Discovering Communities in Organizational Social Network through Hierarchical Clustering

Paweł Szyman[0000−0003−3353−5522] and Dariusz Barbucha[0000−0001−7906−0740]

Department of Information Systems
Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
p.szyman@wznj.umg.edu.pl, d.barbucha@umg.edu.pl

Abstract. Organizational social network is a kind of social network where people or units of an organization are represented in the form of vertices, and relationships or information flow between these people or units are represented in the form of edges between the vertices. By applying a methodology of social network analysis it is possible to analyze different properties of such network. The paper focuses on organizational social network of public organization located in Poland, where the structure of the network is based on digital forms of communication between people within an organization. The main contribution of the paper is to detect communities in such constructed network by applying dedicated hierarchical clustering method. The results of the computational experiment allowed one to validate detected communities and compare them with real structure of the organization.

Keywords: Organizational social network · Community detection · Hierarchical clustering · Similarity

1 Introduction

Social network analysis offers a number of interesting approaches to the analysis of different aspects of an organization and its activity. It assumes that people or units of the organization are represented in the form of vertices in the network, and relationships or information flow between these people or units are represented in the form of edges between the vertices. Although an organizational social network may have different forms depending on a source and a form of information used to its construction, one of the most representative is a social network based on different digital forms of business communication within an organization, especially using electronic mail channel. In such network the edge between vertices represents interaction in terms of messages sent or received [3].

This paper focuses on the organizational social network of one of the public organizations, which has been constructed by extracting the anonymized data referring to email communication between persons employed in this organization. The main goal of the paper is to detect a number of cohesive communities in the network, where a community can be generally defined as a group of vertices

similar to each other [5]. As a part of the analysis, the resulting communities have been compared with real structure of the organization.

The paper is organized as follows. Section 2 defines a problem of community detection and presents a hierarchical approach to solve the community detection problem in organizational social network. Section 3 presents the results of computational experiment carried out to validate the proposed approach. Section 4 concludes the paper and presents directions of the research planned in the near future.

2 Community detection in organizational social network

2.1 Community detection problem

Let $G = (V, E)$ be an undirected graph, representing a social network, where V represents a set of vertices, and E - set of edges between vertices. A subset of vertices within the graph $C \subset G$ such that connections between the vertices are denser than connections with the rest of the network is called *community*. The problem of *community detection* aims at dividing the vertices of a network into some number k of groups, while maximizing the number of edges inside these groups and minimizing the number of edges established between vertices in different groups [4]. Detailed review of the methods used to solve the problem of community detection, the reader can find for example in [5].

2.2 Hierarchical clustering approach to the community detection in organizational social network

The proposed approach (HC) to solve the community detection problem is based on agglomerative hierarchical clustering [5]. The starting point of any hierarchical clustering method is the definition of the similarity measure between vertices. In case of social networks the similarity must be inferred from the adjacency relationships between vertices. In the proposed approach the similarity measures the overlap between the vertices $i, j \in V$ given by the ratio between the intersection and the union of the neighborhoods, i.e. $s_J(i,j) = |\Gamma(i) \cap \Gamma(j)|/|\Gamma(j) \cup \Gamma(j)|$, where $\Gamma(i)$ and $\Gamma(j)$ denote the set of neighbors of vertices i and j, respectively. As a consequence, having the similarity calculated for each pair of vertices, no matter if they are connected or not, a new similarity matrix S_J is created. Due to the fact that proposed hierarchical clustering approach belongs to the agglomerative algorithms group, it has been also essential to define a measure that estimates how similar clusters are, out of the similarity matrix S_J. After a preliminary tests with different measures, the average linkage has been adapted.

3 Computational experiment

Computational experiment has been carried out in order to discover communities in the social network based on e-mail communication between employees

of the public institution through hierarchical clustering and compare detected communities with the organizational structure of the institution.

The experiment was divided into four steps. The first step was to select the observation period (six months in the experiment), a number of investigated departments (single department with the biggest number of employees - 91) and then extraction and collection of the data from organization's e-mail server logs. Total number of identified messages between the employees was equal to 10 364.

The second step was focused on building a network, where employees were represented by vertices and communication between them - by edges. It has been assumed that an edge between two vertices exists if two employees represented by these vertices exchanged at least a single message within observed period. In order to ensure the security of employees' personal data, each e-mail address was anonymized by assigning to it an individual number from 1 to 91.

