

Analysis of Community Structure, Affinity Data and Research Activities using Additive Fuzzy Spectral Clustering

Boris Mirkin¹ and Susana Nascimento²

¹Department of Computer Science
Birkbeck University of London
London WC1E 7HX, UK

²Computer Science Department and Centre for Artificial Intelligence (CENTRIA)
FCT, Universidade Nova de Lisboa
Caparica, Portugal

August 28, 2009

1 Introduction

This work is motivated by the problem of clustering research topics within a Computer Science research organization according to the similarity between topics derived on the basis of the efforts by researchers engaged in them simultaneously. Such a cluster would naturally represent an elementary unit of the department's research as a whole, so that it can be used for placement of the organization within a taxonomy of the field [8] such as the ACM Computing Classification System (ACM-CCS) [1]. We develop a tool for measuring similarity between research topics by asking researchers to indicate those topics that are subject of their current research, along with the proportions of the total effort devoted to them. Then the similarity between topics according to an individual can be defined as the product of these proportions, thus leading to an inner-product-like measure of total similarity according to the organization's research. Such a measure makes it natural to consider the similarity a result of additive action of fuzzy clusters representing the main directions of the organization's research. The additive model is a natural extension of the additive clustering model [16, 15], which itself is an extension of the principal component analysis based on the spectral decomposition of square matrices. Therefore, we develop a sequential fitting method [6] by using the spectral clustering approach. Before application of the method to real-world data on similarity between ACM-CCS research topics, we explore its capabilities as a similarity clustering method on two types of data that are subject of ongoing research efforts by various authors: (a) community structure graphs and (b) affinity similarity data derived from feature based information, and show that it performs well. The method has a number of links to some published methods but differs in that it is model based, includes innate stopping criteria, and utilizes meaningful data normalization options.

2 Similarity Between Research Topics and Additive Fuzzy Clusters

Consider a set of V individuals ($v = 1, 2, \dots, V$), engaged in some of research topics $t \in T$ where T is a prespecified set of scientific subjects. The level of research effort by individual v over topic t is evaluated by a real f_{tv} , which is greater than 0 but smaller than 1. The value f_{tv} may reflect the proportion of research efforts given by v to t and, as such, can represent some summary estimate based on the body of work including such items as published papers, grants obtained and projects completed. Then, for each of the items, the extent of its relation to the corresponding topics can be evaluated. The estimates f_{tv} can be derived from the body of documents posted on web, though this method can be applied only to organizations whose members do post English-written documents of their research on the Internet.

Then the similarity $a_{tt'}$ between topics t and t' can be defined as the inner product of vectors of scores $f_t = (f_{tv})$ and $f_{t'} = (f_{t'v})$, $v = 1, 2, \dots, V$, so that $a_{tt'} = \langle f_t, f_{t'} \rangle = \sum_{v=1}^V f_{tv} f_{t'v}$.

Since our survey tool makes all the individual scores sum up to unity, $\sum_{t \in T} f_t = 1$, the scores of individuals bearing more topics tend to be smaller than those of individuals engaged in fewer numbers of topics. To make up for this, we utilize a natural weighting factor, the ratio of the number of topics marked by individual v , n_v , and n_{max} , the maximum n_v over all $v = 1, 2, \dots, V$.

$$a_{tt'} = \sum_{v=1}^V \frac{n_v}{n_{max}} f_{tv} f_{t'v}, \quad (1)$$

Table 1 represents scores by four individuals to six research topics. The individual weights n_v are in the bottom row.

	i_1	i_2	i_3	i_4
A	0.6			0.2
B	0.4		0.2	0.2
C		0.2	0.4	0.2
D		0.3	0.4	0.2
E		0.5		0.2
F				
n_v/n_{max}	2	3	3	5

Table 1: A sample of membership values for six subjects A–F assigned by four individuals. Each individual is assigned with weight reflecting the number of topics with a positive score (bottom line).

The following decomposition exhibits equation (1) individual-wise so that contributions of individuals to the final similarity are easily seen:

$$\begin{aligned}
& \frac{2}{5} \times \begin{bmatrix} 0.36 & 0.24 & 0.0 & 0.0 & 0.0 \\ 0.24 & 0.16 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} + \frac{3}{5} \times \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.04 & 0.06 & 0.1 \\ 0.0 & 0.0 & 0.06 & 0.09 & 0.15 \\ 0.0 & 0.0 & 0.1 & 0.15 & 0.25 \end{bmatrix} + \\
& \frac{3}{5} \times \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.04 & 0.08 & 0.08 & 0.0 \\ 0.0 & 0.08 & 0.16 & 0.16 & 0.0 \\ 0.0 & 0.08 & 0.16 & 0.16 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} + \frac{5}{5} \times \begin{bmatrix} 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \end{bmatrix} = \\
& = \begin{bmatrix} \mathbf{0.184} & \mathbf{0.136} & 0.040 & 0.040 & 0.040 \\ \mathbf{0.136} & \mathbf{0.128} & 0.088 & 0.088 & 0.040 \\ 0.040 & 0.088 & \mathbf{0.160} & \mathbf{0.172} & \mathbf{0.100} \\ 0.040 & 0.088 & \mathbf{0.172} & \mathbf{0.190} & \mathbf{0.130} \\ 0.040 & 0.040 & \mathbf{0.100} & \mathbf{0.130} & \mathbf{0.190} \end{bmatrix}
\end{aligned}$$

The final matrix exhibits an obvious two-cluster structure, which is made easily discernible by putting all figures of 0.100 and greater in bold font.

The similarity measure (1) has the following properties:

- The similarity between two topics can be positive if and only if there is at least one researcher that is engaged in both.
- The similarity matrix is definite semipositive.
- The greater the individual membership values, the greater the similarity.
- Given a pair of topics, the greater the number of researchers engaged in them, the greater the similarity.

Assume that the individual researchers' scores are but manifested expressions of some hidden trends within the organization that can be represented by fuzzy clusters in exactly the same manner as the manifested scores in the definition of the similarity 1. We assume that a fuzzy cluster of topics is represented by a membership vector $u = (u_t)$, $t \in T$, such that $0 < u_t < 1$ for all $t \in T$, and an intensity $\mu > 0$ that expresses the extent of significance of the trend represented by the cluster, within the organization under consideration. The introduction of the intensity, applied as a scaling factor to u , makes the product μu to be a solution rather than its individual co-factors. That is, given a value of μu_t , it becomes impossible to separate which part of it is μ and which u_t . To resolve this issue, we follow the conventional scheme: let us impose a constraint on the scale of the membership vector u to be fixed at a constant level, for example, by a condition such as $\sum_t u_t = 1$ or $\sum_t u_t^2 = 1$, then the remaining scaler value will define the value of μ .

Our additive fuzzy clustering model involves K fuzzy clusters that reproduce the similarities up to small errors according to the following equations:

$$a_{tt'} = \sum_{k=1}^K \mu_k^2 u_{kt} u_{kt'} + e_{tt'}, \quad (2)$$

where $u_k = (u_{kt})$ is the membership vector of cluster k , and μ_k its intensity.

The product $\mu_k^2 u_{kt} u_{kt'}$ expresses the contribution of cluster k to the similarity $a_{tt'}$ between topics t and t' , which depends on both the cluster's intensity and the membership values. The value μ^2 summarizes the contribution of intensity and will be referred to as the cluster's weight.

Taking the product of values μu_t and $\mu u_{t'}$ to express the extent of similarity between t and t' reflects the probabilistic interpretation of memberships and, also, makes it mathematically simpler. However, a different definition, utilizing operations of maximum or minimum can also be taken without much changing the computational structure of our approach.

The problem of fitting the model (2) can be formalized by using the least-squares approach: given matrix $A = (a_{tt'})$, find K and fuzzy clusters u_k along with their intensities μ_k to minimize the sum of squares of the errors, $\sum_{t,t'} e_{tt'}^2$.

As is well known, provided that A is definite semi-positive, the first K eigenvalues and corresponding eigenvectors form the solution to the problem if no constraints on vectors u_k are imposed. On the other hand, if vectors u_k are constrained to be just 1/0 binary vectors, the model (2) becomes of the so-called additive clustering [16, 5]. A simplified version of model (2), involving a constant, not cluster-specific, intensity, was considered in [15] – the authors proposed to use the Newton's descent method, thus involving many initialization parameters that need to be pre-specified, which is not what an innocent user would be willing to do.

