Computer and Information Science; Vol. 10, No. 2; 2017
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

A New Model for Rating Users' Profiles in Online Social Networks

Mohammad Malli[1], Nadine Said[1] & Ahmad Fadlallah[2]

[1] Faculty of Computer Studies, Arab Open University, Lebanon

[2] Faculty of Arts and Sciences, University of Sciences and Arts in Lebanon, Lebanon

Correspondence: Ahmad Fadlallah, Faculty of Arts and Sciences, University of Sciences and Arts in Lebanon, Lebanon. Tel: 961-1-453003. E-mail: a.fadlallah@usal.edu.lb

Received: December 29, 2016 Accepted: January 29, 2017 Online Published: April 25, 2017

doi:10.5539/cis.v10n2p39 URL: http://doi.org/10.5539/cis.v10n2p39

Abstract

Profiling users in Online Social Networks (OSNs) is of great benefit in multiple domains (e.g., marketing, sociology, and forensics). In this paper, we propose a new model for rating user's profile (i.e., low, medium, high, and advanced) in an OSN community by embedding it into clusters located at predefined range of radius in a low-dimensional Cartesian space. The orthogonal coordinates of the profile are estimated using Principle Component Analysis (PCA) applied on a vector of metrics formulated as a set of attributes of interest (i.e., qualitative and quantitative) mined from the user's profile to characterize his/her level of participation and behavior in the community. The experimentations are conducted on 3000 simulated profiles of three OSNs (Facebook, Twitter and Instagram) by embedding them in three Cartesian spaces of three corresponding communities (Religion, Political and Lifestyle). The results show that we are able to estimate accurately the profile rates by reducing the vector of metrics to a low-dimensional space whittle down to 3-D space.

Keywords: Online Social Networks, OSN, Profiling, PCA

1. Introduction

Online Social Networks (OSNs) have gained a large popularity and became an integral part of our daily activity. Profiling OSN users has been frequently practiced for different purposes (e.g., price discrimination, targeted servicing, fraud detection, and extensive social sorting) in various fields (e.g., marketing, financial, sociology, and forensic science) despite the numerous concerns that have been raised (e.g., security, privacy, ethics, and liability).

In this work, we define a new model for rating user's profile in a community (e.g., political, religion, lifestyle) using a computerized algorithm in order to deal with huge and complex amount of profile's data. Then, the estimated rates are used for positioning the profiles in the clusters of each community, which are classified as low, medium, high, and advanced. To the best of our knowledge, there is no current solution for rating the profiles of OSN users in each community in such a way.

In order to test the accuracy of the proposed model, we experimented the clustering of 3000 profiles in religion, political and lifestyle communities (Note 1) of three social networks (i.e., Facebook, Twitter and Instagram). Each case study consists in embedding the profiles of an OSN community in an independent Cartesian space to observe their distribution in the clusters. Our results show that we are able to estimate accurately the profile rates by reducing the vector of metrics to a low-dimensional space whittle down to 3-D Cartesian space.

The rest of the paper is organized as follows. Section 2 discusses the related works. The clustering method is presented in Section 3. Then, Section 4 details three case studies conducted to validate our model. Finally, Section 5 concludes the paper and presents the future works.

2. Related Works

At the first stages, embedding and networking concepts within technology and society domains present inter-relationships that researchers depended on for identifying or categorizing profiles that belong to individuals or groups. A group profile refers to a category of people (e.g., radical, moderate) that does not necessarily form a community (e.g., religion, politic), but are found to share previously unknown patterns of behavior or other characteristics.

Recent works in this domain have focused on the presentation of user's information and on the analysis of this information to discover a correlation or pattern. Such proposals have presented the users' profiles in graphs and classified them into communities by relying on their attributes. These attributes are mainly extracted using data mining tool from the users' profiles and sometimes natural language processing techniques have been used to interpret data.

In this context, community detection algorithms depending on well-defined metrics are being developed such as the study that differentiates between intra-centrality and inter-centrality metrics, to characterize nodes in communities.

Network vertices are often divided into groups or communities with dense connections within communities and sparse connections between communities.

On the other hand, researchers propose some possible models to visualize networks with the objective of exposing their community structures based on modularity maximization, which can be used as a useful tool for community detection algorithms and for graph layout methods (Li, 2015). Another approach states a node-similarity based mechanism with the intention of exploring the formation of modular networks by applying the concept of hidden metric spaces of complex networks. Likewise, others initialize community attractiveness matrix to initiate a graph clustering algorithm based on the concept of density and attractiveness for weighted networks, considering node and edge weights.

Previous research on OSN communities has also defined factors that are responsible for the member's activity; for in- stance identity-based attachment showed a high impact on information sharing. Studies on the first popular OSNs illustrate a measurement framework to observe user activity in order to address two key issues of online social networks, which are characterization of user activities and usage patterns in these OSNs. Consequently, new algorithms are used for the purpose of distinguishing community formation by detecting a number of communities.

Studies on OSN's communities demonstrated that community structures across different online social networks are similar and are related to users' locations (Fan & Yeung, 2015). Relationships are measured according to mutual interest through two predefined properties which are the reachability that measures the ability of any node to reach out members of community, and the "isolability" that measures the ability of any community to isolate itself from the rest of the network.

Users can form or join existing groups on the basis of shared interests or because dense social connections exist among group members (Meo, Messina, Rosaci, & Sarné, 2014). On the other hand, community-based supervised learning process to detect the set of attributes in a user profile for which it is expected to see a correlation among their attributed values (e.g., job and salary) (Bahri, Carminati, & Ferrari, 2014). Besides, new socially and economically platforms that are drawn recently and may be a useful platform package for managing text, image, audio and video-based analysis modules to detect inappropriate content or high risk behavior.

