A quantitative methodology to identify relevant users in social networks

Claudia Canali, Sara Casolari, Riccardo Lancellotti
University of Modena and Reggio Emilia
Department of Information Engineering
{claudia.canali, sara.casolari, riccardo.lancellotti}@unimore.it

***Abstract*—Social networks are gaining an increasing popularity on the Internet, with tens of millions of registered users and an amount of exchanged contents accounting for a large fraction of the Internet traffic. Due to this popularity, social networks are becoming a critical media for business and marketing, as testified by viral advertisement campaigns based on such networks. To exploit the potential of social networks, it is necessary to classify the users in order to identify the most relevant ones. For example, in the context of marketing on social networks, it is necessary to identify which users should be involved in an advertisement campaign. However, the complexity of social networks, where each user is described by a large number of attributes, transforms the problem of identifying relevant users in a needle in a haystack problem. Starting from a set of user attributes that may be redundant or do not provide significant information for our analysis, we need to extract a limited number of meaningful characteristics that can be used to identify relevant users.**

We propose a quantitative methodology based on Principal Component Analysis (PCA) to analyze attributes and extract characteristics of social network users from the initial attribute set. The proposed methodology can be applied to identify relevant users in social network for different types of analysis. As an application, we present two case studies that show how the proposed methodology can be used to identify relevant users for marketing on the popular YouTube network. Specifically, we identify which users may play a key role in the content dissemination and how users may be affected by different dissemination strategies.

I. Introduction

The increasing importance of social networks in the context of the Internet is testified by their growth in terms of users and content shared on these networks. To understand the potential role of social networks, we may consider that two thirds of the world's Internet population visit a social network site weekly [18], and that Facebook alone has more than 500 millions of active users spanning throughout the world, according to the official site [9]. In a similar way, the amount of messages and content shared among social network users is growing at unprecedented rates. For example, YouTube reached an upload rate beyond 24 hours of video every minute [25] and is currently the largest video sharing site on the Internet, accounting for approximately 60% of the videos watched online [10]. Similarly, Facebook recently become the largest network repository for images [22].

From a business point of view, the popularity of social networks represents a great opportunity to reach a large worldwide audience for marketing purposes, as it is testified by recent viral marketing campaigns and by brand-related social network users groups.

The data available in a social network represent a huge set with tens of millions of users, each described by tens of attributes [13], [17]. This amount of data results into an information overload that does not provide useful information for the identification of the most relevant users for specific analysis. For example, it is unfeasible to identify which users can be exploited for marketing in a social network starting from the entire set of user attributes. A partial solution to the problem is to discard some attributes that are intuitively non-relevant with respect to the specific analysis. However, even after this preliminary selection, we face the following issues:

- some attributes may provide redundant or limited information;
- multiple attributes need to be combined in order to identify relevant users for a specific analysis.

The main contribution of this paper is the proposal of a quantitative methodology that can support social network analysis for identifying relevant users. A qualifying point of our proposal is the use of Principal Component Analysis (PCA) to select and combine user attributes into characteristics that are meaningful for the analysis. To the best of our knowledge, this is the first study that copes with the problem of reducing the complexity of the data available for social network analysis using PCA.

As an example, we apply the proposed methodology to the social component of the popular YouTube site, considering a marketing-oriented analysis. We demonstrate that this methodology allows us to identify relevant users that are more likely to successfully disseminate popular contents and users that are likely to be targets of the content dissemination. Furthermore, we show that target users can be classified in two categories depending on their activities, thus providing an insight on how these users can react to different dissemination strategies.

The remaining of this paper is organized as follows. Section II presents the proposed methodology, describing the principles of PCA and how they are applied to the context of social networks. Section III and IV present two case studies where the proposed methodology is applied to identify relevant users for a marketing analysis on the YouTube network. Finally, Section V describes the related work, and Section VI provides some concluding remarks.

II. METHODOLOGY

In this section we describe the proposed quantitative methodology to identify relevant social network users starting from the set of user attributes.

We should consider that popular social networks may have tens of millions of participating users, each of them characterized by tens of attributes of different nature. The typical user information that may be collected from a social network may be divided into the following categories:

- **Social links**: user social relations including number of incoming links, outgoing links, bidirectional links.
- **Content accesses**: user accesses to the contents on the social network.
- **Uploaded contents**: information on contents uploaded by the user, including number of uploads and received visualizations.
- **Activities**: user activities including number of comments, ratings, marks as favorite.
- **Personal data**: information on the user including age, nationality, language, job.
- **Personal preferences**: user preferences including tags and interests.

The meaning of the above attributes is quite intuitive and needs no additional discussion, with the exception of those referring to the user social links. We should consider that the social networks are represented as directed graphs, where the users are the nodes and the social links are the edges. When a user invites in a social relation other users, he generates outgoing links; on the other hand, the invitations received by other users represent the incoming links. Finally, a social relation that has to be accepted by the invited user to be established determines a bidirectional link.