3 Fitting the Additive Fuzzy Clustering Model with a Spectral Method

We are going to apply the one-by-one principal component analysis strategy for finding one cluster at a time. The strategy works for principal components, because they are mutually orthogonal, but it brings some advantages to the non-orthogonal solutions as well. Found clusters have good tightness properties and are assigned with additive proportions of the data scatter explained by them [6, 9].

Specifically, at each step, we will be considering the problem of minimization of criterion

$$E = \sum_{t,t' \in T} (w_{tt'} - \xi u_t u_{t'})^2 \quad (3)$$

with respect to unknown positive ξ weight (so that the intensity μ is the square root of ξ) and fuzzy membership vector $u = (u_t)$, given similarity matrix $W = (w_{tt'})$.

At the first step, W is taken to be equal to A . Each found cluster changes W by subtracting the contribution of the found cluster (which is additive according to model (2)), so that the residual similarity matrix for obtaining the next cluster will be $W - \mu^2 u u^T$ where μ and u are the intensity and membership vector of the found cluster. In this way, A indeed is additively decomposed according to formula (2) and the number of clusters K can be determined in the process.

To arrive at the spectral clustering framework, let us specify an arbitrary membership vector u and find the value of ξ minimizing criterion (3) at this u . Obviously, criterion (3)

is a convex function of λ so that the first-order condition of optimality should solve the problem:

$$\frac{\partial E}{\partial \xi} = -2 \sum_{t,t' \in T} (w_{tt'} - \xi u_t u_{t'}) u_t u_{t'} = 0.$$

This implies that

$$\xi = \frac{\sum_{t,t' \in T} w_{tt'} u_t u_{t'}}{\sum_{t \in T} u_t^2 \sum_{t' \in T} u_{t'}^2}$$

In matrix terms the optimal ξ is

$$\xi = \frac{\mathbf{u}' W \mathbf{u}}{(\mathbf{u}' \mathbf{u})^2} \quad (4)$$

which is obviously non-negative if matrix W is semi-positive definite.

By putting this ξ in equation (3), one can easily derive that

$$E = \sum_{t,t' \in T} w_{tt'}^2 - \xi^2 \sum_{t \in T} u_t^2 \sum_{t' \in T} u_{t'}^2 = S(W) - \xi^2 (\mathbf{u}' \mathbf{u})^2,$$

where $S(W) = \sum_{t,t' \in T} w_{tt'}^2$ is the similarity data scatter.

Let us denote the last item by

$$G(u) = \xi^2 (\mathbf{u}' \mathbf{u})^2 = \left(\frac{\mathbf{u}' W \mathbf{u}}{\mathbf{u}' \mathbf{u}} \right)^2, \quad (5)$$

so that the similarity data scatter admits the decomposition

$$S(W) = G(u) + E \quad (6)$$

into part $G(u)$ explained by cluster (μ, u) and part E remaining unexplained by the cluster. This allows us to express the contribution of the cluster to data scatter as the proportion $G(u)/S(W)$.

Turning to the problem of finding an optimal cluster, one can see from (6) that it is to maximize the explained part $G(u)$ in (5) or its square root

$$g(u) = \xi \mathbf{u}' \mathbf{u} = \frac{\mathbf{u}' W \mathbf{u}}{\mathbf{u}' \mathbf{u}}, \quad (7)$$

which is the celebrated Rayleigh quotient, whose maximum value is the maximum eigenvalue of a symmetric matrix W , which is reached at the corresponding eigenvector of matrix W in the unconstrained problem.

This shows that the spectral clustering approach is appropriate in our problem. According to this approach, one should find the maximum eigenvalue λ and corresponding normed eigenvector z for W , $[\lambda, z] = \Lambda(W)$, and take its projection to the set of admissible fuzzy membership vectors. In a simplest case of unconstrained fuzzy memberships, the projection operator $\mathcal{P}(z)$ works as follows:

$$\mathcal{P}(z) = \begin{cases} 0, & \text{if } z \leq 0; \\ z, & \text{if } 0 < z < 1; \\ 1, & \text{if } z \geq 1. \end{cases}$$

If, however, one requires that the fuzzy clusters form a fuzzy partition so that $\sum_{k=1}^K u_{kt} = 1$ for each $t \in T$, then after each cluster extraction step k , $k = 1, 2, \dots, K - 1$, the cumulative belongingness $\alpha_{kt} = \sum_{l=1}^k u_{lt}$ should be taken into account. For each t , the unity in the definition of $\mathcal{P}(z)$, should be changed for $1 - \alpha_{kt}$ – this will warrant that $\sum_{k=1}^K u_{kt} \leq 1$.

There are a number of properties of the outlined procedure for sequential extraction of fuzzy clusters along with their weights that will be taken into account in the follow up:

1. The residual matrix W can get negative eigenvalues even if the initial matrix A is semi-positive definite, which may bring the procedure to the end because the formula (4) would lead to a negative ξ .
2. The matrix A can be equivalently substituted by its symmetrized version $\tilde{A} = (A + A')/2$.
3. The cluster contributions are additive so that each can be expressed as a proportion of the initial similarity data scatter $S(A)$.
4. The cluster contributions tend to decrease at each step, but they do not necessarily form a monotonely decreasing sequence, because the spectral cluster does not necessarily minimize criterion (2).
5. The procedure converges so that the scatter of the residual data decreases after each step.

Most items among the above are based on common sense and experimental evidence, but a couple can be proven in a mathematically rigorous way as follows.

Assertion 1. *At any u , the value of the criterion $g(u)$ does not change if $\tilde{W} = (W + W')/2$ is put instead of W .*

Proof: Indeed, when $w_{tt'}$ is changed for $w_{tt'}^s$, the only items affected are in the numerator, the sum $w_{tt'}u_t u_{t'} + w_{t't}u_{t'} u_t$. But $w_{tt'}^s u_t u_{t'} + w_{t't}^s u_{t'} u_t = (w_{tt'} + w_{t't})u_t u_{t'} + (w_{t't} + w_{tt'})u_{t'} u_t)/2 = w_{tt'}u_t u_{t'} + w_{t't}u_{t'} u_t$, which proves the statement.

Thus, the optimizers of $g(u)$ do not change if W^s is used in (7). Therefore, in the remainder, it is always assumed that the data matrix W has been symmetrized with the transformation $\tilde{W} = (W + W')/2$, which guarantees that no complex-valued eigenvalue may appear in our computations.

Assertion 2. *The contributiona of consecutive clusters to the original similarity data A are additive so that equation*

$$S(A) = G_1(u_1) + G_2(u_2) + \dots G_K(u_K) + E_K, \quad (8)$$

where $G_k(u_k)$ are values of the contribution (5) at the residual W before the k -th cluster is extracted, and E_K is the scatter of the final residual matrix.

Proof: We prove the formula for $K = 2$, the other K values can be treated by induction. Indeed, the scatter of the residual matrix $W_1 = W - \mu_1^2 u_1 u_1'$ after subtraction of the first cluster is but the value of E_1 in equation (6) for this case: $S(W) = G_1(u_1) + E_1$. That means that the similar decomposition $S(W_1) = G_2(u_2) + E_2$ for the second cluster can be put in the former equation that becomes $S(W) = G_1(u_1) + G_2(u_2) + E_2$, which proves the statement.

There can be a number of criteria for halting the process of sequential extraction of fuzzy clusters according to its properties:

1. The optimal value of ξ (4) for the spectral fuzzy cluster is negative.
2. The contribution of a single extracted cluster becomes too low, less than a prespecified $\tau > 0$ value (for example, for a network like Karate club of about 30 members, a cluster should contribute at least as much as an average entity, so that $\tau = 1/30$ should be considered a fair choice in this problem).
3. The residual scatter E becomes smaller than a prespecified ϵ value, say less than 5% of the original similarity scatter.
4. A prespecified number K_{\max} of clusters is reached – in many real-world problems such a number is easy to set,

The material above allows us to use the following one-by-one fuzzy cluster extraction spectral algorithm, referred to as ADDI-FS as an extension of the ADDI-S crisp clustering algorithm described in [5]:

ADDI- FS Algorithm

Set $k = 0$, $W = A$, $\epsilon > 0, \tau > 0$;

Repeat

$$k = k + 1;$$

$$[\xi, z] = \Lambda(W);$$

If $\xi > 0$

$\mathbf{u}_k = \mathcal{P}(z)$ or $\mathbf{u}_k = \mathcal{P}(-z)$ depending on which leads to a larger G_k ;

$$\xi_k = \frac{\mathbf{u}_k' W \mathbf{u}_k}{(\mathbf{u}_k' \mathbf{u}_k)^2};$$

$$G_k = \xi_k^2 (\mathbf{u}_k' \mathbf{u}_k)^2;$$

$$W = W - \xi^k \mathbf{u}_k \mathbf{u}_k';$$

$$S(W) = Tr(W'W);$$

Else: Halt.

