# CSIC5011 Project 1: Explore NIPS papers dataset

Zhixian CHEN * Yue WU†

## Abstract

NIPS papers dataset collects papers from various fields accepted by the Conference and Workshop on Neural Information Processing Systems (NIPS), a machine learning and computational neuroscience conference from 1987 to 2016. With the development of computer technology and the coming of big data era, artificial intelligent developed rich branches including cognitive science, psychology, computer vision and so on. At the same time, the fields represented at NIPS also changed a lot during the past decades. In our project, we want to figure out how the diversity of fields represented at NIPS changed in the three decades by clustering analysis. First, we do the clustering analysis on Euclidean Space using two classical dimensional reduction methods: PCA and MDS. We visualize the clustering results respectively into 2D and 3D space. Then we do the clustering analysis on manifold. In the second case, we use two methods to construct the papers graph: k-nearest neighbor graph and paper-author-paper bipartite graph and reduce dimension to 2 based on three manifold learning methods: ISOMAP, LLE, and LE.

## 1 Motivation

Given high-dimensional data, there are various methods to deal with it. The most common way is dimensionality reduction so we can visualize and conceptualize the data.

There are mainly two types of structure in the high dimensional space. One is clumps such as clustering and density distribution. The other type is low dimensional linear or nonlinear manifolds.

## 2 Problem

**Try different dimensionality reduction methods to handle the data and analyze the results:**

Does the data live in a low dimensional subspace?

Or does the data live in a low dimensional submanifold?

**Try to find true structure of the high-dimensional data:**

Clusterings or manifolds?

## 3 Data description

In the reprt, we use NIPS papers dataset. Neural Information Processing Systems (NIPS) is one of the top machine learning conferences in the world. It covers topics ranging from deep learning and

*Contribute to the codes.

†Contribute to the report.

Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

computer vision to cognitive science and reinforcement learning. This NIPS papers dataset collects titles, authors, abstracts, and extracted text for all NIPS papers during 1987-2016.

To be specific, the file *authors.csv* contains information of authors' ids and names, the file *paper authors.csv* contains papers' ids and the corresponding authors' ids which form a sparse matrix of paper coauthors. And the file *papers.csv* contains more specific information of papers including years, titles and several texts.

| | id | name |
|---|---|---|
| 0 | 1 | Hisashi Suzuki |
| 1 | 10 | David Brady |
| 2 | 100 | Santosh S. Venkatesh |
| 3 | 1000 | Charles Fefferman |
| 4 | 10000 | Artur Speiser |

Figure 1: first five lines in *authors.cvs*

| | id | paper_id | author_id |
|---|---|---|---|
| 0 | 1 | 63 | 94 |
| 1 | 2 | 80 | 124 |
| 2 | 3 | 80 | 125 |
| 3 | 4 | 80 | 126 |
| 4 | 5 | 80 | 127 |

Figure 2: first five lines in *paper authors.cvs*

| | id | year | title | event_type | pdf_name | abstract | paper_text |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 1987 | Self-Organization of Associative Database and ... | NaN | 1-self-organization-of-associative-database-an... | Abstract Missing | 767\n\nSELF-ORGANIZATION OF ASSOCIATIVE DATABA... |
| 1 | 10 | 1987 | A Mean Field Theory of Layer IV of Visual Cort... | NaN | 10-a-mean-field-theory-of-layer-iv-of-visual-c... | Abstract Missing | 683\n\nA MEAN FIELD THEORY OF LAYER IV OF VISU... |
| 2 | 100 | 1988 | Storing Covariance by the Associative Long-Ter... | NaN | 100-storing-covariance-by-the-associative-long... | Abstract Missing | 394\n\nSTORING COVARIANCE BY THE ASSOCIATIVE\n... |
| 3 | 1000 | 1994 | Bayesian Query Construction for Neural Network... | NaN | 1000-bayesian-query-construction-for-neural-ne... | Abstract Missing | Bayesian Query Construction for Neural\nNetwor... |
| 4 | 1001 | 1994 | Neural Network Ensembles, Cross Validation, an... | NaN | 1001-neural-network-ensembles-cross-validation... | Abstract Missing | Neural Network Ensembles, Cross\nValidation, a... |

Figure 3: first five lines in *papers.cvs*

# 4 Method

**Linear dimensionality reduction methods: PCA and MDS.**

PCA seeks the most accurate data representation in a lower dimensional space. The good choice of directions or subspace to use is the one which keeps the largest variance. But it is limited to linear projections.

