Evaluation of organizational structure through cluster validation analysis of email communications

Veselka Boeva¹ · Lars Lundberg¹ · Sai M. Harsha Kota¹ · Lars Sköld²

Received: 4 May 2018 / Accepted: 24 July 2018 / Published online: 6 August 2018
© The Author(s) 2018

Abstract

In this work, we apply cluster validation measures for analyzing email communications at an organizational level of a company. This analysis can be used to evaluate the company structure and to produce further recommendations for structural improvements. Our evaluations, based on data in the forms of email logs and organizational structure for a large European telecommunication company, show that cluster validation techniques can be useful tools for assessing the organizational structure using objective analysis of internal email communications, and for simulating and studying different reorganization scenarios.

Keywords Cluster validation measures · Data analysis · Human capital management · Internal communication · Organizational structure

Introduction

Finding an efficient organizational structure is a key aspect in human capital management. In addition, communication within an organization plays an important role for its success. Most of the work that people do in organizations requires some degree of active cooperation and communication with others [18]. Namely, individual members of groups need to communicate with each other to accomplish their

✉ Veselka Boeva
veselka.boeva@bth.se

Lars Lundberg
lars.lundberg@bth.se

Sai M. Harsha Kota
sakf16@student.bth.se

Lars Sköld
lars.skold@telenor.se

¹ Computer Science and Engineering Department, Blekinge Institute of Technology, Karlskrona, Sweden

² Telenor, Sverige, Stockholm, Sweden

production and social functions and, within organizations, groups need to communicate with other groups. On the organizational level, communication can be divided into internal and external communication. Internal communication is communication among employees [29], whereas external communication is the one focusing on the audiences outside of the organization [27]. Kim has examined in [15, 16] the direct and indirect influences of organizational structure and internal communication on employee–organization relationships. The results of these studies have shown that the organizational structure and internal communications are associated with employee–organization relationships. In addition, recent studies in public relations have shown that organizational structure has an effect on internal communication [13, 20, 21, 23]. Therefore, it would be useful to be able to perform objective comparisons and evaluations of different organizational structures based on reliable metrics of the effectiveness of internal communication.

In this paper, we pursue an analysis of the organizational structure of a company with respect to the internal communication at the organizational level by applying different cluster validation techniques. The internal communications can take many forms, such as face-to-face conversations, formal meetings, phone calls, and emails, i.e., each company has a complex social network. This social network can be used to evaluate the company structure and, based on this evaluation, one can recommend some changes that are expected to improve the company structure. In our considerations, the internal communications are represented by email interactions between employees. A study based on data in the forms of email logs and organizational structure for a large European telecommunication company is presented for evaluation and validation of the purposed approach.

Notice that the work presented in this paper is an extended study based on the initial results published in [3]. In the current work, we have studied and validated additional reorganization scenarios. In addition, the proposed approach has been evaluated and validated on a new richer dataset representing internal email communication in the company. We have also applied different pre-processing techniques on the email communication data in order to obtain a more realistic representation of the intensity of the employee interaction. The bibliography and related work section have also been extended with more recent works on the studied problem.

Background and problems

Reorganizations are done to improve quality and efficiency and to reflect the need from the outside world. To control and maximize the use of an organization's human capital, one would like to provide objective measures of how well a certain organization works. One important aspect for obtaining good use of human capital is that the formal organization in terms of divisions, departments, groups, etc. reflects the structure needed to perform the tasks in the organization (see Sect. 3 for more details), i.e., one should be able to handle most of the work tasks within the organizational units.

Performing tasks in a modern organization normally involves email communication between the employees. This means that the email communication graph, where

the nodes are employees and the weights of the arcs reflect the frequency of the email communication between the employees, provides a good estimate of clusters of people performing tasks together. By analyzing and comparing such communication clusters (the informal and invisible organizational structure) with the formal organizational structure in terms of divisions and departments, we can see how well the formal organizational structure maps to the informal structure reflected by the email communication. Such analysis may be important to extract some problems in the company, like employees within some organizational units do not communicate between each other. This may also allow the management to answer the question about proper employee allocation in the organization they are responsible for. There are a number of cluster validation metrics that can be calculated to get a quantitative and objective measure of the fit between the formal and informal organizational structure, e.g., Silhouette Index [25] and Connectivity [8] measures. This means that these kinds of cluster validation measures can be used for comparing the quality of different organizations.

Calculating such cluster validation metrics can be useful in a number of situations, e.g., when:

- comparing the quality of two organizations, e.g., the new and old organization after a reorganization;
- comparing the current organizational structure with a simulated partitioning of employees that optimizes the metrics (this gives an estimate for the possible room for improvement);
- checking the discrepancy between the organizational formal and informal structures;
- fine tuning an existing organizational structure by moving a limited number of employees or a whole team of employees from one organizational unit to another.