Until ($\xi_k \leq 0$ or $G_k \leq \tau$ or $S(W) \leq \epsilon$ or $k == K_{\max}$)

Before proceeding to the analysis of data on similarity between ACM-CCS research topics, one should ask whether the algorithm proposed is competitive with respect to such bench-mark clustering problems as finding community structure and the analysis of affinity similarity index produced from feature based data. These two applications involve different types of data, which, in the authors' view, should be addressed differently.

4 Application to Finding Community Structure

The research in finding community structure in ordinary graphs has been revitalized recently by M. Newman and others, with the usage of the so-called modularity criterion and putting it into spectral analysis frameworks (see, for example, [12, 11, 18, 4]). The graph with a set of vertices T is represented by the similarity matrix $A = (a_{tt'})$ between graph vertices such that $a_{tt'} = 1$ if t and t' are connected by an edge, and $a_{tt'} = 0$, otherwise. Then matrix A is symmetrized by the transformation $(A + A')/2$ after which all diagonal elements are made zero, $a_{tt} = 0$, for all $t \in T$. The modularity criterion is not necessarily monotone over the number of clusters, which some claim can be used for determining the number of clusters K (see, for example, [18, 19]).

The spectral relaxations suggested in the literature involve some transformation of the graph similarity matrix $A = (a_{tt'})$ using the within-row summary values $d_t = \sum_{t' \in T} a_{tt'}$. One of them is subtraction from $a_{tt'}$ an "expected value", which is usually taken to be proportional to the product $d_t d_{t'}$ [11]. Another transformation is dividing $a_{tt'}$ by $\sqrt{d_t d_{t'}}$ [13]. This approach was extended to fuzzy clustering in the space of the first eigenvectors in [19].

Our approach allows for a straightforward application of ADDI-FS algorithm to the network similarity matrix A . First of all, we count on the eigenvector z_1 corresponding to the maximum eigenvalue λ_1 of A . As is well known, this vector is non-negative, thus forming the first fuzzy cluster itself, when conventionally normed. Vector z_1 bears information of the network interaction much beyond that of the vector (d_t) of incidence frequencies. This allows us to hope that subtracting $\lambda_1 z_{1t} z_{1t'}$ from $w_{tt'}$ will make the community structure much clearer. We also expect the matrix A to be rather "thin" with respect to the number of positive eigenvalues, which should allow for a natural halting the cluster extracting process when there are no positive eigenvalues at the residual matrix W .

We first apply ADDI-FS algorithm to Zachary karate club network data, which serves as a prime test bench for community finding algorithms. This simple graph consists of 34 vertices, corresponding to members of the club and 78 edges between them - the data and references can be found, for example, in [12, 19]. The members of the club are divided according to their loyalties towards the club's administrator or instructor. Thus the network is claimed to consist of two communities, with 18 and 16 differently loyal members respectively.

Applied to this data, ADDI-FS leads to three fuzzy clusters to be taken into account. Indeed, the fourth cluster accounts for only 2.4%, which is smaller than the inverse of the number of entities $\tau = 1/34$ suggested above as a natural threshold value. Some characteristics of the found solution are presented in Table 2.

Table 2: Characteristics of Karate club clusters found with ADDI-FS.

Cluster	Contribution, %	λ_1	Weight	Intensity
I	29.00	3.36	3.36	1.83
II	4.34	2.49	1.30	1.14
III	4.19	2.00	0.97	0.98

All the membership values of the first cluster are positive - as mentioned above, this

is just the first eigenvector; the positivity means that the network is well connected. The second and third ADDI-FS clusters match the claimed structure of the network: they have 16 and 18 positive components, respectively, corresponding to the two observed groupings.

This rather perfect matching is at odds with the results of fuzzy clustering method developed in [19]: the authors report of three fuzzy clusters, two of them representing the groupings but with a substantial overlap between them, and the third, smaller, cluster consisting of members 5,6,7,11,17 of just one of the groupings – see [19], p. 487. In the current authors’ opinion this latter cluster may have come as of the members with the largest numbers of connections in the network.

To see the performance of ADDI-FS algorithm on a larger scale, we devised an experiment in randomly drawing a community network. This network comprises two communities, each consisting of a random number of members from 6 to 15; the connecting edges are drawn uniformly randomly with probability p within each community and probability q between the communities. Although the uniform distributions do not necessarily reflect those in real world networks [12, 11], this seems an appropriate device for testing a general clustering algorithm such as ADDI-FS.

After a network is generated, ADDI-FS is run; then the first membership vector is discarded, and the following two types of errors are recorded over the two entity sets corresponding to the positive membership values in membership vectors 2 and 3, after identifying that of the generated communities they correspond to:

- the confusion error, which is the number of entities wrongly assigned between the two clusters, related to the total number of entities generated;
- the omission error, which is the number of entities not assigned to clusters 2 and 3 at all, related to the total number of entities generated;

Table 3 represents averages and standard deviations of each of these two error values over 1000 data generation runs. Each cell in it corresponds to a pair (p, q) , $p = 0.6, 0.7, 0.8, 0.9$ and $q = 0.1, 0.2, 0.3, 0.4$; each of the mean values is accompanied with the standard deviation after slash. Within any cell, the confusion error is on the top with the omission error underneath.

	0.4	0.3	0.2	0.1
0.6	0.329/0.110 0.173/0.112	0.237/0.136 0.146/0.120	0.160/0.140 0.114/0.119	0.166/0.151 0.140/0.131
0.7	0.227/0.132 0.123/0.118	0.141/0.129 0.088/0.110	0.103/0.125 0.082/0.117	0.128/0.145 0.121/0.138
0.8	0.110/0.111 0.064/0.103	0.072/0.100 0.051/0.102	0.061/0.096 0.050/0.104	0.098/0.135 0.089/0.131
0.9	0.043/0.069 0.019/0.067	0.031/0.059 0.014/0.058	0.036/0.071 0.025/0.082	0.074/0.125 0.062/0.131

Table 3: The average confusion and omission errors of ADDI-FS clusters, along with their standard deviations, at different probabilities of the within community links (in rows) and between community links (in columns) resulting from a thousand data generation runs.

As one can see, the errors are rather high at $p = 0.6$, reaching its minimum of total 27.4 % at $q = 0.2$. At each q , the errors decrease with the growth of p . A non-trivial feature of the error, that has been always observed over many series of 1000 runs of the data generation routine, is the lack of monotonicity of the errors with respect q . It appears, the error is always smaller at $q = 0.2$ than at $q = 0.1$. Moreover, with the growth of p the minimum error moves to even greater q values. For example, at $p = 0.9$, the error, totaling to 4.5

Overall, the results show some consistency in the method, and in fact, its efficiency in discovering two-community structure. Its performance when there are more communities in the graph remains to be tested.

5 Application to the Analysis of Affinity Data

The affinity data is similarity obtained from a feature based dataset with a semi-positive definite kernel, usually the Gaussian one. Specifically, given an $N \times V$ matrix $Y = (y_{tv})$, $t \in T$ and $v = 1, 2, \dots, V$, non-diagonal elements of the similarity matrix W are defined by equation

$$w_{tt'} = \exp\left(-\frac{\sum_{v=1}^V (y_{tv} - y_{t'v})^2}{2\sigma^2}\right),$$

with the diagonal elements made equal to zero, starting from founding papers [17, 13]. To apply the spectral approach, this matrix is pre-processed into the so-called Laplacian matrix. First, an $N \times N$ diagonal matrix D is defined, with (t,t) element equal to $d_t = \sum_{t' \in T} w_{tt'}$, the sum of t 's row of W . Then unnormalized Laplacian and normalized Laplacian are defined with equations $L = D - W$ and $L_n = D^{-1/2}LD^{-1/2}$, respectively. Both matrices are semipositive definite and have zero as the minimum eigenvalue. The minimum non-zero eigenvalues and corresponding eigenvectors of the Laplacian matrices are utilized then as relaxations of combinatorial partition problems [17, 13, 19, 4]. Of comparative properties of these two normalizations, the normalized Laplacian, in general, is considered superior [4].

