Association for Information Systems
AIS Electronic Library (AISeL)

Wirtschaftsinformatik Proceedings 2009

Wirtschaftsinformatik

2009

CONTENT-BASED COMMUNITY DETECTION IN SOCIAL CORPORA

Annette Bobrik
TU Berlin

Matthias Trier
TU Berlin

CONTENT-BASED COMMUNITY DETECTION IN SOCIAL CORPORA

Annette Bobrik, Matthias Trier[1]

Abstract

Electronic communication media are a widespread means of interaction. They effect network relationships among people. Such networks provide connectivity but are often structured in clusters. Current cluster analysis in social corpora is mainly based on structural properties. This paper extends existing approaches with content-based cluster identification and community detection in social corpora. Following a design science methodology, we demonstrate our approach using a corporate e-mail dataset. After analyzing relationships between structural and content-based groups we conclude that our method contributes to detecting online communities, especially for large structural or smaller but dispersed topical groups.

1. Introduction

Electronic communication media have become the most widespread means of interacting in a company. In a recent study, 98 percent of all employees with internet access collaborate via e-mail at their workplace, most of them several times a day. Three quarters used it to improve their teamwork and to increase the number of people they actively communicate with [5]. E-Mail is thus a strong force for connecting people at the workplace. It has been shown to complement formal work networks and provide more diverse, participative and less formally aligned relations [2]. Such effects are not limited to the domain of e-mail communication: The internet sports new combined forms of interaction that blend contents and communication (e.g. comments to stories) [21].

All these interactions of people captured over time form a social network structure via communication [11]. One of the primary methods for studying the resulting electronic communities is Social Network Analysis (SNA) [24]. For this paper, the collection of all (electronic) traces of interrelated communication relationships is defined as a social corpus.

Research in online communication networks is directed at understanding the structures and properties of these complex systems to identify structural patterns (e.g. [10]). One important interest is the location of groups in such networks. For this objective, SNA offers a series of clustering and grouping algorithms that find widespread application in practice. For example, Tyler et al. applied a graph-theoretic betweenness centrality algorithm to automatically identify communities in a research department of 400 employees at Hewlett-Packard [22] solely based on studying e-mail interactions. They detected 66 distinct groups reaching from 2 up to 57 members. A subsequent manual validation has shown that about 60 percent of the persons found the automatically identified

[1] TU Berlin, D-10587 Berlin, Franklinstrasse 28/29

groups to be completely correct. This example shows that networks simultaneously allow flows of knowledge and provide connectivity but are also often structured in clusters of actors [12]. Actors tend to cluster in homogenous groups that are only loosely connected to other groups [17]. However such research and the current algorithmic implementations in common SNA tools like Pajek or UCINet only apply structural analysis of community groups.

The domain of content analysis with means of information retrieval has not yet been accommodated in the methodological framework of SNA. Previous related research in content analysis and topic mining is focused on algorithms for generating keyword or topic descriptors in semi-formal text [3]. In the few studies of contents of social corpora, the focus was on expert profiling and on subsequent assessment of search strategies to search for persons with appropriate expertise in online networks [26]. The authors applied conventional text mining and indexed all the messages of a person and created a keyword vector, in which a keyword is weighted by its term frequency-inverted document frequency [26]. A precursor to the work in this paper is our content-based analysis and exploration concept called 'Social Search' [21]. It combines network visualization and the ability to create topic oriented sub-networks of an online discourse. Despite such initial approaches, these and similar studies mostly use content analysis to profile and search for individuals in social corpora. A comprehensive method of applying content-based analysis to study the group level of social networks is missing. Consequently there is not much known about how topical profiles are spread in social corpora, how virtual groups reflect topics, or if topical analysis would yield divergent group segmentation. With such a gap, every conventional structure-based method will not be able to convincingly detect communities in a multi-contextual network with multiple topics that are simultaneously shared among actors. Such communities are not only structural phenomenon but have been defined by seminal literature as networks *about something* [25].