Then, researchers made a comparison of communities' detection algorithms for multiplex that is considered as a set of graphs on the same vertex set because the wrong choice of algorithm can deviate from its intended purpose (Loe & Jensen, 2015). Some studies discuss the topological characteristics of legitimate users, including the formation of tightly knit communities because they consider that it is a promising approach, but they need to devise efficient techniques for identifying hackers, personal attack (FIDIS, 2016) and spammers along with attackers. Other works discussed the problem of community detection in complex networks and defined a vulnerability set and value for each of the communities within these networks.

Other studies designed a search algorithm to find users with the specified keywords in their profile attributes. Notably, it is based on a linear combination of topological distance and trust metrics. It is also dynamic in nature such that it adapts itself for each individual node during the search process. Additional research tracks the evolution of community structure and revise the effect of community-based immunization strategy on epidemic spreading.

Cross-System User Data Discovery has proven to be a successful algorithm in retrieving profiles that may belong to the searched user, correlate them, aggregate the discovered data and return them to the searcher (Carmagnola, Osborne, & Torre, 2014). Other researchers depended on local community neighborhood ratio function as a useful community detection algorithm.

3. Clustering Method

We propose a new model for profiling OSN users in a computerized way in order to deal with the huge and

complex amount of user profile's data. The model consists in calculating a rate for the user's profile in an OSN community.

The proposed clustering method consists in inferring a Cartesian space for each community where each cluster is located at predefined radius range. Then, the user's profile is represented by a point in the space in order to characterize its corresponding cluster.

Basically, the profile of an OSN user in each community is expressed as a vector of metrics that are set of attributes of interest (i.e., qualitative and quantitative) extracted from the user's profile using data mining tools to characterize his/her level of participation and behavior in a community. Then, the vector of metrics is transformed into a vector of uncorrelated and normalized coordinates using Principal Component Analysis (PCA). The transformed vector of independent coordinates (Note 2) becomes the new representation of the user's profile. In this way, the user's profile can be represented as a point in a Cartesian space where the coordinate of each axis can be obtained from the transformed vector. The rate of the user's profile (i.e., low, medium, high, and advanced) is estimated according to its position in the corresponding cluster of the Cartesian space.

The presentation of users' profiles in low-dimensional Cartesian space allows to easily infer the rate of the user's profile (i.e., low, medium, high, and advanced) through the length of its vector that fits within one of the predefined clusters in the Cartesian space. Besides, representing the user's profile in few metrics instead of having very large amount of complex data, leads to discover the most basic dimensionality of data since the "original" profile' metrics are correlated among each other. Furthermore, such approach provides better visualization the clustered users' profiles in low-dimensional Cartesian space given that it is hard to illustrate them in high-dimensional non-orthogonal space.

The problem can be formulated as follows:

Let $(\overrightarrow{x_i} \in \Re^m)$ be the vector of metrics of user i where the jth component of $(\overrightarrow{x_i})$ represents a qualitative or quantitative metric derived from the user's profile as implicit or explicit data, for j = 1...m.

Then, we need to map $(\overrightarrow{x_i} \in \Re^m)$to a reduced vector$(\overrightarrow{y_i} \in \Re^r)$ that contains independent coordinates since there are some correlations between the metrics derived from the user's profile.

To this end, we used the Principal Component Analysis (PCA) to determine the Eigen vectors of each community as the orthogonal axes of the Euclidean space modeling this community. This is achieved through the following steps:

a. Calculate the mean $\overrightarrow{m}$ and Covariance matrix *C* of the dataset containing *N* profiles:

$$\overrightarrow{m} = (\frac{1}{N})\sum_{i=1}^{N} \overrightarrow{x_i} \quad (1)$$

$$C = (\frac{1}{N-1})\sum_{i=1}^{N}[(\overrightarrow{x_i} - \overrightarrow{m})(\overrightarrow{x_i} - \overrightarrow{m})^T] \quad (2)$$

b. Derive the eigenvectors $\overrightarrow{v_1}, \overrightarrow{v_2} \ldots \overrightarrow{v_e}$ of the Covariance matrix

c. Form the orthogonal matrix $E_r = (\overrightarrow{v_1}^T, \overrightarrow{v_2}^T \ldots \overrightarrow{v_r}^T)$ by selecting the *r* eigenvectors $(\lambda_1, \lambda_2 \ldots \lambda_r)$ having the largest eigenvalues.

d. Map each profile vector $(\overrightarrow{x_i})$ of dimension m to a lower dimensional representation $(\overrightarrow{y_i})$ of dimension r with $r < m$ in the orthogonal space of the selected eigenvectors by projecting the vector to the matrix E_r through the following equation:

$$\overrightarrow{y_i} = E_r(\overrightarrow{x_i} - \overrightarrow{m}) \quad (3)$$

where$(\overrightarrow{x_i})$ and$(\overrightarrow{y_i})$ are treated as column vectors of dimensions *m* and *r* respectively.

The obtained coordinates of node *i* (the components of $\overrightarrow{y_i}$) can be considered as the independent and most basic metrics of the user's profile that are used to position it in the Cartesian space and infer its rate. This means that

our model infers easily the profile's rate through a simple process obtained by matrix-vector multiplication.