Yet the Laplacian normalizations cannot be used in our approach straightforwardly, because ADDI-FS relies on maximum rather than minimum eigenvalues. To overcome this issue, the authors of [13] utilized a complementary matrix $M_n = D^{-1/2}WD^{-1/2}$ which relates to L_n with equation $L_n = I - M_n$ where I is the identity matrix. This means that M_n has the same eigenvectors as L_n , whereas respective eigenvalues relate to each other as λ and $1 - \lambda$, so that the matrix M_n can be utilized for our purposes as well. Yet we prefer using the pseudoinverse $L_n^- = \tilde{Z}\tilde{\Lambda}^{-1}\tilde{Z}'$ where $\tilde{\Lambda}$ and \tilde{Z} are defined by the spectral decomposition $L_n = Z\Lambda Z'$ of matrix L_n . First, set T' of indices of elements corresponding to non-zero elements of Λ is determined, after which the matrices are taken as $\tilde{\Lambda} = \Lambda(T', T')$ and $\tilde{Z} = Z(:, T')$. The choice of the pseudoinverse can be explained by the fact that the maximum eigenvalue of L_n^- is the inverse of the minimum non-zero eigenvalue λ_1 of L_n , corresponding to the same eigenvector. The inverse of a λ near 0 could make the value of $1/\lambda$ quite large and greatly separate it from the inverses of other near zero eigenvalues of L_n . Consider, for example, $\lambda_1 = 0.1$ and $\lambda_2 = 0.2$ so that their complements to unity are 0.95 and 0.8 while the inverses are 20 and 5 – the difference in gaps between the values is impressive indeed. The latter suits the ADDI-FS one-by-one

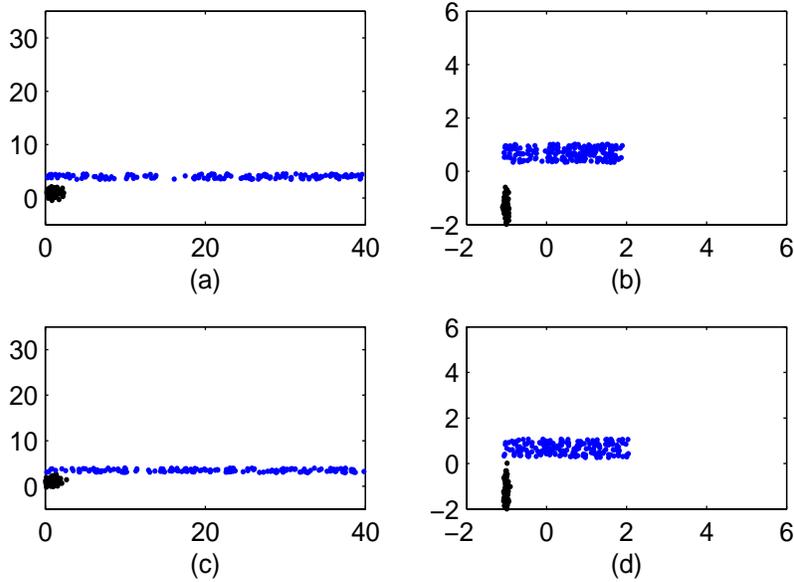


Figure 1: Two versions of clusters of different shapes: that on (a) corresponds to distance 3.5 between them over y-axis, and on (c), distance 0.5 over y-axis. Figures (b) and (d) present the clusters after z-scoring of the data.

approach much better.

To see how this approach works, we adapt an example from [10]: two 2D clusters are generated, one from a normal distribution corresponding to a small ball whereas the other from a uniformly distributed strip, which is much longer. By changing the distance between these clusters, one can test consistency of a clustering method. Specifically, a hundred points are randomly generated by using Gaussian distribution $N(1,0.5)$ over both axes to make cluster 1, and two hundred points are generated uniformly randomly in a strip taking the fragment of x-axis from 0 to 50 while over y-axis the width is one. This is illustrated on Figure 1: the strip is put at $y = 3.5$ on part (a) and at $y = 0.5$ on part (c) of it. Then the data are standardized with conventional z-scoring: by subtracting grand means from each of the coordinates and dividing the results by the feature's standard deviation.

It should be pointed out that the general similarity information has been cleaned out of the data by using the pseudo-inverse Laplace transformation, which implies that the very first fuzzy cluster should correspond to a meaningful grouping in data.

Table 4 presents the averaged results of of ADDI-FS algorithm over a hundred runs of data generation at three different ratios of the cluster sizes. The number of points generated is always 300, but 100 of them belong to the ball in the left column, 150 in the middle column, and 200 in the column on the right. The clusters are defuzzified at 0 level and compared with those generated. The same two types of errors that have been defined in section 4 are registered here: the error of confusion, when a point belonging to one

y	100/200	150/150	200/100
0.5	0.209/0.051	0.135/0.040	0.099/0.059
	0.064/0.116	0.069/0.111	0.249/0.172
1.0	0.151/0.027	0.080/0.024	0.040/0.040
	0.049/0.030	0.072/0.046	0.152/0.099
1.5	0.087/0.033	0.042/0.013	0.037/0.045
	0.034/0.018	0.021/0.015	0.052/0.081
2.0	0.016/0.010	0.010/0.006	0.020/0.041
	0.005/0.007	0.003/0.004	0.018/0.088
2.5	0.001/0.002	0.003/0.004	0.010/0.028
	0.000/0.001	0.000/0.001	0.007/0.066
3.0	0.000/0.001	0.001/0.002	0.002/0.003
	0.000/0.000	0.000/0.000	0.000/0.001
3.5	0.000/0.000	0.000/0.000	0.001/0.001
	0.000/0.000	0.000/0.000	0.000/0.000

Table 4: Average confusion and omission errors, along with their standard deviations (after slash), after a hundred of data generation runs at each of the different values of the y coordinate of the strip cluster, from $y = 0.5$ and $y = 3.5$ – the cases presented on Figure 1. The columns refer to different ratios of the cluster cardinalities.

cluster is identified by the algorithm as belonging to the other, and the error of omission, when a point belongs to neither of the two first clusters.

In general, the errors are consistent with the expectations, almost disappearing at $y = 3.0$ or greater. However the character of the monotonicity at different size ratios is different. At 100/200 ratio, the error is high at $y = 0.5$, but almost disappears at $y = 2.5$; and moreover, the error of omission is rather low here. But at 200/100 ratio, the error of omission is rather high while the confusion error is relatively small, at small y values, and the errors keep appearing even at higher degrees of separation.

Yet if we change the threshold of defuzzification, the errors significantly decrease. Specifically, at the threshold of defuzzification 0.2, the confusion errors entirely disappear, while omission errors are rather low at small y values and become zero at larger y values, as clearly seen in Table 5.

6 Application to the Similarity Between Research Topics

The authors developed a publicly available tool ESSA for e-surveying of members of Computer Science Research organizations (see <https://copsro.di.fct.unl.pt/>). This tool is utilized to obtain a data table whose columns correspond to a set of V individuals or project teams in the organization ($v = 1, 2, \dots, V$), and rows to (some of) research topics taken to be leaves of the ACM-CCS taxonomy ([1]). The total set of research topics involved in the answers is denoted by T so that the table has $|T|$ rows. The (t, v) entry in the table is the score f_{tv} given by member v to the topic t , to express the share of their total research effort devoted to topic t ; f_{tv} is greater than 0 but smaller than 1, and the column v sums

y	100/200	150/150	200/100
value	th=0.2	th=0.2	th=0.2
0.5	0.000/0.000	0.000/0.000	0.000/0.000
	0.047/0.078	0.110/0.135	0.225/0.175
1.0	0.000/0.000	0.000/0.000	0.000/0.000
	0.058/0.035	0.081/0.039	0.154/0.082
1.5	0.000/0.000	0.000/0.000	0.000/0.000
	0.034/0.022	0.019/0.012	0.05/0.077
2.0	0.000/0.000	0.000/0.000	0.000/0.000
	0.005/0.007	0.003/0.004	0.006/0.006
2.5	0.000/0.000	0.000/0.000	0.000/0.000
	0.000/0.000	0.000/0.000	0.001/0.002
3.0	0.000/0.000	0.000/0.000	0.000/0.000
	0.000/0.000	0.000/0.000	0.000/0.001
3.5	0.000/0.000	0.000/0.000	0.000/0.000
	0.000/0.000	0.000/0.000	0.000/0.000

Table 5: Average confusion and omission errors, along with their standard deviations (after slash), after the defuzzification at threshold 0.2. The rows correspond to different values of the y coordinate of the strip cluster, from $y = 0.5$ and $y = 3.5$ – the cases presented on Figure 1. The columns refer to different ratios of the cluster cardinalities.

up to unity - a property which suggests a specific normalization weight assigned to each of the columns, as will be seen later in this section. The topic-to-topic similarities, according to the table supplied by an individual department, have been computed as described in section 2, after which the pseudo-inverse Laplacian L_n^- has been taken as the initial similarity matrix W to be modeled by our model of ADDI-FS. Original similarity matrices corresponding to two real-world Computer Science organizations, one a University department, the other a research center, are supplied in the Appendix.