2. Research Objectives

Based on the recognition of the above shortcomings, this paper is introducing an approach for content-based community detection in social corpora based on content clustering techniques. With this method, current structural insights about groups in social corpora can be extended with a topical perspective. Identified groups can be related to structural patterns in order to identify topic communities which are inefficiently spread across the interaction structure or which are isolated from other people. With this method, we want to contribute to research on better algorithms to identify group-level properties of electronic networks. This will enable more sophisticated automated awareness services (i.e. to answer questions like who is related to my context, or who is a relevant contact?), the support of large overlapping global teams, or other collaborative community services. Our research is driven by the following two research questions:

1. How can content-based clustering be applied for the detection of topical communities in social corpora?
2. To what extend do groups identified by existing methods of structural clustering correlate with content-based communities?

We employ a design science research methodology as our focus is less on empirical insight about factor relationships but on the design and evaluation of a method (a design artifact) [23]. Consequently, the remainder of the paper is organized as follows: In Section 3 the related approaches and basic foundations for content-based clustering in social corpora are given. In Section 4, we introduce our proposed approach to find communities with similar context/collaborative content. We then demonstrate and evaluate the new approach using a corporate e-mail dataset. The corresponding data source is described in Section 5. Finally, we compare our results with subgroups obtained from structural clustering to discuss the general relationship of these two perspectives in Section 6. A conclusion of our findings is presented in Section 7.

3. Methods and Foundations for Community Detection in Social Corpora

In this section we briefly introduce related research and corresponding foundations for our approach of content-based community detection. They include the concepts of Social Network Intelligence, event-driven network analysis, measures of topic importance, and algorithms for structural clustering.

3.1. Event-driven Network Analysis Approach

The methods presented in this paper relate to the underlying concept of Social Network Intelligence (SNI), initially introduced in [19]. This approach aims at extending conventional network analysis with a dynamic analysis on the longitudinal processes in a social network (network dynamics) and a content-based analysis of social networks. The approach allows studying the change and evolution of network structures, the impact of events, or processes of topic dissemination in networks. A comprehensive account of the resulting advantages is presented in [20]. As SNI requires extensive computational effort, the corresponding software implementation Commetrix has been introduced. Commetrix is a java-based tool constructed for event-based dynamic network analysis and attempts to address the limitations of current approaches [19]. The development of this tool yielded a comprehensive set of software-based methods for exploratory static and dynamic visualization with integrated analysis of social network measures. The methodological core of SNI and Commetrix is event-based network analysis: Relationships are not directly considered but their constituting timed events are captured. In communication network analysis such relational events are created by exchanging messages with others. From these events, relationships can be aggregated. For each event several properties are captured. For example, a time stamp or a topical descriptor of each message event can recorded as a message property. Hence, the sequence of messages and the change in relationship structure or strength is represented as a series of relational events in the data model. In addition to these important changes in capturing relationships, the actual actors are modeled together with their actor properties.

3.2. Measure of topic importance

In information retrieval the *term frequency-inverted document frequency* (tf.idf) weighting scheme is often used together with the cosine similarity to determine the similarity between two documents [18]. The importance of a keyword t_k increases proportionally to the number $n_{k,j}$ the keyword appears in the document d_j but decreases with the number of documents of the document collection D it appears in. The keyword occurrence in one document is normalized by the total number of keywords t_k in the document. As social networks can interact on different types of social content and keywords can be retrieved on different levels of abstraction as single words, or word collections, or concepts, *keywords*, *documents* and *document collections* can be generalized as *topics*, *content objects* and *content collections*.