The large set of attributes that describe a user in a social network leads to a data overload from which it is unfeasible to extract useful information to classify relevant users. Indeed, the set of attributes may contain data that do not add any useful contribution for the user classification. Moreover, we may have attributes showing common patterns; this means that these attributes give basically the same informative contribution about users and are, therefore, redundant.

To address these issues, we propose a methodology consisting of the following steps, as outlined in Figure 1:

1) Qualitative selection of attributes of interest to discard attributes that are intuitively non-relevant for the analysis.
2) Aggregation of attributes with similar informative contribution to remove redundancies.
3) Definition of user characteristics to combine into metrics the aggregated attributes.
4) Identification of the relevant users.

Let us now describe in details the steps of the proposed methodology. The first step aims to reduce the number of user attributes through a qualitative selection. The basic idea is that some attributes may be immediately discarded because their are likely to be useless for the given analysis. For example, in an analysis aiming to identify users who can distribute content with high visibility we are not interested in the user physical height. However, this qualitative selection is not sufficient in order to identify which users are relevant for a specific analysis for a twofold reason. First, we may still have redundant or low-contributing attributes in the remaining set; second, we have to cope with multiple attributes that we need to combine into user *characteristics* to classify the social network users.

The second step of the methodology aims to aggregate the correlated attributes in order to remove data redundancies. To this aim, we exploit PCA [1] that is a statistical data analysis technique able to transform the original space of possibly correlated data into a new space of uncorrelated variables. Other approaches for coping with redundant data exist, such as the Independent Component Analysis (ICA) [11], however we choose to apply PCA to the social network analysis because it is considered as the baseline technique for this kind of data analysis and it can provide another view of the original data capable to show the main characteristics of social network users.

Let p be the number of attributes of each user, n the number of users in the social network and X the $n \times p$ measurement matrix, that is the input of PCA. If considering X, each column i denotes the i-th user attribute and each row j represents a user. We refer to individual columns of a matrix as X_i and the matrix X can be written as $(X_1, X_2, \ldots, X_p)$. The PCA maps the original data onto a new set of axes, that are called *dimensions*. Each dimension is a linear combinations of the original attributes and has the property that it points in the direction of maximum *variation* or *energy* (with respect to the Euclidean norm) remaining in the data.

Hence, PCA transforms the p attributes sets $(X_1, X_2, \ldots, X_p)$ into p dimensions sets $(Z_1, Z_2, \ldots, Z_p)$. Each dimension Z_i is equal to:

$$Z_i = \sum_{j=1}^{p} a_{ij} X_j \tag{1}$$

where a_{ij} is the correlation value between the j-th variable and the i-th dimension [1]. A coefficient a_{ij} that is close to $+1$ or -1 impacts on the dimension Z_i, while a coefficient a_{ij} that is close to 0, implies no impact. The PCA is very effective in the case of data with statistical properties that greatly differ among the different variables, such as the social network user attributes. The PCA is able to identify the attributes having the greatest variance values, that will be dominant in the characterization of the entire data set and of the single dimensions, while the other attributes will be less significant. Moreover, PCA identifies redundancy in case of data with a low intrinsic dimensionality, as it happens with social network user attributes that are likely to follow common patterns. The transformation operated by PCA has two main properties:

- Z_1 contains the most information and Z_p the least, that means $Var[Z_1] \geq Var[Z_2] \geq \cdots \geq Var[Zp]$;
- there is no information overlap between the dimensions, that means $Cov[Z_i, Z_j] = 0$, $\forall i \neq j$.

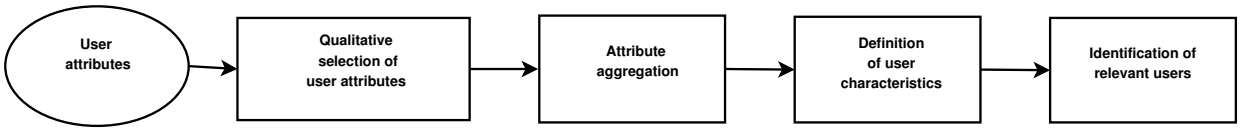

Fig. 1. Methodology steps

As a consequence of the first property, we can retain only the first q dimensions, discarding the dimensions with the lowest variability, while conserving a large amount of information on the original data set [2]. To choose the number q of retained dimensions, there are several methods [1] including the visual method, that can be based on the scree plot of the percentage of variation of each dimension or on the scree plot of the eigenvalue of each dimensions, and the Kaiser criterion, that takes into account the eigenvalue related to the dimensions. For every retained dimension, we are able to select the attributes providing a higher contribution of information, that are characterized by the higher values of the correlation a. We aggregate together the user attributes that show the higher correlation indexes in the same dimension. In this way, we have reduced the dimensionality of the data from p possibly correlated user attributes to q uncorrelated dimensions, with $q << p$.