Related work and contribution

The term organizational structure refers to the formal configuration between individuals and groups regarding the allocation of tasks, responsibilities, and authority within the organization [4, 5]. Some studies in public relations have shown that organizational structure has an effect on internal communication. For instance, the influence of organizational structure on organizational communication was well illustrated in [6, 7, 10, 32]. Holtzhausen's survey research [10] conducted in a large South African organization found that structural changes in process implementation led to improved information flow and face-to-face communication. More specifically, the research showed that addressing the internal communication process from a strategic perspective with subsequent structural changes to enhance that strategy provided practitioners with a tool to improve information flow and change communication behavior in organizations.

The social network of a company can also be analyzed to increase the company effectiveness through discovering its hidden potential. As it was demonstrated in [13, 22], the discovered knowledge may lead to various positive effects in organization

management and architecture. Some more recent studies have used a social network extracted from organizational email communication to evaluate company structure and, based on this evaluation, recommend some changes that have to be made within a company to improve its structure [20, 21, 23]. In [20], the authors have applied some of the most popular metrics used in social network (*e.g.*, in-degree centrality, out-degree centrality, centrality betweenness) to compare ranks of social network position to organization structure. The ideas proposed in [21, 23] describe the effectiveness of matching organizational structure and social network, extracted from email communication, and further possibilities to use those results as a base to redesigning company structure. Their analysis is based on the differences between employees' social scores within each division. Both studies used Enron email logs for evaluation and validation of their approaches. The Enron case became famous worldwide in 2001 due to a financial manipulation scandal and the Federal Energy Regulatory Commission made the Enron email dataset public during its investigation. Although the Enron official hierarchy is still not fully publicly available, there are some sources that provide information of job positions of the selected employees together with their division. The latter led to a number of studies which focused on analyzing the relationship between formal and informal structures [11, 26]. Bar-Yossef et al. proposed an interesting cluster ranking algorithm based on the strength of the clusters and they also applied it to the Enron dataset [2].

In our work, we do not use social metrics to examine the email communication network of a company, but we rather analyze the organizational structure with respect to the internal email communication applying different cluster validation techniques. Each research study is usually limited by the data availability. Namely, in order to evaluate the organization by analyzing its email communication network we also need organizational details. Organizations are not always opened to supply this information. In the current study, we use data in the forms of email logs and organizational structure for a large European telecommunication company to evaluate and validate the proposed approach. We show that cluster validation techniques can be useful tools for assessing the organizational structure using objective analysis of internal email communications and for simulating and studying different reorganization scenarios.

Methods

Cluster validation techniques

One of the most important issues in cluster analysis is the validation of clustering results. Essentially, cluster validation techniques are designed to find the partitioning that best fits the underlying data and should, therefore, be regarded as a key tool in the interpretation of clustering results. The data-mining literature provides a range of different cluster validation measures, which are divided into two major categories: external and internal [12]. External validation measures have the benefit of providing an independent assessment of clustering quality, since they evaluate the clustering result with respect to a pre-specified structure.

However, previous knowledge about data is rarely available. Internal validation techniques, on the other hand, avoid the need for using such additional knowledge, but have the problem that they need to base their validation on the same information used to derive the clusters themselves. Additionally, some authors consider a third approach of clustering validity, which is based on the relative criteria [9]. The basic idea is to evaluate the clustering result by comparing clustering solutions generated by the same algorithm but with different input parameter values. A number of validity indices have been defined and proposed for each of the above approaches. Furthermore, internal measures can be split with respect to the specific clustering property they reflect and assess to find an optimal clustering scheme: compactness, separation, connectedness, and stability of the cluster partitions. Compactness evaluates the cluster homogeneity that is related to the closeness within a given cluster. Separation demonstrates the opposite trend by assessing the degree of separation between individual groups. The third type of internal validation measure (connectedness) quantifies to what extent the nearest neighboring data items are placed into the same cluster. The stability measures evaluate the consistency of a given clustering partition by clustering from all but one experimental condition. The remaining condition is subsequently used to assess the predictive power of the resulting clusters by measuring the within-cluster similarity in removed experiment. A detailed summary of different types of validation measures can be found in [8, 31].

Since none of the clustering algorithms perform uniformly best under all scenarios, it is not reliable to use a single validation measure, but instead to use a few that reflect various aspects of a partitioning. In this sense, we have selected to use two internal validation measures that cover the first three categories from the above classification, for analyzing the internal communication at the organization level: Silhouette Index (SI) for assessing compactness and separation properties and Connectivity for assessing connectedness.

Silhouette Index [25] is a cluster validity index that is used to judge the quality of any clustering solution $C = \{C_1, C_2, \dots, C_k\}$. Suppose a_i represents the average distance from object i to the other objects of the cluster to which the object is assigned, and b_i represents the minimum of the average distances from object i to objects of the other clusters. Then, the Silhouette Index of object i can be calculated by

$$s(i) = (b_i - a_i) / \max\{a_i, b_i\}.$$

The overall Silhouette Index for clustering solution C of m objects is defined as:

$$s(C) = (1/m) \sum_{i=1}^m (b_i - a_i) / \max\{a_i, b_i\}.$$

The values of Silhouette Index vary from -1 to 1 and higher value indicates better clustering results.

