Analysis of Organizational Structure through Cluster Validation Techniques

Veselka Boeva, Lars Lundberg, Sai M. Harsha Kota, Lars Sköld

Blekinge Institute of Technology
Karlskrona, Sweden
veselka.boeva@bth.se

DSHCM 2017, ICDM 2017 Workshop on Data Science for Human Capital Management (New Orleans, November 18, 2017)

Use 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.

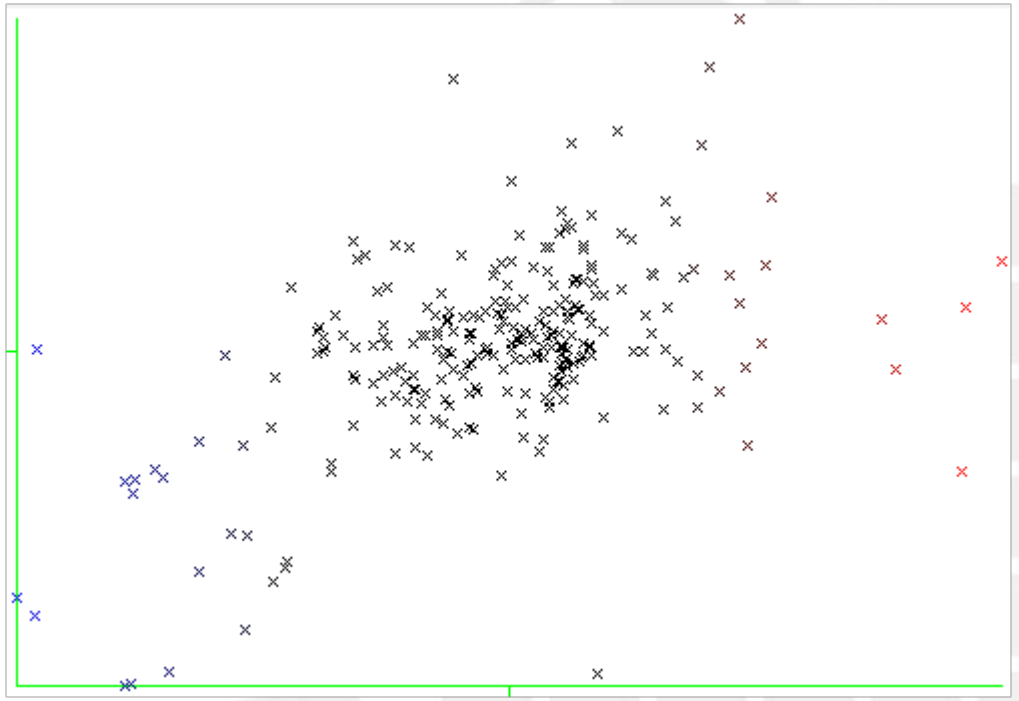
Using Cluster Validation Techniques for Analysis of Organizational Structure

- 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);
- fine tuning an existing organizational structure by moving a limited number of employees from one organizational unit to another.

Clustering Analysis

- **Clustering analysis** is a process that partitions a set of objects into clusters in such a way that objects from the same cluster are similar and objects from different clusters are dissimilar.

- K-means clustering
  - Decompose the data set into ***k*** **disjoint clusters minimizing** the within-cluster sum of distances
  - The cluster center is the mean data vector averaged over all objects in the cluster.



Cluster Validation Techniques

- **Cluster validation techniques** are designed to find the partitioning that best fits the underlying data.
- Cluster validation measures are divided into two major categories: ***external*** and ***internal***.
- **External** validation measures evaluate the clustering result with respect to a pre-specified structure.
- **Internal** measures base their validation on the same information used to derive the clusters themselves.