MDS attempts to preserve the pairwise distances and construct a configuration of $n$ points in Euclidian space by using the information about the distances between the $n$ patterns.

**Nonlinear dimensionality reduction methods: ISOMAP, LLE and LE.**

ISOMAP is an extension of MDS, where pairwise Euclidean distances between data points are replaced by geodesic distances, computed by graph shortest path distances.

LLE is motivated by the idea that global information about geodesic distance might not be accurate while requires expensive computational cost. Because when points are close enough, they are similar, while points are far, there is no faithful information to measure how far they are. It assumes that any point in a high dimensional ambient space can be a linear combination of data points in its neighborhood. This is a local method as it involves data points in local neighbors and hence a sparse eigenvector decomposition.

LE shares the similar idea with LLE that points close should stay close and ignore global information. LE algorithm defines the distance matrix directly using some kernels.

In this report, we try these methods to analyze the NIPS papers dataset.

# 5 Visualization and analysis

**Part 1. Dimensional Reduction and visualization on Euclidean Space**

We use PCA and MDS to realize dimensional reduction and visualization of the dataset.

First we analyze the paper-coauthor sparse matrix contained in the file *paper authors.csv* directly. We classify the papers into three groups "1987-1997", "1997-2007", "2007-2017" according the year.

**Step 1**: Use PCA and MDS to the whole matrix and visualize the data(Fig.4, Fig.5);
Visualize the data by different groups(Fig.6, Fig.7) from the overall results.

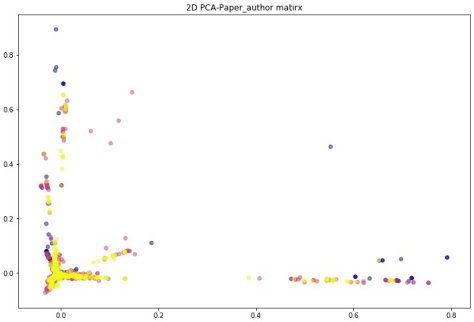


Figure 4: 2D-PCA on the overall dataset.

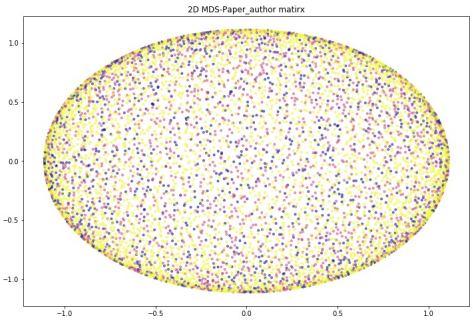


Figure 5: 2D-MDS on the overall dataset.

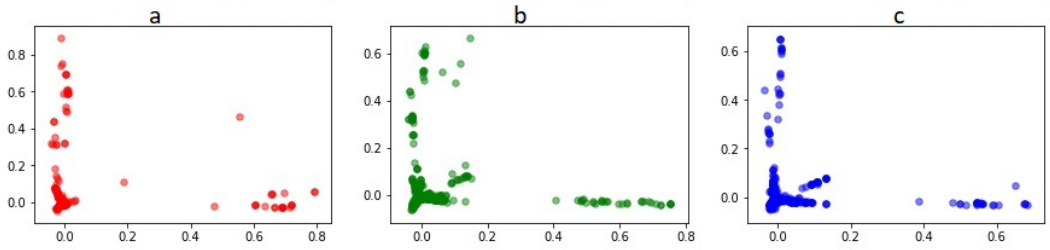


Figure 6: PCA results of three periods separated from Fig.4, (a):1987-1997, (b):1997-2007, (c):2007-2017.

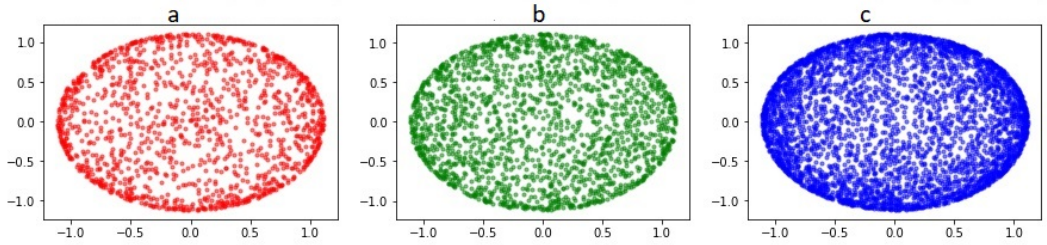


Figure 7: MDS results of three periods separated from Fig.5, (a):1987-1997, (b):1997-2007, (c):2007-2017.