The potential single distinction of the genuine similarity data from the affinity data is in handling the diagonal. It is made zero at the affinity data so that the entire focus is on the relations between the entities. Yet with the genuine similarity data the diagonal may bear an important distinction between the entities, which may affect the results.

Consider, for example, a typical genuine similarity dataset in Table 6 of the frequency of human confusion between different segmented numerals (such as presented in Figure 2). The diagonal dominates the data and shows, for example, that humans tend to identify 1 and 0 better than 8 and 9.

The diagonal dominates the data and, indeed, zeroing it does change the results, because it changes the summary values d_t (leaving L in the denominator unaffected).

The defuzzification results at the threshold 0.3 are shown for the first five ADDI-FS membership vectors in Table 7. One can see that the two meaningful clusters at which both clusterings agree, correspond to groupings of numerals $\{1,4,7\}$ and $\{6, 8, 0\}$, which have been found by a hierarchical aggregation algorithm based on maximizing the hi-squared coefficient of the aggregate table in [7]. The other clusters are rather different, except

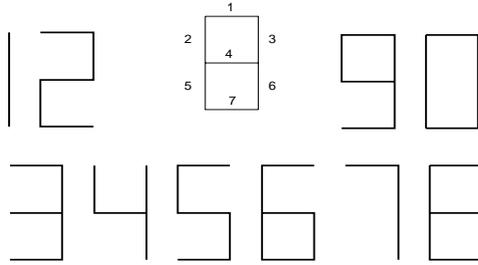


Figure 2: Digits: Styled digits formed by segments of the rectangle.

Stimulus	Response									
	1	2	3	4	5	6	7	8	9	0
1	877	7	7	22	4	15	60	0	4	4
2	14	782	47	4	36	47	14	29	7	18
3	29	29	681	7	18	0	40	29	152	15
4	149	22	4	732	4	11	30	7	41	0
5	14	26	43	14	669	79	7	7	126	14
6	25	14	7	11	97	633	4	155	11	43
7	269	4	21	21	7	0	667	0	4	7
8	11	28	28	18	18	70	11	577	67	172
9	25	29	111	46	82	11	21	82	550	43
0	18	4	7	11	7	18	25	71	21	818

Table 6: The Keren and Baggen (1981) data on confusion of the segmented numeral digits in an identification experiment, reprinted from [7].

perhaps the clusters $\{3, 5, 9\}$ at z , and $\{3,5,6,9\}$ at u , both closely resampling cluster $\{3,5,9\}$ from [7]. Yet some may claim that these differences are not quite important because of low contributions to the data scatter.

The results at the similarity between research topics are much closer showing no differences at the first three decimals at all.

Because one of the organizations, A, is a research center whereas the other, B, is a university department, one should expect the total number of research topics in A is smaller than that in B, and, similarly, the number of clusters in A should be less than that in B. Indeed, research centers are usually created for a limited research goal, whereas university departments must cover a wide range of topics in teaching which necessarily affects the research efforts. Indeed, the number of ACM-CCS topics scored in A is 46 versus 54 in B. Also, the number of clusters in A is two, whereas in B it is four. In contrast to other datasets analyzed in this paper, both research topic data sets have brought the clustering process to a halt not because of low contributions but because the continuation of the process became impossible: the next spectral cluster membership vector gives a negative value to the weight (4).

The cluster membership values sorted in the descending order are given in Table 8 for research center A and in Tables 9 and 10 for university department B. For each of

Table 7: ADDI-FS results at Digits data with the diagonal unchanged or zeroed (defuzzified at 0.3).

Numeral	Cluster1		Cluster2		Cluster 3		Cluster 4		Cluster 5	
	u	z	u	z	u	z	u	z	u	z
1	+	+								
2					+			+		+
3						+	+			+
4	+	+							+	
5						+	+	+		
6			+	+			+	+		
7	+	+								
8			+	+						
9						+	+			
0			+	+						+
Contribution, %	20.3	23.2	8.3	9.5	8.3	5.2	2.3	1.0	3.5	0.6
Intensity	2.38	1.38	1.90	1.04	1.89	0.90	1.38	0.60	1.53	0.52

the topics, we present both its ACM-CCS code and the string attached to it according to ACM-CCS. We maintain the ACM-CCS string style in that those attached to the second level nodes, like J.2 rather than J.2.3, are printed in capital.

The contributions total to about 50% for organization A, and 60% for organization B, which is a good result for clustering. It should be remembered that the contribution weights reflect the tightness of clusters rather than their priority standing.

The clusters in Table 8 may reflect the following. Cluster 1 is of pattern recognition and its applications to physical sciences and engineering including images and languages, with offshoots to general aspects of information systems. In cluster 2, all major aspects of computing mathematics are covered, with an emphasis on reliability and testing, and with applications in the area of life sciences.

The four clusters found at the university department B also have a more or less clear meaning. The first of them is of Software Engineering and Distributed Systems with a strong offshoot to the History of computing. Cluster 2 is of computing applications including various phases such as model development and project management. The subject of Cluster 3 is data storage with offshoots to general issues of computing. Cluster 4 comprises knowledge representation and related aspects (like Mathematical logics and Data structures) including many offshoots to general aspects of computer intelligence.

Overall these results are consistent with the informal assessment of the research conducted in each of the research organizations. Moreover, the sets of research topics that have been chosen by individual members of the centers follow the cluster structure rather closely, falling mostly within one of them.

Yet one can feel that some members are given more weight than the others - such is the case of cluster 3 in Table 10, which is based on the grouping supplied by just one member of the department whereas choices of some other members, not related to the choices of the others, have not been reflected in the clustering. This suggests that the method, probably, can be further improved if more attention is given to the spectral decompositions of the

Table 8: ADDI-FS results at data of similarity of research topics in research center A.

Cluster 1		
Eigenvalue	46.50	
Contribution	35.2%	
Intensity	5.57	
Weight	31.04	
Membership	Code	Topic
0.69911	I.5.3	Clustering
0.3512	I.5.4	Applications in I.5 PATTERN RECOGNITION
0.27438	J.2	PHYSICAL SCIENCES AND ENGINEERING (Applications in)
0.1992	I.4.9	Applications in I.4 IMAGE PROCESSING AND COMPUTER VISION
0.1992	I.4.6	Segmentation
0.19721	I.2.6	Learning
0.17478	H.5.2	User Interfaces
0.17478	I.6.4	Model Validation and Analysis in I.6 SIMULATION AND MODELING
0.16689	I.2.7	Natural Language Processing
0.16689	I.5.1	Models in I.5 PATTERN RECOGNITION
0.14453	I.5.2	Design Methodology (Classifiers)
0.13646	H.5.0	General in H.5 INFORMATION INTERFACES AND PRESENTATION
0.13646	H.0	GENERAL in H. Information Systems
0.13646	H.4.0	General in H.4 INFORMATION SYSTEMS APPLICATIONS
0.02867	I.2.11	Distributed Artificial Intelligence
Cluster 2		
Contribution	15.2%	
Eigenvalue	32.90	
Intensity	4.52	
Weight	20.41	
Membership	Code	Topic
0.46756	J.3	LIFE AND MEDICAL SCIENCES (Applications in)
0.40619	I.2.8	Problem Solving, Control Methods, and Search
0.34435	F.2.1	Numerical Algorithms and Problems
0.32681	F.4.1	Mathematical Logic
0.30067	G.1.6	Optimization
0.25967	D.3.3	Language Constructs and Features
0.23748	G.2.2	Graph Theory
0.18722	G.3	PROBABILITY AND STATISTICS
0.17359	B.2.3	Reliability, Testing, and Fault-Tolerance
0.17359	B.7.3	Reliability and Testing
0.17203	I.2.0	General in I.2 ARTIFICIAL INTELLIGENCE
0.1537	G.1.0	General in G.1 NUMERICAL ANALYSIS
0.11827	I.2.3	Deduction and Theorem Proving
0.10195	G.1.7	Ordinary Differential Equations
0.06175	K.2	HISTORY OF COMPUTING
0.00726	D.1.6	Logic Programming

Table 9: ADDI-FS results at data of similarity of research topics in university department B.