Connectivity captures the degree to which objects are connected within a cluster by keeping track of whether the neighboring objects are put into the same cluster [8]. Define m_{ij} as the j th nearest neighbor of object i , and let $x_{im_{ij}}$ be zero

if i and m_{ij} are in the same cluster and $1/j$ otherwise. Then for a particular clustering solution $C = \{C_1, C_2, \dots, C_k\}$ of m objects and a neighborhood size n_r , the Connectivity is defined as:

$$\text{Conn}(C) = \sum_{i=1}^m \sum_{j=1}^{n_r} x_{im_{ij}}.$$

The connectivity has a value between zero and mH_{n_r} and should be minimized.

External validation measures can be two types: unary and binary [9]. Unary external evaluation measures take a single clustering result as the input and compare it with a known set of class labels to assess the degree of consensus between the two. Comprehensive measures like the F -measure provide a general way to evaluate this [30]. In addition to unary measures, the data-mining literature also provides a number of indices, which assess the consensus between a produced partitioning and the existing one based on the contingency table of the pairwise assignment of data items. Most of these indices are symmetric and are, therefore, equally well suited for the use as binary measures, i.e., for assessing the similarity of two different clustering results. Probably, the best known such index is the Rand Index [24], which determines the similarity between two partitions as a function of positive and negative agreements in pairwise cluster assignments.

We plan to use the F -measure as an external validation measure to match the clustering solutions generated by k -means (or other clustering algorithms) on the data presenting the internal communication at the organization level (i.e., the informal organizational structure) to the existing organizational structure that partitions the employees into working units.

The F -measure is the harmonic mean of the precision and recall values for each cluster [19]. Let us consider two clustering solutions $C = \{C_1, C_2, \dots, C_k\}$ and $C' = \{C'_1, C'_2, \dots, C'_l\}$ of the same data set. The first solution is a known partition of the considered data set while the second one is a partition generated by the applied clustering algorithm. The F -measure for a cluster C'_j is then given as:

$$F(C'_j) = \frac{2|C_i \cap C'_j|}{|C_i| + |C'_j|},$$

where C_i is the cluster that contains the maximum number of objects from C'_j . The overall F -measure for clustering solution C' is defined as the mean of cluster-wise F -measure values, i.e., $F(C') = 1 / \sum_{j=1}^l F(C'_j)$. For a perfect clustering, when $l=k$,

the maximum value of the F -measure is 1.

Analysis of organizational structure

Suppose we have a database D that contains email interactions among employees in a company for a given time period. Assume that n employees in total work for the company. Using the email database, we initially create an interaction matrix, where entry e_{ij} ($i, j = 1, 2, \dots, n$) represents the number of emails sent by employee i to employee j . The interaction matrix is normalized by dividing each e_{ij} ($i, j = 1, 2, \dots, n$) with the total number of emails, denoted by e_i , sent by employee i , i.e., each value en_{ij} in the normalized interaction matrix is equal to e_{ij}/e_i ($i, j = 1, 2, \dots, n$). Next, using the normalized interaction matrix we can construct an undirected graph, where vertices represent employees and edges represent communication between two employees. The graph can be represented by an adjacency matrix, where each entry of the matrix is a value $d_{ij} \in [0, 1]$ ($i, j = 1, 2, \dots, n$) expressing the distance (dissimilarity) between employees i and j . The distance d_{ij} ($i \neq j$) can be defined by the following equality

$$d_{ij} = 1 - 0.5(en_{ij} + en_{ji}). \quad (1)$$

In addition, $d_{ii} = 0$ by default, for all $i = 1, 2, \dots, n$. It can easily be shown that $d_{ij} \in [0, 1]$ for all $i, j = 1, 2, \dots, n$.

Notice that in [3], the distance matrix has been normalized by the maximum number of exchanged emails between two employees in the company. However, this can be misleading if there are two employees who exchange quite a lot emails in comparison with the others. This will reduce the distance for all employees. Therefore, in the current study we normalize each row i ($i = 1, 2, \dots, n$) of the interaction matrix by the total number of emails sent by employee i (e_i). Evidently, then the normalized value en_{ij} represents the relative intensity (weight) of communication of employee i with employee j with respect to the other employees in the company, i.e., $en_{i1} + en_{i2} + \dots + en_{in} = 1$ ($i = 1, 2, \dots, n$). In this way, in the normalized interaction matrix employees with very few email communications will have very close weights to ones of employees with many email communications. The latter is the necessary condition for our further analysis of the email communication data of the company, since we need to compare any two employees with respect to their relative level of interaction with the others.

Next suppose the employees are partitioned into k main working units (divisions, departments), i.e., the existing formal organizational structure determines a cluster partition. This cluster partition is considered as a known partitioning of the employees in the company and it is denoted by $C = \{C_1, C_2, \dots, C_k\}$, where C_i ($i = 1, 2, \dots, k$) presents an organizational division. The cluster validation measures introduced in the foregoing section can be used to evaluate whether this known partitioning of the employees into k divisions reflects a clustering (informal) structure present in the email interaction database D . For example, we can study whether the divisions are of a high quality, i.e., the “within” communications are high in comparison with the “between” communications. In addition, it would be useful to have an evaluation of the overall organizational structure. We can further investigate which employees appear to be well assigned (classified), which ones are wrongly assigned

(misclassified), and which ones lie in between divisions. We can also try to obtain an idea about the number of “natural” divisions that are present in the email interaction database.