Cluster Validation Techniques

- **Silhouette Index** assesses compactness and separation of a clustering solution: $s(C) = (1/m)\sum_{i=1}^{m}(b_i - a_i)/\max\{a_i, b_i\}$.
  - value varies from -1 to 1 and should be **maximized**
- **Connectivitv** assesses connectedness of a clustering solution: $Conn(C) = \sum_{i=1}^{m}\sum_{j=1}^{n_r} x_{im_{ij}}$.
  - value varies from 0 to infinity and should be **minimized**
- **F-measure** can be used to match two clustering solutions (e.g., the formal and informal organizational structure).
  - value varies from 0 to 1 and should be **maximized**

Organizational Email Communication Graph

- An **undirected graph**, where the nodes are employees and edges represent communication between two employees.
- Build a **distance matrix**, where each entry is a value d_{ij} expressing the distance between employees *i* and *j*.
- The distance can be calculated by taking into account the intensity of email interaction between employees, e.g.,

$$d_{ij} = (e_{max} - e_{ij})/e_{max} ,$$

e_{ij} = the number of exchanged emails between *i* and *j*.
e_{max}= the maximum number of exchanged emails between two employees.

Analysis of Email Communications

- evaluate the overall organizational structure.
- evaluate whether the "within" communications are high in comparison with the "between" communications.
- investigate which employees appear to be well-allocated, which ones are wrongly-allocated, and which ones lie in between divisions.
- try to obtain an idea about the number of "natural" divisions that are present in the email interaction database.

Analysis of Email Communications at the Division Level

- Compare the performance of the different divisions w. r. t. the effectiveness of their email communication by using SI and Connectivity.
- Analyze further the worst performing divisions by computing individual SI of their employees.
- Find for each employee with negative SI the second-best division, *i.e.,* the closest competitor.
- It could be recommended to allocate these employees to their closest division.

