the Web was not only used to display information or connect pages, it was also created to be a distance communication tool. In fact, since the last decades of the twentieth century, human interactions have been increasingly impacted by the presence of technological tools that allow people to communicate independently of their physical proximity.

Online social networks (OSNs) have become pervasive, supplying the need for a communication system open to all audiences with massive membership across the globe. Such popularity has made the understanding of interactions in social networks an object of interest in many areas of knowledge. Much has been researched and published with the objective of understanding patterns of behaviour and psycho-social characteristics of users of social networks, as well as establishing relations of propagation and information flow in digital environments.

One of the theoretical approaches adopted for this purpose is the Social Network Analysis (SNA), which has been gaining interest in the academic community (Su and Chan 2017), especially in the Social and Behavioural Sciences. A great part of this interest is associated

to the identification and modelling of relationships between involved entities and, mainly, to discovering behaviour profiles in those environments (Wasserman and Faust 1994). From the SNA point of view, an environment can be characterised through patterns and regularities in the relations between interacting entities. From a social network perspective, its analysis depends critically of its structure (Tripathy, Sishodia, and Jain 2014).

Since the structure of a social network is composed by a finite set of social entities, and a relation or set of relations of those entities, therefore it is naturally possible to model a social network through graphs (Tripathy, Sishodia, and Jain 2014). The relations established by the users network characterise the network in many ways.

According to Peng et al. (2018), an SN can be treated as a complex network composed of individuals in society and their relationships among individuals. This type of complex social structure plays an important role in the dissemination and diffusion of information. Easley and Kleinberg (2010) also state that several characteristics of a network can be extracted using complex networks metrics, to better understand the behaviour of users in the network by observing their interactions and relative

positioning in the graph. Through its structure, it is possible to identify groups formed by a set of vertices that are more connected amongst themselves than to other vertices. Sets of vertices relatively more cohesive, or densely connected, form regions, also called clusters, which can reflect the existence of groups in the social network (Hansen, Shneiderman, and Smith 2010).

OSNs, classified by Peng et al. (2018) as a type of topology of network, are virtual communities that allow people to connect and interact with each other, corresponding to a social structure of individuals who share a common space of interests, needs and similar goals, to communicate, collaborate and share information (Chen 2014).

Amongst the current OSNs, Facebook is considered the most accessed worldwide.<sup>1</sup> According to data published about Facebook, in November 2016, the tool reached a daily average of 1.19 billion active users around the world, and 1 billion people access the social network through a mobile device monthly (Facebook 2016). In addition, in August 2016, the average number of items shared daily by users on Facebook was 4.75 billion (Fu, Wu, and Cho 2017).

In Ruas, Nobre, and Cardoso (2014), the authors propose the existence of three interaction profiles for Facebook users. The authors applied a questionnaire in which the respondents could answer to which of the three profiles they identified with the most and observed that 48% declared *participants*, whose typical behaviour is to interact, comment and share the content published by their contacts in the network. 34% said they felt they were *content producers*; this is a user that mainly produces new content in the social network. 18% said that they were *viewers*, the user whose predominant behaviour is to see what happens in the social network. These profiles were conceptually defined through the analysis of an online questionnaire, that was available for 35 days and got 296 answers from Facebook users. Each respondent user reported to how they saw themselves in the social network: someone who is predominantly a spectator of what happens, someone who interacts actively in the network, or a user that produces content for the social network.

This work considers the results presented by Ruas, Nobre, and Cardoso (2014) and proposes quantitatively to characterise the identified Facebook user groups: *viewer*, *participant* and *content producer*. In other words, opposed to what was evaluated by Ruas, Nobre, and Cardoso (2014), in which the authors applied questionnaires in which users declared themselves to be one of a certain interaction profile, this work aims to analyse if these user profiles do exist on Facebook based on a collection of interactions collected directly from this social

network. That is, the approach proposed in this work seeks to evaluate the users' interactions directly from their interactions in the network, through an automatic process of data collection of the network. This avoids the bias of the research that may have been given in the questionnaire, since users self-reported their profiles, in addition to increasing the representativeness of the sample.

For this, we used grouping techniques, which belong to unsupervised learning, and which describe hidden patterns in non-labeled data, since a priori we do not have the classification of users regarding the interaction profile. The attributes considered will be the interactions collected from the Facebook, represented by the commenter, liker, Post author and user tagged attributes, performed and received by the users. We also used complex network metrics obtained from these interactions. In total, 65,707 unique records were collected on Facebook and sixteen attributes, eight related to network interactions and eight resulting from complex network metrics.

Besides, from the groups generated by the grouping techniques, this work aims to characterise each of these groups, inferring a function of the now labeled training data. For this, it is intended to identify the classification rules, that is, the attributes that really characterise a user in a particular interaction profile. It was also verified if these groups are mutually exclusive, or if there is a possibility of users behaving only one or more groups. However, our goal is to find rules that allow us to identify the behaviour of an ideal type of user, according to Max Weber's methodological proposal.

According to Weber's theory (Weber 1981), in order to analyse a given situation or social relation, especially when it comes to generalisations, it is necessary to create an 'Ideal Type', which is an instrument that will guide an investigation, such as a species of parameter. The ideal type refers to a mental construction of reality, in which the researcher selects a certain number of characteristics of the object under study, to construct a 'tangible whole', that is, a type.

The construction of the ideal type fulfills two basic functions: (1) providing a limiting case with which social phenomena can be contrasted; generating an unambiguous concept that will facilitate the classification and comparison of social situations; (2) assisting as a scheme for type generalisations. Starting from the concept of ideal type, it is possible to analyse several real facts as deviations from the idealised or constructed.

An ideal type creates a selective model of social organisation that can then be explored analytically and empirically by specialists. The ideal type is not intended to be a general representation of a phenomena category; but

rather a heuristic model that allows the exploration and extension of some of the characteristics of the institutions and concrete social behaviours that it partially represents.

The groups were identified and evaluated using the following clustering methods: two based in prototypes (*K*-Means and Self-Organising Maps – SOM), and one density-based (DBSCAN). The reason to utilise algorithms based on distinct approaches (prototypes and density) is that we do not know how the data are distributed. The DBSCAN algorithm can work with clusters of arbitrary sizes and shapes, because it utilises a density-based group definition. However, when the groups present different densities, the DBSCAN performance may be hindered, which does not happen for *K*-means and SOM (Tan, Steinbach, and Kumar 2006). As a quality measure for the clusters, we considered the silhouette index metric (Rousseeuw 1987).

As the main contribution of this work we describe the profiles of the main actors of the social network (*viewer*, *participant* and *content producer*) through interaction data as *Liker*, *Commmenter* and *Post author* actions and complex network metrics that allow characterising the distance of each profile in relation to the centrality of the network.

This article is organised as follows: Section 2 presents related work. In Section 3, we present the methodology adopted for the development of the research. Section 4 presents the experiments realised with the *K*-Means, SOM and DBSCAN clustering algorithms, to group the data collected from Facebook, and the characterisation of the profiles found in the social network. Finally, Section 5 presents the conclusions, final considerations and limitations of our work.

## 2. Related work

There are many researches carried out in the context of matching the user profiles. Some of the findings are worth being mentioned here below:

In Ruas, Nobre, and Cardoso (2014), the authors identified the profiles of users interacting on the social network Facebook. There were three categories of users: *viewer* (user who predominantly observes what is happening in the social network, without necessarily interacting with the content), *participant* (user interacting by enjoying, commenting and/or sharing content in the social network) and, finally, the *content producer* (user whose main activity is to publish content on the network itself). The authors noted that the majority of users, 48%, said they were *participants*, 34% are *viewers* and 18% are *content producers*. These data were obtained from a questionnaire answered by 296 respondents,

made available on Facebook itself. In this work, Ruas, Nobre, and Cardoso (2014) wanted to know the predominant profile of users, placing the options of choice as mutually exclusive. That is, users had to choose, among the three options available, which best describes the behaviour of themselves in the social network.

