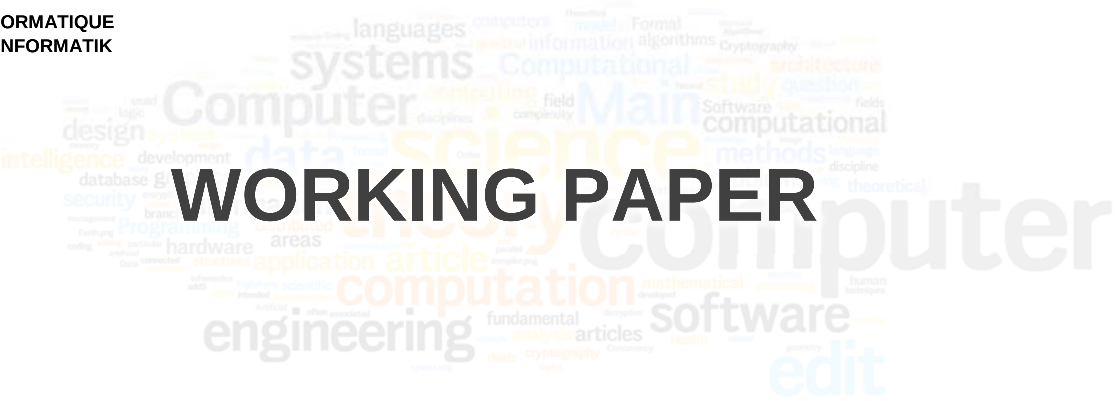
UNI
FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

DEPARTEMENT D’INFORMATIQUE
DEPARTEMENT FÜR INFORMATIK
Bd de Pérolles 90
CH-1700 Fribourg

www.unifr.ch/inf



WORKING PAPER

Exploring Nuances of User Privacy Preferences on a Platform for Political Participation

A. Kaskina

Internal working paper no 17-02

November 2017

Exploring Nuances of User Privacy Preferences on a Platform for Political Participation

Aigul Kaskina

University of Fribourg,
Bd. de Pérolles 90, Switzerland, CH-1700
aigul.kaskina@unifr.ch

Abstract. A problematic gap between existing online privacy controls and actual user disclosure behavior motivates researchers to focus on a design and development of intelligent privacy controls. These intelligent controls intend to decrease the burden of privacy decision-making and generate user-tailored privacy suggestions. To do so, at first it is necessary to analyze user privacy preferences. Previous studies have shown that user privacy profiles tend to have a multidimensional structure, which in turn might bring issues of an inexact user classification. This paper proposes to apply a fuzzy clustering approach, where fuzzy membership degree values can be used for the calculation of more precise personalized privacy suggestions. Based on the real-world dataset collected from a political platform [1], the fuzzy c-means algorithm was applied to demonstrate the multidimensionality and the existence of imprecise user privacy profiles, where a user simultaneously possesses features inherent in several clusters.

Keywords: User Privacy Preferences, Fuzzy Clustering, Data Analysis

1 Introduction

In the digitalised world people are disclosing more and more personal information in online platforms. Researchers delineate a number of problems related to one's privacy. The concept of "privacy paradox" explicates that people's actual disclosure behaviour in most cases diverge with their inner privacy attitudes. Different factors plays an important role in a such paradoxical behaviour. As an example, people could choose to explicitly disclose or share information about themselves, their opinions, and their activities as means of declaring their loyalties or differentiating themselves from others [1].

People are also confronted with privacy compromises and trade-offs [2], [3]. With that, it becomes harder for people's minds to estimate implying risks in their disclosure behaviour. Moreover, due to the complexity of privacy controls in online platforms, it becomes more difficult to precisely express one's privacy

[1] The platform used in the present research is Participa Inteligente (https://participacioninteligente.org)

decision, which may lead to an uncertainty in privacy decision-making [4]. Therefore, studies heavily focus on analysing user's desired/actual disclosure behavior in online platforms, how to quantify the user's privacy preferences and other factors which might impact their desired/actual disclosure behaviour.

This work investigates the question on *how to detect and quantify an underlaying uncertainty in user privacy preferences on online platform for political participation.* The reminder of the paper is organised as follows: first, Section 2 gives a short literature review related to the user disclosure behaviour in online platforms. Then Section 3 explains the method applied for an exploration of user privacy preferences. In particular, the dataset of user privacy preferences is described in Section 3.1, fuzzy clustering analysis presented in Section 3.2, and the results are discussed in Section 4. Finally, Section 5 summarises concluding remarks and outlook for the future work.

