Revisiting Local Walking Based on Social Network Trust (LWSNT): Friends Recommendation Algorithm in Facebook Social Networks

W Nurkarim[1], **A W Wijayanto**[1,2]

[1] Politeknik Statistika STIS, Jl. Otto Iskandardinata No.64C, Jakarta, Indonesia
[2] BPS-Statistics Indonesia, Jl. Dr. Sutomo 6-8, Jakarta, Indonesia

*Corresponding author's e-mail: 221810650@stis.ac.id, ariewahyu@stis.ac.id

Abstract. In the last decades, the internet penetration rate and online social network users have grown very fast. Online social network, such as Facebook, is a platform where one can find friends without having to meet face to face. A social network is represented by a large graph because it involves many participants. Hence, it is hard to find potential friends who have the same thoughts and interests. The Local Walking Based on Social Network Trust (LWSNT) algorithm is one of the popular algorithms for social friend recommendation. This study re-examines whether the correlation between attributes gives un-match ranks in different cases (cases with and without correlation). We assess the performance of LWSNT in Facebook networks under the supervised manner by comparing its F-score against similar methods. By using Kendall's tau correlation, the results show that the correlation of attributes has no significant effect on the order of friend recommendations. In addition, the LWSNT performance is quite inferior against the Common Neighbors algorithm and Jaccard index.

1. Introduction

In recent years, the internet has grown rapidly to connect people across different countries and continents. People not only live in the real world but can also be in the virtual world. Social media is a platform where one can find friends without having to meet face to face. Likewise in the real world, the virtual world also allows someone to connect with each other. Thus, if we draw a line that represents the relationship among internet users, a social network can be created in graph form.

A social network comprising many participants is represented by a large graph [1], in which suitable friends (potential edges) are difficult to identify. FOF (friend of a friend), one of the most popular friend recommendation algorithms, selects friends based on the number of mutual friends among users by focusing on the relationship between users and ignoring the effect of the attributes of each user. Yu et al [2] introduced a new algorithm that combines a social graph with the influence of the attributes of each user, which is called the LWSNT (Local Walking on Social Network Trust) algorithm. It utilizes each user's attribute information to find potential friends with similar interests.

However, the LWSNT algorithm assumes that each attribute is mutually independent and ignores any correlation between attributes. For example, a woman tends to have a hobby in cooking, which means the age and gender attributes tend to determine a person's interests. Hence, a crucial question arises, "How the combination of attributes that correlate will affect when we calculate the user's trust index?". In this paper, we re-examine the LWSNT algorithms in two ways. First, we examine whether the correlation between attributes has a significant effect on the order of top-N friends who will be recommended to the target user. Second, we also investigate the performance in the real-world Facebook social network against competitor algorithms under the supervised settings. We compare the accuracy of LWSNT with the two popular methods: the Common Neighbors and Jaccard index.

2. Related Works

Murad, et al [1] introduced a new algorithm to help someone find a suitable friend on the friends list using the graph colouring method. For instance, a manager will select a candidate for a job based on the characteristics of the applicants. When a manager knows the characteristics of a person, then the manager will decide which one will be the best candidate to be selected. The algorithm of Chin et al [3] selects friends based on the proximity (physical) and interest features. They found that the choice of a friend was influenced by previous meetings and encounters.

A social network can grow into a huge size. User's personal information can be used as an indicator to see the interest of the user. Lee & Brusilovsky [4] shows that social connections indicate similarities between users depending on the strength of the connection between these users. This similarity is greater when users are directly connected and decreases with the distance between users. In social networks, trust is an effective way of assessing human interaction both offline and online [5]. Gambetta, a sociologist, states that trust is a subjective assessment of one agent to judge another agent [6]. Yu et al [2] introduced an algorithm to recommend friends by combining social graphs with user attribute information to calculate the degree of trust of social network users. From this study, the results obtained that 1) In the Sina Weibo microblogging network, the LWSNT algorithm is generally superior to the Common Neighbors (CN) and Jaccard algorithms, 2) The performance of this algorithm is affected by the number of friends of the target user, more friends used, the better the performance of this algorithm.

3. Research Methodology

3.1. Research Framework

Recall that in this study, we want to find an attribute correlation effect on the order of friend recommendations. Based on the notation of Berndtsson [7] and Polancik [8], the following diagram shows the research framework of such goal.