However, the tf.idf measure has some disadvantages when analyzing social content. In particular, we do not compare single content objects but content collections of varying size. Thus, we propose a variation of the tf.idf measure for comparing social content collections. Similar to the tf.idf measure the *importance* of topic t_k for actor *a* increases proportionally to the number of content objects c_j in the content collection of the actor, C_a, containing the topic but decreases with the number of content objects of the overall content base *C* it appears in. The importance measure can be formalized as:

$$\mathrm{importance}_k(a) = \frac{\frac{\{c_j \in C_a : t_k \in c_j\}}{|C_a|}}{\ln \frac{|C|}{\{c_j \in C: t_k \in c_j\}}}$$

3.3. Structural Clustering

A cluster analysis performed on the network structure seeks to detect densely connected subgroups. A widely used clustering algorithm for community detection is the divisive hierarchical *edge betweenness* algorithm by Newman and Girvan [6]. The algorithm iteratively removes edges from the network until it disaggregates into smaller subgroups. As a result a hierarchy of nested partitions is obtained ranging from all actors in one large group to each actor in a group by his own[2]. From this hierarchy the partition is chosen which maximizes the *modularity* measure [16]. The maximum of the modularity measure indicates that the actors are maximally connected with their group members and minimally connected with other groups.

4. A New Approach of Content-based Clustering

In contrast to the edge betweenness algorithm which is working merely on the structure of the network we propose a content-based clustering procedure which takes the network context into account. It is basically an iterative partitioning clustering procedure which assigns objects to clusters due to the similarity of their features. While clustering social corpora the actors are the objects to be clustered and the topics in their content collections are the features the clustering is performed on. The similarities of the topics are calculated from the importance of the topics in each content collection. We assume that collaborative content approximates the context of the network activities and actors are assigned to subgroups with similar context expressed by the content objects they are related to. With the Enron dataset each message is assigned to the content collection of its sender and receiver. In a preliminary semi-automatic text mining procedure the body text and subject of each message is reduced to a set of topics weighted by their importance (see Section 3.2).

Each iterative partitioning clustering procedure is characterized by the initial partition, the proximity index, the type of cluster representative, the type of pass and the statistical criterion used [8]. Topic similarities are usually calculated from the cosine similarity [18]. We use the cosine similarity with the K-means clustering method by McQueen [13]. The K-means clustering method is an inclusive, combinatorial K-means pass through the data which is iteratively optimized by the square error criterion. The square error criterion measures the dissimilarity of the members of a group to its cluster representative. The cluster representative is either an object with average feature values (centroid) or the member of the group which is on average most similar to all other members (medoid) [9]. Centroids do not necessarily have an expression in the data objects. For this reason we use medoids as cluster representatives. The results and the performance of a partitioning clustering procedure strongly depend on the chosen parameters. Especially the choice of the initial partition will influence the outcome of the analysis. There are several ways to select the initial partition [1,

[2] This holds only for connected graphs. Unconnected graphs already start with their actors assigned to several unconnected subcomponents which are then further divided.

8]: a) randomly assign the objects to K clusters, b) select K seed points from the data, e.g. randomly or K well-separated objects, e.g. the centroid of the whole data and then selecting successive seed points which are at least a certain distance away from those already chosen, or c) use a pre-computed initial partition, e.g. from a preliminary hierarchical cluster analysis. With the first two methods the optimal number of clusters in the data has to been known, or estimated, before. With the third method using a hierarchical clustering algorithm the optimal number of clusters can be derived from the clustering hierarchy applying an appropriate stopping rule (see [15]). We use an initial partition obtained from a hierarchical clustering procedure using cosine similarity, the weighted average linkage rule by McQuitty [14] and Hartigan's stopping rule which minimizes the change of the within-cluster sum of squares to the representative of the cluster for two subsequent partitions [7].

Our content-based cluster analysis approach is not intended to compete with the structural clustering but to provide additional insights in group formation processes within collaborative networks. Each clustering procedure will identify different subgroups in a network due to the structural or content-based similarity of the actors. Depending on the network structure they belong to a set of densely connected members that are rather loosely connected to other members. Depending on the network context approximated by the content of their interactions they belong to a set of actors with similar context which are not necessarily densely connected to each other. Besides a visual comparison of the results from the two clustering procedures we evaluate some structural and content-based properties. The structural properties are the density, the average clustering coefficient, the average relationship strength, and the number of unconnected components. The density measures how many possible relationships are established whereas the clustering coefficient measures how many established relationships form relationships themselves. The average relationship strength provides information about the communicational activity within the cluster. The content dissimilarity is measured by the square error value which is the sum of the within-cluster dissimilarities of the content collections of the group members to the content collection of their medoid.