2 Literature Review

Privacy is a tricky and hazardous topic. As it is a "faceless issue"[5], people are often not aware about risks and implications of information they are disclosing online. A well-known phenomena of a "privacy paradox", shows that people with serious privacy attitudes are still revealing quite intimate information about their lives for trivial rewards [5]. The gap between a privacy attitude and an actual disclosure behaviour is influenced by different types of rewards and benefits. Hui *et al.* [6] indicated that aside the popular monetary saving or time saving benefits, various types of benefits like social adjustment, or altruism, when used appropriately can also motivate users to engage in online disclosure.

Nevertheless, benefits and rewards are not the only factors that impact users' disclosure behaviours. Early studies have investigated the relationship of personality and privacy preferences in offline environment. Marshall [7] and Pedersen [8] identified highly similar set of privacy dimensions, and described how personality determines peoples' privacy preferences. Marshall found a correlation between person's introversion and his total privacy score, while Pedersen showed that low self-esteem was associated with solitude and anonymity.

Considering privacy disclosure in online environment, Quercia *et al.* [9] using a Big Five personality measurement classified users into privacy conscious and pragmatic majority types. They found that privacy conscious users are correlated with traits as openness and extraversion. In contrast, Egelman and Peer [22] argued that personal traits such as decision-making and risk-taking attitudes are much stronger predictors for privacy attitudes than traditional Big Five personality model. In contrast to [12], [13], [14], where a summated composite score represents a disclosure behaviour, Knijnenberg et al. [11] argued that disclosure behaviours are in fact multidimensional. They suggested that privacy disclosure classification should move from the "one-size-fits-all" approach while estimating user disclosure behaviour. They showed that people can be classified into distinct groups which show very different behaviours along privacy dimensions.

Classifying users' disclosure behaviour can further be used for the development of intelligent privacy controls. However, the issue of an oversimplification may occur if to to rely on a uni- or even multidimensional classification of disclosure behaviors. This work shows that user disclosure profiles should be considered not only as multidimensional, but also as fuzzy. To the best of our knowledge, this is a first attempt to interpret the multidimensionality and uncertainty of user privacy preferences using a fuzzy logic approach. This work demonstrates to what extent the nuances of user privacy behaviours can be captured with the help of fuzzy c-means clustering; thus a further fuzzification of constituent attributes of user privacy profiles might contribute to a better design and development of intelligent privacy controls.

3 Method

3.1 Dataset Collection

Differently from the previous research presented in [15], where data was collected from a survey, the dataset of this work contains real user privacy profiles collected from the platform for a political participation. This platform was developed for the presidential election processes in Ecuador held in February 2017. Users' privacy settings were collected during the 4-month period of December 2016 – March 2017. After a cleaning and preparation steps, the final dataset consisted of 391 user profiles. Among them, women are 131, men are 253, and 7 users who did not provide a gender information. The major age of users is between 23 and 36 (median age is 28).

The privacy control of the platform was designed according to the user privacy framework described in [15]. It consists of the *data type* and the *audience* blocks through which a user can express visibility preferences for a particular data type. Accordingly, a user profile in the dataset is represented as a set of users $U = \{\overrightarrow{u_1}, \overrightarrow{u_2}, ..., \overrightarrow{u_N}\}$. Users' privacy settings belongs to a set $S = \left\{s_{(1,1)}, s_{(2,2)}, ..., s_{(i,k)}\right\}$, where $i \in I$ is a set of user's *data types ("MyActivity, ContactMe, MyRelations, MyTopics, PersonalInfo, VoteIntention")*, and $k \in K$ is a set of the user's *visibility preferences* related to *("OnlyMe, Friends, FriendsOfFriends, Public")* with numeric values equal to [1, 2, 3, 4], correspondinly.