The Silhouette Indices discussed in Sect. 4.1 can easily be constructed in the considered context and can be used to assess whether we have compact and clearly separated organizational units. In addition, the Connectivity scores can be used to evaluate whether employees are well connected within a division. In order to calculate SI and Connectivity scores, we only need the partitioning of the employees into the divisions and the collection of all distances between employees, i.e., the distance matrix. Thus, we can compare the performance of the different divisions with respect to the effectiveness of their email communication using SI and Connectivity validation measures.

The worst performing divisions (the ones with lowest SI and highest Connectivity scores) can further be analyzed by computing individual Silhouette Indices of their employees. The computed Silhouette Indices for such a division, e.g., division C_p , can be depicted in a plot. In addition, for each employee $i \in C_p$ we can find the second-best (or best in case of negative SI) choice (e.g., division C_r), i.e., if he/she could not be allocated to this division, which division would be the closest competitor. We can use the plot built for division C_p and look initially for employees whose SI scores are negative. Let us consider an extreme case when there is an employee $i \in C_p$ whose Silhouette Index value [denoted by $s(i)$] is close to -1 . This means that employee i is much closer to division C_r than to C_p , i.e., this person communicates much more with employees from division C_r than with people from his/her own division C_p . Therefore, it would have seemed much more natural to assign employee i to division C_r . In general, all such employees can be found in the worst performing divisions, and it could be recommended to allocate these employees to their closest division.

We have a different situation when $s(i)$ of employee $i \in C_p$ is about zero. Then, it is not clear whether employee i should have been assigned to C_p or C_r . If such situations are predominating in the analyzed divisions (i.e., most employees have silhouette indices that are close to zero), this could be an indication that the current organizational structure generally needs to be improved. This is especially valid when the overall Silhouette Index for the current organizational structure is close to zero or negative.

We can also examine the effect of migration of groups of employees on organizational performance. For instance, we can move groups of employees from the worst performing divisions to other available divisions and evaluate how this affects communication behavior of the divisions. In this way, we are able to perform the above discussed analysis but on a different level of organizational granularity; in this case, the organization is considered as built by the different working groups (teams) instead by the individual employees. In [3], we have not studied and validated this reorganization scenario.

A validation approach, which is based on the relative criteria and does not involve statistical tests, is discussed in [9]. Its general idea is to choose the best clustering scheme of a set of defined schemes according to a pre-specified criterion. We can apply this idea to our context in order to get some clue about the optimal number

of organizational units, i.e., the number of units that is supported by the underlying email communication database. This can be implemented by trying to generate clustering results for a range of different numbers of clusters (divisions) and subsequently assess the quality of the obtained clustering solutions. For example, SI and Connectivity measures can be used as validity indices to identify the best clustering scheme. A number of other principled ways, such as AIC [1] and BIC [28], can also be used to determine the number of clusters apart from the above method.

We then can run some clustering algorithm (e.g., k -means) in order to partition the employees into the found optimal number of divisions. The F -measure introduced in the foregoing section can further be applied to match the generated cluster partition with the current organizational structure and assess to what degree this partition differs from the existing one. This might indicate whether to recommend a future redesigning of the company structure, e.g., if the F -measure score is below a given threshold.

Let us suppose that we have two databases D_1 and D_2 that contain email communications among employees before and after a company reorganization, respectively. In this context, it would be useful if we can compare the two organizational structures to determine which is better. We would like to know whether the implemented structural changes have led to improved information flow and email communication. We can evaluate and compare the two cluster partitions (informal organizational structures) determined by the two email communication databases D_1 and D_2 with respect to SI and Connectivity measures. In addition, the Rand Index mentioned in the foregoing section can be used for assessing the similarity of the formal and informal organizational structure before and after the reorganization, i.e., two similarity scores can be generated. For example, if the former score (the Rand Index calculated before the reorganization) outperforms the latter one this can be a sign that the implemented structural changes have not improved the communication behavior in the organization.