To measure the stability in group memberships between the structural and the content-based clusters we use two different measures: The first stability measure is the percentage of structural group members belonging to the same content-based cluster. The second stability measure is based on a version of the Dice coefficient [4] for actor-oriented group membership stability. It measures how stable the group membership of an actor is in terms of constant group members and structural as well as content-based group size. Suppose X_a and Y_a are the sets of group members of actor a in the structural and the content-based clustering. $X_a \cap Y_a$ is then the overlap between both groups. The two *actor-oriented group membership stability* measures can be formalized as follows:

$$\text{GMS\#1}(a) = \frac{|X_a \cap Y_a|}{|X_a|} \quad \text{and} \quad \text{GMS\#2}(a) = \frac{2|X_a \cap Y_a|}{|X_a| + |Y_a|}$$

To calculate the *group membership stability* of a structural cluster the average values of both actor-oriented group membership stability measures are used, namely AGMS#1(a) and AGMS#2(a).

5. Data Source

Our analysis is based on a subsample of the publicly available corporate Enron e-mail dataset. The Enron Corporation was an American energy company which went bankrupt in late 2001. The subsample sorely consists of internal messages exchanged between non-isolated Enron employees on management level from September 1th 2000 to March 31th 2001. On the one hand this was done to prevent the analysis to be overcrowded with too many peripheral actors with minimal impact on the network and to focus on the internal collaborative relationships (focus on Enron employees). On the other hand preliminary studies on the Enron dataset and its context have shown that within the selected period of time important communication threads can be found as it covers the California's electricity crisis which led to the downfall of the Enron company (focus on selected period of time). Two additional reasons for the subsample selection had to be considered. First, the selected period is more stable in the presence and activity of the network participants than earlier or later periods. Second, as the analysis of subgroups is more interested in larger collections of actors than in single individuals a window size has to be chosen where large and dense network structures can emerge. A much shorter period would yield to networks with many unconnected, stringy components where the results from the structural as well as content-based clustering procedures are obvious.

Within the subsample 112 actors formed 441 relationships with average relationship strength of 17.39 messages. The density of the network is 7.09 percent. The diameter only amounts to a path length of 5 steps and the average path length between every pair of at least indirectly connected actors is 2.76 steps. However, there are two unconnected components in the network, one large component with 106 actors and one small component with 6 actors. The core group of active people, which together accumulate 80 percent of all network activity, has a share of 18.75 percent. The maximum degree is 24 contacts. The most active actor has sent 826 and received 234 messages from contacts in the sample. Interestingly, the most active actor and the most connected actor are not the same person. Moreover, the most connected actor has a much smaller network activity with 49 messages sent and 60 messages received, which is an almost perfect reciprocal relation. The clustering coefficient of the subsample, that is how many contacts of all actors are connected themselves, is 35.20 percent.

6. Results and Interpretation

The structural as well as the content-based clustering procedure has been applied to the subsample of the Enron dataset described in the previous Section. Figure 1 provides the visual representation of the cluster memberships for both clustering procedures. Table 1 and Table2 contain additional information about the clusters. With the structural clustering (A) most of the clusters are well separated from the others. Only in the extremely dense center of the network the clusters seem to intermingle. In comparison, the content-based clusters (B) are more distributed over the entire network. The members of cluster C4 even belong to two unconnected components of the network. The overall content dissimilarity of each actor to its group members is expressed by the square error value: the sum of squared dissimilarities calculated from the cosine similarities of all actors’ content collections to the content collection of the medoid of their cluster. As the content-based clustering procedure identifies some smaller clusters and two large clusters with 30 (C5) and 48 (C6) members one would expect a larger square error. However, comparing the results from (A), 3.47, and (B), 2.77, the content-based clustering solution identifies more similar clusters.