Table 1. Dataset description

User	**Variables (data types)**					
	MyActivity	*ContactMe*	*MyRelations*	*MyTopics*	*PersonalInfo*	*VoteIntention*
1	2	3	2	3	3	2
2	1	3	1	1	1	1
3	4	3	3	2	1	1
...	...	...	...	...	...	...
391	4	4	4	2	2	1

3.2 Fuzzy clustering

Fuzzy clustering allows objects to be associated with many clusters according to their membership degree value and is based on the fuzzy set theory introduced by Zadeh [16]. The main goal of the fuzzy clustering is to compute the similarity that an object shares with each cluster using a membership function. The membership function calculates the membership degree of each object in every cluster with values in the range of [0,1]. A high degree of similarity between the object and a cluster is assigned when a membership value is close to 1, whereas values close to 0 imply a low similarity [17]. In this work, the fuzzy c-means algorithm was used with Euclidean distance similarity measure in a vector space to execute the partitioning of objects into clusters. The algorithm was calculated within the R environment using package *fclust* [24].

To determine an optimal number of clusters an evaluation of cluster validity indexes of clustering is a necessary step. The optimal partition can be determined by the point of the extrema of the validation indexes in dependence of the number of clusters. In addition, the weighting exponent m is an important parameter as it influences the quality of inferences that can be done about the further validity of clustering results. The weighting exponent m=*2* was used in our data analysis, as the best recommended value for calculations of fuzzy clustering [20].

Table 2. Validation indexes of fuzzy c-means clustering

# of clust	**PC**	**PE**	**MPC**	**SIL**	**SIL.F**	**XB**
2	0.7717	0.3612	0.5434	0.6209	0.7319	0.1565
3	0.6423	0.6116	0.4634	0.5154	0.7368	0.299
4	0.5696	0.7967	**0.4261**	**0.4409**	0.7721	0.7664
5	0.5416	0.9134	0.427	0.4647	0.7738	1.5684
6	0.5331	0.9879	0.4397	0.5053	0.7723	0.3972
7	0.5336	1.037	0.4558	0.5181	0.7808	0.3527
8	0.5505	1.0411	0.4863	0.4958	0.7689	0.9501
9	0.5653	1.0449	0.5109	0.5263	0.7887	0.2165
10	0.5734	1.0564	0.5259	0.5618	0.7918	0.1583
11	0.5701	1.0947	0.5271	0.556	0.8047	2.7169E+16
12	0.5741	1.1144	0.5353	NA	NA	2.25733E+18
13	0.571	1.1501	0.5352	NA	NA	1.91666E+18
14	0.5677	1.1853	0.5345	NA	NA	5.50123E+18

There are three validity measures which are exclusively based on the membership values: partition coefficient (PC), partition entropy (PE) [18] and modified partition coefficient (MPC) [19]. The maximum value of PC, MPC and minimum value of PE indicates a good partition in the meaning of a more sharp partition result and inverse values for a fuzzy partitioning result. The most popular validity index that measures both compactness and separation of clusters is Xie and Beni (XB) index. Minimized value of XB index [20] suggests the best partition of the dataset.

A common practice in the cluster validation is to run clustering the algorithm with different values of c-centers of clusters on a given dataset, and calculate corresponding validation indexes per each execution of clustering algorithm. On our dataset, the fuzzy c-means clustering algorithm was executed for 2 – 14 cluster centers. The results of validation indexes are displayed in Table 1.

If to consider validation indexes in the meaning of the sharp clustering result, all validation indices agreed on 2 clusters, except of the SIL.F index that suggested the existence of 11 clusters. In terms of the fuzzy partitioning result, all indices showed different results, where only both MPC and SIL indexes signified the agreed best value with 4 clusters. According to cluster validity indexes, the fuzzy c-means clustering algorithm was applied to our dataset with 4 cluster centroids for fuzzy partitioning of users.

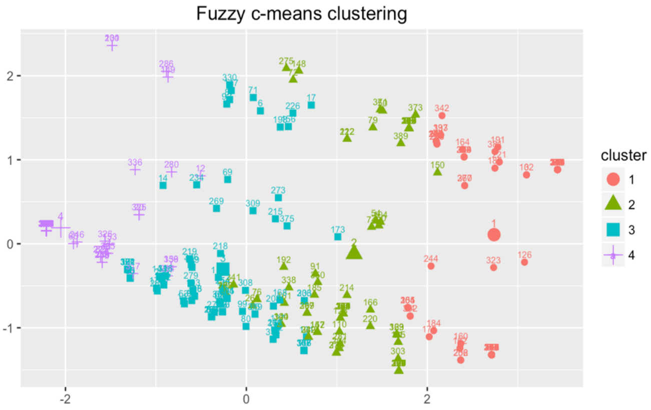


Fig. 1. Fuzzy c-means clustering with 4 clusters

Table 3. Cluster centroid Euclidean similarity

	Cluster 1	Cluster 2	Cluster 3
Cluster 2	2.141966		
Cluster 3	3.425611	1.347993	
Cluster 4	6.061216	4.026885	3.052340