We now have a set of attributes, X_i^*, for each of the dimension Z_i, where $i \in [1, q]$. The attributes belonging to each dimension Z_i identify a user *characteristic*, Z_i^*, that may be used for the relevant user classification. To this aim, we define a metric for each characteristic that combines the previously identified attributes as following:

$$Z_i^* = \sum_{j \in X_i^*} a_{ij} X_j \tag{2}$$

where a_{ij} is the correlation value obtained through PCA that corresponds to the weight used for the attribute combination.

The last step of the methodology is the user classification according to the user characteristics. For each characteristic Z_i^* we rank the users accordingly. We can thus select the users that are relevant for the specific analysis as the top fraction of the ranked users list. Furthermore, we can analyze the relationship among different user characteristics using scatter plots. This study allows us to verify if the sets of relevant users selected according to different characteristics are overlapped or if they are disjoint, providing additional insight on the relevant user population.

III. Case study: Identification of source and target users in YouTube

We now apply the proposed methodology to the user attributes available in the YouTube network with the goal to identify and classify relevant users for marketing-oriented content dissemination.

The choice of YouTube is motivated by the large number of users, close to 50 millions (according to Google search data) and by the growing amount of shared contents, that has reached a volume of 24 new hours of video that are uploaded every minute of the day [25]. The impressive growth of users and contents makes YouTube an interesting case study for marketing-oriented social networks analysis, as testified by the increasing trend to exploit this network to disseminate contents like film trailers, advertisements and viral marketing videos.

In the following of the section we describe the collected data set and the application of the methodology to identify the relevant users from a marketing point of view.

To collect the YouTube data we crawl the social network graph exploiting the public APIs provided by YouTube. This approach is commonly used in literature, and gives us access to large data sets [17].

The implemented crawler collects data about YouTube users and their social links following a Snowball approach [17]. Crawling starts from a list of randomly selected users. Then, for each step, the crawler explores the outgoing links of a not yet visited user to retrieve the list of his/her contacts, that are added to the list of users to be visited in the next crawling step. The crawling has been carried out in the second half of 2009 when we collected data on 1,708,414 users and more than 12,935,561 social links.

Table I reports the user attributes $(X_1, \ldots, X_{11})$ obtained from YouTube and a short explanation of each of them. Additional attributes are not collected by crawler as they are not considered of interest in the context of selecting relevant users for marketing. The choice of the considered attributes corresponds to the first step of the proposed methodology, that is the qualitative selection of the user attributes of interest.

The attributes in Table I represent two types of information: data about the user social links (subscribers, subscriptions and friends) and data about the user interaction with the shared content (views, downloads, uploads, favorites). It is worth to note that the subscribers to the user channel represent in YouTube the incoming links of the user; the subscriptions to channels of other users represent the outgoing links; the friends represent a bidirectional connection among two YouTube users. There are two versions of the subscribers and subscriptions attributes, called crawl and total. The crawl version refers to the numbers obtained by counting the user links retrieved during the crawling operation, while the total version refers to summary information available on the home page of each visited user. Basically the total attributes represent exact information offered by the YouTube site, while the list of links in the crawl versions may be truncated by the limitations of the YouTube APIs.

The data set obtained from crawling consists of eleven attributes for nearly two million users. Due to the high number of attributes it is not possible to use the data set directly to identify which users are relevant for marketing on a social

TABLE I
YOUTUBE USER ATTRIBUTES

Attribute		Description
X_1	Subscribers (crawl)	Number of users subscribed to the user channel (obtained by collecting and counting links in the social network)
X_2	Subscribers (total)	Number of users subscribed to the user channel (obtained from the user profile)
X_3	Subscriptions (crawl)	Number of subscriptions of the user (obtained by collecting and counting links in the social network)
X_4	Subscriptions (total)	Number of subscriptions of the user (obtained from the user profile)
X_5	Friends	Number of friends of the user in the social network
X_6	Views	Number of views to videos uploaded by the user
X_7	Downloads	Number of videos watched by the user
X_8	Uploads	Number of videos uploaded by the user
X_9	Favorites	Number of videos marked as favorite
X_{10}	Comments	Number of comments added by the user
X_{11}	Rates	Number of videos rated by the user

network. To this aim, we exploit the methodology based on PCA described in Section II. It is worth to note that, since we work on heterogeneous data, we had to normalize the data set to zero mean and unit variance, as suggested in [14], to provide a reliable aggregation of user attributes into dimensions. To aggregate the user attributes into a minimal number of dimensions, we use the visual method based on the scree plot considering the percentage of variance of each dimension. This method is widely used in different contexts [14], [1] because it provides reliable results and allows an easy and fast interpretation. We compute the percentages of variance of the dimensions and we sort them in decreasing order. This operation allows us to visualize the dimensions in order of significance and to ignore the dimensions with lower variability. Indeed, these ordered values often show a clear "elbow" that separates the most important dimensions (characterized by higher percentage of variance) from the least important ones (that have lower values). In Figure 2, we show the results of the scree plot of the percentage of variance related to dimensions ordered accordingly: the curve begins with a sharp decrease, and levels off after the third dimension. For this reason, we select the first three dimensions (Z_1, Z_2, Z_3) for describing the characteristics of users of the YouTube network. These dimensions accounted for 29%, 19% and 12% of the total variance, respectively, and explain 60% of the total variance of the data.