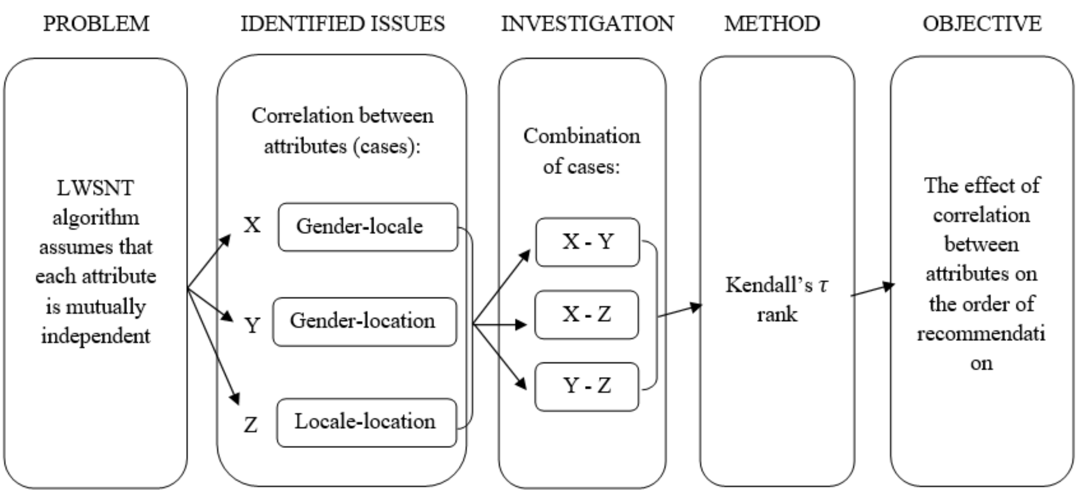


Figure 1. Research Framework

3.2. Research Methods

Relationships between users can be represented using graphs. When the relationship is connected in two directions, the graph used is undirected [9,10]. Meanwhile, when there exist possibilities of one-way relationship (a user may not follow the other), then a directed graph is used. In this study, we use undirected graphs in which every relationship is symmetric, that is A is related (or not) to B in exactly the same way as B is related to A, where nodes represent the member of a network and an edge indicate relationships. In the following, we describe the research methods used in this study.

3.2.1. Breadth First Search (BFS) Algorithm. The BFS algorithm is a search algorithm that explores the nodes in a wide manner and exploring all possible actions at one level. This algorithm will be used in the LWSNT algorithm to find friends who will be recommended (potential friends). The search is carried out from the target user node, then continued with the neighbouring node. Such mechanism is contradict to other search algorithm, such as the Depth First Search (DFS) as shown in Figure 2.

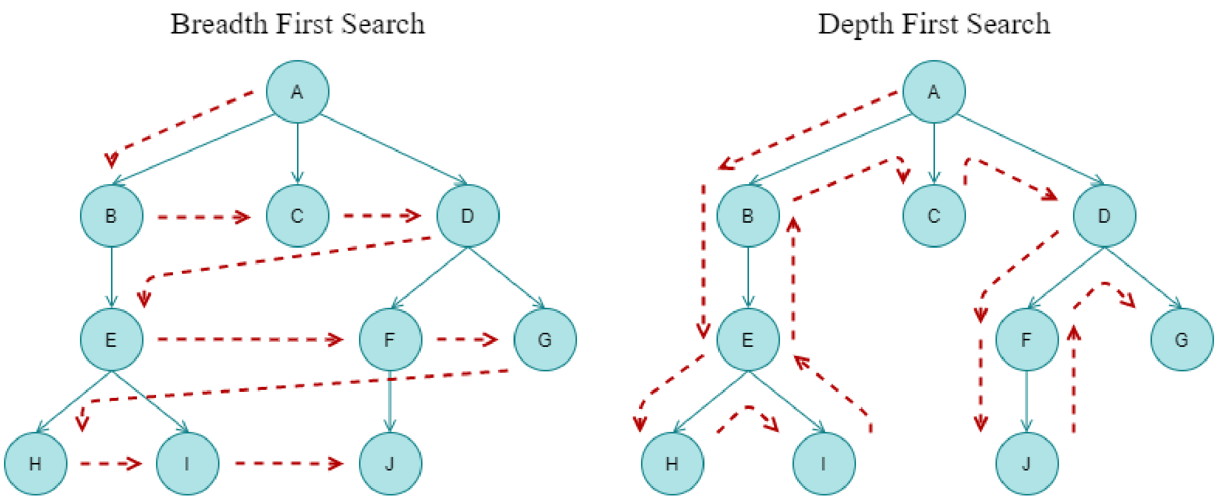
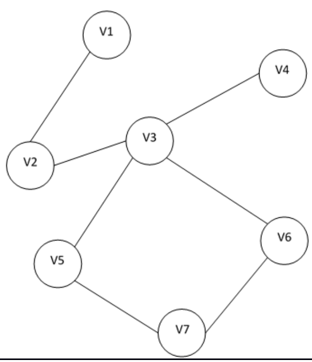