Figure 1 displays the result of a distribution of observations and Table 3 shows the Euclidean distances among cluster centroids. A greater distance between clusters correspond to a greater dissimilarity. Observations in the cluster 1 have opposite privacy preferences with observations in the cluster 4, because centroids of those clusters have the lowest similarity value. In contrast, the smallest distance is between the cluster 2 and the cluster 3. In the next section the intra-cluster characteristics are discussed in detail.

4 Discussion of Results

4.1 Cluster characteristics

Table 4 outlines the cluster characteristic related to a cluster size, user characteristic such as gender, and percentage of visibility preference values across all six data types per cluster. It can be clearly seen that users in the cluster 1 are highly privacy-preserved keeping their profile's data types private and unshared, while cluster 4 has the biggest size, where users' privacy profiles distinguished as totally public. The cluster 3 is the second largest cluster, where users are still sharing their data types mostly to public, but also 20% of privacy decisions occurred to be private. The cluster 2 is the smallest one, and users have 53% of preferences sharing to friends within this cluster.

Interesting to observe the gender representation within each clusters. One point must be taken into consideration that the initial dataset had almost a doubled amount of male users compared to a number of female users (253 males against 131 females). Another point is that the size of the cluster 4 outweighs the size of other clusters, therefore the majority of gender presentation is shown in the cluster 4. However, it can be seen that a majority of females prefer to have private privacy settings (cluster 1), and a minority of females appears in cluster 3. In contrast, a majority of males appears to be in cluster 3, while a minority of males prefer to have private profiles. Those people who did not provide the information about their gender appear in clusters 1 and 2.

Table 4. Cluster characteristics

	size	**Female**	**Male**	**NA**	**Public**	**FoFs**	**Friends**	**OnlyMe**
Cluster 1	82	37	41	4	8%	1%	24%	66%
Cluster 2	77	31	46	0	23%	3%	53%	21%
Cluster 3	89	25	61	3	55%	7%	17%	20%
Cluster 4	143	38	105	0	95%	2%	3%	0%

Additionally, users in the cluster 1 allowing to only friends the possibility to contact them while keeping the rest of their data types private. Some users in cluster 4 tend to restrict their personal information to friends and to keep other data public. In turn, clusters 2 and 3 have users with various privacy preferences per each data type. Majority of users in cluster 2 set their privacy settings to friends. Users in cluster 3 prefer to keep private personal information and vote intention private, and other data to set up visible to public. The graphical representation of each cluster centroid, as well as the example of fuzzy user privacy profile is presented in the next section. The advantage of fuzzy clustering is that it shows to what extent the user posses intrinsic features of each cluster. This infomation can improve the accuracy in the classification of multidimensional user privacy profiles and avoid a discriminative sharp classification.

4.2 Fuzzy user privacy profile

As mentioned before, fuzzy clustering can detect differentiated inclination of users' privacy preferences per data type. Figure 2 depicts the vector of each calculated *cluster centroid* and vectors of two user privacy profiles. As it shown below, fuzzy clustering algorithm assigned user-181 to the cluster 4 with the highest membership degree value m = 0.99. The privacy profile vector of the user-181 is perfectly aligned with the vector of the corresponding centroid, meaning that user-181 agrees across all dimensions with cluster 4 centroid, which has visibility preferences set to public. On the other side, the user-139 was also assigned to the cluster 4, but with the highest membership degree of m = 0.36, because he also belongs to the cluster 3 with the membership degree of m = 0.27, and to cluster 2 with m = 0.24.

As it is seen from the user-139 privacy profile, he does not agree with the cluster 4 on the privacy decision related to "MyActivity". In that case, cluster 4 would suggest to open this data types to public, whereas cluster 3 would recommend to open it only to friends of friends. Moreover, cluster 3 has more restrictive visibility preferences with regard to "VoteIntention" data type compared to the initial user's privacy decision. Based on that, one of the privacy suggestion might be that the user-139 can be recommended either to share "MyActivity" data to public according to cluster 1 and to restrict "VoteIntention" visible only to friends according to cluster 3.