To associate the user attributes to the three dimensions previously identified, we exploit the correlation values shown in Table II. Higher correlation values correspond to stronger relationship between the attribute and the considered dimension; on the contrary, lower is the correlation value, less is the impact of the attribute on the dimension. For each attribute we mark with a bold font the dimension showing the highest correlation.

Let us now analyze the attributes belonging to the three dimensions and identify the corresponding characteristics:

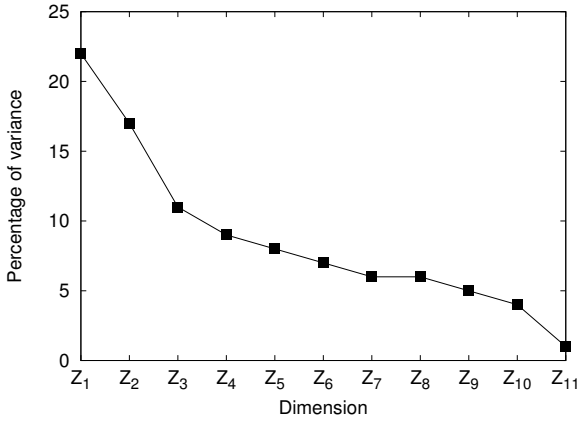

Fig. 2. Scree plot of the user attributes

TABLE II
CORRELATION VALUES OF USER ATTRIBUTES AND DIMENSIONS

User Attributes		Dimension		
		Z_1	Z_2	Z_3
X_1	**Subscribers (crawl)**	**0.498245**	0.083614	0.468016
X_2	**Subscribers (total)**	**0.877944**	-0.052234	-0.271755
X_3	**Views**	**0.857427**	-0.043848	-0.259509
X_4	**Downloads**	0.188238	**0.526520**	0.305601
X_5	**Friends**	**0.710871**	-0.029308	-0.166952
X_6	**Uploads**	0.407461	0.159767	**0.527366**
X_7	**Favorites**	-0.036045	**0.678520**	-0.142709
X_8	**Subscriptions (crawl)**	-0.139144	**0.606619**	-0.122505
X_9	**Subscriptions (total)**	0.015911	**0.423055**	0.139663
X_{10}	**Comments**	0.007330	**0.468037**	-0.186424
X_{11}	**Rate**	0.006757	**0.589447**	-0.210042

Z_1. This dimension includes the number of subscribers to the user channel, views to the videos uploaded by the user and friendship relations. These aggregated attributes identify the characteristic of being *popular* users of the social network.

Z_2. This dimension includes the video downloads, subscriptions to other user channels, markings of content as favorites, video ratings and comments added by the user. These attributes characterize the user as a *consumer* of contents.

Z_3. This dimension includes only the number of video uploaded by the user, hence it characterize of a user as a *producer* of contents within the social network.

We observe that Z_1 characterizes the popular users, who may play a critical role as sources of information to rapidly spread content within the social network. We observe that the aggregation of some attributes in this dimension is consistent with results in literature. In particular, the correlation between the views to the user uploaded content and the subscribers, which suggest that a significant fraction of users navigation in social networks follows the so-called social links, has been proved in [16], [15]. Furthermore, we observe that the attributes related to the number of subscribers (crawled and

total) are correctly assigned to the same dimension. On the other hand, Z_2 identifies users as consumers of contents, that means they are potential targets from a marketing point of view. Even in this case the two attributes related to the number of subscriptions (crawled and total) are aggregated into the same dimension. Finally, Z_3 identifies the users as producers of contents, but without any insight on the future visibility and on the marketing potential of the supplied content. For this reason, in the rest of the analysis we will not consider this third dimension, but we focus on Z_1 and Z_2 that allow us to identify preferential sources and targets for content dissemination.

In order to identify the relevant users for marketing on a social network, we now have to combine the attributes belonging to the first two dimensions into two metrics, Z_1^* and Z_2^*, that we use to measure the associated user characteristics.

We compute these metrics as follows:

$$Z_1^* = \sum_{j \in X_1^*} a_{ij} X_j \tag{3}$$

where $X_1^* = (X_1, X_2, X_3, X_5)$, that is the set of the attributes belonging to Z_1, and:

$$Z_2^* = \sum_{j \in X_2^*} a_{ij} X_j \tag{4}$$

where $X_2^* = (X_4, X_7, X_8, X_9, X_{10}, X_{11})$, that is the set of the attributes belonging to Z_2.

We sort the user list according to Z_1^* and Z_2^*. In Figure 3 and Figure 4, we show the cumulative distribution function of the metrics values Z_1^* and Z_2^*, respectively. For uniformity of presentation, the values of the metrics are normalized in the range $[0, 1]$.