The content dissimilarity as well as the average relationship strength of the structural clusters does not depend on the group size. Cluster S4 has about half the number of members as cluster S7 but the same content dissimilarity. That means the members of the larger cluster S7 have more similar content collections than those in the smaller cluster S4. Clusters S4 and S5 have almost the same size but the members in cluster S5 communicate over more different topics than those in cluster S4. Interestingly, the communicational content of the smaller clusters S1, S2 and S3 is far less similar than that of the larger clusters S4, S6 and S7. The clustering coefficient is high for medium sized clusters and decreases with cluster size. But even in the largest, least dense cluster, S9, almost half of all contacts are connected themselves.

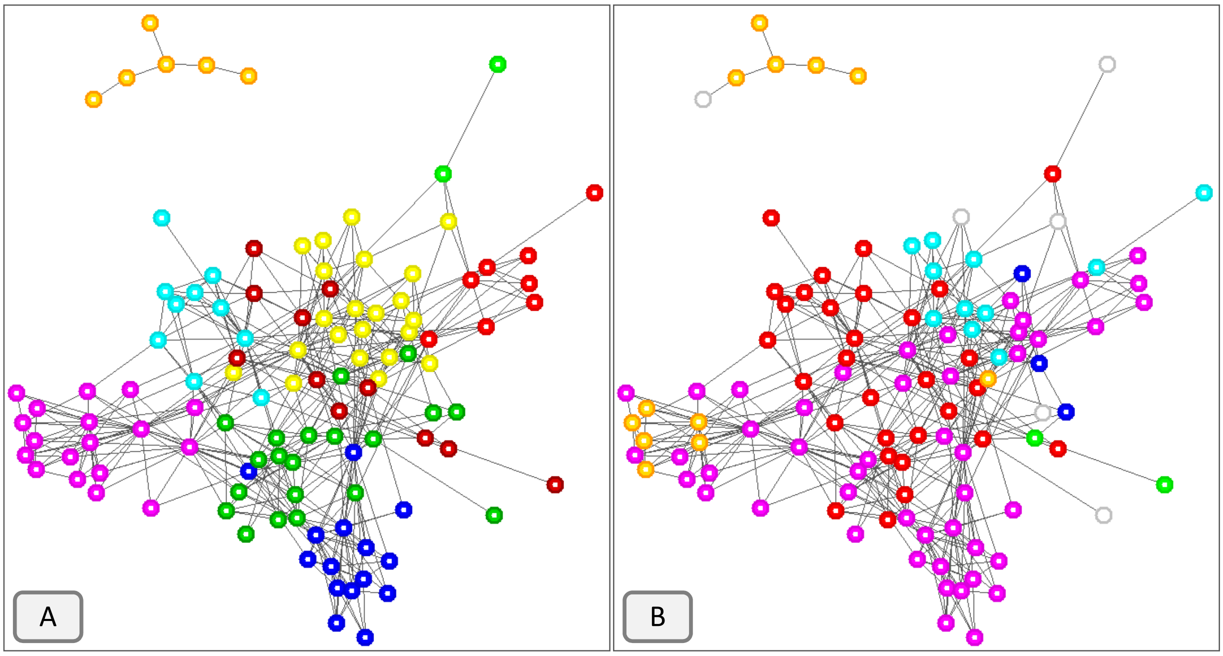

Figure 1: Structural versus content-based community detection. (A) Structural clustering: 9 clusters, square error: 3.47; (B) Content-based clustering: 12 clusters, 6 singleton clusters (light-grey), square error: 2.77.

In contrast to the structural clustering obtained from the edge betweenness clustering algorithm the content-based clusters can contain more than one connected component and sometimes, as with clusters C4 and C6, also include entirely unconnected actors (singletons). As the larger content-based clusters are spread over the entire network the group density decreases with the group size whereas the communicational activity tends to increase. As with the structural clustering the clustering coefficient is high for medium sized clusters and then decreases. But again, even in the largest cluster many contacts are also connected themselves. The content dissimilarity uniformly increases with cluster size but the relation of content dissimilarity to cluster size decreases. Thus, the group members become more similar with increasing group size.