4. Model Validation

4.1 Overview

We recall that our model consists in profiling OSN users in the space of each community. For testing the accuracy of our model, we have tested the clustering of 3000 profiles in three communities of interest (Religion, Political and Lifestyle) of three social networks (Facebook, Twitter and Instagram). Each case study consists in embedding the profiles of an OSN community in an independent Cartesian space to observe their distribution in the clusters.

We started by defining the vector of metrics representing the users profile in each community as a set of attributes of interest (i.e., qualitative and quantitative) to characterize his /her level of participation and behavior in this community. Then, we applied our clustering model on three case studies:

- Community of Religion in Facebook.
- Community of Politics in Twitter.
- Community of Lifestyle in Instagram.

4.2 Vector of metrics

A large vector of qualitative and quantitative metrics can be defined and extracted from the user's profile using data mining tools. Then, Principle Component Analysis can be applied for dimensionality reduction to find the uncorrelated and normalized coordinates. In our experiments, we define the following vector of nineteen metrics for the profiles of Facebook, Twitter and Instagram as illustrated in Table 1:

Table 1. Selected Metrics for user's profile

Metric	Description
M1	Number of friends (Facebook) or Followers (Twitter or Instagram)
M2	Degree of relevance of profile name to the community: not relevant (0), low relevance (1), medium relevance (2), high relevance (3)
M3	Degree of relevance of profile photo to the community: not relevant (0), low relevance (1), medium relevance (2), high relevance (3)
M4	Degree of relevance of profile status to the community : not relevant (0), low relevance (1), medium relevance (2), high relevance (3)
M5	Number of posted relevant4 Books
M6	Number of relevant user's Page
M7	Number of relevant activities/events (to which user participated)
M8	Number of relevant groups (joined by user)
M9	Number of relevant surveys (filled by user)
M10	Number of relevant keywords (used by user) in previous search through Search Engine Optimizer (SEO).
M11	Number of received likes on relevant content(text, image, and video)
M12	Number of sent likes on relevant content.
M13	Number of received comments (Facebook/Instagram) or tweets (Tweeter) on relevant content
M14	Number of sent comments (Facebook/Instagram) or tweets (Tweeter) on relevant content
M15	Number of received posts containing a relevant content.
M16	Number of sent posts containing a relevant content.
M17	Number of received emails containing a relevant content
M18	Number of sent emails containing a relevant content
M19	Number of relevant content shared

4.3 Case Studies – Datasets

In order to test the accuracy of our model, we have generated 1000 fake profiles for each of the three case studies by assigning random values of the metrics for each profile given that it is of difficulty to collect large number of real OSN accounts. It is worth to mention that the fact that the profiles used for testing are fake and not real does

not have any negative impact on the validity of our testing. A computer program generates random values after restricting the range of values of each metric according to Table 2.

Table 2. Metrics Values Conditions

Metrics	Range	Metrics	Range
M1	0-1000	M11	0-1000
M2	0-3	M12	0-1000
M3	0-3	M13	0-200
M4	0-3	M14	0-200
M5	0-100	M15	0-200
M6	0-100	M16	0-200
M7	0-100	M17	0-100
M8	0-100	M18	0-100
M9	0-100	M19	0-100
M10	0-100		

4.4 Clusters Boundaries

We have generated 3000 OSN profiles vectors to be embedded in three Cartesian spaces representing the three selected communities which are religion, politics and lifestyle (one space per community). Once the dimensionality reduction is applied using PCA on the vector of metrics of all the OSN profiles of our dataset, we obtain the cartesian coordinate system $\mathbb{R}r$ of each community. Afterward, we are able to derive the orthogonal coordinates of each profile and its position in the corresponding space of its community. Once the points representing the profiles are distributed in the Cartesian space, a clustering method can be applied that optimizes an objective function such as:

- Hierarchical clustering algorithm: Hierarchical clustering algorithms (Jeon & Yoon, 2015) (Shepitsen, Gemmell, Mobasher, & Burke, 2008) can be applied based on Agglomerative Hierarchical clustering algorithm or Divisive Hierarchical clustering algorithm which are reverse to each other. Divisive Hierarchical clustering is a top-down approach starting by grouping all the points in one cluster. Then, it consists of splitting the cluster into two sub-clusters which are in turn divided into sub-clusters iteratively for being presented in a dendrogram graph. Hierarchical clustering is not required in our case since each point can be leveled through the magnitude of its vector in the Cartesian space where the distance between points can be easily calculated as well.
- K-means clustering algorithm: K-means is a popular algorithm used for clustering by grouping points into K clusters (Kanungo, et al., 2002). This can be applied on our dataset for determining the K clusters by selecting K random points as centers of the clusters. Then, the rest of points can be assigned to the cluster of the closest center. Afterward, the K centers should be re-determined by identifying in each cluster the point that minimizes the summed delay with the other points in the same cluster. This is repeated until reaching an unchangeable set of centers. Although the K-means is an interesting method for minimizing an objective function known as squared-error function, it is not useful in our case where our aim is to define clusters that represent the profiles rates and not only the proximity between the profiles point
- EMST based clustering algorithm: A Euclidean minimum spanning tree (EMST) can be used to detect clusters in the Cartesian space with irregular boundaries without assuming a spherical shaped clustering structure (Zahn, 1971). Clusters are detected to achieve some measure of optimality, such as minimum intra-cluster distance or maximum inter-cluster distance which is not our scope for profiling OSN users.
- Density-based clustering algorithm: Density-based clustering algorithm groups nearby points together that form a dense region (Ester, Kriegel, Sander, & Xu, 1996). A cluster is constructed through the selection of a new random point and a neighbor point is retrieved to be grouped in the cluster if it has sufficiently many neighbor points in turn. Despite the importance of this algorithm for many applications, we cannot rely on it for rating users' profiles in each community.