Cluster 1		
Eigenvalue	37.44	
Contribution	26.7%	
Intensity	5.26	
Weight	27.68	
Membership	Code	Topic
0.43055	K.2	HISTORY OF COMPUTING
0.39255	D.2.11	Software Architectures
0.35207	C.2.4	Distributed Systems
0.3412	I.2.11	Distributed Artificial Intelligence
0.3335	K.7.3	Testing, Certification, and Licensing
0.30491	D.2.1	Requirements/Specifications in D.2 Software Engineering
0.27437	D.2.2	Design Tools and Techniques in D.2 Software Engineering
0.24126	C.3	SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS
0.19525	D.1.6	Logic Programming
0.19525	D.2.7	Distribution, Maintenance, and Enhancement in D.2 Software Engineering
Cluster 2		
Contribution	13.4%	
Eigenvalue	26.65	
Intensity	4.43	
Weight	19.60	
Membership	Code	Topic
0.66114	J.1	ADMINISTRATIVE DATA PROCESSING
0.29567	K.6.1	Project and People Management in K.6
0.29567	K.6.0	General in K.6 MANAGEMENT OF COMPUTING AND INF. SYSTEMS
0.29567	H.4.m	Miscellaneous in H.4 INF. SYSTEMS APPLICATIONS
0.29567	J.7	COMPUTERS IN OTHER SYSTEMS
0.2696	J.4	SOCIAL AND BEHAVIORAL SCIENCES
0.16271	J.3	LIFE AND MEDICAL SCIENCES
0.14985	G.2.2	Graph Theory
0.14593	I.5.3	Clustering
0.12307	I.6.4	Model Validation and Analysis
0.10485	I.6.5	Model Development

residual matrices.

Table 10: ADDI-FS results at data of similarity of research topics in university department B.

Cluster 3		
Contribution	18.9%	
Eigenvalue	24.31	
Intensity	4.83	
Weight	23.31	
Membership	Code	Topic
0.613	E.2	DATA STORAGE REPRESENTATIONS
0.55728	I.0	GENERAL in I. Computing Methodologies
0.55728	H.0	GENERAL in H. Information Systems
Cluster 4		
Contribution	3.7%	
Eigenvalue	19.05	
Intensity	3.20	
Weight	10.26	
Membership	Code	Topic
0.35713	I.2.4	Knowledge Representation Formalisms and Methods
0.35636	F.4.1	Mathematical Logic
0.29495	F.2.0	General in F.2 ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY
0.28713	I.5.0	General in I.5 PATTERN RECOGNITION
0.28169	I.2.6	Learning
0.25649	K.3.1	Computer Uses in Education
0.24848	I.4.0	General in I.4 IMAGE PROCESSING AND COMPUTER VISION
0.24083	F.4.0	General in F.4 MATHEMATICAL LOGIC AND FORMAL LANGUAGES
0.18644	H.2.8	Database Applications
0.17707	H.2.1	Logical Design
0.17029	I.2.3	Deduction and Theorem Proving
0.15727	E.1	DATA STRUCTURES
0.15306	I.5.3	Clustering
0.14976	F.2.2	Nonnumerical Algorithms and Problems
0.14809	I.2.8	Problem Solving, Control Methods, and Search
0.14809	I.2.0	General in I.2 ARTIFICIAL INTELLIGENCE

7 Conclusion

A fast spectral fuzzy clustering algorithm is proposed within the framework of an additive fuzzy clustering model. The algorithm extracts cluster membership and intensity values one-by-one using the maximum eigenvalue and corresponding eigenvector of a matrix, obtained by subtracting from the initial similarity matrix those similarities that are due to the previously found clusters. The algorithm has a number of natural criteria for halting the process of extraction, thus appropriately defining the number of clusters. It is suggested that three typical data types – community structure graphs, affinity data derived from distances, and genuine similarity data – are to be differently normalized, according

to the expected structure of their eigenvalues; the latter two involving pseudo-inverse of the normalized Laplacian. The algorithm correctly clusters benchmark data and shows consistency over experimentally generated datasets.

We argue that some different types of data should get differently normalized, at least when applying ADDI-FS. Specifically, community data should get just symmetrization and diagonal removal, but not necessarily a Laplacian transformation. When we applied the normalized Laplacian transformation of the random test data with two communities, the confusion error was, on average, about 30% and the omission error about 25%. Explanation of this effect remains a task for the future. Yet it is the inverse Laplacian transformation of the original similarity data that is considered the similarity index to which our ADDI-FS model applies.

The motivation for the additive fuzzy clustering model comes from the analysis of research conducted by a University department or another research organization as a whole - we specifically target the Computer Science organizations. Similarity clusters of research topics are suggested to represent the elementary units of the Computer Science research according to the department's research efforts that can be further mapped to an integral portrayal of computer Sciences such as the ACM Classification of Computing Subjects (ACM-CCS). According to our analyses of several real-world Computer Science organizations in the UK and Portugal, all meaningful clusters of research topics do involve some core head subjects fitting in the ACM-CCS taxonomy nodes, but also may considerably extend them across the ACM-CCS tree topology. The following are among the issues remaining to be addressed in this regard: (a) how the search through web documents posted by individual researchers and project teams can contribute to the evaluation of similarity between research topics within an organization; (b) how to interpret the similarity clusters within a taxonomy; (c) how to integrate the information of clusters found in different organizations; (d) how this can be meaningfully applied to evaluation of the contribution by a research organization to the field of sciences.

8 Acknowledgements

The materials for this paper have been developed within the framework of COPSRO project (grant PTDC/EIA /69988/2006) funded by the Portuguese Foundation for Science & Technology. The authors express their gratitude to L. Moniz Pereira who initiated the work on the analysis of similarity between research topics, and Igor Guerreiro for developing the code for the ESSA e-survey tool. The help of members of CENTRIA, UNL, Lisboa, Portugal, the Department of Computer Science and Information Systems, Birkbeck University of London, London UK in collecting the data on their research activities is appreciated.

References

- [1] *The ACM Computing Classification System* (1998), url= <http://www.acm.org/class/1998/ccs98.html>.
- [2] Bezdek, J., Keller, J., Krishnapuram, R., Pal, T. (1999) *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers.

- [3] K. Inoue and K. Urahama (1999) Sequential fuzzy cluster extraction by a graph spectral method, *Pattern Recognition Letters*, 20, 699-705.
- [4] U. von Luxburg (2007) A tutorial on spectral clustering, *Statistics and Computing*, 17, 395-416.
- [5] B. Mirkin (1987) Additive clustering and qualitative factor analysis methods for similarity matrices, *Journal of Classification*, 4, 7-31; Erratum (1989), 6, 271-272.
- [6] B. Mirkin (1990) Sequential fitting procedures for linear data aggregation model, *Journal of Classification*, 7, 167-195.
- [7] B. Mirkin (1996) *Mathematical Classification and Clustering*, Kluwer,
- [8] B. Mirkin, S. Nascimento and L. Moniz Pereira (2008) Representing a Computer Science research organization over ACM Computing Classification System, In in P. Eklund, O. Haemmerlé (Eds.) Proc. 16th International Conference on Conceptual Structures, ISSN 1613-0073 (CEUR-WS.org), Toulouse, France, 57-65.
- [9] B. Mirkin (2008) The iterative extraction approach to clustering, In A.N. Gorban, B. Kegl, D.C. Wunsch, and A. Zinovyev (Eds.) *Principal Manifolds for Data Visualization and Dimension Reduction*, LNCSE 58, Springer-Verlag, 151-177.
- [10] B. Nadler, M. Galun (2007) Fundamental limitations of spectral clustering, *Neural Information Processing Systems Conference-06*, Vol. 19.
- [11] M.E.J. Newman (2006) Finding community structure in networks using the eigenvectors of matrices,
- [12] M. Newman and M. Girvan (2004) Finding and evaluating community structure in networks, *Physical Review E*, 69, 026113.
- [13] A. Ng, M. Jordan and Y. Weiss (2002) On spectral clustering: analysis and an algorithm, In T.G. Ditterich, S. Becker, Z. Ghahramani (Eds.) *Advances in Neural Information Processing Systems*, 14, MIT Press, Cambridge Ma., 849-856.
- [14] M. Roubens (1978) Pattern classification problems and fuzzy sets, *Fuzzy Sets and Systems*, 1, 239-253.
- [15] M. Sato, Y. Sato and L.C. Jain (1997) *Fuzzy Clustering Models and Applications*, Physica-Verlag, Heidelberg, 122 p.
- [16] R.N. Shepard and P. Arabie (1979) Additive clustering: representation of similarities as combinations of overlapping properties, *Psychological Review*, 86, 87-123.
- [17] J. Shi and J. Malik (2000) Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 888-905.
- [18] S. White and P. Smyth (2005) A spectral clustering approach to finding communities in graphs, *SIAM International Conference on Data Mining*

- [19] S. Zhang, R.-S. Wang and X.-S. Zhang (2007) Identification of overlapping community structure in complex networks using fuzzy c-means clustering, *Physica A*, 374, 483-490.