To identify the users that represent the best sources for content distribution we consider the top fraction of users sorted according to Z_1^*. For example, if we look at the top 20% best users for content dissemination we can simply consider users that have a value of Z_1^* higher than the 80-th percentile of the cumulative distribution function, that is $Z_1^* > 0.6$. In a similar way, the users with $Z_2^* > 0.54$ (that is the 80-percentile of Z_2^* cumulative distribution) are the 20% of the social network population representing preferential targets for content dissemination.

In order to classify users on the basis of the selected characteristics, in Figure 5 we show the scatter plot of the users that represent potential sources and targets for marketing-oriented content dissemination in YouTube. We can outline four main areas in the graph. The two areas on the top of the figure represent users that are characterized by high values for Z_2^*. These users represent potentially good targets for content dissemination through YouTube. Both the areas on the right part of the graph are characterized by high values of Z_1^* and contain users that are highly popular. Such users are most suitable to disseminate contents. The area in the rightmost top part represents the users with high values for both Z_1^* and Z_2^*, that are at the same time popular and active content consumers, while users in the leftmost lower area are of no interest for our marketing analysis.

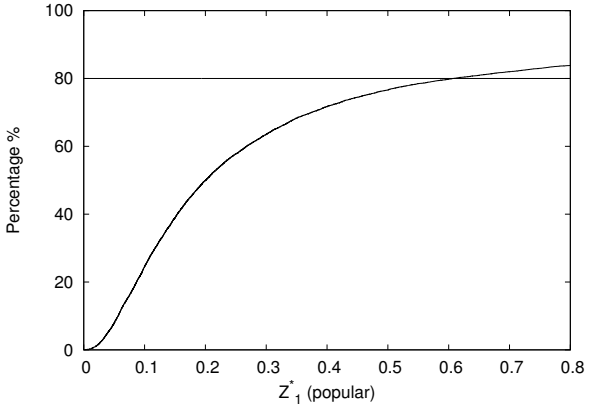

Fig. 3. Cumulative distribution of Z_1^*

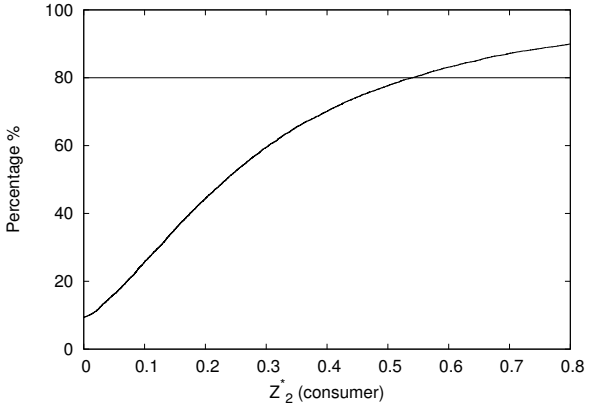

Fig. 4. Cumulative distribution of Z_2^*

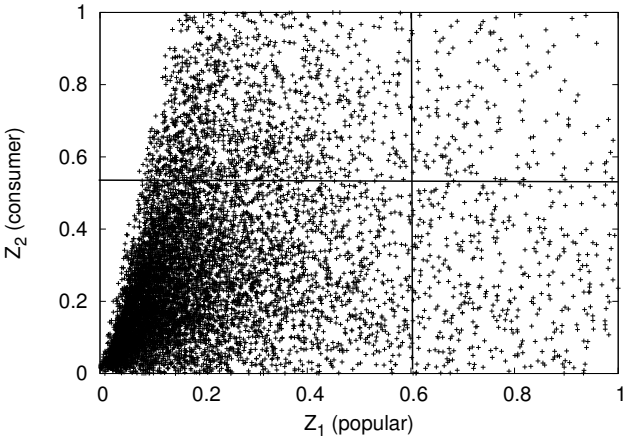

Fig. 5. Scatter plot of the user characteristics

IV. Case study: Analysis of target population in YouTube

The proposed methodology is now applied to the second case study considered in this paper. The goal for this case study is to analyze the population of target users in the YouTube network to evaluate if they represent an homo-

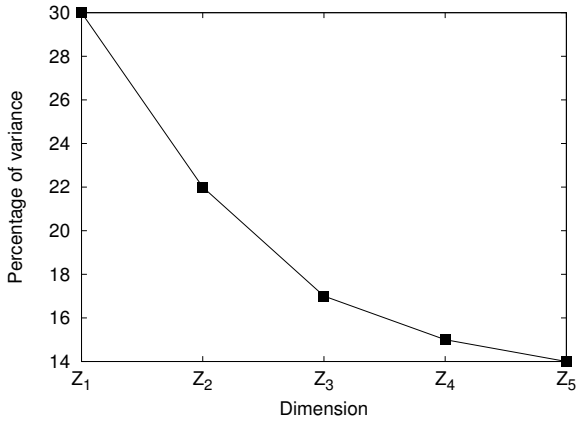

Fig. 6. Scree plot of the user activities

geneous group or if there are targets that may be reached through different dissemination strategies. For each target user identified in the previous case study we consider a set of five attributes, $(X_1, \ldots, X_5)$, describing the user activities over the time window of the crawling period, as shown in Table III. Specifically, we exploit the YouTube activity feeds to collect information about the number of subscriptions issued by each users, number of friendship created in the considered period, number of comments, rates, and marks as favorite carried out by users over the shared contents.