Figure 2. Comparison of Breadth First Search (BFS) and Depth First Search (DFS) node exploration and searching algorithm

3.2.2. Local Walking Based on Social Network Trust (LWSNT) Algorithm. The LWSNT algorithm is a friend recommendation algorithm that combines a social graph with the method of calculating the trust degree of the target user which is calculated using attribute information, such as age, gender, hobbies, and so on. According to [2], this algorithm has the following stages.

Table 1. Step of LWSNT Algorithm

Step	Procedures
Step 1	Create a social graph based on existing relationships/edges. An example of a simple social graph can be seen in Figure 3. 

Step	Procedures
	Figure 3. Simple social graph
Step 2	Calculate the trust index of each attribute of the target user (yi) using the attribute information used with equation (1): $$y_i(x_i = value) = \frac{s_i(x_i = value)}{\lvert F(v)\rvert} \quad (1)$$ where y_i is the trust index at each node, x_i is the value of each variable, s_i is the number of friends of the target user who have the i-th attribute value, |F(v)| : the total number of friends of the target user. For example, target user A has 10 male friends and 30 female friends, then y_i(x = female) = 30/40 = 0.75. And so does x = male.
Step 3	Explore each node using the Breadth First Search algorithm to find candidate friends to recommend. Then calculate the trust index for each candidate (Yid) with equation (2) : $$Y_{id} = \alpha \text{ x } \sum_{1}^{c}\left(z_i \text{ x } \frac{1}{\Sigma z_i} \text{ x } \frac{s_i(x_i = value)}{\lvert F(v)\rvert}\right) \quad (2)$$ With α : $$\alpha = \frac{k - t_{id}}{k} \quad (3)$$ where Y_{id} is the trust index of each candidate, α is the trust attenuation factor, k is the level value used in node browsing (according to the six-dimensional theory, the value k ranges from 2 to 6), t_{id} : level of walking steps each candidate friend to target user, and z_i is the weight given to each attribute with default value = 1
Step 4	Order Yid from largest to smallest. So that the elements in the first N are top N friends who will be recommended to the target user.

3.2.3. Cramer's V. Cramer's V and Phi is the most popular of Chi-Square-based measure of two categorical variables association. The difference between the two lies in the size of the contingency table used. Phi represents the association between two dichotomous variables, while Cramer's V is used to measure the association between two categorical variables when there is more than a 2 x 2 contingency table. Cramer's V can be calculated using the following formula: [12]

$$V = \sqrt{\frac{\chi^2}{n(M-1)}} \quad (4)$$

where V is the Cramer's V coefficient, χ^2 is calculated Chi-Square, n is the sample size, and M is the minimum number of rows or columns. Chi-Square test statistic is calculated as follows:

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(f_{bj} - f_{ej}\right)^2}{f_{ej}} \quad (5)$$

where χ^2 is the Chi-Square statistic, k is the number of cells, f_b is the observed absolute frequency within cell j, and f_e is the expected absolute frequency within cell j.

Cramer's V can be generalized for varying sizes because it is not affected by sample size. It can be useful in comparing multiple χ^2 test statistics which mean comparing several correlations between variables with each other. The value of Cramer's V ranges $0-1$. Interpretation of Cramer's V test are shown in the following Table 2 [12].