We can calculate the individual SI for the employees within the organization and rank the employees in decreasing order w.r.t. the calculated SI scores. In addition, it is possible to rank the employees with respect to their hierarchical position in the organization. The hierarchical position (HP) is a measure that denotes the importance of an employee within the organization [17]. The HP of each employee i can be calculated as follows:

$$HP(i) = \frac{\sum_{j \neq i} D(i, j)}{n - 1}, \quad (2)$$

where $D(i, j)$ is a hierarchical difference between employees i and j , and n is a number of employees within the organization. The hierarchical difference $D(i, j)$ is computed as follows:

$$D(i, j) = \begin{cases} 1, & \text{if } i \text{ is higher in the hierarchy than } j \\ 0, & \text{if } i \text{ and } j \text{ are at the same level of the hierarchy} \\ -1, & \text{if } i \text{ is lower in the hierarchy than } j \end{cases}.$$

Kendall's rankings comparison method can further be used to compare two rankings [14]. It compares the nodes (employees) in pairs, i.e., the positions of pair employees within both rankings. If the position of employee i is related to the position of employee j in both rankings monotonically in the same direction, then this pair is well correlated. It is assumed that when the level in hierarchy is the same within the pair, then it does not matter whether they are in different positions in the ranking generated based on SI scores. Kendall's rank correlation coefficient is a value in the range $[-1, 1]$, where 1 means that two rankings are perfectly correlated and -1 means that they are completely different (in the opposite order). By comparing the two ranking, we can analyze whether employees who are high in the official hierarchy (ordered by HP) are also important (well-integrated) in the social network of the organization (ordered by SI). This analysis can be performed not only globally, but also locally at department level. The evaluation scenario discussed in this paragraph has not been studied and validated in our initial work published in [3].

Experiments

Telecommunication company

We have used data in the forms of email logs and organizational structure for a large European telecommunication company for an evaluation of the ideas discussed in Sect. 4.2. The employee data from the telecommunication company are provided in two separate files: `profiles.csv` and `events.csv`.

The profiles file contains an anonymized identity number for each employee and the division that the employee works at. There are six divisions: Broadband and TV, Business, Consumer Marketing, IT, Network and Stab (management). The profiles file also contains information about groups and hierarchical positions. The events file consists of the email messages exchanged between the employees during 1 week. That file has three columns: from-email, to-email and number-of-emails.

In this work, we study and evaluate two different reorganization scenarios. Namely, we analyze the email communication data both at an individual employee level and at a group level, i.e., two different organizational granularity levels are considered.

As it was mentioned above, the analyzed company consists of six different divisions. In addition, the employees of the organization are grouped into a number of different working teams (groups). The employees in each group belong to the same division. Each group might contain different number of employees.

There are 1852 employees of the organization in the `profiles.csv` file. Out of those 1852 employees, 436 employees have neither sent nor received any emails during the studied time period (1 week). These employees have been removed, which means that 1416 employees are left.

There are 332,331 emails exchanged between the employees in the `events.csv` file. Out of those, only 161,626 messages have been exchanged between the considered 1416 employees. The other messages are exchanged with persons outside the organization (external communication); those external emails have been removed.

Table 1 Divisions, number of groups, number of employees

Divisions	# groups	# employees
Broadband and TV	23	104
Business	27	271
Consumer marketing	102	471
IT	32	242
Network	25	208
Stab	27	120

Notice that we have not considered and treated specially the emails sent simultaneously to multiple employees. It may worth to study in our future work whether filtering out such emails from the email communication data of the company will change the generated evaluation results.

Thus, we consider in total 1416 employees who have shared 161,626 email messages. These employees have 236 managers. These 1416 employees can be grouped into 236 working teams based on their manager (see Table 1).

Based on the 161,626 email messages exchanged by the 1416 employees, we have initially created a 1416×1416 interaction matrix, where each entry is the number of the messages employee i sent to employee j ($i, j = 1, 2, \dots, 1416$). The interaction matrix is normalized as explained in Sect. 4.2. Then, we have used the normalized interaction matrix in order to build a distance matrix according to Eq. 1 given in Sect. 4.2. This matrix is used to generate k -means clustering solutions on the set of 1416 employees as well as to compute SI and Connectivity scores discussed in the following section.

Results and discussion

We have initially evaluated the communication behavior of the telecommunication company at an individual employee level. We have calculated Silhouette Index (SI) and Connectivity scores of the six divisions and the entire organization.¹ The obtained results are depicted in Figs. 1, 2. All divisions have positive SI values (see Fig. 1). However, the SI value of consumer marketing division is very close to zero. In addition, the Connectivity score of consumer marketing division is the highest one (see Fig. 2), i.e., the consumer marketing division can be considered as the worst performing one with respect to both measures. This may be due to the fact that this division is much larger than the other five divisions. There are 471 employees divided into 102 groups in this division (see Table 1).

In view of the above, we have further analyzed the communication behavior of the employees in the consumer marketing division. We have calculated the individual SI for the 471 employees in the division and the corresponding SI values are plotted in increasing order in Fig. 3. There are 65 employees with negative SI

¹ N.B. the Connectivity score of the entire organization is simply the sum of the Connectivity scores of the divisions, see the corresponding definition in Sect. 4.1.

Fig. 1 SI values of the six divisions and the entire organization (at an individual level)

Fig. 2 Connectivity values of the six divisions and the entire organization (at an individual level)

scores. This means that these 65 persons communicate more with employees in one of the other five divisions than with employees in his/her own division.