TABLE III
USER ACTIVITY ATTRIBUTES

Attribute		Description
X_1	Favorites	Number of videos marked as favorite by the user
X_2	Comments	Number of comments added by the user
X_3	Rates	Number of videos rated by the user
X_4	Friends	Number of friends of the user in the social network
X_5	Subscriptions	Number of subscriptions of the user

Figure 6 shows the scree plot based on the percentage of variance of all the dimensions that describe the user activities $(Z_1, \ldots, Z_5)$. These values are characterized by a slow decay that is not sufficient to isolate the most significant dimensions using only the visual method. For this reason, we exploit the Kaiser criterion, that identifies the most important dimensions selecting only those characterized by eigenvalues > 1, as suggested in [1]. In Table IV, we report the eigenvalues of the dimensions related to the user activities. Only the first two dimensions have eigenvalues > 1 and for this reason they can be considered sufficient to describe the target user population.

TABLE IV
EIGENVALUE OF THE USER ACTIVITIES

Dimension	Eigenvalue
Z_1	**1.5104751**
Z_2	**1.1233025**
Z_3	0.8749225
Z_4	0.7784665
Z_5	0.7128338

Now we analyze the attributes belonging to the two dimensions and identify the corresponding characteristics. The correlation values between the user activity attributes and the dimensions (Z_1, Z_2) are shown in Table V, where we outline using a bold font for each user activity attribute the dimension showing the highest correlation. The two dimensions can be described as:

Z_1. This dimension includes marks as favorite, comments and ratings. These attributes identify the user characteristic of preferring content-oriented activities.

Z_2. This dimension includes subscriptions to other user channels and creation of friendship relations. These attributes identify the user characteristic of performing a high number of social-oriented activities.

TABLE V
CORRELATION VALUES OF USER ACTIVITIES AND DIMENSIONS

Activity Attributes		Dimension	
		Z_1	Z_1
X_1	**Favorites**	**0.6198760**	0.0097078
X_2	**Comments**	**0.6454538**	-0.2643319
X_3	**Rates**	**0.6510846**	-0.3628119
X_4	**Friends**	0.1510087	**0.7869264**
X_5	**Subscriptions**	0.5127412	**0.5499556**

From a marketing point of view, the dimension Z_1 characterizes the users that devote a significant effort to enrich content metadata as ranking and comments. Due to this behavior, we may expect that these users tend to exploit metadata also to search and access contents within the social network. For a similar reason, the dimension Z_2 characterizes the users that are likely to rely mostly on social navigation to access contents on the social network, meaning that these users follow their social links to find interesting contents: this is the typical case of a user that periodically checks the recent uploaded content of his/her friends or subscribed user channels.

We now have to classify the target users from a marketing point of view according to their characteristics. To reach this purpose, we need to estimate the metrics Z_1^* and Z_2^* and for each of them to compute the cumulative distribution function of the normalized values, shown in Figure 7 and Figure 8, respectively. Considering the 80-th percentile as example of reference value to identify the relevant users for marketing, we select the users with $Z_1^* > 0.23$ or $Z_2^* > 0.7$. Figure 9 is a scatter plot of users based on the two identified characteristics. Users with $Z_1^* > 0.23$ (rightmost part of the graph) are more interested in accessing contents on the basis of their metadata, such as tags, received ratings, etc. Such users may be targets for viral marketing campaign where content propagation is driven by the content inherent features, for example because a video provides a funny story that attracts the user attention. The second set of users is characterized by $Z_2^* > 0.7$ (top part of the graph). These users are preferential targets when the advertisement aims to a customer fidelization with respect to the content source user, for example when a brand-related message has to be disseminated. It is worth to note that Figure 9 does not show users that have a high value for

both characteristics. This suggests that there are two separate classes of target users. As a consequence any attempt to exploit social network for marketing-oriented content dissemination must tailor contents and strategies according to the intended target user population.