Maia, Almeida, and Almeida (2008) presented an analysis of the Youtube social network regarding production, publication and visualisation of content in the network. To achieve that, 1,467,003 user profiles were collected, with the following attributes: number of uploads, number of views, number of channel accessed, data of registration on the network, age, clustering coefficient, reciprocity, input and output degrees in the graph representing the network. For the clustering process, they utilised *K*-Means with the BetaCV<sup>2</sup> as a metric for evaluating cluster quality. The authors identified five groups in the OSN: small community member (users that form highly interconnected small communities in the network), content producer (older users in the network who produce videos, representing 23% of the collected data), content consumer (typical users who consume content in the network, meaning they watch more than they produce), producer & consumer (a group that has both characteristics, representing 48% of the collected users), and lastly, the others cluster (a group with a low value to every attribute considered).

O'Donovan et al. (2013) performed an analysis of the Facebook social network, as well as of the propagation of information in the network. To achieve that, they collected data from 1327 Facebook users through a web service, with the authorisation of the participants. In total, 489,929 posts, 244,809 comments, 289,313 photos and 4903 videos published in the social networks have been collected. After this process, the data was clustered using *K*-Means with a goal to identify groups of similar behaviours in the network. They utilised 12 attributes in the clustering process: average of publications per day (posts, comments and videos published), number of friends, typical hour for publishing, ratio of public and private posts in the network, relative frequency of status updates, topic diversity in text posts, number of text posts, number of video posts, number of photos or album posts, average length of captions in videos, photos and albums, and diversity of topics in video, photos and album publications (using captions and description). The authors identified five groups in the social network: multimedia and engaged users (a group of users with a high number of posts and multimedia publications, with a majority of female users), users with low engagement (users that are not too involved with Facebook and have few publications), private broadcasters (a small group of users that had a high amount of private

publications), self-engaged users (the group with the most users, that acts similarly to the first group, however have a greater interest in text posts than on multimedia posts), and finally, multimedia expert (a small group that posts predominantly multimedia content, also with a female majority).

Benevenuto et al. (2010) performed an analysis of the UOL (in Portuguese, 'Universo Online') video service, and presented a characterisation of its users' sessions, server requests, and browsing profiles. One of the results of the study was the discovery of different profiles of users that access the system. According to the authors, that information could be used by the system administrators to personalise services. The authors observed users changing their profiles from spectators to content producers in the network. They also proposed a method to categorise users according to their browsing profile: one of these profiles, named Viewers, predominantly visualises videos, while other example of profile, Searchers, are users who usually begin their sessions by looking for specific content.

In Vasconcelos et al. (2012), the authors analysed how Foursquare<sup>3</sup> users explore three resources available in the platform: *Tips*, hints (evaluations or recommendations) posted by other users, reflecting their positive or negative impressions on a given place; *Dones*, markings done by users agreeing with tips, as a feedback; and *ToDos*, lists where users add locations for visiting. Through the analysis of these features, users were clustered into four distinct behaviour profiles on this social network. Two of these profiles correspond to regular users, differentiating in terms of their engagement, such as frequency of posting tips. Another group are influential users, which post several tips on the network (some of those user accounts are run by companies or brands). The last user profile are spammers, characterised by posts not related to the location of the tips, typically with links in their posts.

In this way, our proposal approaches the work presented by Ruas, Nobre, and Cardoso (2014), in which the authors found three profiles of Facebook users regarding aspects of interaction such as likes, comments and shares made by the users in the social network. However, in order to achieve the objectives, the authors applied questionnaires with a very small sample size (N=296), in addition to the self-named users belonging to one of the three groups presented by the authors, *participants*, *content producers* and *viewer*.

Thus, the proposal presented in this paper investigates the existence of these three groups using an automatic data collection of the interactions between users in 6 years (between 2009 and 2014) of Facebook use, totalling 65,707 unique records of Facebook. The problem was

modelled as a graph, where the vertices are the users and the edges are the interactions between them.

In addition, from the groupings performed, we also perform a characterisation, using classification rules, of the profiles found, besides verifying that these groups may not be necessarily exclusive.

Unlike the other aforementioned studies, which focus on the user's profile regarding published content, our proposal aims to analyse the users' profile in relation to the type of interaction performed with the different actors of this social network.

Therefore, our results can be used to direct policies of service customisation, as well as to provide content recommendations for users of a given profile. In addition, they can help guiding the development of tools and functions for SNSs which cover the particularities of each profile, increasing user engagement and frequency use of the SNS.

### 3. Materials and methods

Our interest in this research was to understand the users profile of Facebook. So the research was guided by the following questions:

Q1: The existence of Facebook interaction profiles: *viewer*, *participant* and *content producer*, found by Ruas, Nobre, and Cardoso (2014), is confirmed when we adopt an automatic interaction collection methodology?

Q2: Which rules characterise each of the identified profiles: viewer, participant, or content producer?

Q3: Are these profiles exclusive?

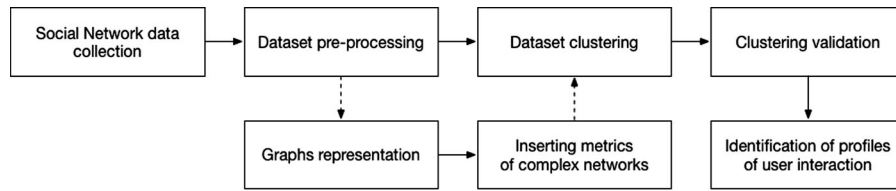
To answer the questions, our work was developed according to the steps presented in Figure 1. The following sections detail each of these steps.

#### 3.1. Social network data collection and description

For the identification of the different Facebook user profiles, regarding their interaction behaviour, we performed an automated collection of Facebook data based on the method called snowball. The data were collected through the Graph API tool, currently disconnected, provided by Facebook.<sup>4</sup> With this API, it is possible to collect information referring to the profile of Facebook users (such as name, age, gender, nationality, relationship status, etc.), to their public posts (respecting the Facebook use policy), and to the interactions performed in those posts.

From this API, you can also collect data concerning friends (in up to two levels of distance) and interactions with those users. In the first step, the search was initiated





**Figure 1.** Methodology adopted.

through a user seed (unavoidably) and, from that user, the collection was performed storing all the interactions that the user seed and his friends in the social network made (since published as public). In the second step, the search was expanded to 'friends of friends', that is, all the (public) interactions of the second level of the network were also collected regarding the user seed. Thus, all the public interactions carried out in the delimited period of time in the acquisition process in up to two levels of distance of the user used as seed were selected and stored for the identification of online interaction profiles.

It should be noted that the data collected on the social network are derived from interactions (likes, comments, tags, etc.) in publications made during the established six-year period (from January 2009 to December 2014). The data collection was started in 2009 because it is the year the social network started to popularise in Brazil (country of the seed user). The records collected were modelled in a directed graph, being the vertices of this graph the users of the network, and the edges their interactions through social network content.

To automatise the data collection, the NodeXL API<sup>5</sup> has been used to store the social network data. This API collects the information provided by the Graph API, converting and storing those in an electronic spreadsheet.

Table 1 presents the amount of unique users obtained (vertices) and their interactions (edges). Figure 2 presents the amount of interactions (likes, comments, user tags and posts) collected in the social network during the six-year period.