The third step focused on applying the proposed hierarchical approach to discover communities in the social network built in the previous step. The proposed algorithm has been implemented using *R* software package with the *iGraph* and *LSA* libraries. In the fourth step, the obtained results (presented in Table 1 and Fig. 1) were analyzed and the communities detected in the network by HC algorithm were compared with the structure of the organization.

The result of hierarchical clustering in the form of a dendrogram is presented in Fig. 1. One can observe that 5 communities were detected in the network. In order to confirm this observation, it has been also decided to calculate the value of the silhouette index for different number of clusters. The best value (0.3265) has been obtained for 5 clusters. What is important, these results correctly refer to real organizational structure, where 5 units form the investigated department.

In the next step, the structure of detected clusters has been compared with the structure of organizational departments. Table 1 presents all detected clusters by HC algorithm with information about the number of vertices belonging to

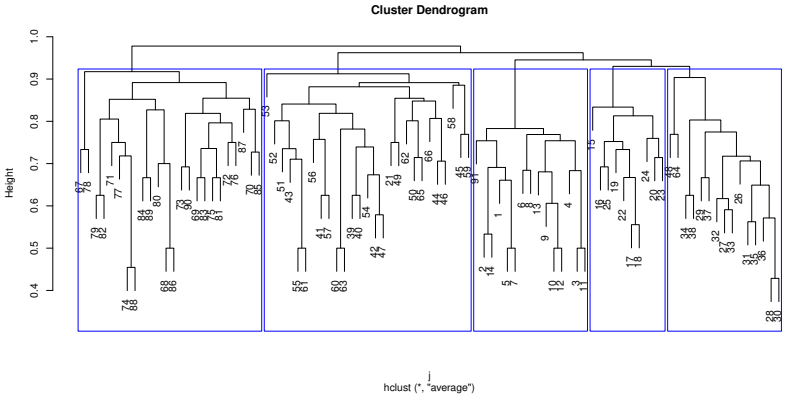

Fig. 1. Dendrogram showing clusters detected by HC algorithm.

them (in brackets) with a detailed list of these vertices. The number of vertices correctly assigned to clusters (when compare to the structure of the department) was 87 out of 91 (96%). In case of 4 out of 91 vertices (4%) they have been assigned incorrectly (presented in bold in Table 1 - 22, 49, 65, 68).

Table 1. Division of the network into communities by HC algorithm.

Clusters	Allocation of vertices to the communities
C1 (15)	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
C2 (10)	16 17 18 19 20 21 23 24 25 26
C3 (15)	27 28 29 30 31 32 33 34 35 36 37 38 39 **49 65**
C4 (27)	**22** 40 41 42 43 44 45 46 47 48 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 66 67
C5 (24)	**68** 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91

In the last step of the analysis, results obtained by HC algorithm have been compared with the results obtained by Louvain method [2]. In one of the last Authors' paper [1], it has been shown that Louvain method can efficiently detect communities in studied social network. It turned out that division the network into communities obtained by both methods is similar. Moreover, the value of modularity index calculated for communities detected by both algorithms is almost the same (Louvain method - 0.504965, HC - 0.491054).

4 Conclusions

Hierarchical clustering method with dedicated similarity measure has been used to solve the problem of community detection in organizational social network based on e-mail communication. Computational experiment allowed one to discover communities which in most cases refer exactly to the structure of the organization. An interesting direction of the future work will be an investigation of different similarity measures between vertices as a fundamental part of different clustering methods dedicated to community detection problems.

References

1. Barbucha, D., Szyman, P.: Detecting Communities in Organizational Social Network Based on E-mail Communication. In: Czarnowski, I. et al. (eds.) Intelligent Decision Technologies. Smart Innovation, Systems and Technologies, vol 238, pp. 15–24. Springer, Singapore (2021).
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**, P10008 (2008).
3. Christidis, P.: Intensity of Bilateral Contacts in Social Network Analysis. Information **11**, 189 (2020).
4. Coscia, M., Giannotti, F., Pedreschi, D.: A classification for community discovery methods in complex networks. Stat. Anal. Data Min. 4 (5), 512-546 (2011).
5. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Phys. Rep. **659**, 1-44 (2016).