**Step 2**: Use PCA and MDS to the data by different groups respectively and visualize the data(Fig.8,Fig.9);
Visualize the whole matrix by putting different groups together(Fig.10, Fig.11).

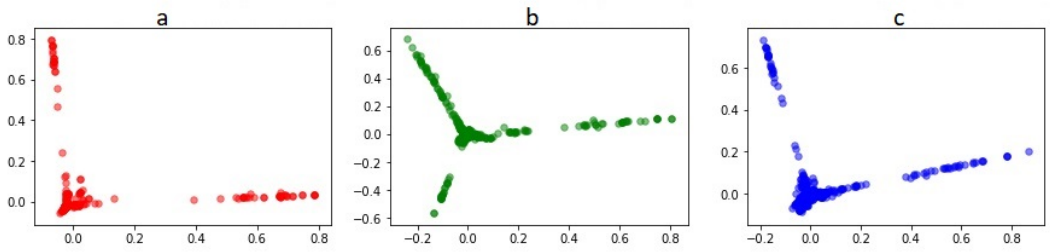


Figure 8: Separate the data into three groups by time and do 2D-PCA respectively. Here are the PCA results of three period groups, (a):1987-1997, (b):1997-2007, (c):2007-2017.

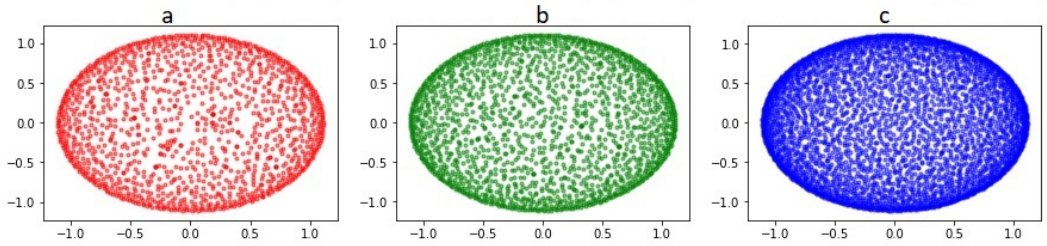


Figure 9: MDS results of three period groups respectively, (a):1987-1997, (b):1997-2007, (c):2007-2017.

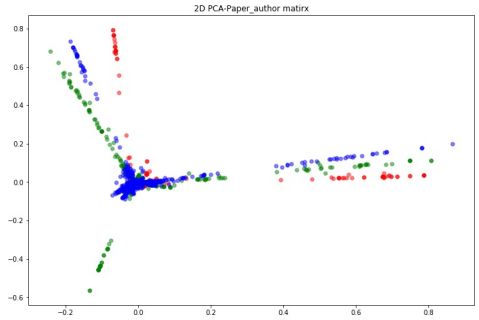


Figure 10: PCA result from integrating three groups in Fig.8 together.

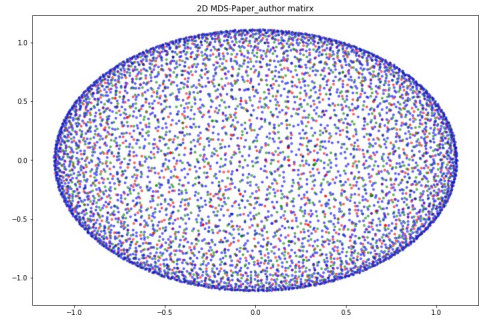


Figure 11: MDS result from integrating three groups in Fig.9 together.

**Interpretations:**
(1) From Fig.4, we can see there might be some clusters but the number of clusters is not obvious. Combined with Fig.6, It shows the concentration of three groups are overlapped.

(2) From Fig.8 and Fig.10, we can see classification by time might be helpful. Papers in the different groups has different PCA results but their PCA directions are close. It can be interpreted by the fact that these papers are in the related field(computer science) but have their own focus in each period.