While these clustering methods are very useful for grouping points into clusters, another method is needed for

defining clusters that represent predefined rates so that the profiles embedded in each cluster can be easily rated. In this case, we propose that clusters are defined as specific range of radius in an (r-1)-sphere. This can be done by defining the boundary between each two adjacent clusters through subjective model by consulting experts in the domain of social media. In this way, they have defined subjective values of the vectors modeling the profiles of ten OSN users that could be rated on each boundary. Then, the transformation of these vectors to the Cartesian space provides ten points on each boundary where their average magnitude defines the radius of the boundary. Take the case where there are four clusters rated as low, medium, high, and advanced. In this case, the radius of the boundaries separating these clusters can be estimated as follows:

- For the boundary between the low and medium clusters, we have started by finding the profiles that could be rated at this location of the space. Therefore, let $\overrightarrow{A_i}, i = 1...10$ be the set of vectors that are modeling profiles of ten OSN
- Users having subjective values of metrics that are rated in-between low and medium. The transformation of $\overrightarrow{A_i}$ to the Cartesian space provides respectively the set of vectors $\overrightarrow{B_i}$ having an average length B_m. Then, we consider Bm as the radius of the *(r-1)*-sphere S_1 that is separating the low and medium clusters in the space.
- As for the second boundary separating the medium and high clusters, we have defined C_i' as the set of vectors that are modeling profiles of ten OSN users having subjective values of metrics that are rated in-between medium and high. The transformation of $\overrightarrow{C_i}$ to the Cartesian space provides respectively the set of vectors $\overrightarrow{D_i}$ having an average length D_m. Then, we consider D_m as the radius of the *(r-1)*-sphere S_2 separating the medium and high clusters in the space.
- Finally, the boundary between the high and advanced clusters is inferred by defining $E_i'$the set of vectors that are modeling profiles of ten OSN users having subjective values of metrics that are rated in-between high and advanced.
- The transformation of E_i' to the Cartesian space provides respectively the set of vectors F_i' having an average length F_m. Then, we consider F_m as the radius of the *(r-1)*-sphere S_3 separating the high and advanced clusters in the space.

Tables 3 and 4 summarize the defined original vectors for clusters boundaries and the i mapping to the orthogonal space.

Table 3. Vector Mapping

Vectors	Average Vectors	Profiles i	Vectors
$\overrightarrow{A_m}$	$\overrightarrow{A_i}$	i=1...10	[a1,a2...a19]
$\overrightarrow{C_m}$	$\overrightarrow{C_i}$	i=1...10	[c1,c2...c19]
$\overrightarrow{E_m}$	$\overrightarrow{E_i}$	i=1...10	[e1,e2...e19]

Table 4. Vectors Presentation

Real Vectors	Transformed Vectors	Average Magnitude	Radius
$\overrightarrow{A_i}$	$\overrightarrow{B_i}$	B_m	r_1
$\overrightarrow{C_i}$	$\overrightarrow{D_i}$	D_m	r_2
$\overrightarrow{E_i}$	$\overrightarrow{F_i}$	F_m	r_3

This means that the four clusters are defined as the following:

- Cluster 1: inside the (r-1)-sphere S1
- Cluster 2: between the (r-1)-sphere S1and S2.
- Cluster 3: between the (r-1)-spheres S2and S3.
- Cluster 4: outside S3

The position of an OSN user in a cluster in the Cartesian space should reflect the level of participation and behavior in the corresponding community. These positions can classify the users' profiles in the community according to the following mapping:

- Cluster 1: Low profile.
- Cluster 2: Medium profile.
- Cluster 3: High profile.
- Cluster 4: Advanced profile

4.5 Results

In order to validate our model, we rely on the four clusters characterized in the (r-1)-spheres of the Cartesian Space R^r using the subjective vector of metrics. Then, we evaluate the total error and matrix reconstruction error for the different values of r to evaluate the maximum reduction that can be tolerated for our dataset. First, the original vector ($\overrightarrow{x_i}$) is reconstructed from ($\overrightarrow{y_i}$) through equation 4:

$$\overrightarrow{X_i} = \overrightarrow{m} + E_r^T \overrightarrow{y_i} \tag{4}$$

Then, the total error over all the profile vectors is estimated as follows (Equation 5):

$$TE_r = (N-1) \sum_{j=r+1}^{d=19} \lambda_j \tag{5}$$

where (λ_j) are the eigenvalues discarded in the dimensionality reduction.

Figure 1 presents the relative total error ($TE_r / sum_T E$) for the different dimensionality reduction (r=1...19). The figure shows that the number of orthogonal dimensions can be reduced down to 3 dimensions with a relative error less than 0.06. When the number of dimensions is reduced down to 2 and 1, the relative total error increases considerably to reach 0.22 and 0.39 respectively. This means that the greatest proportion of eigenvalues hold for the first three principle components and the rest of them is relatively negligible.