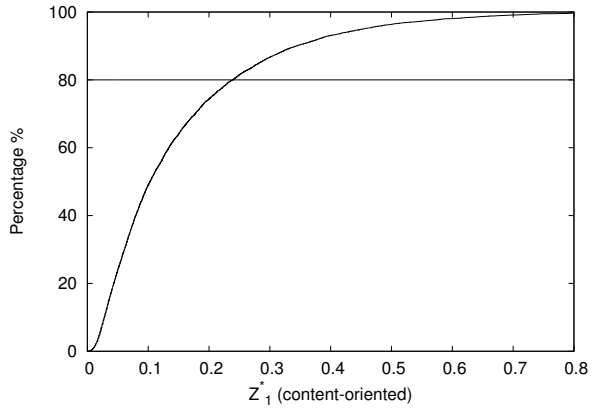

Fig. 7. Cumulative distribution of Z_1^*

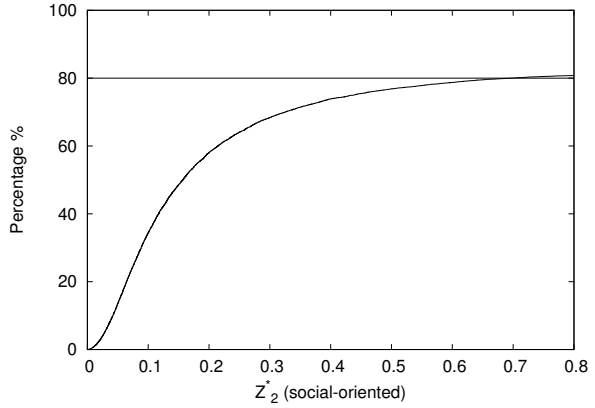

Fig. 8. Cumulative distribution of Z_2^*

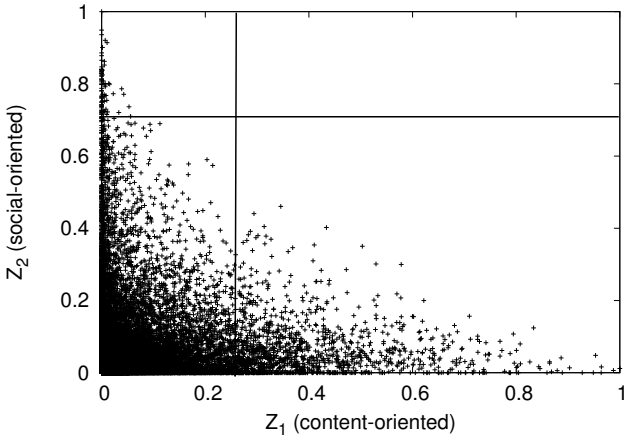

Fig. 9. Scatter plot of the characteristics

V. Related Work

Most studies on social networks focus on characterization. In particular, these studies may be divided on two main areas [13]: static characteristics, such as social network structure [5], [3], [17], and dynamic characteristics, such as workload [10], [8], [7] and effects of social navigation [16], [24], [15].

A characterization of the social network structure was initially proposed in [5], [3], that analyze the evolution of one social network over time, and in [17], that compares the static properties of multiple social network. The study in [5] is about user groups formation and evolution in LiveJournal; models for group evolution are also proposed. The authors in [3] analyze Cyworld, that is the largest Korean social networking site. Such studies aim to capture power law in the network characteristics, but no clear effort is devoted to cope with the high number of user attributes. Furthermore, these studies failed to define a methdology to identify relevant users in the social network.

The most relevant studies on workload characterization refer to YouTube, where serving video content represents a challenge due to resource size and image quality requirements. The studies in [10], [8] track YouTube transactions and consider content access patterns and video characterization in terms of size and category. Cha *et al.* [7] present an extensive analysis of the video distribution, age, and content duplication in YouTube. Finally, [6] describes the behavior of social network users by means of a stochastic finite state automata. However, these studies fail to capture how the social network structure may affect the workload characteristics. Furthermore, in these studies the heterogeneity of social network users is not considered and no user classification is provided.

More recently, several studies have been devoted to the impact of social navigation on content popularity in social networks. For example Lerman and Jones [16] consider the Flickr site and show the correlation between the number of incoming links of a user and the popularity of the uploaded images in terms of number of views. Also the study in [24] concerns the Flickr network and analyzes the users behavior in relation to the age of the uploaded contents. While these efforts may provide hints on how business may exploit social network, the studies do not address the issues of how the complexity of data collected on a social network can be addressed to identify the relevant user.

Classification problems characterized by data described by a high number of attributes are typically addressed by exploiting techniques as the Principal Component Analysis (PCA) [1] and the Independent Component Analysis (ICA) [11]. These techniques have the scope to transform a set of data characterized by a high number of attributes in a new set able to identify the main data characteristics that should be useful to their classification. However, while PCA has been always used as a popular method in data processing, ICA was originally developed for separating mixed audio signals into independent sources. Only recently ICA has been applied to data processing, but it is

not clear which technique offers the best performance since previous studies give contrasting results. ICA certainly can only be applied when the independent components are non-Gaussian, but in the context of social networks we can not be sure of the nature of the data distributions. Another technique that is not directly applicable to our context is the Binary Logistic PCA [20], derived from PCA. This technique, indeed, can only be used with binary data, while we have to analyze user attributes expressed as integer values.