It is noticeable in Figure 2 that, with the exception of the 'Liker' attribute, there is a decrease in the number of records throughout time. The amount of comments made in the social network reached its peak value in 2012 and, from that year forward, the records dropped a total 6.17% when comparing 2012–2014. In contrast,

for the 'Liker' attribute, which is the most utilised function in Facebook, the records increased considerably yearly. Between 2011 and 2012, there has been an increase of 284.86% in the number of likes for example.

To obtain the clusterings, we utilised only data from 2011 to 2014, because as shown in Figure 2, data from 2009 and 2010 describe the formation of the social network in Brazil. Both first years of data collection present lower values for the interaction attributes, hence the discrepant values for the complex network metrics when compared to the other years. The exclusion of this period resulted in a reduction of 2.6% (3.096) from the total of unique users in the dataset.

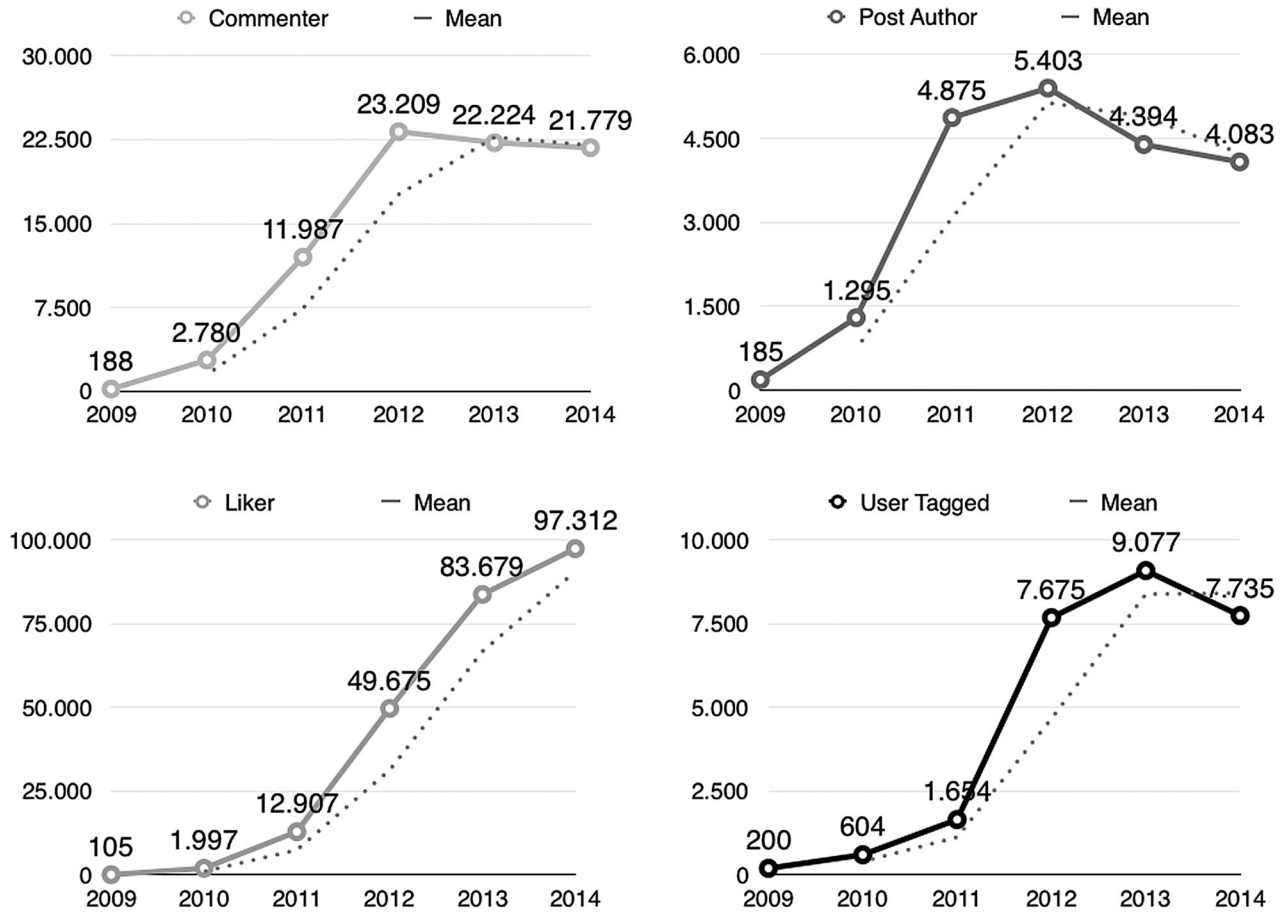
### 3.2. Dataset pre-processing

After the data collection process, it was necessary to perform a pre-processing of the dataset. One of the objectives of our research is to characterise the way Facebook users interact, aiming to find behavioural patterns regarding their interactions. Therefore, the attribute 'friendship' (a variable that indicates whether users  $X$  and  $Y$  are friends in the network) has been disregarded, since this type of relation does not characterise an interaction in the social network. Simply put, the fact that two users are friends in the network does not imply that there is going to be any interaction between them.

Another necessary task in this step was the junction of repeated interactions. Facebook data collected through Graph API correspond to interactions of liking, commenting, authoring posts and tagging users, which happened in a given period of time. That means that a user  $X$  can present several interactions with user  $Y$ , which results in a repetition of the same record of interaction between them, regarding different contents. Hence, we merged every record that had repeated  $X \rightarrow Y$  interactions, assigning a counter for each type of interaction, to convey how many times those had happened.

**Table 1.** Facebook data collected.

	2009	2010	2011	2012	2013	2014	Accumulated
Records	146.060	161.984	195.686	280.374	337.561	378.523	1.500.190
Vertices	411	2.685	12.871	37.506	58.763	66.948	119.235
Edges	678	6.676	31.423	85.962	111.374	130.909	288.752



**Figure 2.** Interactions collected between 2009 and 2014.

The automated Facebook data collection generated a dataset of 65,707 registers (unique users) and 8 attributes referring to social interactions on Facebook. A description of each attribute, their value range, averages and standard deviations (SD) are presented in Table 2.

From the modelling of the problem through graphs, eight complex networks metrics (Easley and Kleinberg 2010) were calculated<sup>6</sup> and inserted into the dataset (Table 3). These metrics were added to reveal, for example, which users were responsible for linking different unconnected components in the graph, which users had a higher importance in the social network, what was the most centred node (user) in the social network in a determined time, amongst others. Therefore, the inserted metrics proportionate new information on the users and their interactions in the social network.

The resulting dataset has eight characteristics relating to interactions performed on Facebook, automatically collected by the Graph API, and eight complex networks metrics extracted from the graph using Gephi. The attribute vector of the dataset is composed by the variables described in Equations (1) and (2). Table 3 presents a

description of the complex networks metrics considered.

$$\text{Interactions} = \{\text{Commenter In}, \text{Commenter Out}, \text{Liker In}, \text{Liker Out}, \text{Post Author In}, \text{Post Author Out}, \text{User Tagged In}, \text{User Tagged Out}\} \quad (1)$$

$$\text{Metrics} = \{\text{Degree}, \text{Hub}, \text{Authority}, \text{Clustering Coefficient}, \text{Local Clustering Coefficient}, \text{Closeness Centrality}, \text{Betweenness Centrality}, \text{PageRank}\} \quad (2)$$

In order to find patterns of interactions in the collected data, we consider that some information related to the network topology would add information relevant to the characterisation of these online behaviour patterns. For this reason, we use the complex network metrics to highlight the characteristics of the users in relation to the structure of which they are part.