We have also simulated a reorganization scenario for fine tuning of the existing organization by moving a small number of employees (the ones with the lowest SI scores) from the consumer marketing division to some other division (its closest competitor). Namely, we have migrated 2% (8 employees), 4% (16 employees), 6% (24 employees), 8% (32 employees), and 10% (40 employees), respectively, from the consumer marketing division to their second-best division. The SI values for the consumer marketing division and the entire organization are recalculated after these migrations and the increases in the corresponding SI values are given in Fig. 4. The figure shows that the SI values for the consumer marketing division as well as for the

Fig. 3 SI scores of employees in the consumer marketing division

Fig. 4 SI values of the entire organization benchmarked to SI of the consumer marketing division (at an individual level)

entire organization increase when employees with low SI scores are migrated. The increase is largest when 40 employees with the lowest (negative) SI scores (10%) are migrated. Logically, the SI scores for the entire organization are less affected than ones for the consumer marketing division.

In order to get an idea about the optimal number of divisions, we have run the *k*-means clustering algorithm for different number of clusters, i.e., the set of 1416 employees is clustered by applying *k*-means for all values of *k* between 2 and 20. Then, we have used the SI and Connectivity measures as validity indices to

Fig. 5 SI values generated by *k*-means on the set of 1416 employees for different number of clusters

Fig. 6 Connectivity values generated by *k*-means on the set of 1416 employees for different number of clusters

identify the best partitioning scheme. The SI and Connectivity scores produced on the generated clustering solutions are depicted in Figs. 5, 6, respectively. Notice that both indices increase as the number of clusters increases. Therefore, we search for the values of k at which a significant local change in value of the index occurs [8]. These values are different for the considered validity indices. For example, the optimal number of divisions supported by the SI measure is 10, 12 and 16 (see Fig. 5), while 6 and 10 are the ones specified by the Connectivity measure (see Fig. 6). Consequently, based only on these results we can consider

Table 2 The F -measure values generated on the clustering solutions produced by k -means for all values of k between 6 and 10 (at an individual level)

Number of divisions	6	7	8	9	10
F -measure	0.442	0.362	0.216	0.181	0.349

The highest value produced by the F -measure is marked in bold

10, since it is supported by both measures, as the optimal number of divisions for the studied company.

We can further compare the SI and Connectivity values generated by k -means on the set of 1416 employees for $k=6$ (see Figs. 5, 6, respectively) with the corresponding values for the entire organization (which has six divisions) given in Figs. 1, 2. As it can be seen, the latter values are worse than the ones in Figs. 5, 6, i.e., there is room for improvement in the current organizational structure.

The F -measure introduced in Sect. 4.1 can be used to match the cluster partitions generated by k -means clustering algorithm on the set of 1416 employees with the current organizational structure and assess to what degree these partitions are close to the existing structure. Table 2 shows the F -measure values generated on the clustering solutions that have been produced by k -means on the set of employees for all values of k between 6 and 10. The highest F -measure score is obtained for $k=6$. Combining this finding with the results reported in Figs. 5, 6, we could conclude that the optimal number of divisions is six. In addition, the partition of employees into six working units proposed by the k -means clustering solution could be used as a direction to future redesigning of the company structure.

As it was mentioned above, we analyze the email communication data of the company both at an individual employee level and at a group level. In order to perform the latter analysis, we use email messages exchanged by 236 working teams (groups) of the company. We have initially created a 236×236 interaction matrix, where each entry is the number of the messages the employees of group i sent to the employees of group j ($i, j=1, 2, \dots, 236$). The interaction matrix is normalized as explained in Sect. 4.2. We have used the normalized interaction matrix in order to build a symmetric distance matrix according to Eq. 1 given in Sect. 4.2. This distance matrix is used to compute SI and Connectivity scores presented in Figs. 7, 8, respectively. It is interesting to notice that the consumer marketing division is again the worst performing one with respect to both cluster validation measures.

At the group level, we have also simulated a reorganization scenario for fine tuning of the company, but instead of migrating employees individually we have moved groups of employees (the groups with the lowest SI scores) from the consumer marketing division to its closest competitor. There are 102 groups of employees in this division. We have calculated the individual SI for these 102 groups of employees in the consumer marketing division. There are 22 groups of employees with negative SI scores. Then, we have migrated 2% (2 groups=7 employees), 4% (4 groups=18 employees), 6% (6 groups=25 employees), 8% (8 groups=37 employees), and 10% (10 groups=43 employees) from the consumer marketing division to their

Fig. 7 SI values of the six divisions and the entire organization (at a group level)

Fig. 8 Connectivity values of the six divisions and the entire organization (at a group level)

second-best division, respectively. After these migrations, SI scores for the consumer marketing division and the entire organization have been recalculated and the corresponding increases in SI values are plotted in Fig. 9. We can observe that the migration of groups of employees has produced a higher increase in the SI scores than the migration of individuals (see Fig. 4). This might be due to the fact that by migrating groups of employees the working teams of the organization are not separated between a few divisions, but instead they are kept together in one and the same division.