Appendix: Similarity Tables

Table 11: Similarity between research topics in research center A.

B.2.3	B.2.3/1.00, B.7.3/1.00, D.3.3/1.00, F.4.1/1.00, G.1.6/3.00, I.2.8/3.00
B.7.3	B.2.3/1.00, B.7.3/1.00, D.3.3/1.00, F.4.1/1.00, G.1.6/3.00, I.2.8/3.00
D.1.6	D.1.6/9.00, I.2.0/4.50, I.2.11/1.50, I.2.3/1.50, I.2.4/12.00, I.2.6/1.50
D.3.1	D.3.1/0.83, F.3.2/1.67, H.2.0/0.83, I.2.4/4.17, I.2.5/0.83
D.3.3	B.2.3/1.00, B.7.3/1.00, D.3.3/2.88, F.2.1/1.25, F.4.1/1.00, G.1.6/3.00, I.2.0/1.25, I.2.8/6.75, J.3/4.37
F.2.1	D.3.3/1.25, F.2.1/13.33, G.2.2/7.50, I.2.0/0.83, I.2.8/2.50, J.3/7.92
F.3.2	D.3.1/1.67, F.3.2/3.33, H.2.0/1.67, I.2.4/8.33, I.2.5/1.67
F.4.1	B.2.3/1.00, B.7.3/1.00, D.3.3/1.00, F.4.1/33.83, G.1.6/3.00, G.3/8.33, I.2.3/8.75, I.2.8/3.00, K.2/1.75
G.1.0	G.1.0/1.33, J.3/5.33
G.1.6	B.2.3/3.00, B.7.3/3.00, D.3.3/3.00, F.4.1/3.00, G.1.6/9.00, I.2.8/9.00
G.1.7	G.1.7/3.13, G.3/3.13, I.2.0/6.25
G.2.2	F.2.1/7.50, G.2.2/4.50, J.3/3.00
G.3	F.4.1/8.33, G.1.7/3.13, G.3/11.46, I.2.0/6.25
H.0	H.0/4.17, H.4.0/4.17, H.5.0/4.17, I.2.6/4.17
H.2.0	D.3.1/0.83, F.3.2/1.67, H.2.0/0.83, I.2.4/4.17, I.2.5/0.83
H.3.0	H.3.0/2.67, I.4.0/2.67, I.5.0/5.33, I.6.0/2.67
H.4.0	H.0/4.17, H.4.0/4.17, H.5.0/4.17, I.2.6/4.17
H.5.0	H.0/4.17, H.4.0/4.17, H.5.0/4.17, I.2.6/4.17
H.5.2	H.5.2/0.17, I.5.3/2.67, I.6.4/0.17, J.2/0.33
H.5.4	H.5.4/6.25, I.3.8/2.50, I.7.4/1.25, J.5/7.50, K.3.1/2.50, K.8.0/5.00
I.2.0	D.1.6/4.50, D.3.3/1.25, F.2.1/0.83, G.1.7/6.25, G.3/6.25, I.2.0/15.58, I.2.11/0.75, I.2.3/0.75, I.2.4/6.00, I.2.6/0.75, I.2.8/2.50, J.3/2.92
I.2.11	D.1.6/1.50, I.2.0/0.75, I.2.11/8.50, I.2.3/0.25, I.2.4/12.00, I.2.5/2.00, I.2.6/1.25, I.2.7/1.00, I.5.1/1.00, I.5.2/0.75, I.5.4/1.00
I.2.3	D.1.6/1.50, F.4.1/8.75, I.2.0/0.75, I.2.11/0.25, I.2.3/3.38, I.2.4/2.00, I.2.6/0.25, K.2/0.63
I.2.4	D.1.6/12.00, D.3.1/4.17, F.3.2/8.33, H.2.0/4.17, I.2.0/6.00, I.2.11/12.00, I.2.3/2.00, I.2.4/49.33, I.2.5/6.67, I.2.6/2.00
I.2.5	D.3.1/0.83, F.3.2/1.67, H.2.0/0.83, I.2.11/2.00, I.2.4/6.67, I.2.5/1.33
I.2.6	D.1.6/1.50, H.0/4.17, H.4.0/4.17, H.5.0/4.17, I.2.0/0.75, I.2.11/1.25, I.2.3/0.25, I.2.4/2.00, I.2.6/8.42, I.2.7/4.00, I.5.1/4.00, I.5.2/3.00, I.5.4/4.00
I.2.7	I.2.11/1.00, I.2.6/4.00, I.2.7/4.00, I.5.1/4.00, I.5.2/3.00, I.5.4/4.00
I.2.8	B.2.3/3.00, B.7.3/3.00, D.3.3/6.75, F.2.1/2.50, F.4.1/3.00, G.1.6/9.00, I.2.0/2.50, I.2.8/16.50, J.3/8.75
I.3.8	H.5.4/2.50, I.3.8/1.00, I.7.4/0.50, J.5/3.00, K.3.1/1.00, K.8.0/2.00
I.4.0	H.3.0/2.67, I.4.0/2.67, I.5.0/5.33, I.6.0/2.67
I.4.6	I.4.6/2.67, I.4.9/2.67, I.5.4/6.67, J.2/1.33
I.4.9	I.4.6/2.67, I.4.9/2.67, I.5.4/6.67, J.2/1.33
I.5.0	H.3.0/5.33, I.4.0/5.33, I.5.0/10.67, I.6.0/5.33
I.5.1	I.2.11/1.00, I.2.6/4.00, I.2.7/4.00, I.5.1/4.00, I.5.2/3.00, I.5.4/4.00
I.5.2	I.2.11/0.75, I.2.6/3.00, I.2.7/3.00, I.5.1/3.00, I.5.2/2.25, I.5.4/3.00
I.5.3	H.5.2/2.67, I.5.3/42.67, I.6.4/2.67, J.2/5.33
I.5.4	I.2.11/1.00, I.2.6/4.00, I.2.7/4.00, I.4.6/6.67, I.4.9/6.67, I.5.1/4.00, I.5.2/3.00, I.5.4/20.67, J.2/3.33
I.6.0	H.3.0/2.67, I.4.0/2.67, I.5.0/5.33, I.6.0/2.67
I.6.4	H.5.2/0.17, I.5.3/2.67, I.6.4/0.17, J.2/0.33
I.7.4	H.5.4/1.25, I.3.8/0.50, I.7.4/0.25, J.5/1.50, K.3.1/0.50, K.8.0/1.00
J.2	H.5.2/0.33, I.4.6/1.33, I.4.9/1.33, I.5.3/5.33, I.5.4/3.33, I.6.4/0.33, J.2/1.33
J.3	D.3.3/4.38, F.2.1/7.92, G.1.0/5.33, G.2.2/3.00, I.2.0/2.92, I.2.8/8.75, J.3/33.54
J.5	H.5.4/7.50, I.3.8/3.00, I.7.4/1.50, J.5/9.00, K.3.1/3.00, K.8.0/6.00
K.2	F.4.1/1.75, I.2.3/0.63, K.2/0.13
K.3.1	H.5.4/2.50, I.3.8/1.00, I.7.4/0.50, J.5/3.00, K.3.1/1.00, K.8.0/2.00
K.8.0	H.5.4/5.00, I.3.8/2.00, I.7.4/1.00, J.5/6.00, K.3.1/2.00, K.8.0/4.00

Table 12: Similarity between research topics in university department B.