However, it is possible to find in the literature that some metrics such as clustering coefficient, closeness, betweenness, PageRank are better suited to analyse the complete network structure (Akhtar 2014). However,

**Table 2.** Interaction data collected on Facebook.

Attribute	Description	Range	Average (SD)
<i>Commenter In</i>	Number of comments received	0 to 63 comments	4.16 (9.27)
<i>Commenter Out</i>	Number of comments given	0 to 244 comments	1.07 (5.68)
<i>Liker In</i>	Number of likes received	0 to 10907 likes	6.01 (107.25)
<i>Liker Out</i>	Number of likes given	0 to 10906 likes	6.16 (115.80)
<i>Post Author In</i>	Number of posts in their own page	0 to 75 posts	5.90 (15.85)
<i>Post Author Out</i>	Number of posts in other user pages	0 to 157 posts	0.42 (3.45)
<i>User Tagged In</i>	Number of tags received	0 to 19 tags	0.38 (1.02)
<i>User Tagged Out</i>	Number of tags given	0 to 74 tags	0.26 (1.54)

this was not possible due to the limited data collection of the social network, as previously described. In this way, the intention was to use complex network metrics to characterise network interaction patterns, and not to use the metrics to characterise the topology.

**Table 3.** Complex network metrics considered.

Metric	Definition
<i>Degree</i>	The degree of a vertex $v$ is the number of incident edges of that vertex.
<i>Hub</i>	Defines the value of the links coming out of a vertex $v$ towards those it is connected to. The more components the vertex $v$ connects, the greater its hub value will be in the network.
<i>Authority</i>	The authority value is calculated by adding every hub value of the vertices connected to a vertex $v$ . This can be interpreted as a node having greater authority in the network as it connects directly to nodes with a high hub value. This value is dependent on which nodes a node is connected to.
<i>Clustering coefficient</i>	The clustering coefficient of a vertex $v$ is defined by the probability of two random friends of $v$ being friends amongst themselves, that being, the fraction of pairs of friends of vertex $v$ that are connected through an edge.
<i>Local Clustering Coefficient</i>	The local clustering coefficient of the vertex $v$ is the fraction of pairs of neighbouring vertexes of $v$ that are connected to every pair of neighbours of $v$ .
<i>Closeness Centrality</i>	In connected graphs, there is a natural distance metric between every pair of nodes, which is defined by the length of their shortest paths. The separation of a node $s$ is defined as the sum of its shortest path to every other node, and its proximity is defined as the inverse of the separation. Therefore, as a node gets more central in the network, its total distance to every other nodes decreases.
<i>Betweenness Centrality</i>	Is an indicator of the centrality of a vertex $v$ in the network. It is equal to the number of shortest paths, from every node in the network to every other node, that cross the node. A vertex with a high betweenness centrality value has a great influence on the transport of items through the network, considering that information transference, for example, goes through shortest paths.
<i>Page Rank</i>	In general, it is a value, calculated by an algorithm, to measure the 'importance' of each node on the network, based on the structure of the network connections.

Thus, Lee, Kim, and Jeong (2006) makes a comparison between three different sampling approaches (called node sampling, link sampling and snowball sampling) and analyzes the obtained values to the Degree distribution and average path length, Betweenness centrality, Assortativity e Clustering coefficient metrics. Considering the sampling approaches used in our methodology (snowball sampling), the authors note that the lower the collected sample with the snowball method, the lower the Degree distribution and average path length, Betweenness centrality, Assortativity obtained values will be, with the possibility of having the found values to vary for the Clustering coefficient (in comparison to the real value considering the whole network).

However, Costenbader and Valente (2003), after analysing the correlation between the complex network metrics based on a sample and the results obtained considering the complete network, states that the results indicate a relatively high correlation, although, in some cases, substantial differences were detected between the real properties of the network and those calculated in randomly selected sub-samples for some network metrics. This indicates that, in some circumstances, researchers may still use network data for which some data is missing to study network properties. In other words, researchers who do not interview all members of a community or the complete collection of users of a network can still take advantage of some aspects of complex network theory and techniques

In order to verify the real relevance of these complex networks metrics, we applied a Pearson correlation test amongst every attribute in the dataset (Equations (1) and (2)), after normalising their values between 0 and 1. Every value in the correlation matrix, where  $\rho < 0.5$ , meaning we could not find a significant correlation between the variables.

### 3.3. Cluster identification and validation

In this stage of the methodology, the dataset was clustered in order to verify the existence of behaviour patterns regarding interactions in the social network. Our objective is to verify if the methodology proposed in this work in obtaining information directly from the social network, without necessarily asking the user about his profile, confirms the existence of the profiles identified by Ruas, Nobre, and Cardoso (2014). That is, our problem is characterised by being unsupervised, since we do not have the classification of the collected users. To achieve that, three distinct clustering algorithms were considered:  $K$ -means, SOM and DBSCAN, and the input attributes were the data automatically collected from Facebook and complex networks metrics.



The  $K$ -means algorithm is one of the most frequently utilised for clustering problems (MacQueen 1967). The ‘ $K$ ’ refers to the fact that the algorithm considers a priori a fixed number of clusters to find. The algorithm utilises an iterative greedy approach to find clusters that minimise the sum of distances of each register to its cluster centroid. The algorithm initialises the cluster centroids by randomly generating  $K$  points in the data space, usually by creating an uniformly random value inside the valid interval for each dimension considered. Each  $K$ -means iteration consists basically of two steps: (1) assigning an register to a cluster and (2) updating the centroid (Zaki and Meira 2014).

The neural network Self-Organising Map (SOM) is a clustering and data visualisation technique based on a neural grid of competitive learning. On this grid, the neurons compete amongst themselves to be activated, resulting in only a single neuron being active at a time. It works similarly to the strategies based on centroids which best represent the registers in a set (Tan, Steinbach, and Kumar 2006). In the context of the SOM neural network, one neuron is assigned for each cluster centroid. The output structure of SOM can be an one-dimensional array, a two-dimensional matrix, as well as more complex structures in 3D (Kantardzic 2011). Utilising a distinct approach from  $K$ -means, the SOM algorithm imposes a topographic ordination on its centroids, so that centroids closer in the grid are more closely related amongst themselves than to the farther centroids (Tan, Steinbach, and Kumar 2006). On the tests we executed, we utilised 0.2 as the neighbourhood weight value, for training and refining of the neurons in the network.

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm belongs to a category of density-based clustering techniques, and it was proposed by Ester and Kriegel (1996). The algorithm requires the definition of three parameters: a set of points ( $D$ ), the maximum allowed distance to incorporate new registers ( $\varepsilon$ ), and the minimum number of points allowed in the neighbourhood of a point ( $MinPts$ ). The algorithm selects a point in the dataset as the seed and starts expanding the cluster according to the connectivity of the points. The major limitation of DBSCAN is its sensibility to its parameters  $\varepsilon$  and  $MinPts$  (Zaki and Meira 2014, pp. 377). There is no simple or automated way to obtain the best values for these parameters. However, a heuristic presented by Ester and Kriegel (1996) obtains a good estimate. This heuristic aims to determine the best values for  $\varepsilon$  and  $MinPts$  to the most sparse cluster in the dataset, so that the neighbourhood  $\varepsilon$  of the points in the remaining clusters will have a number of points larger than  $MinPts$ .

Regarding the DBSCAN parameters used in our work, for the estimation of  $\varepsilon$  and  $MinPts$ , we employed the aforementioned heuristic, which utilises the distance to the closest neighbour of each point. To estimate the best  $MinPts$  value, we considered four possible choices: 2500, 5000, 6000 e 7000. As the data distribution in the  $n$ -dimensional space was unknown, these values were chosen to verify the possible cluster densities, where: 2500 represents a more sparse cluster, and 7000 a denser one. To estimate the  $\varepsilon$  value, we used the  $K$ -distances function (*KNNDistancesSampler*), implemented on the ELKI<sup>7</sup> data mining tool, which calculates the distances of each point to its closest neighbour (Schubert et al. 2015).