The SI and Connectivity values generated at a group level (see Figs. 7, 8, respectively) are benchmarked with the corresponding values for the six divisions and the entire organization given in Figs. 1, 2 (at an individual level). As one can observe

Fig. 9 SI values of the entire organization benchmarked to SI of the consumer marketing division (at a group level)

Fig. 10 SI values generated by k -means on the set of 236 groups of employees for different number of clusters

the latter values are worse than the ones in Figs. 7, 8, i.e., the email communication behavior of the organization at a group level is better than that at an individual employee level.

At the group level, we have also clustered the set of 236 groups of employees by applying k -means for all values of k between 2 and 20 in order to get an idea about the optimal number of units. Then, we have used the SI and Connectivity measures as validity indices to identify the best partitioning scheme. The SI and Connectivity scores produced on the generated clustering solutions are depicted in Figs. 10, 11, respectively. Interestingly, in the considered scenario $k=6$ is not supported by

Fig. 11 Connectivity values generated by k -means on the set of 236 groups of employees for different number of clusters

Table 3 The F -measure values generated on the clustering solutions produced by k -means for all values of k between 6 and 10 (at a group level)

Number of divisions	6	7	8	9	10
F -measure	0.348	0.214	0.296	0.204	0.230

The highest value produced by the F -measure is marked in bold

any of the applied measures. In fact, the optimal number of divisions according to the SI measure is 18 (see Fig. 10), while 10, 13 and 18 are the number of units supported by the Connectivity measure (see Fig. 11). Evidently, the 236 groups of the company can be grouped into 18 different divisions based on the results presented in Figs. 10, 11, respectively. It is also interesting to notice that the SI score for $k=18$ in Fig. 10 is higher than the SI values for the entire organization given in Fig. 1 (six divisions), Fig. 5 ($k=10, 12$ and 16) and Fig. 7 (six divisions).

We have further studied k -means clustering solutions generated on the set of 236 groups for values of k between 6 and 10. We have used F -measure to match these cluster partitions on the set of 236 groups with the current organizational structure and assess to what degree these partitions are close to the existing structure. Table 3 shows the F -measure values generated on the clustering solutions that have been produced by k -means on the set of groups of employees for all values of k between 6 and 10. The highest F -measure score is again produced for $k=6$.

Finally, we have calculated the Kendall's rank correlation coefficients between hierarchical position (HP) and Silhouette Index values of the employees of the six divisions and the entire organization. The obtained results are depicted in Fig. 12. The studied telecommunication company has 236 employees as managers. The data (profiles.csv file) provided by the company contain information of a manager of each employee together with his/her division. This information is enough to calculate the

Fig. 12 Kendall's rank correlation coefficients of the six divisions and the entire organization

HP scores of the employees of the six divisions and the entire organization. These are calculated as it was described in Sect. 4.2 (see Eq. 2). Kendall's rank correlation coefficient [14] has a value in the range $[-1, 1]$ and the positive value (close to 1) means that two rankings are well correlated. We have compared the two rankings in order to get insight whether employees who are high in the official hierarchy are also important in the social network of their division. The analysis shows the Kendall's rank coefficients of all divisions are positive with values in the range $[0.06, 0.25]$ (see Fig. 12). The two rankings are quite similar for the Broadband and TV division with the highest Kendall's rank and are not so well correlated for the Business division with the Kendall's score very close to zero. It is interesting to notice that the Broadband and TV division also outperform the other divisions w.r.t. SI and Connectivity measures (see Figs. 1, 2), i.e., it overall demonstrates a good internal communication climate. However, the communication behavior of the Business division may need further analysis in order to find the reason for its comparatively low Kendall's rank correlation coefficient. In addition, we can further analyze and compare the two rankings separately for female and male employees in the company. However, the telecommunication company data used in this study do not provide information about gender of the employees. It would be interesting to study in our future work whether experiments on richer data will confirm the results reported on Enron data in [20]. Namely, employees who are low in the hierarchy, but high in social network metrics, are mostly women.

Conclusion

In this paper, we have performed a study that explores the use of cluster validation measures for analysis of the organizational structure of a company with respect to the internal communication at the organizational level. Namely, we have studied

how one, based on internal email communication, can analyze the organizational structure through different cluster validation techniques. We have used data in the forms of email logs and organizational structure for a large European telecommunication company for an evaluation and validation of our ideas. We have calculated a number of cluster validation metrics (e.g., SI, Connectivity, and F -measure) on the telecommunication company data and based on this have analyzed the internal communication behavior of the company and discussed changes that have to be made to its structure in order to improve the communication.

Most companies do frequent reorganizations, and a lot of time and resources are put into such activities. However, there are no well-established objective measures regarding how successful reorganizations are. In this paper, we have proposed an approach based on cluster validation techniques that makes it possible to put quantitative and objective measures on the performance of an organization. We believe that this is an important step towards being able to evaluate and compare different organizational structures, and objectively evaluate the success of a reorganization.

For future work, we aim to pursue further evaluation and validation of the proposed approach on richer data presenting internal email communication of a company for longer time periods. In addition, we plan to study some graph-based and hierarchical clustering algorithms and their corresponding evaluation metrics. Our future plans also involve integrating additional information into the informal organizational structure supplied by company experts, such as whether some divisions exchange emails more regularly than others.