(3) The results from MDS(Fig.5, Fig.11) show the data points are nearly evenly distributed in the embedding coordinates. And the number of papers is increasing as time. However, we can not obtain any useful information further from MDS.

**Part 2. Using Manifold Learning to explore the data**

Now we turn to the manifold learning methods. Here we ultilize two methods to construct a paper graph:

- $K$ **nearest neighbor graph.** In *sklearn.manifold*, Isomap, Local Linear Embedding and Laplacian Eigenmap algorithms are implemented by conscturcting the k-nearest neighbor graph. In this report, $K = 5$.

- **Graph based on common authors.** We consturct the connections between papers based on their common authors: if paper $i$ and paper $j$ have common author, we add a link between $i$ and $j$ with the number of the common author as the corresponding weight and set the reciprocal of number of the common author as the new weight when calculating the shortest path distance of $i$ and $j$.

**Step 1:** Based on $K$ nearest neighbor graph, use ISOMAP, LLE.

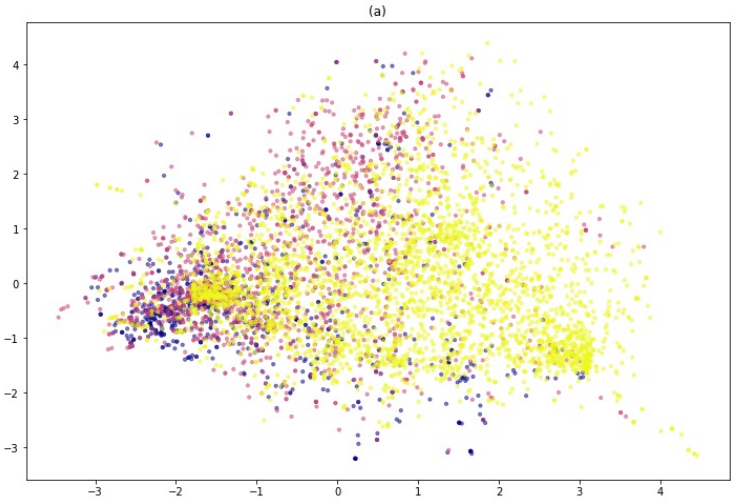


Figure 12: ISOMAP result based on KNN graph. blue: 1987-1997, pink: 1997-2007, yellow: 2007-2017. We can see the transition of the data in different group. The first coordinate follows the order of years.

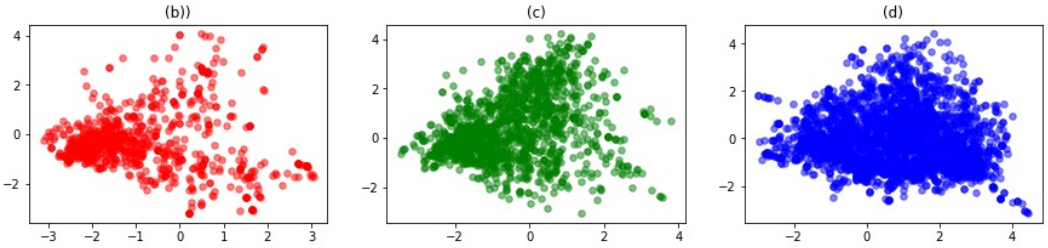


Figure 13: ISOMAP results of different time groups separated from Fig.12, (a):1987-1997, (b):1997-2007, (c):2007-2017.

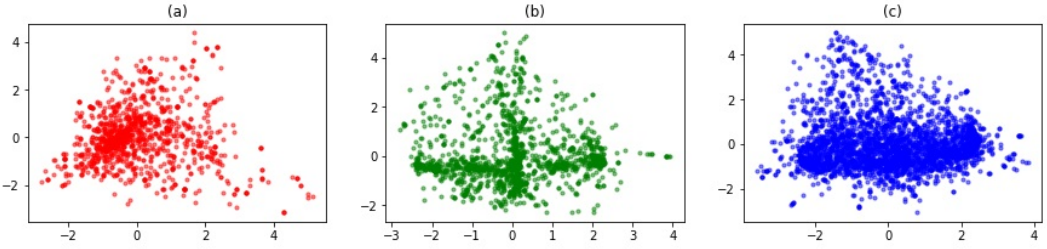


Figure 14: Based on KNN graph, ISOMAP results applied to three period groups, (a):1987-1997, (b):1997-2007, (c):2007-2017.