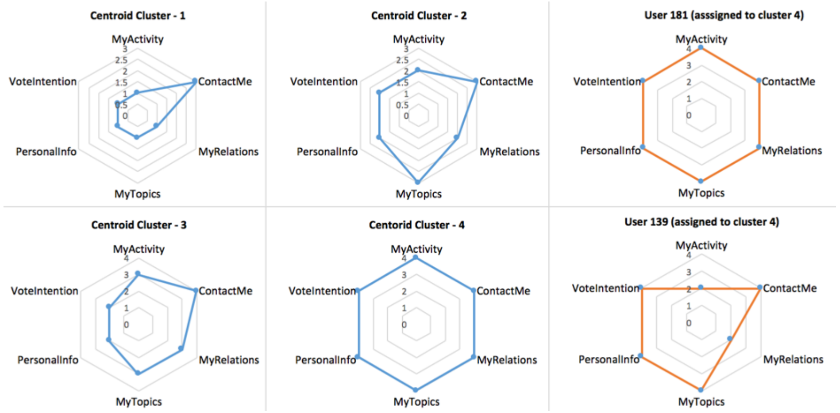


Fig. 2. Centroids' and users' privacy profile vectors

Knowing this additional insights about users' privacy preferences gives more options for the design of privacy suggestions as well as increasing its precision. Using sharp classification, saying, for example, that users in the cluster 4 are "privacy liberals", in our case makes an exact classification for the user-181, but simultaneously discriminates the user's-139 privacy opinion regarding other data types, which are more private. Thus, nuances of user privacy preferences can be well detected with the fuzzy clustering.

5 Conclusions

This work made several contributions. The advantage of having the real-world dataset of user privacy preferences allowed to unveil the existence of complexity of privacy decisions. Even in the 6-dimensional space, user privacy profile very rarely expressed an agreed privacy decision per different dimension. This demonstrates the multidimensionality of the user privacy behaviours. The multidimensionality itself entails to have a higher risk while trying to assign class labels. It becomes very hard to define a classification for the multidimensional privacy profiles, and if it is labelled based on a sharp classification, there is a risk of missing additional data, thus increasing the loss of classification accuracy. Therefore, the main contribution of this work is by applying a fuzzy data analysis to solve the issues of classification precision. It helps to detect the variance of the user privacy profile, as a result providing more options on privacy suggestions and avoiding a discriminative labelling.

Several limitations of this work should be accounted. First of all, the context of the system played an important role on final results of the data analysis. Users privacy preferences on the political data might be more cautious and, therefore, more restrictive compared to privacy behaviors within social networking platforms. It was noticed that users who did not share their vote intention tend to hide also their personal information and topics of political interests. Instead, users prefer to be contacted by email rather than make public their political profiles. Secondly, our dataset was limited by the traditional “sharing matrix" in which users decide what data to be shared with whom. Apart from the sharing matrix behaviour, there could be analysed other different user privacy behaviours such as “selective sharing", “friend management", etc. [21]. Thirdly, the data analysis was conducted on 391 real user profiles, however a larger dataset may contain additional information to be inferred. In addition, the dataset represented the population of only Ecuador citizens, which could also influence the results of the data analysis. From the technical perspective, the fuzzy c-means clustering has been conducted using only Euclidean distance metric. Though, a comparison of fuzzy clustering results based on different metrics would give additional analysis insights on user privacy preferences.

The future work will imply the design of the engine that will generate user privacy suggestions, in particular, based on the fuzzy classification rules. We plan to implement a fuzzy inference system, that will calculate privacy suggestions for the multidimensional user privacy profiles. After, a user-centric evaluation framework will be developed.

Acknowledgments. The author would like to thank Prof. Dr. Andreas Meier, Prof. Dr. Luis Teran, Marcel Wehrle for the collaboration during the development of the platform. Special thank to Dr. Bart Knijnenburg, Abhilash Menon, Oleksii Tyshchenko and Lorena Recalde for providing with useful insights and constructive feedbacks that improved the manuscript.