The studies that exploit PCA for data classification can be divided in two broad categories: application of PCA to time series and application of PCA to non-temporal data. The studies focused on temporal data, include analysis of network traffic flows [4] and workload characterization for utility computing [2]. Application of PCA to non-temporal data classification has been proposed for face recognition [12], brain imaging [23], meteorology [19], and fluid dynamics [21]. However, to the best of our knowledge, the PCA analysis has never been applied for the social network analysis.

VI. Conclusions

Social networks are becoming an extremely popular media: the users population and the amount of content exchanged represent an interesting opportunity for marketing and advertisement. However, the number of social network users and of attributes describing them hinders the exploitation of the social networks potential. The need to collect huge amounts of data and the inherent complexity of analyzing them to extract useful information require novel mechanisms to identify which users are relevant for the goals of the analysis.

We propose a quantitative methodology based on PCA that can support social network analysis for identifying relevant users starting from a data set where each user is described by a large number attributes. We apply this methodology to data crawled from YouTube to demonstrate the effectiveness of our proposal in a marketing context. We analyze YouTube users to identifying the most relevant users for marketing. We identify sources and targets for content dissemination. Furthermore, we classify the target population into users that mainly exploit social links to look up for contents and users that are more interested in content tags, rank and other metadata. This insight can provide an additional marketing advantage to Internet-based business because it allows us to find users that are most responsive to different types of dissemination strategies, such as viral marketing or brand-based campaign.

Acknowledgements

The authors acknowledge the support of FP7–SEC–2009-1 project VIRTUOSO "Versatile Information Toolkit for End-Users Oriented Open-Sources Exploitation".

References

[1] H. Abdi and L. Williams. Principal component analysis. *Computational Statistics*, 2010. in press.

[2] B. Abrahao and A. Zhang. Characterizing application workloads on cpu utilization in utility computing. *Technical Report HPL-2004-157, Hewllet-Packard Labs*, 2004.

[3] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. of the 16th International Conference on World Wide Web (WWW'07)*, May 2007.

[4] L. Anukool, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. *Joint International Conference on Measurement and Modeling of Computer Systems*, pages 61–72, 2004.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD'06)*, Aug. 2006.

[6] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing User Behavior in Online Social Networks. In *Proc. of Usenix/ACM Internet Measurement Conference 2009*, Nov. 2009.

[7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. of the 7th ACM SIGCOMM conference on Internet measurement (IMC'07)*, San Diego, CA, Oct 2007.

[8] X. Cheng, C. Dale, and J. Liu. Statistics and Social Network of YouTube Videos. In *Proc. of 16th International Workshop on Quality of Service (IWQoS'08)*, Enschede, The Netherlands, Jun. 2008.

[9] FaceBook Statistics, 2010. http://www.facebook.com/press/info.php?statistics.

[10] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: A view from edge. In *Proc. of Internet Measurement Conference (IMC'07)*, Oct. 2007.

[11] A. Hyvärinen and E. Oja. Independent Component Analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[12] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, Jan. 1990.

[13] B. Krishnamurthy. A measure of Online Social Networks. In *Proc. of 1st. Int'l Conference on COMmunication Systems and NETworkS (COMSNETS)*, Jan. 2009.

[14] A. Lakhina, K. Papagiannaki, and M. Crovella. Structural analysis of network traffic flow. In *Proc. of Sigmetrics 2004*, Jun. 2004.

[15] K. Lerman. Social Information Processing in News Aggregation. *IEEE Internet Computing*, 11(6):16–28, 2007.

[16] K. Lerman and L. Jones. Social Browsing on Flickr. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'07)*, Mar. 2007.

[17] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of the 7th ACM SIGCOMM Conference on Internet measurement (IMC'07*, Oct. 2007.

[18] NielsenWire. Social Networking's New Global Footpront, 2009. Nielsen Online Report.

[19] R. W. Preisendorfer. *Principal component analysis in meteorology and oceanography*. Elsevier, 1988.

[20] A. I. Schein, L. K. Saul, and L. H. Ungar. A Generalized Linear Model for Principal Component Analysis of Binary Data. In *Proc. of the 9th International Workshop on Artificial Intelligence and Statistics*, Jan. 2003.

[21] L. Sirovich, K. S. Ball, and L. R. Keefe. Plane waves and structures in turbulent channel flow. *Phys. Fluids*, pages 2217–2226, 1990.

[22] Technorati. Facebook Looking to Improve Photo Tagging, Jul. 2010. Research Report.

[23] D. Ts'o, R. D. Frostig, E. E. Lieke, and G. A. Functional organization of primate visual cortex revealed by high resolution optical imaging. *Science*, pages 417–420, 1990.

[24] M. Valafar, R. Rejaie, and W. Willinger. Beyond friendship graphs: a study of user interactions in Flickr. In *Proc. of the 2nd ACM Workshop on Online Social Networks (WOSN'09)*, Aug. 2009.

[25] YouTube Statistics, 2010. http://www.youtube.com/t/fact_sheet.