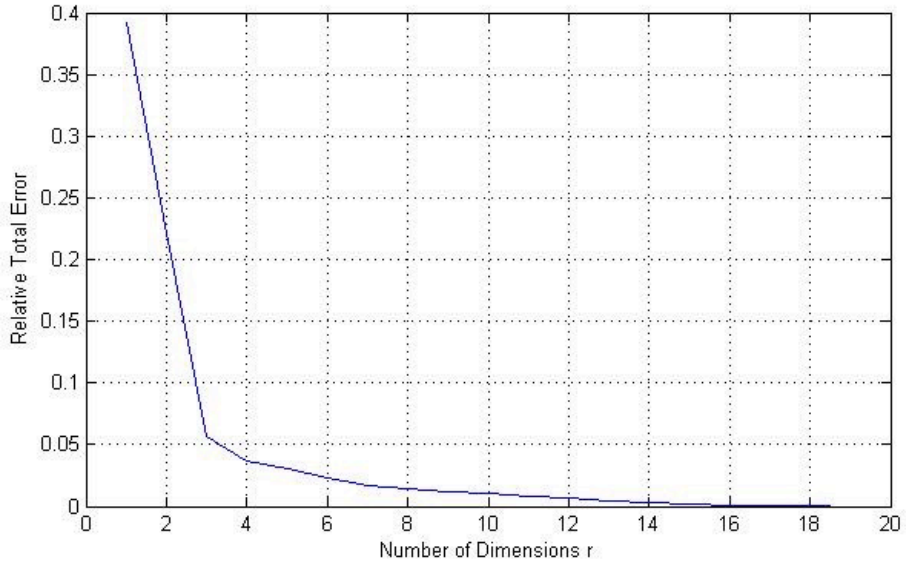

Figure 1. Relative Total Error

In addition, the reconstruction error of the reconstructed vector with respect to the original vector is calculated as follows (Equation 6):

$$error_r = \sum_{i=1}^{N} (\overrightarrow{x_i} - \overrightarrow{X_i})^2 \tag{6}$$

Figure 2 illustrates the relative reconstruction error ($error_r / error_1$) for the different possibilities of dimensionality reduction (r=2...19).

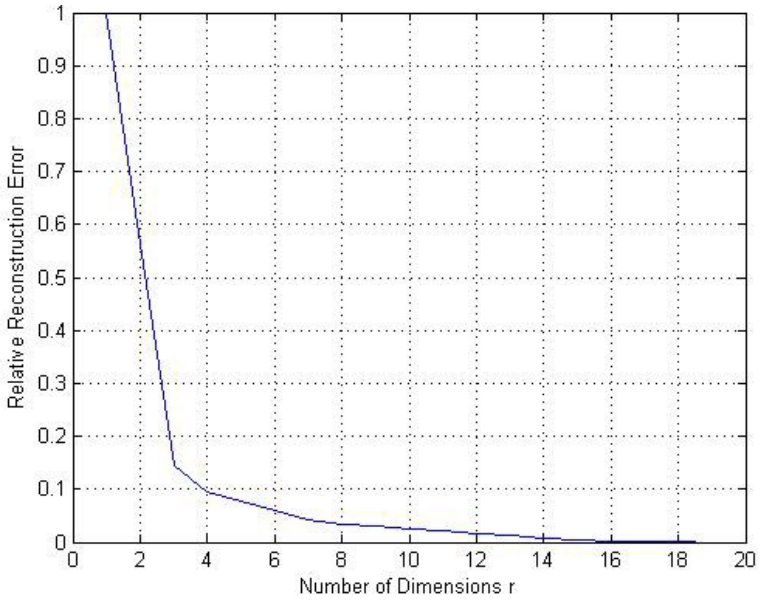

Figure 2. Relative Reconstruction Error

The figure shows that the number of dimensions can be reduced down to 4 with a relative errorless than 0.1 while this relative error increases slightly (equal to 0.15) when the number of dimensions r is equal to 3 and considerably when r is equal 2, to reach 0.55.

To test the impact of this dimensionality reduction on the accuracy of the clustering, we have conducted the following experiment. Basically, our clustering method is considered accurate if any two profiles that are "close" to each other in the original metric space have typically similar differences among each other in the Cartesian space. Thus, two nearby points in the metric space must have very similar coordinate vectors, and so must be mapped to nearby points under this embedding system to be located in the same cluster of the Cartesian space. To simplify the problem, this is validated as the following:

- Low Profiles: For all profiles i (Note 3) satisfying $||\overrightarrow{y^i}|<r_1$

Let l_i= number of match cases satisfying $x_j^i \leq a_j, j = 1...19$

The fraction of match cases for low profile i is: $al_i = l_i/19$

- Medium Profiles: For all profiles i satisfying $r_1<||\overrightarrow{y^i}||<r_2$

Let m_i = number of match cases satisfying $a_j<x_j^i \leq c_j, j = 1...19$

The fraction of match cases for Medium profile i is: $am_i = m_i/19$

- High Profiles: For all profiles i satisfying $r_2<||\overrightarrow{y^i}||<r_3$

Let h_i = number of match cases satisfying $c_j<x_j^i \leq e_j, j = 1...19$

The fraction of match cases for High profile i is: $ah_i = h_i/19$

- Advanced Profiles: For all profiles i satisfying $||\overrightarrow{y^i}||>r_3$

let d_i = number of match cases satisfying $x_j^i>e_j, j = 1...19$

The fraction of match cases for advanced profile i is: $ad_i = d_i/19$

The results show a complete matching when the number of dimensions r is reduced down to four. The reduction to three is illustrated in Figures 3, 4, 5 and 6 the Cumulative Distribution Function (CDF) that evaluates the following matching vectors of the profiles belonging to each cluster of the 3-D space: al_i, am_i, ah_i and ad_i.