Table 1: Overview of structural clusters and structural group membership stability.

Cluster	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**
Color	light green	orange	red	cyan	dark red	dark blue	magenta	green	yellow
Members	2	6	8	10	11	15	18	20	22
Square Error	0.23	0.29	0.24	0.18	0.52	0.20	0.18	0.76	0.87
Density [%]	100.0	33.33	60.71	55.56	43.64	49.52	40.52	30.53	7.83
Clustering Coeff. [%]	-	0.00	68.18	66.39	67.92	62.88	49.04	44.21	46.34
∅ Rel. Strength	1.00	10.80	10.00	71.60	21.08	18.94	10.94	9.97	30.30
AGMS#1(a)	0.00	0.67	0.57	1.00	0.67	1.00	0.53	0.34	0.25
AGMS#2(a)	0.00	0.42	0.17	0.47	0.36	0.55	0.35	0.24	0.28

Comparing the AGMS#1(a) values some structural group members belong almost entirely to the same content-based cluster: cluster S2 is mostly included in cluster C4, cluster S5 mostly in cluster

C5, clusters S3 and S7 mostly in cluster C6. Some structural groups belong entirely to the same content-based cluster, namely cluster S4 to cluster C5 and cluster S6 to cluster C6. The two largest structural clusters S8 and S9 dissolve in several content-based clusters: cluster S8 in clusters C5, C6 and some singletons, cluster S9 in clustersC2, C6, C3 and some singletons. These two clusters have also a higher content dissimilarity. Comparing the AGMS#2(a) values even if most or all members stick together they now belong to larger groups with several new members. The average group membership stability for the whole network is 0.56 for GMS#1(a) and 0.34 for GMS#2(a). That means, for each actor more than half of the structural group members are in the same content-based cluster and about one third of the structural clustering structure (group size and group membership) stays the same.

Table 2: Overview of content-based clusters (non-singletons).

Cluster	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Color	light green	dark blue	cyan	orange	red	magenta
Members	2	3	11	12	30	48
Components (Singletons)	1	1	2	3 (1)	1	2 (1)
Square Error	0.00	0.13	0.29	0.37	0.86	1.12
Density [%]	100.00	66.67	34.55	24.24	20.69	12.50
Clustering Coeff. [%]	-	0.00	58.33	71.43	45.89	40.93
∅ Rel. Strength	2.00	1.50	13.89	8.31	29.77	23.69

In summary, the content-based clusters of the subsample of the Enron dataset still tend to cover well connected and somewhat separated regions of the network, like cluster S6, if they are established through similar content. But at the same time the clustering procedure groups together those actors from different parts of the network, like cluster C4, that are involved into similar topics without regarding their structural relationships.