Acknowledgements This work is part of the research project “Scalable resource-efficient systems for big data analytics” funded by the Knowledge Foundation (Grant: 20140032) in Sweden.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
2. Bar-Yossef, Z. et al. (2006). Cluster ranking with an application to mining mailbox networks. In *Sixth international conference on data mining ICDM'06* (pp. 63–74). IEEE Computer Society.
3. Boeva, V., Lundberg, L., Kota, S. M. H., & Sköld, L. (2017). Analysis of organizational structure through cluster validation techniques. *DSHCM 2017, ICDM 2017 workshop on data science for human capital management* (pp. 170–176). New Orleans: IEEE.
4. Galbraith, J. R. (1987). Organization design. In J. W. Lorsch (Ed.), *Handbook of organizational behavior*. Englewood Cliffs: Prentice Hall.
5. Greenberg, J. (2011). *Behavior in organizations* (10th ed.). Upper Saddle River: Prentice Hall.
6. Grunig, L. A. (1992). How public relations/communication departments should adapt to the structure and environment of an organization...And what they actually do. In J. E. Grunig (Ed.), *Excellence in public relations and communication management* (pp. 467–481). Hillsdale: Lawrence Erlbaum Associates.

7. Grunig, L. A., Grunig, J. E., & Dozier, D. M. (2002). *Excellent public relations and effective organizations: A study of communication management in three countries*. Mahwah: Lawrence Erlbaum Associates.
8. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17, 2(3), 107–145.
9. Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21, 3201–3212.
10. Holtzhausen, D. (2002). The effects of a divisionalized and decentralized organizational structure on a formal internal communication function in a South African organization. *Journal of Communication Management*, 6, 323–339.
11. Hossain, L. (2009). Effect of organizational position and network centrality on project coordination. *International Journal of Project Management*, 27(7), 680–689.
12. Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs: Prentice Hall.
13. Kazienko, P., Michalski, R., & Palus, S. (2011). Social network as a tool for improving enterprise architecture. In J. O'Shea et al. (Eds.) *LNAI* (vol. 6682, pp. 651–660).
14. Kendall, M. G., & Gibbons, J. D. (1990). Rank correlation methods. In E. Arnold (Ed.), *A Division of Hodder & Sloughton* (5th ed.). London: Oxford University Press.
15. Kim, H.-S. (2005). Organizational structure and internal communication: An organizational-level analysis. In *8th international public relations research conference "the impact of PR in creating a more ethical world: why can't we all get along?"* (pp. 196–207).
16. Kim, H.-S. (2005). Organizational structure and internal communication as antecedents of employee-organization relationships in the context of organizational justice: A multilevel analysis. *PhD in Philosophy, University of Maryland, College Park, MD 20742–7011*, 301, 314–1328.
17. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal ACM*, 46(5), 604–632.
18. Kraut, R., et al. (2002). *Informal communication in organization: Forms, function, and technology* (p. 07962). Bellcore: Morristown.
19. Larsen, B., Aone, C. (1999). Fast and effective text mining using linear time document clustering. In *KDD-99* (pp. 16–29).
20. Michalski, R., Palus, S., & Kazienko, P. (2011). Matching organizational structure and social network extracted from email communication. *LNBIP*, 87, 197–206.
21. Michalski, R., & Kazienko, P. (2014). *Social network analysis in organizational structures evaluations. Encyclopedia of social network analysis and mining* (pp. 1832–1844). New York: Springer.
22. Palus, S., Brodka, P., & Kazienko, P. (2010). How to analyze company using social network? *CCIS*, 111, 159–164.
23. Palus, S., Brodka, P., & Kazienko, P. (2011). Evaluation of organization structure based on email interactions. *Journal of Knowledge Society Research*, 2(1), 1–13.
24. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850.
25. Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational Applied Mathematics*, 20, 53–65.
26. Rowe R. et al. (2007). Automated social hierarchy detection through email network analysis. In *9th WebKDD 2007 and 1st SNAKDD 2007* (pp. 109–117). New York: ACM.
27. Saunders, M. (1999). Linking external communication and organizational effectiveness. *Organizational Development Journal*, 17(4), 35–40.
28. Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
29. Smidts, A., Pruyn, A. D., & Van Riel, C. B. M. (2001). The impact of employee communication and perceived external prestige on organizational identification. *ACAD Manage Journal*, 44(5), 1051–1062.
30. van Rijsbergen, C. J. (1979). *Information retrieval*. USA: Butterworth-Heinemann Newton.
31. Vendramin, L., Campello, R. J. G. B., & Hruschka, D. T. R. (2010). Relative clustering validity criteria: A comparative overview, 2010 Wiley Periodicals, Inc. *Statistical Analysis and Data Mining*, 3, 209–235.
32. Widhiastuti, H. (2012). The effectiveness of communications in hierarchical organizational structure. *International Journal of Social Science and Humanity*, 2(3), 185–190.