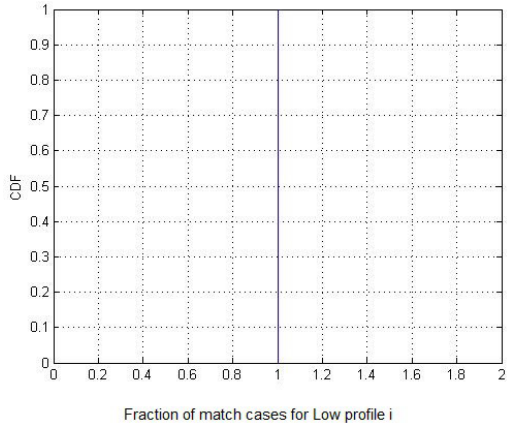

Figure 3. CDF of al_i

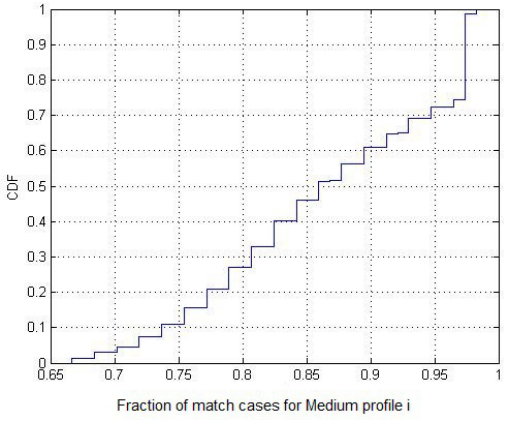

Figure 4: CDF of am_i

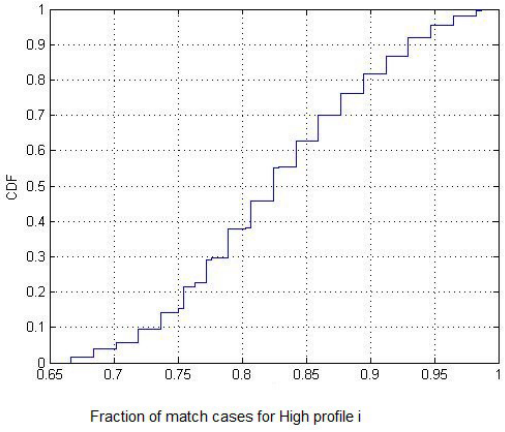

Figure 5. CDF of ah_i

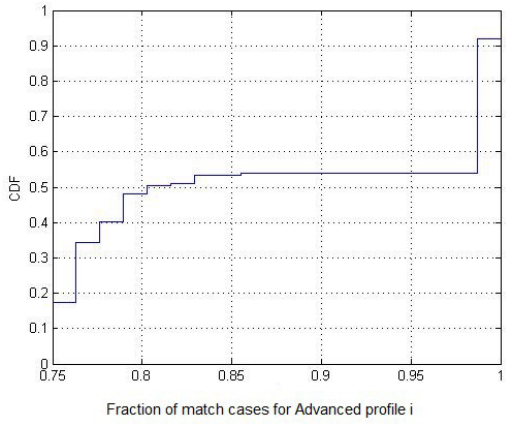

Figure 6: CDF of ad_i

These figures show the following percentages of OSN profiles having metrics within the pre-defined ranges:

- Profiles clustered in Cluster 1: All profiles have 100%.
- Profiles clustered in Cluster 2 and Cluster 3: 90% of these profiles have more than 70% to 75% matching.
- Profiles clustered in Cluster 4: All these profiles have more than 75 % matching.

It is noticed that the low percentage of non-matching cases of metrics belonging to a specific profile should not lead to profiling the corresponding user in another cluster. This is mainly because this low percentage of metrics is not considered as principal component in PCA analysis when the number of directions is reduced to three major components (3-D). It is obvious that better results can be also obtained if the number of principle components is reduced to greater numbers (>3). The results prove also that the major principle components in our dataset are the first four and the error increases slightly when the number of dimensions is reduced to three.

To limit the number of orthogonal directions without affecting the accuracy of the clustering, we have compared the impact of the dimensionality reduction on clustering when the number of the principle components is limited to 3 and 4 respectively. Figure 7 plots the CDF of the relative variation in length of every point embedded by comparing its magnitude inferred in both the 3-D and 4-D spaces as the following (Equation 7):

$$\frac{|(Length\ Vector\ i\ in\ 3D) - (Length\ Vector\ i\ in\ 4D)|}{Length\ Vector\ i\ in\ 3D} \tag{7}$$

The figure shows that around 85% of the points representing the profiles have variation in length less than 6%.

Then, we have checked the impact of this low percentage of variation on clustering and we have found that none of the points modeling the users' profiles change its cluster when reducing the number of dimensions to 3. One may think that it is still probable in another data set that a point positioned very close to the boundary of an adjacent cluster may be erroneously clustered due to the error (even if small) resulted from the reduction of the number of principle components to three. Thus, one can reduce the number of dimensions to three to simplify the presentation of profiles in 3-D space but this tolerates that some points positioned very close to the boundaries of clusters may be not precisely clustered and in this case further analysis is required to be applied on such points if necessary.