The  $\varepsilon$  and  $MinPts$  values used in our experiments belong to the  $[0.01 - 0.52]$  and  $[2.500 - 7.000]$  intervals, respectively.

We performed experiments (using the ELKI data mining tool and SPSS) for each pair of  $MinPts$  and  $\varepsilon$  values, with normalised data indexed by the tool. The values were normalised in the  $[0,1]$  interval, and the indexation was made through a  $R^*$  tree structure (Beckmann et al. 1990), which is recommended in Ester and Kriegel (1996). The noisy data found were treated according to the approach utilised in Moulavi et al. (2014). Therefore, all the experiments ignored noise when calculating evaluation metrics, for those represent inconsistent data regarding cluster evaluation.

In our work, we utilised silhouette index (Rousseeuw 1987) as a metric to evaluate cluster quality. The silhouette index evaluates the cluster through the similarity between points inside a cluster, its cohesion, and also through the separation between clusters. For each point  $p$ , we calculate the distance averages for every point in its own cluster ( $a$ ) and the distance average for points in the nearest neighbour cluster ( $b$ ). Thus, ( $a$ ) represents the cohesion of point, and ( $b$ ) representsis is separation. The silhouette index value of a point  $p$  is calculated according to Equation (3):

$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}. \quad (3)$$

The overall sillhouette index is defined as the average value  $\bar{S}_i$  between every point in the dataset, given by Equation (4):

$$Silhouette\ Index = \frac{1}{n} \sum_{i=e}^n S_i. \quad (4)$$

The values  $\bar{S}_i$  are defined in the  $[-1,1]$  interval. A value close to 1 indicates that a point (register in the dataset)  $x_i$  is closest to the points in its own cluster and farthest from the neighbouring clusters. A value close to 0

means  $x_i$  is close to the borders of two clusters. Finally, a value close to  $-1$  indicates that  $x_i$  is close to the points that belong to a different cluster (Zaki and Meira 2014).

### 3.4. Characterising and validating the user interaction profiles

After the clustering and validation step, we performed a characterisation of the interaction profiles of the Facebook social network users. The goal was to define particular features of each cluster, in order to understand their profiles. The interaction profile were identified through the application of the C4.5 decision-tree classification algorithm (Quinlan 1993), from which rules of implication were extracted to describe these profiles. To generate the decision tree, we utilised the absolute values of the attributes, and the J48 classifier. This classifier is an implementation of C4.5 available on the Weka data mining tool (Hall et al. 2009). The algorithm was executed with the default parameters and a 10-fold cross validation (Kohavi 1995).

Three performance metrics were used to evaluate the classification results: precision (PR), recall (RE) and F-Measure (FM).

Precision measures the ratio of correct classifications amongst the total classifications, for one label (Equation (5)):

$$PR = \frac{TP}{TP + FP} \quad (5)$$

Recall measures the ratio of correct classifications considering one label (Equation (6)):

$$RE = \frac{TP}{TP + FN} \quad (6)$$

The F-measure consists in a harmonic mean between precision and recall (Equation (7)):

$$FM = \frac{2(PR)(RE)}{(PR) + (RE)} \quad (7)$$

where TP is the number of True Positives, TN is the number of True Negatives, FP is the number of False Positives, and FN corresponds to the number of False Negatives.

## 4. Experiments and analysis of the results

To identify and describe the interaction profiles of Facebook users, in this section we describe the results of the clustering process of the dataset described in Section 3.2, considering the three clustering techniques: K-Means, SOM and DBSCAN.

Initially, the number of groups  $K$  (needed for K-means) and the number of neurons for the SOM neural network were altered according to the quality of the clusterings found, using the silhouette index metric. The goal of this process was to find the number of clusters that best divided the dataset, meaning that every cluster should have a high average silhouette index value, which would indicate cohesive and well separated clusters. According to (Rousseeuw 1987), an iterative process to find the ideal  $K$  number of clusters can be used, by increasing the number of clusters and observing their average silhouette index value. The author states that when the average overall silhouette index for  $z$  clusters is smaller than the obtained with  $z-1$  clusters, that means  $z-1$  is the adequate number of clusters for that dataset.

Figure 3 presents the clustering results for K-means and SOM, for 2, 3 and 4 clusters, observing the average silhouette index value.

The highest average silhouette index value was reached by the SOM algorithm. Moreover, the best cluster number was exactly three clusters, as expected in Ruas, Nobre, and Cardoso (2014), with an average silhouette value of 0.89 for SOM, and 0.54 for K-means. It is apparent that K-means had a negative value for cluster 3, indicating a cluster that should not exist (registers in that cluster are more similar to register in the neighbour cluster than to those in their assigned cluster).

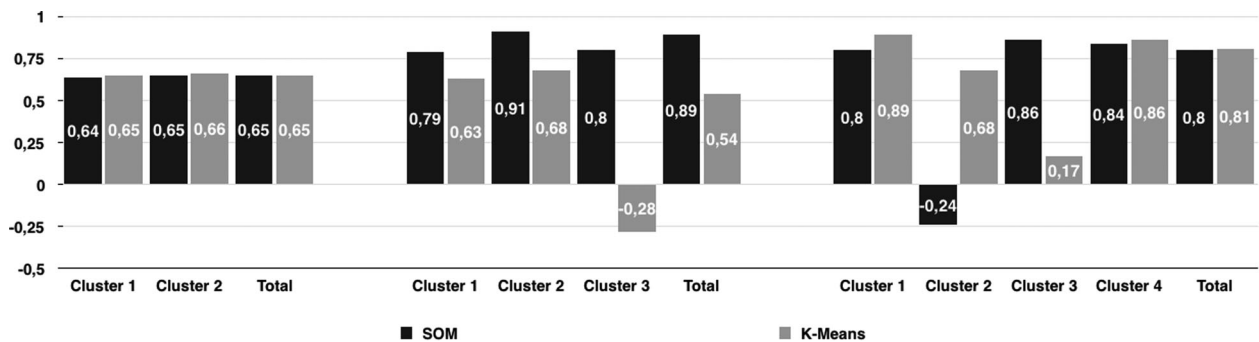


Figure 3. Results of silhouette index obtained for K-Means and SOM clusterings.

**Table 4.** Evaluation of the clusters obtained by DBSCAN.

MinPts	$\varepsilon$	Clusters	Silhouette index	Noise
2500	0.01	2	0.2197	60.73%
2500	0.05	4	0.4811	18.16%
2500	0.18	2	0.7926	2.76%
2500	0.26	2	0.7925	1.83%
5000	0.04	2	0.3801	34.17%
<b>5000</b>	<b>0.13</b>	<b>3</b>	<b>0.8396</b>	<b>5.03%</b>
5000	0.19	2	0.7928	2.70%
5000	0.26	2	0.7925	1.85%
5000	0.46	1	N/A	0.38%
6000	0.08	1	N/A	27.07%
6000	0.13	2	0.7558	14.41%
6000	0.19	2	0.7892	3.16%
6000	0.25	2	0.7924	2.04%
7000	0.04	1	N/A	44.41%
7000	0.08	1	N/A	27.08%
7000	0.15	2	0.7560	14.28%
7000	0.20	2	0.7553	12.32%
7000	0.52	1	N/A	0.35%

Notes: There is no silhouette value for  $K=1$ , as the metric analyses the distance between different clusters.

Contrastingly, all clusters found by SOM had positive values greater or equal to 0.79, which indicates a strong clustering structure for the artificial neural network results.

The DBSCAN experiments showed clusterings with a satisfactory silhouette index value. The experiment results, along with the parameters for each test, are presented in Table 4.