Table 2. Interpretation of Cramer's V

Estimate values	Interpretation
$\mathbf{0-0.1}$	Negligible
$\mathbf{0.1-0.2}$	Weak
$\mathbf{0.2-0.4}$	Moderate
$\mathbf{0.4-0.6}$	Relatively strong
$\mathbf{0.6-0.8}$	Strong
$\mathbf{0.8-1}$	Very strong

3.2.4. Kendall's τ Correlation. The Kendall's τ correlation is a non-parametric statistic to test the relationship between two variables with data in the form of scores that can be ranked, at least ordinal measurement variable. The null hypothesis on this test is that there is no relationship between the two variables [13].

$$\tau = \frac{S}{n\,(n-1)/2} \tag{6}$$

where τ denotes Kendall's τ correlation coefficient, S is the number of scores in the order of fairness of the data pairs on one of the variables, and n is the number of pairs X and Y (item) being ranked.

Kendall's τ is extension of Spearman's rank, where it be used when the same rank is repeated too many times in a small dataset [14]. Spearman's rank correlation is the also non-parametric method to measures the strength of association between two ranked variables. Compared to the Spearman's rank correlation, Kendall's is used less frequently. Spearman's rank correlation can be measured by the formula:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \tag{7}$$

where ρ denotes Spearman's rank coefficient, d is the difference between the two ranks of each observation, and n is the number of observations.

Both coefficients can measures the correlation in the same type of information. In general, they attain different values. Both of them have sensitivity to detect existence of correlation. In Spearman's rank, the smaller number of samples, the more r_s deviates from the actual value. Kendall's τ does not provide a biased estimate of the true value [11].

3.2.5. Common Neighbors and Jaccard Index Method. As comparison methods, this study uses the methods based on node similarity which takes the common neighbour as the consideration, like Common Neighbors [15] and Jaccard index [16]. Common Neighbors measure the node similarity only by considering the numbers of friends in common, while Jaccard index in addition to the number of friends in common also considers the total number of friends for two users. Each method as shown in Equation below:

Common Neighbors :

$$sim(u, v) = |\rho(u) \cap \rho(v)| \tag{7}$$

Jaccard index :

$$sim(u, v) = \frac{|\rho(u) \cap \rho(v)|}{|\rho(u) \cup \rho(v)|} \tag{8}$$

where $\rho(x)$ is a set of friends from the user *x*.

4. Results and Discussion

This study uses the Facebook network data which is publicly available on the Stanford website: https://snap.stanford.edu/data/ego-Facebook.html. The dataset has 4,039 nodes, 88,234 edges, and 1,406 attributes. For user security protection, this dataset uses the anonymous feature. For instance, the original dataset may have contained a feature "gender = male", then the new data would simply contain "gender = anonymized feature 77". Thus, using the anonymized data it is possible to determine whether particular attribute has the same value for any two members, but not the value of this attribute.

4.1. Evaluation of Friend Recommendation Criteria of LWSNT Algorithm

Trust index is degree of trust between the target user and all his candidate friends. The target user as a user who will be given a friend recommendation is a user with ID 0. Then a friend candidate is explored using the BFS algorithm. The level value used in node browsing (k) is 3, so BFS will browse each node in a wide range from V1 to V3. From a large dataset consisting of thousands of nodes and tens of thousands of edges, we reduce it to small data by only taking the target user's social network, i.e, user ID 0. So that we get the number of 150 nodes of friends (V1), 3 nodes of friends of friends (V2), and 103 nodes V3. In total, 257 nodes used, with 3 attribute information: gender, locale, and location. First, we calculated the correlation of the combination of attributes using Cramer's V. The correlation results are obtained as shown in the Figure 4.