A.0	A.0/0.83, D.2.6/0.83, D.2.8/1.67, E.1/2.50, H.2.8/2.50
C.2.4	C.2.4/16.21, C.3/1.50, D.1.6/5.00, D.2.1/5.50, D.2.11/8.00, D.2.2/1.00, D.2.7/5.00, H.2.2/3.13, H.2.3/3.13, H.2.4/3.13, H.3.3/1.25, I.2.11/3.00
C.3	C.2.4/1.50, C.3/2.25, D.2.1/0.75, D.2.11/4.50, D.2.2/1.50, I.2.11/4.50
D.1.6	C.2.4/5.00, D.1.6/1.88, D.2.1/1.88, D.2.11/1.88, D.2.7/1.88
D.2.1	C.2.4/5.50, C.3/0.75, D.1.6/1.88, D.2.1/2.79, D.2.11/3.38, D.2.2/1.17, D.2.7/1.88, I.2.11/1.50, K.2/3.33, K.7.3/2.00
D.2.11	C.2.4/8.00, C.3/4.50, D.1.6/1.88, D.2.1/3.37, D.2.11/10.87, D.2.2/3.00, D.2.7/1.88, I.2.11/9.00
D.2.2	C.2.4/1.00, C.3/1.50, D.2.1/1.17, D.2.11/3.00, D.2.2/1.67, I.2.11/3.00, K.2/3.33, K.7.3/2.00
D.2.6	A.0/0.83, D.2.6/0.83, D.2.8/1.67, E.1/2.50, H.2.8/2.50
D.2.7	C.2.4/5.00, D.1.6/1.88, D.2.1/1.88, D.2.11/1.88, D.2.7/1.88
D.2.8	A.0/1.67, D.2.6/1.67, D.2.8/3.33, E.1/5.00, H.2.8/5.00
E.1	A.0/2.50, D.2.6/2.50, D.2.8/5.00, E.1/7.50, H.2.8/7.50
E.2	E.2/8.13, F.2.2/0.88, H.0/6.00, H.2.3/1.50, I.0/6.00
F.2.0	F.2.0/7.50, F.4.0/5.00, F.4.1/5.00, I.2.3/2.50, I.2.4/5.00
F.2.2	E.2/0.87, F.2.2/11.29, F.4.1/1.67, G.2.2/2.50, H.2.1/4.17, H.2.3/10.50, I.2.4/6.67, I.6.4/1.50, I.6.5/1.50, J.3/2.50, J.4/1.00
F.4.0	F.2.0/5.00, F.4.0/3.33, F.4.1/3.33, I.2.3/1.67, I.2.4/3.33
F.4.1	F.2.0/5.00, F.2.2/1.67, F.4.0/3.33, F.4.1/20.67, H.2.1/1.67, I.2.3/1.67, I.2.4/6.00
G.2.2	F.2.2/2.50, G.2.2/7.08, I.2.6/0.83, I.5.3/4.17, I.6.4/4.58, I.6.5/3.75, J.3/7.92, J.4/2.50
H.0	E.2/6.00, H.0/4.50, I.0/4.50
H.2.1	F.2.2/4.17, F.4.1/1.67, H.2.1/4.17, I.2.4/6.67
H.2.2	C.2.4/3.13, H.2.2/5.21, H.2.3/5.21, H.2.4/5.21, H.3.3/2.08
H.2.3	C.2.4/3.13, E.2/1.50, F.2.2/10.50, H.2.2/5.21, H.2.3/23.21, H.2.4/5.21, H.3.3/2.08
H.2.4	C.2.4/3.13, H.2.2/5.21, H.2.3/5.21, H.2.4/5.88, H.2.5/4.00, H.2.8/0.67, H.3.3/2.08, J.3/1.33
H.2.5	H.2.4/4.00, H.2.5/24.00, H.2.8/4.00, J.3/8.00
H.2.8	A.0/2.50, D.2.6/2.50, D.2.8/5.00, E.1/7.50, H.2.4/0.67, H.2.5/4.00, H.2.8/10.83, H.3.3/4.00, I.2.6/4.00, I.5.2/2.67, J.3/1.33
H.3.3	C.2.4/1.25, H.2.2/2.08, H.2.3/2.08, H.2.4/2.08, H.2.8/4.00, H.3.3/6.83, I.2.6/6.00, I.5.2/4.00
H.4.m	H.4.m/1.00, J.1/5.00, J.4/1.00, J.7/1.00, K.6.0/1.00, K.6.1/1.00
I.0	E.2/6.00, H.0/4.50, I.0/4.50
I.2.0	I.2.0/1.00, I.2.4/1.00, I.2.6/3.00, I.2.8/1.00, I.5.0/1.00, K.3.1/3.00
I.2.11	C.2.4/3.00, C.3/4.50, D.2.1/1.50, D.2.11/9.00, D.2.2/3.00, I.2.11/9.00
I.2.3	F.2.0/2.50, F.4.0/1.67, F.4.1/1.67, I.2.3/0.83, I.2.4/1.67
I.2.4	F.2.0/5.00, F.2.2/6.67, F.4.0/3.33, F.4.1/6.00, H.2.1/6.67, I.2.0/1.00, I.2.3/1.67, I.2.4/15.00, I.2.6/3.00, I.2.8/1.00, I.5.0/1.00, K.3.1/3.00
I.2.6	G.2.2/0.83, H.2.8/4.00, H.3.3/6.00, I.2.0/3.00, I.2.4/3.00, I.2.6/15.83, I.2.8/3.00, I.5.0/3.00, I.5.2/4.00, I.5.3/4.17, I.6.4/0.83, J.3/1.67, K.3.1/9.00
I.2.8	I.2.0/1.00, I.2.4/1.00, I.2.6/3.00, I.2.8/1.00, I.5.0/1.00, K.3.1/3.00
I.4.0	I.4.0/8.33, I.5.0/8.33
I.4.10	I.4.10/7.50, I.4.5/5.00, I.4.6/2.50, I.4.8/5.00, I.4.9/5.00
I.4.5	I.4.10/5.00, I.4.5/3.33, I.4.6/1.67, I.4.8/3.33, I.4.9/3.33
I.4.6	I.4.10/2.50, I.4.5/1.67, I.4.6/0.83, I.4.8/1.67, I.4.9/1.67
I.4.8	I.4.10/5.00, I.4.5/3.33, I.4.6/1.67, I.4.8/3.33, I.4.9/3.33
I.4.9	I.4.10/5.00, I.4.5/3.33, I.4.6/1.67, I.4.8/3.33, I.4.9/3.33
I.5.0	I.2.0/1.00, I.2.4/1.00, I.2.6/3.00, I.2.8/1.00, I.4.0/8.33, I.5.0/9.33, K.3.1/3.00
I.5.2	H.2.8/2.67, H.3.3/4.00, I.2.6/4.00, I.5.2/2.67
I.5.3	G.2.2/4.17, I.2.6/4.17, I.5.3/20.83, I.6.4/4.17, J.3/8.33
I.6.4	F.2.2/1.50, G.2.2/4.58, I.2.6/0.83, I.5.3/4.17, I.6.4/3.08, I.6.5/2.25, J.3/5.42, J.4/1.50
I.6.5	F.2.2/1.50, G.2.2/3.75, I.6.4/2.25, I.6.5/2.25, J.3/3.75, J.4/1.50
J.1	H.4.m/5.00, J.1/25.00, J.4/5.00, J.7/5.00, K.6.0/5.00, K.6.1/5.00
J.3	F.2.2/2.50, G.2.2/7.92, H.2.4/1.33, H.2.5/8.00, H.2.8/1.33, I.2.6/1.67, I.5.3/8.33, I.6.4/5.42, I.6.5/3.75, J.3/12.25, J.4/2.50
J.4	F.2.2/1.00, G.2.2/2.50, H.4.m/1.00, I.6.4/1.50, I.6.5/1.50, J.1/5.00, J.3/2.50, J.4/2.00, J.7/1.00, K.6.0/1.00, K.6.1/1.00
J.5	J.5/16.67
J.7	H.4.m/1.00, J.1/5.00, J.4/1.00, J.7/1.00, K.6.0/1.00, K.6.1/1.00
K.2	D.2.1/3.33, D.2.2/3.33, K.2/16.67, K.7.3/10.00
K.3.1	I.2.0/3.00, I.2.4/3.00, I.2.6/9.00, I.2.8/3.00, I.5.0/3.00, K.3.1/9.00
K.6.0	H.4.m/1.00, J.1/5.00, J.4/1.00, J.7/1.00, K.6.0/1.00, K.6.1/1.00
K.6.1	H.4.m/1.00, J.1/5.00, J.4/1.00, J.7/1.00, K.6.0/1.00, K.6.1/1.00
K.7.3	D.2.1/2.00, D.2.2/2.00, K.2/10.00, K.7.3/6.00