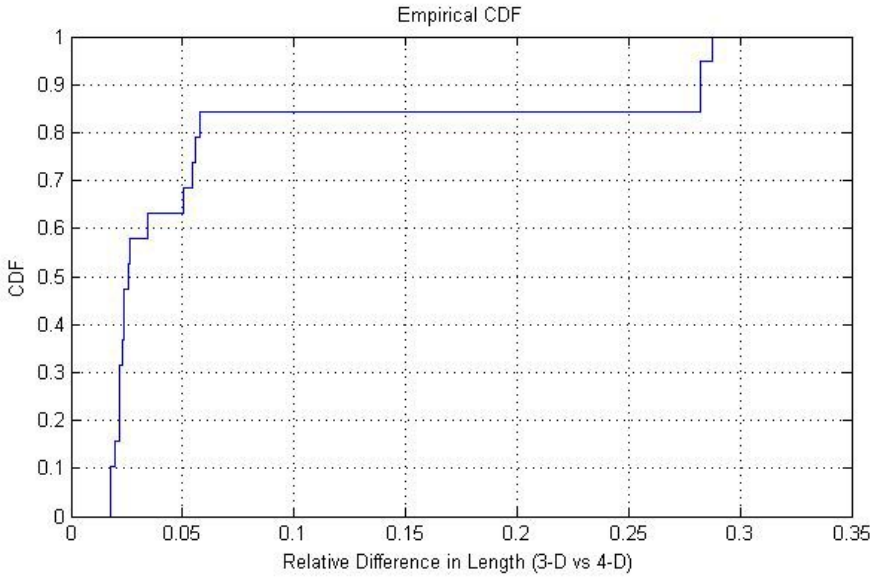

Figure 7. CDF of length variation

5. Conclusions and Perspectives

Our model for clustering OSN users has a number of interesting properties as compared to the traditional mining methods of user's profile. The major advantage is that it allows rating the users' profile (i.e., low, medium, high, and advanced) in each community after presenting it in a low-dimensional Cartesian space embedding the community. The network coordinates of the profile are inferred using PCA after mining the attributes as a set of metrics reflecting its activities and behavior within the community.

The model has been validated by profiling 3000 OSN users in the communities of politics, religion, and lifestyle of three popular social networks. This is achieved by inferring a profile rate in each community that well reflects to the level of participation and tendency of the user. The presented results show valid clustering of users profiles after dimensionality reduction of the vector of metrics whittle down to three dimensions.

Further studies need to continue conducting empirical research to ascertain more factors that contribute to refine and enhance the profiling of users. Particularly, identifying users that are most likely to impress others and direct their attitudes may help for characterizing, preventing and controlling advanced malpractice profiles. On the other hand, future studies must also detect the fabricated profiles that work on manipulating users in a specific community.

References

Briggle, A. C. M. (2009). Embedding and networking: conceptualizing experience in a technosociety. *Technology in Society, 31*(4), 374-383.

Terrettaz-Zufferey, A. L. F. R. (2007). Pattern Detection in Forensic Case Data Using Graph-Theory: Application to Heroin Cutting Agents. *Forensic Science International, 167*(20).

Arab, M., & Afsharchi, M. (2014). Community detection in social networks using hybrid merging of

sub-communities. *Journal of Network and Computer Applications, 40*, 73-84. Retrieved from http://www.sciencedirect.com/science/article/pii/S1084804513001835

Das, B. J. S. (2011). India Social Networking Sites A Critical Analysis of Its Impact on Personal and Social Life. *International Journal of Business and Social Science, 2*(14).

Bahri, L., Carminati, B., & Ferrari, E. (2014, June). Community-Based Identity Validation on Online Social Networks. *2014 IEEE 34th International Conference on Distributed Computing Systems*, (pp. 21-30).

Biswas, A., & Biswas, B. (2015). Investigating community structure in perspective of ego network . *Expert Systems with Applications, 42*(20), 6913-6934. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417415003292

Dwyer, C. S. R. (2007). Trust and Privacy Concern within Social Networking Sites: A Comparison of Facebook and MySpace,. *Americas Conference on Information Systems-AMCIS*.

Hu, C. L. Z. (2015). Achieving self-congruency? Examining why individuals reconstruct their virtual identity in communities of interest established within social network platforms. *Computers in Human Behavior, 50*, 465-475.

Carmagnola, F., Osborne, F., & Torre, I. (2014). User data discovery and aggregation: The CS-UDD algorithm . *Information Sciences, 270*, 41-72. Retrieved from http://www.sciencedirect.com/science/article/pii/S0020025514002175

Chung, N., Nam, K., & Koo, C. (2016). Examining information sharing in social networking communities: Applying theories of social capital and attachment . *Telematics and Informatics, 33*(1), 77-91. Retrieved from http://www.sciencedirect.com/science/article/pii/S0736585315000490

Constine, J. (2014, April). Why Is Facebook Page Reach Decreasing? More Competition And Limited Attention. *Why Is Facebook Page Reach Decreasing? More Competition And Limited Attention.*

Conti, M., Poovendran, R., & Secchiero, M. (2012, August). FakeBook: Detecting Fake Profiles in On-Line Social Networks. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, (pp. 1071-1078).

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., & Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds . *Computers, Environment and Urban Systems, 53*, 47-64. Retrieved from http://www.sciencedirect.com/science/article/pii/S0198971514001306

Boyd, D. N. E. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of computer-Mediated communication, 13*(1), 210-230.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining (Kdd-96), 96*(34), 226-231.

Eustace, J., Wang, X., & Cui, Y. (2015). Community detection using local neighborhood in complex networks . *Physica A: Statistical Mechanics and its Applications, 436*, 665-677. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378437115004598

Schneider, F. D. A. (2014). Understanding Online Social Network Usage from a Network Perspective. *Understanding Online Social Network Usage from a Network Perspective*.