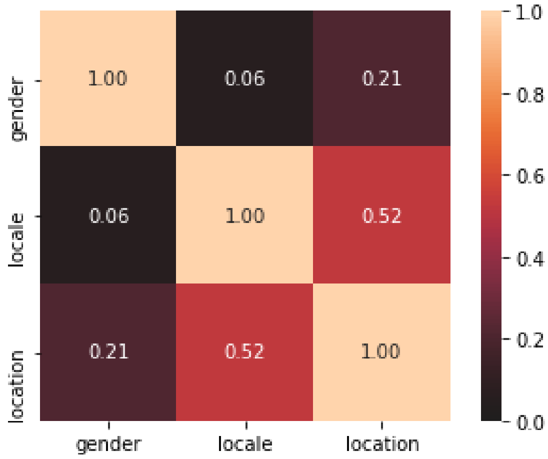


Figure 4. Heat Map of Correlation

X, Y, and Z each represent a combination of attributes: gender - locale, gender - location, and locale - location. Based on interpretation of Cramer's V test are shown in the Table 2, we can see that X has no correlation, Y has a weak correlation, and Z has a relatively strong correlation. Hence, there are three cases to be tested a rank match: X – Y, X – Z, and Y – Z. Then the BFS algorithm is used to explore nodes starting from vertex 0. Each node that is explored at each level is calculated for its trust index against the target user. The trust index is evaluated for each pair of attributes. The following shows some trust index values for each node in each combination of variables.

Table 4. Trust index values for each combination of variables

Node	Combination of variable		
	X	Y	Z
2	0.2575	0.2425	0.025
3	0.59	0.3475	0.4625
7	0.59	0.3475	0.4625
…	…	…	…
566	0.1633	0.053	0.125
569	0.1967	0.0858	0.1242
570	0.1967	0.0792	0.1175
1025	0.1967	0.0867	0.125

Nodes that have the same index value may also have the same features value. That is because index calculation uses the attribute value for each user, multiplied by a given weight. The Kendall Tau correlation test will see the relationship of 3 cases that have been defined previously based on the ranked of the data. The result of Kendall's τ test as shown the table below.

Table 5. Results of the Kendall's test

Combination of cases	P-value	Conclusion	Correlation
X and Y	0.0	Rank match	0.754
X and Z	0.0	Rank match	0.683
Y and Z	0.0	Rank match	0.755

From the above table, the three pairs have shown the same conclusion, i.e., that there is a rank match. When calculating the relationship of ranked data between the uncorrelated X and the weak correlated Y, it found that they both has the same rank. This also applies to Z which has a strong correlation. It means that there is or no a correlation does not affect the order of friend recommendations from the LWSNT algorithm. Therefore, when entering the correlation into the calculation of the trust index or simply assuming that the variables are independent of each other, there is no significant effect on the order of friend recommendations.

4.2. Evaluation of Friend Recommendation Accuracy of LWSNT Algorithm

To measure the accuracy performance of LWSNT algorithm against other algorithms in real-world Facebook network dataset under the supervised setting, we use the weighted accuracy, that is F-score metric as an evaluation criterion. Using data on 68 target users, and we take test data as much as 10% of friends from the target user and the remaining 90% is train data. Then, we run the algorithm to get a list of friends' recommendation results on top-N. The list of recommendation results is then searched for compatibility with the test data that has been defined at the beginning. The following table shows the F-score of each algorithm at a different number of N.

Table 6. The F1 value of three kinds algorithms when the number of $|\boldsymbol{F}(\boldsymbol{v})|>\mathbf{80}$

	N = 1	N = 2	N = 5	N = 10
LWSNT	0.035	0.066	0.126	0.152
Common Neighbors (CN)	**0.154**	**0.289**	0.509	0.622
Jaccard	0.149	0.277	**0.516**	**0.638**

The above table shows the performance of the 3 algorithms when the target user has many friends. The table shows that the more the number of recommendations, the better the performance generated for the three algorithms. Jaccard's performance is the best, little difference from the Common Neighbors. CN is the best algorithm at N very small, whereas Jaccard have best performance at N greater. In contrast, LWSNT has the poorest performance in all setting, because it has the lowest F-score per N compared to the CN and Jaccard algorithms.