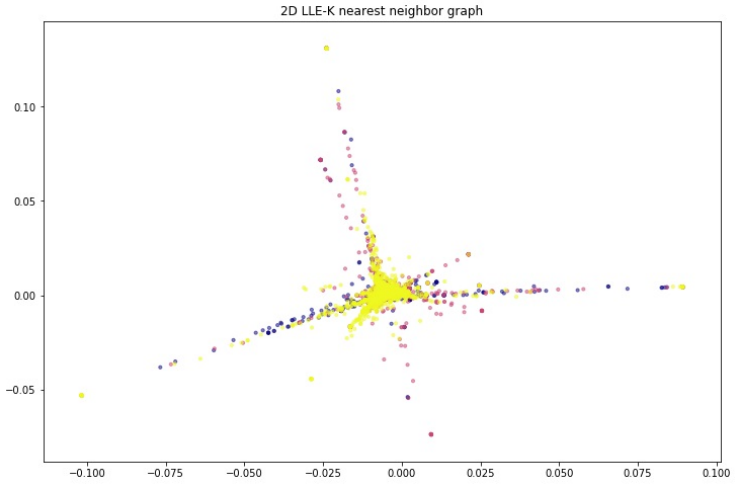


Figure 15: LLE result based on KNN graph. blue: 1987-1997, pink: 1997-2007, yellow: 2007-2017.

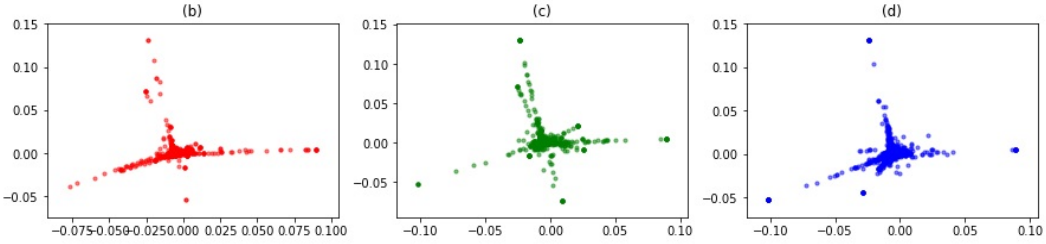


Figure 16: LLE results of different time groups separated from Fig.15, (a):1987-1997, (b):1997-2007, (c):2007-2017.

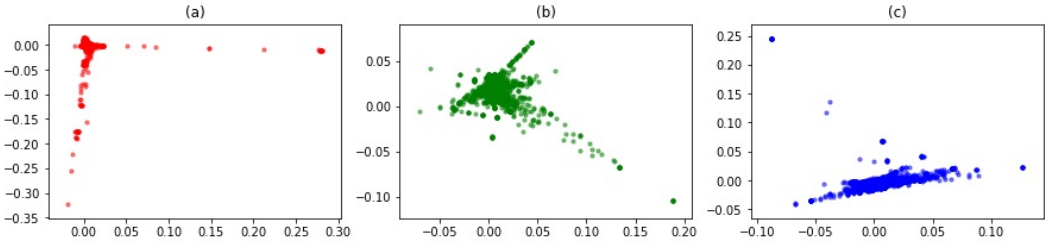


Figure 17: Based on KNN graph, ISOMAP results applied to three period groups, (a):1987-1997, (b):1997-2007, (c):2007-2017.

**Interpretations:**

(1) In Fig.12, 2-D ISOMAP embedding based on KNN graph shows the first coordinate follows the order of years, since the colors blue, pink and yellow represent periods "1987-1997", "1997-2007", "2007-2017" respectively. Hence in these thirty years, the topics keep changing but the main body are overlapped.

(2) From Fig.13 and Fig.14, we can see in different ten years, the focuses and numbers of papers are different.

(3) The results from LLE(see Fig.15. Fig.16, Fig.17), we have the similar phenomenon to the results from ISOMAP.