It is noticeable that most tests clustered the data in two groups. However, the best average silhouette index result (0.8396) was found on a three clusters execution, with a small amount of noise data (5.03%).

Compared to silhouette index results for the prototype-based algorithms, DBSCAN has improved the results as compared to  $K$ -means, with a best value of 0.84 against 0.54. However, it remained below the best SOM clustering results, that got a 0.89 average silhouette index value.

It is observable in Table 5 that the clusters found by DBSCAN are rather similar to those found by SOM. All the users assigned to Clusters 1 and 3 by DBSCAN had the same assignment by SOM. Regarding Cluster 2, there were divergencies for only 178 (140 + 38) users, as 140 were assigned to Cluster 2 by DBSCAN and to Cluster 1 by SOM, and 38 that were assigned to

Cluster 2 by DBSCAN were assigned to Cluster 3 by SOM. All grouped users in Cluster 1 and Cluster 3 by the algorithm DBSCAN received the same grouping by SOM. As for Cluster 2, there were divergencies in the grouping of only 178 (140 + 38) users, given that 140 were in Cluster 2 by DBSCAN and were grouped in Cluster 1 by SOM and 38 which were in Cluster 2 according to DBSCAN were grouped in Cluster 3 by SOM. The other differences are due to the existence of noise, resulting from a limitation of the DBSCAN algorithm.

#### 4.1. User profile characterisation

We determined a confidence interval, with a 5% level of significance, to present the clustered data. Table 6 contains the intervals of values defined for the attributes of our dataset.

Regarding the interaction characteristics (the first eight indicated in Table 6), it is apparent that users assigned to Cluster 3 receive the most comments (*Commenter In*), likes (*Liker In*), and have the most posts in their own pages *Post Author In*, indicating characteristics of the Content Producer role in the network, a profile that holds creation of content as its main characteristic. It is our expectation that Facebook users which have a connection to a Content Producer interact with their content through the actions made available in the social network.

Clusters 1 and 2 had similar interaction value intervals, having a greater variation only for the *Liker In* and *Commenter Out* interactions. Therefore, Cluster 1, considering only the records of interactions performed, was considered representative of the *Participant* profile in the network. This profile is characterised mainly by its participative behaviour in the social network, meaning they mostly interact with content presented to them.

On the other hand, Cluster 2 was deemed as the *Viewer* in the social network. This profile is characterised by visualising what happens in the social network, without necessarily interacting with the content. As shown in Table 6, Cluster 2 presents low interaction values, with the exception of the 'Like' interactions (*Liker In* and *Liker Out*), which is considered the main Facebook feature.

Regarding the complex networks characteristics, it is noticeable that cluster 1 (Participants in the network) got the highest *Authority* and *Hub* values. This occurred due to the amount of interactions this Cluster has in the social network (as shown by the *Degree* attribute). The *Hub* value is defined according to the links a vertex has, and the greater number of components this vertex interacts with, the greater its value. The *Authority* attribute value is exclusively dependent on the *Hub*, since it is

**Table 5.** Comparison of the clusters obtained by SOM and DBSCAN.

		SOM			Total
		Cluster 1	Cluster 2	Cluster 3	
DBSCAN	Cluster 1	7650	0	0	7650
	Cluster 2	140	48550	38	48728
	Cluster 3	0	0	6021	6021
	Noise	511	1430	1367	3308
	Total	8301	49980	7426	65707

**Table 6.** Confidence intervals, grouped by cluster, for the attributes considered.

Variable	Cluster 1		Cluster 2		Cluster 3	
	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Commenter In	1.16	1.24	0.63	1.28	28.88	29.19
Commenter Out	2.07	2.44	0.48	1.13	1.29	1.70
Liker In	2.29	2.43	4.51	6.77	12.45	12.62
Liker Out	5.11	8.14	5.24	7.60	3.15	4.68
Post Author In	0.31	0.36	-0.02	0.63	49.63	49.93
Post Author Out	0.85	1.10	-0.04	0.61	0.54	0.79
User Tag In	0.30	0.33	0.02	0.67	0.64	0.70
User Tag Out	0.54	0.65	-0.12	0.53	0.23	0.31
Degree	6.72	9.03	2.16	2.92	4.19	7.79
Hub	64.94	88.576	26.92	30.73	33.97	51.28
Authority	91.86	125.12	34.05	40.80	37.56	58.44
Clustering Coef.	0.30	5.09	0.79	2.51	-0.75	3.64
Local Clustering Coef.	0.0914	0.0986	-0.2837	0.3685	0.0414	0.0473
Closeness Centrality*	4.20e+16	4.23e+16	4.35e+15	4.38e+15	7.18e+15	8.32e+15
Betweenness Centrality	-4.14e+18	9.03e+19	-8.47e+17	1.37e+19	-1.47e+19	1.85e+20
PageRank	4.97e+10	6.12e+10	2.23e+10	2.51e+10	2.33e+10	3.13e+10

Notes: The value for Closeness Centrality given by Gephi is inverse closeness centrality. Therefore, in the case of inverse closeness centrality the higher the value, the closer to the centre. The inverse centrality is more efficient for the calculation of closeness centrality (Freeman 1978).

the sum of the *Hub* values of every vertices a vertex is connected to. Therefore, these attributes have higher values for Cluster 1 as compared to the other two clusters.

The fact that Cluster 1 has a higher degree value in the graph (which indicates that people assigned to that cluster interact more often in the social network) results in the cluster also having a high value for *Clustering Coefficient* (the probability of two random neighbours of a vertex interacting with each other), *Local Clustering Coefficient* (coefficient that means how much the neighbours of a vertex are connected to other pairs of neighbours), *Closeness Centrality* (length of the shortest paths that go through a vertex), *Betweenness Centrality* (indicator of centrality of a vertex in the network, determined by the number of shortest paths that go through it), and *PageRank* (a type of 'fluid' that circulates in the network, determining how important a vertex is by how much of that 'fluid' accumulates on it).

Therefore, users with a Participant role in the social network are highly important for the network structure. Those users are responsible, for example, for speeding up the transmission of information on the virtual environment, a pivotal function for the Facebook social network. This role was previously expected to fall onto the *Content Producer* profile.

The behaviour analysis were performed considering the values of our attributes as shown in Table 6. However, to better characterise the clusterings we obtained, we submitted the dataset to a classification process through a decision tree (Quinlan 1986), as described in Section 3.4.

To achieve that, the dataset was labelled with the profiles as assigned by SOM and DBSCAN, because those got the best silhouette index results. In this new dataset, the attribute 'Interaction Profile' was added to

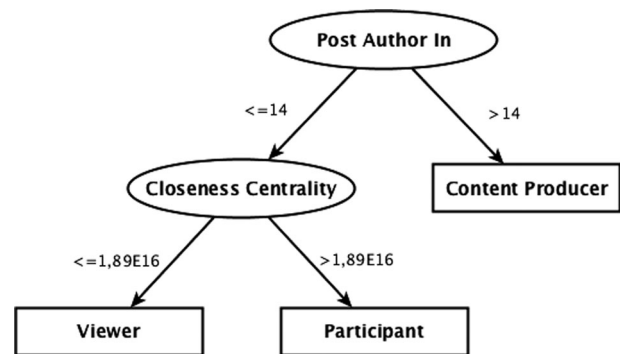
the registers, considering only the coinciding assignments for both SOM and DBSCAN. This means that the 178 users which were assigned to distinct clusters, and those deemed as noise by DBSCAN (Table 5) were not considered. Therefore, the behaviour characterisation of each cluster was performed considering the 7650<sup>8</sup> users of cluster 1 (*Participants*), 48550 users of cluster 2 (*Viewers*), and 6021 users of cluster 3 (*Content Producers*). All the attributes referring to interactions in the social network and complex networks metrics were used in the construction of the decision tree.