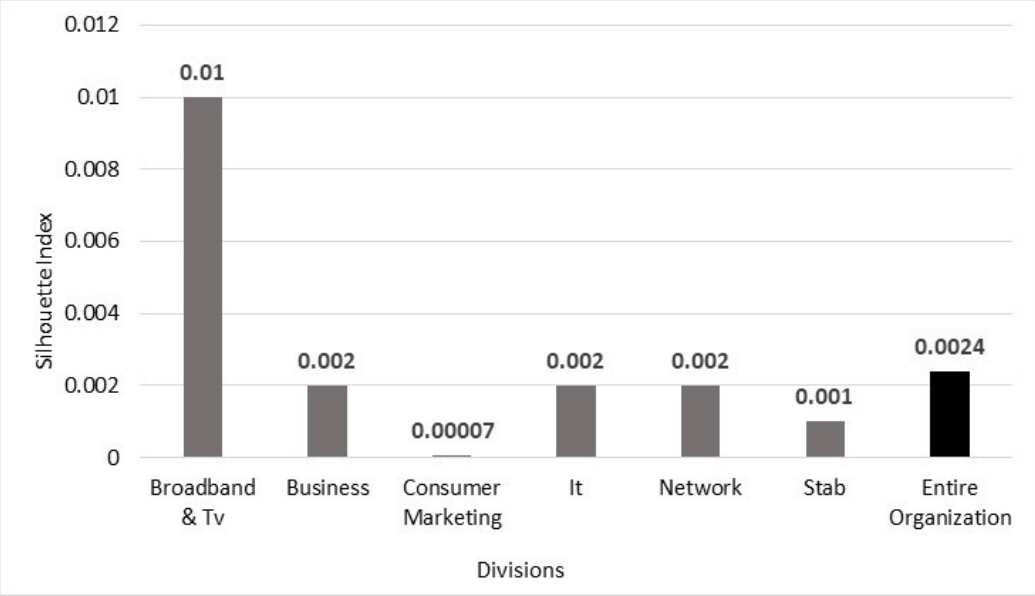
Find the Optimal Number of Divisions

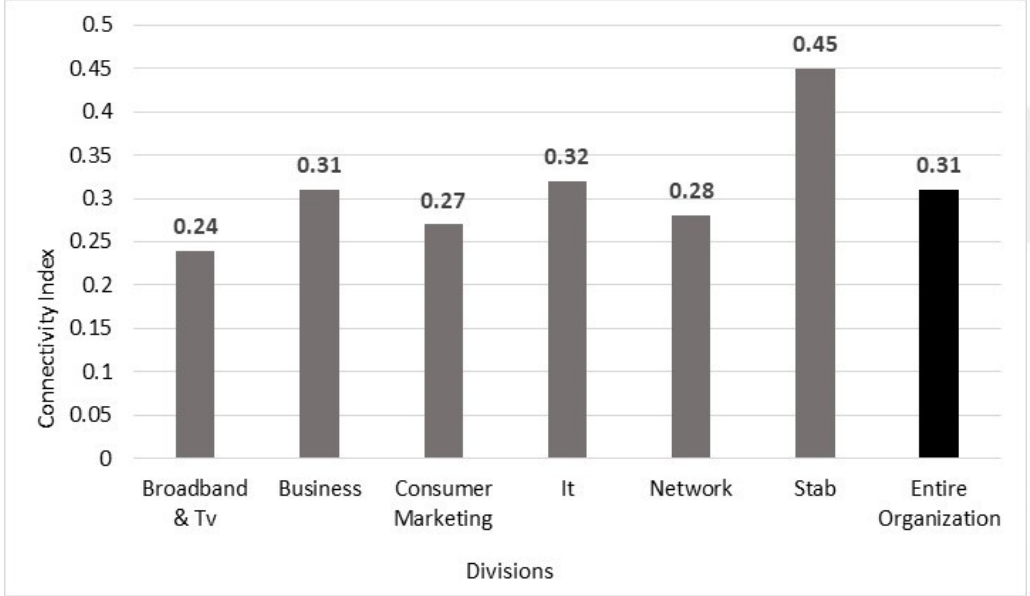
- Generate clustering solutions for a range of different numbers of clusters (divisions).
- Assess the quality of the obtained clustering solutions using SI and Connectivity as cluster validation indices.
- Run a clustering algorithm in order to partition the employees into the found optimal number of divisions.
- Apply F-measure to match the generated cluster partition with the organizational structure and assess to what degree these partitions are different.

Data Pre-Processing & Study Design

- The employees who have neither sent nor received any emails during the studied time period have been removed.
- The messages that are exchanged with persons outside of the organization have been removed.
- Build a communication matrix based on 42366 email messages exchanged by 1061 employees, where each entry is the number of the messages exchanged between two employees.
- Use the communication matrix in order to build a 1061 x 1061 distance matrix.