**Step 2:** Based on common authors graph, use ISOMAP, LLE, LE.

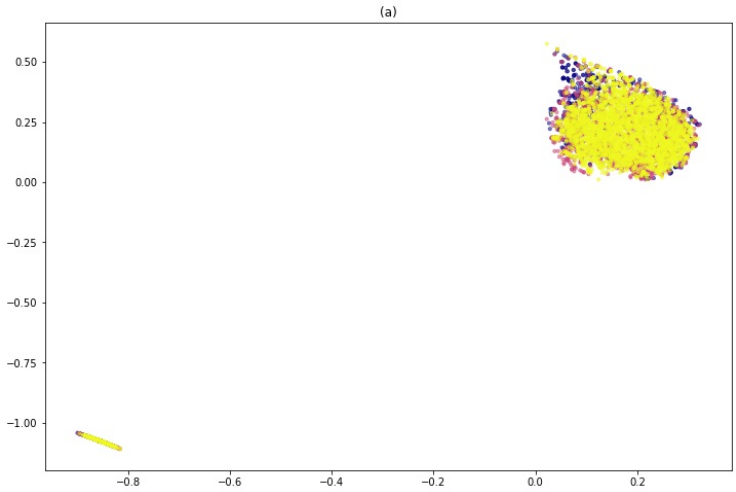


Figure 18: ISOMAP results based on coauthor graph. blue: 1987-1997, pink: 1997-2007, yellow: 2007-2017.

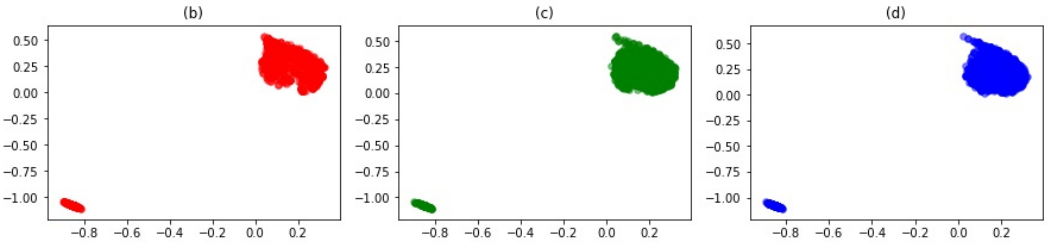


Figure 19: ISOMAP results of different time groups separated from Fig.18, (a):1987-1997, (b):1997-2007, (c):2007-2017.

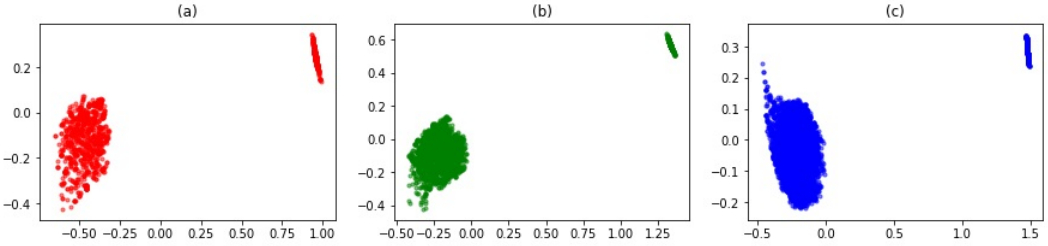


Figure 20: Based on coauthor graph, ISOMAP results applied to three period groups, (a):1987-1997, (b):1997-2007, (c):2007-2017.

**Interpretation:**
In Fig.18, we can see two clusters using 2D ISOMAP based on the coauthors graph. It indicates there are two graphs which means two groups of papers without common authors between them.

# 6 Conclusion

1. From the results of part 1, linear methods of dimensionality reduction such as PCA and MDS don't work well on this dataset. We may speculate the dataset distributed on a manifold rather than a hyperplane.

2. By comparing results from analyzing the whole data and results from analyzing different groups of data, we can see papers from different periods have distinct but close characters. So the focuses of papers in different time is changing and this is reasonable since every period has its hot issues.

3. From the results of part 2, we can see ISOMAP works better beacuse in the result of 2D ISOMAP based on KNN graph, the first coordinate follows the order of years.

4. We also conduct the ISOMAP to the integral data and three time groups. We gain two clusters which don't show if using KNN graph. Then we speculate there might be two groups of authors. In each group, the authors have cooperated with each other.

# Referrence

[1] A Mathematical Introduction to Data Science. Yuan Yao.

[2] Course slides of UNSUPERVISED LEARNING 2011. Rita Osadchy.