The decision tree generated after the group clustering had only 2 of the 16 attributes in the dataset: *Post Author In* and *Closeness Centrality* (Figure 4).

The rules were filtered based on how many registers they labelled correctly. After this filter, we obtained three rules, from which we were able to label 98.8% of all our dataset.

The first rule refers to the *Viewer* profile:

*Rule 1:* If *Post Author In*  $\leq 14$  and *Closeness Centrality*  $\leq 1.89e+16 \rightarrow$  *Viewer*



**Figure 4.** Decision tree obtained through SOM and DBSCAN clusterings.



According to this rule, we consider Viewers the users that publish at the most 14 posts in their pages, and have a closeness centrality value inferior or equal to  $1,89E+16$ . The viewers predominantly observe what happens in the social network, and for that reason the attribute values that characterise this profile are lower, when compared to the other profiles.

The second rule labels the *Participant* profile:

*Rule 2: If Post Author In  $\leq 14$  and Closeness Centrality  $> 1.89e+16 \rightarrow Participant$*

According to this rule, we consider participants the users that publish at the most 14 posts in their pages, but have a closeness centrality value greater to  $1,89e+16$ . This rule agrees with our definition for that profile, because participants have a typical behaviour of interacting with the content presented in the social network. The fact that these users interact more often in the social network results in a higher value for their closeness centrality (compared to viewers), indicating that these users find themselves in more central positions in the network. Therefore, we concluded that the interactions performed by *Participants* speed up the dissemination of information on Facebook.

The final rule refers to the *Content Producers* profile:

*Rule 3: If Post Author In  $> 14 \rightarrow Content Producer$*

The third rule states that *Content Producers* are users that published more than 14 posts in their personal pages in the social network. This profile is characterised by creating content in the social network, so it was expected that they had a greater amount of posts if compared to the other interaction profiles, and that the users connected to them interacted with that content.

The generated decision tree had a precision rate and recall of 100%, 99.99% and 99.83% and of 99.94%, 99.98% and 99.92% for the *Participant*, *Viewer* and *Content Producer* profiles, respectively. It is noticeable that, for every cluster, the model had satisfactory classifying results for the metrics utilised, indicating that the rate of false negatives and false positives is very low for the generated model.

Thus, by observing the results obtained by the clusters and the decision tree generated after the categorisation of the groups found, it is concluded that Facebook has actually three user profiles regarding aspects of interaction. In this way, the proposed methodology contributes to validate these results, since an automatic data collection is used; without, therefore, questioning the user to which group they thought to belong.

This work has also found that only two attributes are sufficient to characterise the three interaction profiles: the attribute *Post Author In*, number of posts which the user posts in their own page, and Closeness Centrality. In a connected graph, the more central a node is, the

closer it is to all other nodes. In this way, this work shows that when the user has high value for this attribute it is because they are influential in the network. Thus, for the generated tree, we observed that the profile user *Participant* is the most influent.

These results are very important because it signals that the most important users of the network are the ones who interact more, participate with comments, share and cause the information in the network to propagate.

In addition, it signals a need for actions and measures that would make users who have few interactions motivated to interact more, thus increasing network success.

#### 4.2. Analysis of user behaviour between different profiles

The last investigation of our work was to verify if these user profiles are exclusive. That is, if the participating user can also be a viewer, for example.

According to Zaiontz (2014), in the existence of more dependent variables it is also possible to execute multiple ANOVAs (Analysis of Variation). However, this practice can increase the chance of errors where null hypotheses that are true are rejected.

For this reason and considering that we have a sample composed of more than two groups, with numerical measurement scale, and with more than one variation factor, the MANOVA (Multivariate Analysis of Variation) technique was used. The factors of variation, analysed together, were: *commenter in*, *commenter out*, *liker in*, *liker out*, *post author in*, *post author out*, *user tagged in* and *user tagged out*. Based on the assumption of three behavioural groups, previously identified, the sample was divided according to Table 7.

From then on, the applied technique aimed to answer the following question:

*Do all behaviours (variation factors) affect together when the subgroup of people analysed is changed?*

According to the data characteristics, it was obtained an  $\chi^2_{critical} = 26.296$  and an  $\chi^2_{calculated} = 59,902.529$  was obtained. Thus, for  $\chi^2_{critical} > \chi^2_{calculated}$ , with a significance of 5%, it was observed that all behaviours are not affected jointly by the different subgroups. This implies that, for example, a user with the *Content Producer* profile can, at some point, perform a user-specific behaviour with

**Table 7.** Distribution of Facebook users, divided by profile.

Profile	Amount
Viewer	50,204
Participant	8,154
Producer	7,349
Total	65,707



the *Viewer* profile, and so on, for each of the identified online behavioural profiles.

## 5. Conclusion

Currently, one of the main interests in social network analysis is to find and characterise patterns and regularities in the relations between interacting entities on an online social environment. In this article, we confirmed the presence of three interaction profiles on the Facebook, proposed by Ruas, Nobre, and Cardoso (2014) (*viewer*, *participant* and *content producer*), through the analysis of automatically collected data referring to over a million records from the social network, concerning the interactions of 65,707 Facebook users.

In academic researches focused on studying OSN user behaviour, it is unusual to find published articles relating the use of the structure of the network and its characteristics to analyse user behaviour. In our research, the aim was to utilise these informations to improve clustering quality, adding different information than those already existing in the dataset.

Furthermore, the complex networks metrics we considered helped us interpret the clusters found, since they brought new and relevant informations referring to the position of a register in the graph and to the interactions of users (vertices) in the social network.

Regarding the clustering methods, the SOM algorithm proved itself more effective in clustering our dataset, once it 'learns' with the data, that being, the neurons in the neural network specialise in recognising certain behaviour patterns, searching for a cohesive clustering where registers clustered together are similar between themselves and dissimilar to those belonging to different clusters (which resulted in a silhouette index close to 1). This factor was evident when comparing the clustering results of SOM, *K*-Means and DBSCAN.

The DBSCAN had a silhouette index value 55% greater than the one we got from *K*-means, and a slightly lower, 6%, comparing to SOM. Besides that, most of the registers were grouped in the same profile as the neural network results. Only 178 (0.27% of the dataset) registers corresponded to a different profile in the SOM results, not considering noise.

We demonstrated the existence and presented the characterisation of the three social network user interaction profiles proposed in Ruas, Nobre, and Cardoso (2014): *viewer*, *participant* and *content producer*, which represent respectively 76.06%, 12.63% and 11.31% of the users in our automatically collected dataset. Regarding users with a participant profile, the fact that these users interact more often in the social network results

in a higher value of closeness centrality, as compared to viewers, indicating that these users speed up the dissemination of content on Facebook. The content producers had a noticeably greater number of publications in their pages, causing a considerably higher degree of input interactions than the other profiles. Therefore, based on the interactions available in the tool, other users recognise content producers as so, and react accordingly.

The confirmation and characterisation of the user profile allow identifying the main actors of a social network, in this case, Facebook. One of the longings in creating a social network for information sharing is to increase its life cycle, and this condition depends on its activity.

We can consider that the life of a network is mainly associated with content producers and the return of participants in the expectation that the viewer can become one of the other previous actors.

Knowing the pattern of each profile can aid in predicting the social network life cycle. This would be done by timing the number of users with the content producer profile. In the same way, the addition and decrease of the activities of participants and viewer (number of tanned ones) that would show the evolution/involution of the social network activity can be observed.

From the knowledge of the pattern that characterises each profile of a social network, it becomes possible to monitor each of its members and propose strategies that could consolidate their actions within the network or even modify them. This induces the proposal of adaptive persuasive technologies more focused on an individual group or member of the social network.