Fan, W., & Yeung, K. (2015). Similarity between community structures of different online social networks and its impact on underlying community detection . *Communications in Nonlinear Science and Numerical Simulation, 20*(3), 1015-1025. Retrieved from http://www.sciencedirect.com/science/article/pii/S1007570414003165

FIDIS. (2016). Future of Identity in the Information Society. *Future of Identity in the Information Society*.

Zeno, G. S. P. (2008). *D6.7c: Forensic Profiling*. Tech. rep., FIDIS Deliverables.

Garg, A., Bhattacharyya, P., Martel, C. U., & Wu, S. F. (2009, Aug). Information Flow and Search in Unstructured Keyword Based Social Networks. *2009 International Conference on Computational Science and Engineering, 4*, 1074-1081.

Geradts, Z. (2006). Forensic implications of identity systems. *Datenschutz und Datensicherheit - DuD, 30*(9), 557-559.

Gyarmati, L., & Trinh, T. A. (2010, September). Measuring user behavior in online social networks. *IEEE Network, 24*(5), 26-31.

Jeon, Y., & Yoon, S. (2015, Sept). Multi-Threaded Hierarchical Clustering by Parallel Nearest-Neighbor Chaining. *IEEE Transactions on Parallel and Distributed Systems, 26*(9), 2534-2548.

Fernandez, K. C. L. (2012). Predicting Social Anxiety From Facebook Profiles. *Social Psychological and Personality Science, 3*(6), 706-713.

Kalaitzakis, A., Papadakis, H., & Fragopoulou, P. (2012, Aug). Evolution of User Activity and Community Formation in an Online Social Network. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, (pp. 1315-1320).

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002, Jul). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 881-892.

Li, W. (2015). Visualizing network communities with a semi-definite programming method . *Information Sciences, 321*, 1-13. Retrieved from http://www.sciencedirect.com/science/article/pii/S002002551500403X

Linial, N. (2002). Finite Metric Spaces: Combinatorics, Geometry and Algorithms. *Proceedings of the Eighteenth Annual Symposium on Computational Geometry* (pp. 63-63). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/513400.513441

Liu, R., Feng, S., Shi, R., & Guo, W. (2014). Weighted Graph Clustering for Community Detection of Large Social Networks. *Procedia Computer Science, 31*, 85-94. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877050914004256

Loe, C. W., & Jensen, H. J. (2015). Comparison of communities detection algorithms for multiplex . *Physica A: Statistical Mechanics and its Applications, 431*, 29-45. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378437115002125

Lu, Z., Sun, X., Wen, Y., Cao, G., & Porta, T. L. (2015, Nov). Algorithms and Applications for Community Detection in Weighted Networks. *IEEE Transactions on Parallel and Distributed Systems, 26*(11), 2916-2926.

Lynn, R., & Witte, J. C. (2015, 01). Do social network sites increase, decrease, or supplement the maintenance of social ties? In *Communication and Information Technologies Annual* (pp. 79-106). Emerald.

Ma, L., Jiang, X., Wu, K., Zhang, Z., Tang, S., & Zheng, Z. (2012). Surveying network community structure in the hidden metric space . *Physica A: Statistical Mechanics and its Applications , 391*(1), 371-378. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378437111006066

Meo, P. D., Messina, F., Rosaci, D., & Sarné, G. M. (2014, Sept). Improving the Compactness in Social Network Thematic Groups by Exploiting a Multi-dimensional User-to-Group Matching Algorithm. *2014 International Conference on Intelligent Networking and Collaborative Systems*, (pp. 57-64).

Morrow, K. (2014, July). Facebook, Sentiment Analysis and Emotional Contagion. *Facebook, Sentiment Analysis and Emotional Contagion*.

Nierhoff, M. (2014, October). The average Facebook page performance for September 2014 can easily compare for that in September 2013. *The average Facebook page performance for September 2014 can easily compare for that in September 2013*.

Popescu, A. C., & Farid, H. (2005, October). Exposing digital forgeries in color filter array interpolated images. *IEEE Transactions on Signal Processing, 53*(10), 3948-3959.

Religion, T. A., & Politics, P. s. (2014, September). More Americans Support Mixing Religion and Politic. *More Americans Support Mixing Religion and Politic*.

Rutkin, A. (2013, October). How Your Facebook Profile Reveals More About Your Personality Than You Know. *How Your Facebook Profile Reveals More About Your Personality Than You Know*.

Claudio, M. R. S., & Ramirez-Marquez, J. E. (2011). Vulnerability metrics and analysis for communities in complex networks . *Reliability Engineering \& System Safety , 96*(10), 1360-1366. Retrieved from http://www.sciencedirect.com/science/article/pii/S0951832011000299

Shepitsen, A., Gemmell, J., Mobasher, B., & Burke, R. (2008). Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering. *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp. 259-266). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1454008.1454048

Vanhove, T., Leroux, P., Wauters, T., & Turck, F. D. (2013, May). Towards the design of a platform for abuse detection in OSNs using multimedial data analysis. *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, (pp. 1195-1198).

Yuan, P., & Tang, S. (2015). Community-based immunization in opportunistic social networks . *Physica A: Statistical Mechanics and its Applications , 420*, 85-97. Retrieved from http://www.sciencedirect.com/science/article/pii/S037843711400942X

Zahn, C. T. (1971, Jan). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers, C-20*(1), 68-86.

Notes

Note 1. The number and type of communities is an experimentation choice; the same experimentation could be run for a higher number and different types of communities.

Note 2. The metrics could be correlated among each other.

Note 3. Among the 3000 experimental profiles.