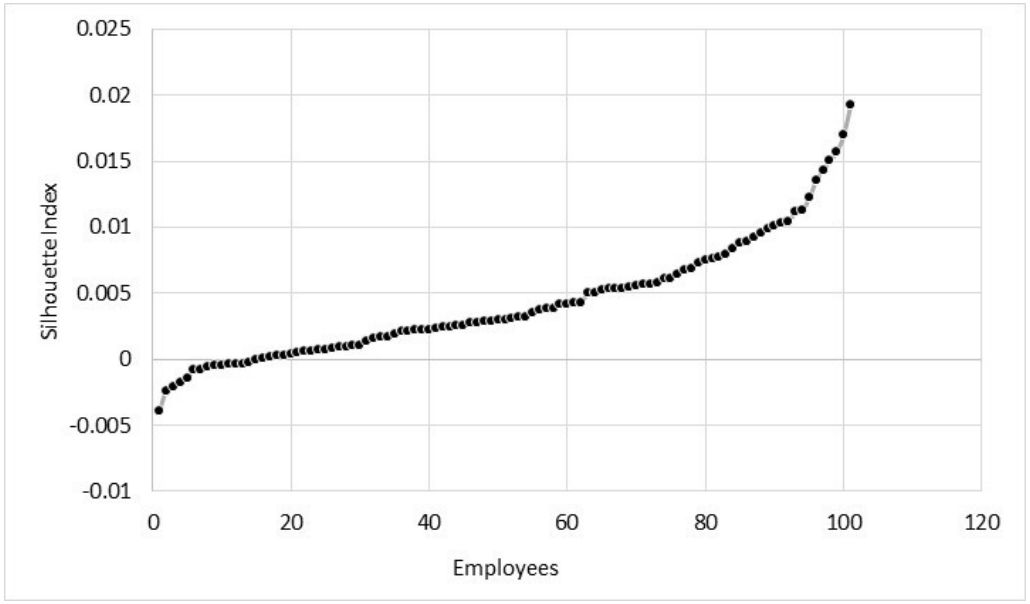
Experimental Results



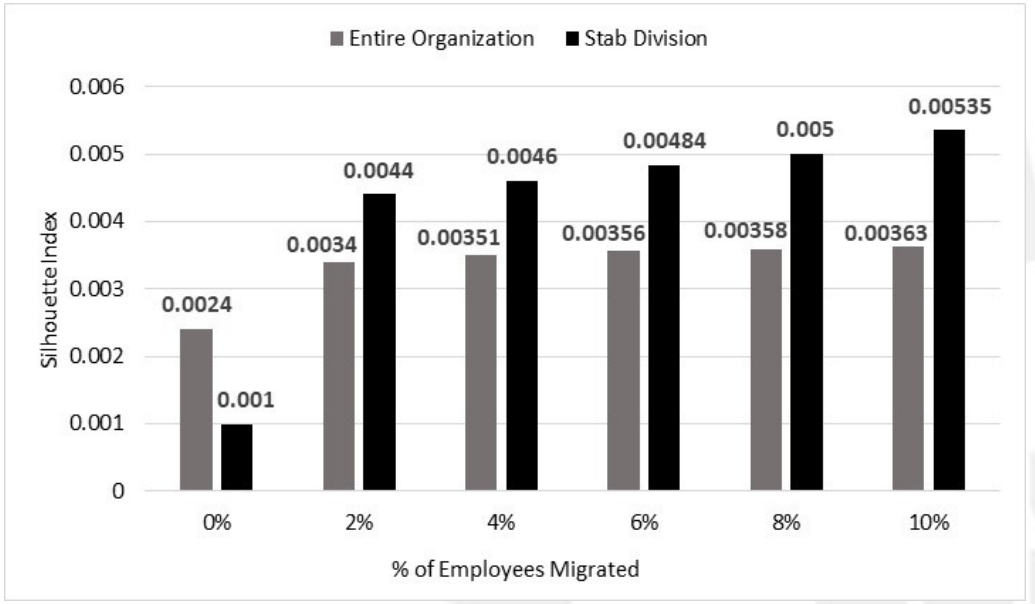


Si and Connectivity values of the six divisions and the entire organization

Experimental Results

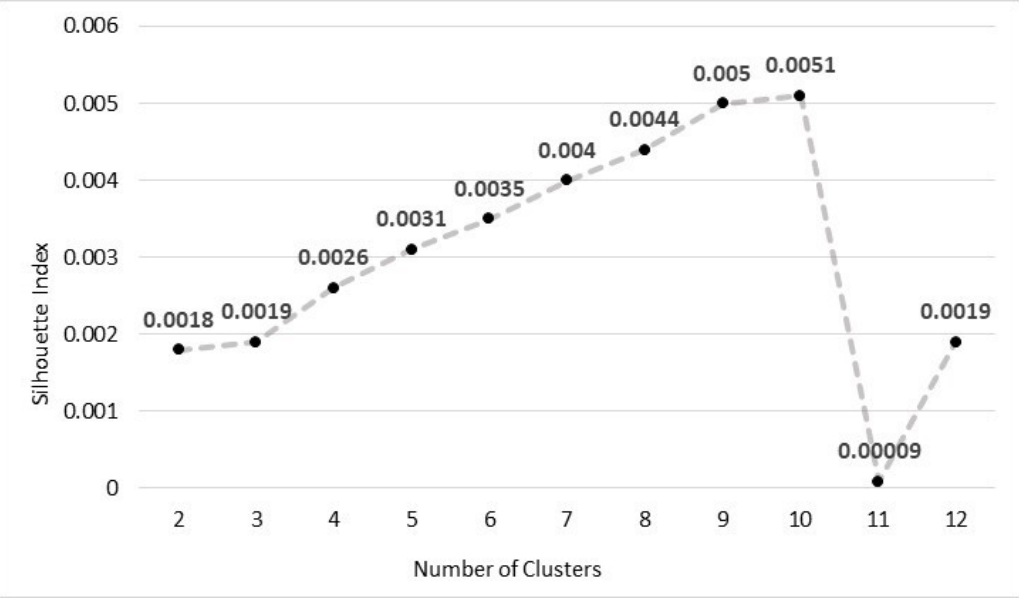


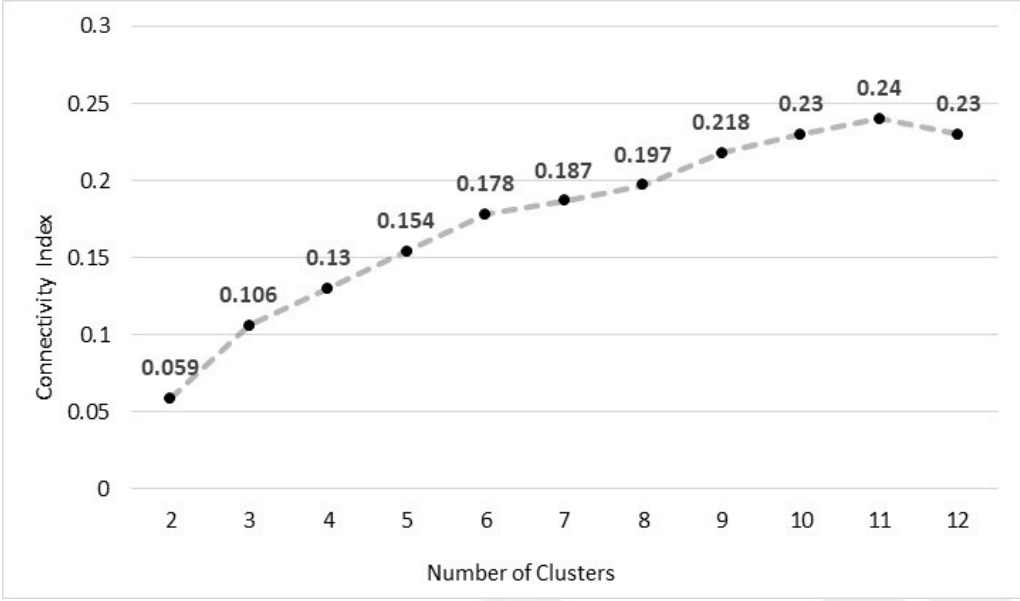
SI scores of employees in the Stab division



SI values of the entire organization benchmarked to SI of the Stab division

Experimental Results





Si and Connectivity values generated by k-means on the set of 1061 employees for different number of clusters

Experimental Results

Number of clusters	6	7	8	9	10
***F*-measure**	0.306	0.308	0.215	0.214	0.165

The *F*-measure values generated on the clustering solutions produced by *k*-means for all values of *k* between 6 and 10

Conclusion & Future Work

- The conducted initial study that explores the use of cluster validation measures for analysis of the internal communication of a company has showed promising results.
- Evaluation and validation on richer data presenting internal email communication for longer time periods.
- Study some graph-based and hierarchical clustering algorithms and their corresponding evaluation metrics.
- Integration of additional information such as whether some divisions exchange emails more regularly than others etc.