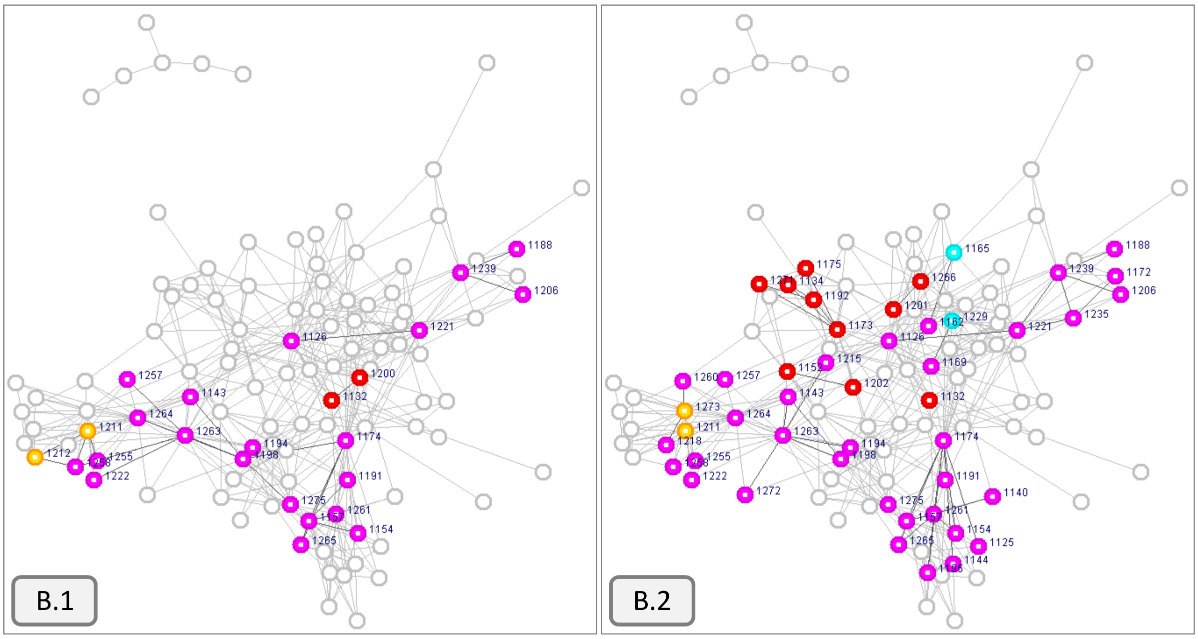

Figure 2: Topic spread in content-based clustering. Node color: content-based cluster membership (light gray: filtered out), node label: database id. (B.1) Topic: ‘load curtailment’; (B.2) Topic: ‘Dynegy’.

Figure 2 provides a visual validation of the content-based clustering solution (B) in Figure 1 for two different topics. Network (B.1) shows all actors whose content collections contain the topic ‘load curtailment’, network (B.2) those whose content collections contain the topic ‘Dynegy’. ‘Load curtailment’ refers to reducing the energy demand peaks by load balancing, whereas ‘Dynegy’ refers to Dynegy Inc. which is a large energy company and has been a rival of Enron in the energy market. The actors not containing these topics are filtered out. For a better localization of the topic subgroups the inactive nodes and edges are presented in light gray. Both topic subgroups are spread over the entire network. Although the majority of actors involved in the topics belong to content-based cluster C6, the first topic is better represented by the clustering solution as more actors from the second topic belong to other clusters (namely C3, C4 and C5). However, a clustering procedure establishes group memberships on properties that help to distinguish objects. Thus, if a topic is communicated by too many actors it is not meaningful enough to distinguish them and the groups are established on lesser used topics. The results from the content-based clustering can be used to classify topics and content objects in terms of relevance and distinction. Comparing the topic subgroup in (B.1) with the entire network and the structural clusters in (A) it becomes clear that a mere structural perspective on communicational activities fails to reveal the complex nature of collaborative content.

7. Conclusion

This paper introduces a content-based approach to clustering social networks. Similar to most previous studies in the domain of clustering, it could be noticed, that a main challenge is the definition of the number of clusters. Using a corporate dataset of e-mail communication, we could show that the method is able to produce a series of content-related clusters and thus to support the detection of content-based online communities. A comparison of structural and content-based clusters yielded the insight that significant differences between the results of both procedures could be identified. Especially if a topic is spread throughout the structural groups resulting into similar but disconnected actors, or if a structural group is very large, content-based analysis can add value to conventional structural analysis. These results motivate further inquiries into content-based detection of contribute to better algorithms to identify group-level properties of electronic networks. However, in content-based communities, better known as communities of practice or communities of interest, actors can be a member of more than one community joining them on basis of their (temporal) interests [25]. Therefore, we intend to focus on employing clustering-algorithms (such as cliques and n-clans) that allow for multiple cluster memberships.

8. Literature

[1] ALDENDERFER, M.S. and R.K. BLASHFIELD, Cluster Analysis, Sage Publications, Newbury Park 1984.

[2] BIKSON, T.K. and J.D. EVELAND, The interplay of workgroup structures and computer support, in: Intellectual teamwork: Social and technological foundations of cooperative work, J. Galagher, R. Kraut and C. Egido (Eds.), p. 243-290, NJ: Erlbaum, Norwood 1990.