Regarding the limitations of this article, it is worthwhile to note that, despite our efforts to collect interactions from as much social network users as possible, the data collection tool imposes a restriction of only using a single seed user. This naturally restricts how many users we can analyse, even though this is a limitation inherent to data collection through the Graph API tool. However, when applying the methodology proposed in this article using a different seed user, we can affirm with 95% of confidence that the values found for the interaction profiles characterisation will fall between the intervals we presented, granted that the study is done with a similarly distributed dataset.

As future works, we suggest:

- 1 Verifying whether the same three profiles can be identified in other online social networks;
- 2 Verifying whether, when making a user from a different nationality the seed, the cultural factor will have an impact in the number of interactions for each profile.

## Notes

1. Data from *Alexa.com*. Accessed on April 7, 2017.
2. The BetaCV is the ratio between the average intra-cluster distance and the average inter-cluster distance. The smaller the BetaCV value, the better the cluster, since it indicates that the intra-cluster distances are smaller than the inter-cluster distances (Zaki and Meira 2014).
3. A location-based online social network based which combines SNSs with geographic information sharing. Available at: <https://pt.foursquare.com/>.
4. Available in: <https://developers.facebook.com/docs/graph-api>
5. Available in: <https://nodexl.codeplex.com>
6. All the complex networks metrics were extracted using the Gephi tool, available in <https://gephi.org/>
7. Available in <https://elki-project.github.io/>
8. See Table 5.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Pedro Henrique B. Ruas  <http://orcid.org/0000-0001-6423-8681>

## References

- Akhtar, Nadeem. 2014. "Social Network Analysis Tools." 04.
- Beckmann, Norbert, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. 1990. The R\*-tree: An Efficient and Robust Access Method for Points and Rectangles. In *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data (SIGMOD '90)*. ACM, New York, NY, USA, 322–331. doi:10.1145/93597.98741
- Benevenuto, Fabrício, Adriano Pereira, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. 2010. "Characterization and Analysis of User Profiles in Online Video Sharing Systems." *Journal of Information and Data Management* 1 (2): 261.
- Chen, Yi-Fen. 2014. "See You on Facebook: Exploring Influences on Facebook Continuous Usage." *Behaviour & Information Technology* 33 (11): 1208–1218. doi:10.1080/0144929X.2013.826737.
- Costenbader, Elizabeth, and Thomas W. Valente. 2003. "The Stability of Centrality Measures when Networks are Sampled." *Social Networks* 25 (4): 283–307. <http://www.sciencedirect.com/science/article/pii/S0378873303000121>.
- Easley, David, and Jon Kleinberg. 2010. *Networks, Crowds and Markets: Reasoning About a Highly Connected World*. Cambridge, PA: Cambridge University Press.
- Ester, Martin, and Hans-Peter Kriegel. 1996. "A Density-Based Algorithm for Discovering Clusters a Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, 226–231. AAAI Press.
- Facebook. 2016. "Facebook News Room." Novembro. Accessed March 10, 2016. <http://br.newsroom.fb.com/company-info/>.
- Freeman, Linton C. 1978. "Centrality in Social Networks Conceptual Clarification." *Social Networks* 1 (3): 215–239.
- Fu, Pei-Wen, Chi-Cheng Wu, and Yung-Jan Cho. 2017. "What Makes Users Share Content on Facebook? Compatibility Among Psychological Incentive, Social Capital Focus, and Content Type." *Computers in Human Behavior* 67: 23–32.
- Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. "The WEKA Data Mining Software: An Update." *ACM SIGKDD Explorations Newsletter* 11 (1): 10–18.
- Hansen, Derek, Ben Shneiderman, and Marc A. Smith. 2010. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. 1. San Francisco, U.S.: Morgan Kaufmann. doi:10.1016/C2009-0-64028-9.
- Kantardzic, Mehmed. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms*. New York: John Wiley & Sons, Inc.
- Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2 (12): 1137–1143. San Mateo, CA: Morgan Kaufmann.
- Lee, Sang Hoon, Pan-Jun Kim, and Hawoong Jeong. 2006. "Statistical Properties of Sampled Networks." *Physical Review E* 73: 016102. <https://link.aps.org/doi/10.1103/PhysRevE.73.016102>.
- MacQueen, J. B. 1967. "Some Methods for Classification and Analysis of Multivariate Observations." *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. 1 (14): 281–297. Berkeley: University of California Press.
- Maia, Marcelo, Jussara Almeida, and Virgílio Almeida. 2008. "Identifying User Behavior in Online Social Networks." *Proceedings of the 1st Workshop on Social Network Systems – SocialNets '08*, 1–6.
- Moulavi, Davoud, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. 2014. "Density-Based Clustering Validation." In *Proceedings of the 14th SIAM International Conference on Data Mining*, 839–847. SIAM.
- O'Donovan, Francis T., Connie Fournelle, Steve Gaffigan, Oliver Brdiczka, Jianqiang Shen, Juan Liu, and Kendra E. Moore. 2013. "Characterizing User Behavior and Information Propagation on a Social Multimedia Network." *Electronic Proceedings of the 2013 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2013*.
- Peng, Sancheng, Yongmei Zhou, Lihong Cao, Shui Yu, Jianwei Niu, and Weijia Jia. 2018. "Influence Analysis in Social Networks: A Survey." *Journal of Network and Computer Applications* 106: 17–32. <http://www.sciencedirect.com/science/article/pii/S1084804518300195>.
- Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1 (1): 81–106.
- Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann.
- Rousseeuw, Peter J. 1987. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis."

- Journal of Computational and Applied Mathematics* 20: 53–65.
- Ruas, Pedro Henrique Batista, Cristiane Neri Nobre, and Ana Maria Pereira Cardoso. 2014. “A Influência das estratégias persuasivas no comportamento dos usuários no Facebook.” In *Proceedings of the 13th Brazilian Symposium on Human Factors in Computing Systems, IHC '14, Porto Alegre, Brazil*, 255–264. Sociedade Brasileira de Computação.
- Schubert, Erich, Alexander Koos, Tobias Emrich, Andreas Züfle, Klaus Arthur Schmid, and Arthur Zimek. 2015. “A Framework for Clustering Uncertain Data.” *Proceedings of the VLDB Endowment* 8 (12): 1976–1979.
- Su, Chris Chao, and Ngai Keung Chan. 2017. “Predicting Social Capital on Facebook: The Implications of Use Intensity, Perceived Content Desirability, and Facebook-enabled Communication Practices.” *Computers in Human Behavior* 72: 259–268.
- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. New York: Pearson Education.
- Tripathy, B. K., M. S. Sishodia, and Sumeet Jain. 2014. “Societal Networks: The Networks of Dynamics of Interpersonal Associations.” In *Social Networking*, 101–127. Vol. 65, Springer, Cham. New York, NY. doi:10.1007/978-3-319-05164-2\_5
- Vasconcelos, Marisa Affonso, Saulo Ricci, Jussara Almeida, Fabrício Benevenuto, and Virgílio Almeida. 2012. “Tips, Dones and Todos: Uncovering User Profiles in Foursquare.” In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, New York, NY, USA*, 653–662. ACM. doi:10.1145/2124295.2124372.
- Wasserman, S., and K. Faust. 1994. “Social Network Analysis: Methods and Applications.” *Structural Analysis in the Social Sciences*. Cambridge University Press, New York, NY. doi:10.1017/CBO9780511815478
- Weber, Max. 1981. “Ensaio de Sociologia.” Ed. Guanabara.
- Zaki, Mohammed J., and Wagner Meira Jr. 2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York, NY: Cambridge University Press.