References

1. Palen L., and Dourish P.: Unpacking privacy for a networked world. Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, (2003)
2. Awad, N.F. and Krishnan, M.S., 2006. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. MIS quarterly, pp.13-28 (2006)
3. Guo, S. and Chen, K., 2012, September. Mining privacy settings to find optimal privacy-utility tradeoffs for social network services. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom) (pp. 656-665). IEEE (2012)
4. Madejski M., Johnson M., and Steven M. Bellovi: A study of privacy settings errors in an online social network. In Pervasive Computing and Communications Workshops (PERCOM Workshops), IEEE International Conference on, pp. 340-345. IEEE, (2012)
5. Leslie J. K.: The consumer psychology of online privacy: insight and opportunities from behavioral decision theory. The Cambridge Handbook of consumer psychology (2016)
6. Hui K.L., Tan CY B., and Goh CY.: Online information disclosure: Motivators and measurements. ACM Transactions on Internet Technology (TOIT), pp. 415-441, ACM (2006)
7. Marshall N. J.: Personality correlates of orientation toward privacy. Proceedings of the 2nd Annual Environmental Design Research Association Conference, Carnegie-Mellon University, Pittsburgh (1970)
8. Pedersen D. M.: Relationship of personality to privacy preferences. Journal of Social Behavior & Personality (1987)
9. Quercia, D., Las Casas, D., Pesce, J.P., Stillwell, D., Kosinski, M., Almeida, V. and Crowcroft, J., 2012, May. Facebook and privacy: The balancing act of personality, gender, and relationship currency. In 6th International AAAI Conference on Weblogs and Social Media. (2012)
10. Kuo, T. and Tang, H.L., 2013, July. Personality's influence on Facebook's privacy settings: A Case of college students in Taiwan. In International Conference on Human Aspects of Information Security, Privacy, and Trust (pp. 127-134). Springer Berlin Heidelberg.(2013)
11. Knijnenburg, B.P., Kobsa, A. and Jin, H., 2013. Dimensionality of information disclosure behavior. International Journal of Human-Computer Studies, 71(12), pp.1144-1162 (2013)
12. Liu K., and Terzi E.: A framework for computing the privacy scores of users in online social networks. ACM Transactions on Knowledge Discovery from Data (TKDD) vol.5, pp.6, (2010)
13. Ghazinour K., Matwin S., and Sokolova M.: YOURPRIVACYPROTECTOR, A recommender system for privacy settings in social networks. arXiv preprint arXiv:1602.01937 (2016).
14. Acquisti A., Leslie K. J., and Loewenstein G.: The impact of relative standards on the propensity to disclose. Journal of Marketing Research vol. 49, pp.160-174, (2012)
15. Kaskina, A. and Meier, A., 2016, March. Integrating privacy and trust in voting advice applications. In eDemocracy eGovernment (ICEDEG), 2016 Third International Conference on (pp. 20-25). IEEE. (2016)
16. Zadeh, L. A.: Fuzzy sets. Information and control vol. 8, no. 3, pp. 338-353.(1965)

17. Bezdek, J.C., Ehrlich, R. and Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Computers Geosciences, 10(2-3), pp.191-203 (1984)
18. Bezdek, J. C.: Cluster validity with fuzzy sets. pp.58-73, (1973)
19. Fan, J.L., Wu, C.M. and Ma, Y.L., 2000. A modified partition coefficient. In Signal Processing Proceedings. WCCC-ICSP 2000. 5th International Conference on (Vol. 3, pp. 1496-1499). IEEE (2000)
20. Pal, N.R. and Bezdek, J.C., 1995. On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy systems, 3(3), pp.370-379 (1995).
21. Wisniewski, P.J., Knijnenburg, B.P. and Lipford, H.R., 2017. Making privacy personal: Profiling social network users to inform privacy education and nudging. International Journal of Human-Computer Studies, 98, pp.95-108 (2017)
22. Egelman S. and Peer, E.: Predicting privacy and security attitudes. ACM SIGCAS Computers and Society, 45(1), pp.22-28 (2015)
23. Schrammel, J., Köffel, C. and Tscheligi, M., 2009, September. Personality traits, usage patterns and information disclosure in online communities. In Proceedings of the 23rd British HCI group annual conference on people and computers: celebrating people and technology (pp. 169-174). British Computer Society. (2009)
24. Ferraro, M.B. and Giordani, P., fclust: an R package for fuzzy clustering.