[3] CASTELLANOS, M., HotMiner: Discovering Hot Topics from Dirty Text, in: Survey of text mining: Clustering, Classification, and Retrieval, M.W. Berry (Ed.), Springer 2003.

[4] DICE, L.R., Measures of the amount of ecological association between species, J. Ecology **26**, p. 297-302 (1945).

[5] FALLOWS, D., Email at work - few feel overwhelmed and most are pleased with the way email helps them do their jobs. PEW Internet and American Life Project, [cited 2002-2002-01-28]; Available from: http://207.21.232.103/pdfs/PIP_Work_Email_Report.pdf (2002).

[6] GIRVAN, M. and M.E.J. NEWMAN, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA **99**, p. 7821–7826 (2002).

[7] HARTIGAN, J.A., Clustering Algorithms, Wiley-Interscience, New York 1975.

[8] JAIN, A.K. and R.C. DUBES, Algorithms for Clustering Data, Pretence Hall, Endelwood Cliffs, New Jersey 1988.

[9] KAUFMAN, L. and P.J. ROUSSEEUW, Finding Groups in Data - An Introduction to Cluster Analysis, Wiley-Interscience, Hoboken, New Jersey 1990.

[10] KOSSINETS, G. and D.W. WATTS, Empirical analysis of an evolving social network, Science **Jan 6**, p. 88-90 (2006).

[11] KRACKHARDT, D., The strength of strong ties: The importance of philos in organizations, in: Organizations and networks: Theory and practice, N. Nohiram and R. Eccles (Eds.), p. 216-239, MA: Cambridge University Press, Cambridge 1991.

[12] LEVINE, S.S. and R. KURZBAN, Explaining Clustering in Social Networks: Towards an Evolutionary Theory of Cascading Benefits, Managerial and Decision Economics **27**(2-3), p. 173-187 (2006).

[13] MCQUEEN, J.B., Some methods of classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, p. 281-297 (1967).

[14] MCQUITTY, L.L., Similarity analysis by reciprocal pairs for discrete and continuous data, Educational and Psychological Measurement **27**, p. 21-46 (1966).

[15] MILLIGAN, G.W. and M.C. COOPER, An examination of procedures for determining the number of clusters in a data set, Psychometrika **50**, p. 159-179 (1985).

[16] NEWMAN, M.E.J. and M. GIRVAN, Finding and evaluating community structure in networks, Physical Review E **69**, p. 026113 (2004).

[17] RAVASZ, E. and A.-L. BARABÁSI, Hierarchical Organization in Complex Networks, Physical Review E **67** (2003).

[18] SALTON, G. and M. MCGILL, Introduction to Modern Information Retrieval, McGraw-Hill, New Yoek 1984.

[19] TRIER, M., IT-supported Visualization and Evaluation of Virtual Knowledge Communities, Faculty of Computing Sciences and Electrical Engineering, Technical University of Berlin: Berlin, p. 270 (2005).

[20] TRIER, M., Towards Dynamic Visualization for Understanding Evolution of Digital Communication Networks, Information Systems Research **19**(3) (2008).

[21] TRIER, M. and A. BOBRIK, Social Search: Exploring and Searching Social Architectures in Digital Networks, Forthcoming in IEEE Internet Computing (2008).

[22] TYLER, J.R., D.M. WILKINSON and B.A. HUBERMAN, Email as spectroscopy: automated discovery of community structure within organizations, in: Proceedings of Communities and Technologies, Kluwer (2003).

[23] VAISHNAVI, V. and W. KUECHLER, Design Research in Information Systems, 2004-01-20, [cited 2006-05-12]; Available from: http://www.isworld.org/Researchdesign/drisISworld.htm (2004).

[24] WASSERMAN, S. and K. FAUST, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge 1994.

[25] WENGER, E., R. MCDERMOTT and W.M. SNYDER, Cultivating Communities of Practice, Harvard Business School Press, Boston 2002.

[26] ZHANG, J. and M.S. ACKERMAN, Searching for expertise in social networks: a simulation of potential strategies, in: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work (2005).