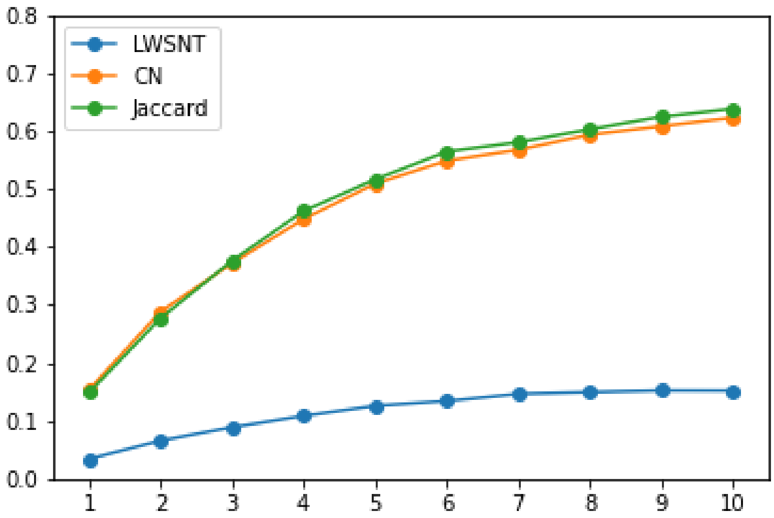


Figure 5. Comparison of F-score between the three algorithms

Through the experimental evaluation, the following remarks can be drawn. First, each algorithm performs better with more recommendations. Second, Common Neighbors (CN) and Jaccard index are superior performance to the LWSNT algorithm in the Facebook network.

5. Conclusion

In this study, we re-examine the performance of LWSNT in real-world Facebook online social network dataset by examining whether the correlation between attributes gives un-match ranks in different cases (cases with and without correlation). To measure the performance of LWSNT in

Facebook networks under the supervised manner, the comparison of F-score metric against similar methods is conducted.

By applying the non-parametric correlation test to 3 difference combinations of attributes, the results show that there is a rank match between the combinations. In other words, the correlation does not give a significantly different order of results. Hence, it can be concluded that when the user has attribute information more than one, where there will be a possible correlation between attribute, it does not have a significant effect on the order of friends' recommendations.

Additionally, the performance of LWSNT algorithm in friend recommendation is greatly affected by the structure and type of graph constructed from the social networks. In the Facebook social network, the LWSNT has inferior performance compared to the node similarity-based methods such as the Common Neighbors (CN) and Jaccard index recommendation. Further improvement to the LWSNT friend recommendation criteria is expected and left to our future work by incorporating the low-dimensional features of graph structure into consideration.

References

[1] Murad, O., Sleit, A., & Sharaiah, A. (2016). Improving friends matching in social networks using graph coloring. *International Journal*, *15*(8).

[2] Yu, L., Yang, Y., & Wu, X. (2016). A Friend Recommendation Algorithm Based on Social Network Trust. *International Journal of Engineering and Technical Research*, *7*(4).

[3] Chin, A., Xu, B., & Wang, H. (2013, May). Who should I add as a" friend"? A study of friend recommendations using proximity and homophily. In *Proceedings of the 4th International Workshop on Modeling Social Media* (pp. 1-7).

[4] Lee, D. H., & Brusilovsky, P. (2010, June). Social networks and interest similarity: the case of CiteULike. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (pp. 151-156).

[5] Lu, L., Yuan, Y., Chen, X., & Li, Z. (2020). A Hybrid Recommendation Method Integrating the Social Trust Network and Local Social Influence of Users. *Electronics*, *9*(9), 1496.

[6] Gambetta, D. (2000). Can we trust trust. *Trust: Making and breaking cooperative relations*, *13*, 213-237.

[7] M. Berndtsson, J. Hansson, B. Olsson, dan B. Lundell. (2008). *Thesis Projects - A Guide for Students in Computer Science and Information System 2nd Edition*. London: Springer.

[8] Polančič, G. (2007). *Empirical research methods poster*. Maribor: Faculty of Electrical Engineering and Computer Science, Institute of Informatics.

[9] Wijayanto, A. W., & Murata, T. (2017). Flow-Aware Vertex Protection Strategy on Large Social Networks. The 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (pp 58-63). ACM.

[10] Wijayanto, A. W., & Murata, T. (2018). Pre-emptive spectral graph protection strategies on multiplex social networks. Applied Network Science, 3 (1), 5. Springer.

[11] Lapata, M. (2006). Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, *32*(4), 471-484.

[12] Rea, L. M., & Parker, R. A. (1992). *Designing and conducting survey research*. San Francisco: Jossey-Bass Publishers.

[13] Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

[14] Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, *18*(3), 91-93.

[15] Golbeck, J. (2005). *Computing and applying trust in web-based social networks*. College Park, Md: University of Maryland.

[16] Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Lausanne: Impr. Corbaz.