Chemometrics and Intelligent Laboratory Systems 191 (2019) 21–29

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Collaboration patterns and network in chemometrics

Chuan-Quan Li [a], Nan Xiao [a,b], Ye Wen [a], Shi-Hui He [c], Yuan-Da Xu [d], You-Wu Lin [e], Hong-Dong Li [f], Qing-Song Xu [a,*]

[a] School of Mathematics and Statistics, Central South University, Changsha, 410083, PR China
[b] Seven Bridges Genomics, 1 Main Street, Cambridge, MA, 02142, USA
[c] Tencent Technology (Shenzhen) Company Limited, Shenzhen, 518057, PR China
[d] Program of Applied and Computational Mathematics, Princeton University Princeton, NJ, 08544, USA
[e] School of Mathematics and Statistics, Guangxi Teachers Education University, Nanning, 530023, PR China
[f] Center for Bioinformatics, School of Information Science and Engineering, Central South University, Changsha, 410083, PR China

ARTICLE INFO

Keywords:
Chemometrics
Collaboration network
Centrality
Community detection

ABSTRACT

In this study, we have selected the papers in the chemometric field that were published between 2001 and 2015 in the six journals (*Chemometrics and Intelligent Laboratory Systems, Journal of Chemometrics, Analytica Chimica Acta*, *Analytical Chemistry*, *Talanta* and *Journal of Chromatography A*) to investigate the collaboration patterns and network. We reveal and visualize the collaboration patterns, publication trends, and research hotspots on chemometrics based on the datasets. The central chemometricians are then determined under different indexes. Finally, we discover that the largest component of the network is clustered into the "Big Europe" and "China-Changsha" community. The rest components in the network are also identified and explained. The results show the features of the collaboration patterns and network in chemometric research and present a new way to explore the zeitgeist in the area.

1. Introduction

In the early 1970s, Professor Wold and his team started to employ *kemometri* in their seminal work on spline functions [1]. Subsequently, he presented the term *chemometrics*, which is equivalent to *kemometri*, and in 1974, formally defined it as "the art of extracting chemically relevant information from data produced in chemical experiments" in an analogy with biometrics and econometrics [2,3]. Meantime, Wold and Kowalski founded the International Chemometrics Society [3,4]. Since then, *chemometrics* has been rapidly developed and widely applied in chemistry and chemical engineering. At the same time, this area has drawn the attention of researchers from different countries worldwide. Therefore, it deserves our effort to explore the perspective on chemometrics, as well as further study the evolving collaboration patterns and the fundamental structure of the collaboration network in chemometrics.

Recently, Newman studied the structure of the scientific collaboration networks of physicists and computer scientists and presented the differences in the patterns of collaborations between the fields [5]. In another paper [6], he analyzed a hybrid coauthorship and citation network of physicists. Ji [7] applied new community detection methods to the two network datasets, namely coauthorship and citation networks, for statisticians. In this paper, we consider the collaboration patterns and network in chemometric research, including the collaboration trends, research topics, network centrality, community detection, and so on.

First, we need to choose the suitable and representative journals. Because chemometrics originated from analytical chemistry and most of chemometrics methods are applied in analytical chemistry [8], we choose those journals with the analytical chemistry category in the Institute for Scientific Information (ISI) system. As to those journals in other chemistry fields, they are not in our consideration even they also publish a small number of chemometric papers. Two prominent journals, *Chemometrics and Intelligent Laboratory Systems* (*ChemoLab*) and *Journal of Chemometrics* (*JChemom*), came into being in the late 1980s. They publish original research, reviews and other types of papers on development of novel statistical, mathematical, computer techniques in chemistry and related disciplines. In addition, a part of the papers on other journals of analytical chemistry, such as *Analytical Chemistry* (AC), *Analytica Chimica Acta* (ACA), *Journal of Chromatography A* (JCA), *Talanta,* is related to chemometric methods and applications. For example, ACA encourages the submission of manuscripts about all aspects of analytical theory and

* Corresponding author.
E-mail address: qsxu@csu.edu.cn (Q.-S. Xu).

https://doi.org/10.1016/j.chemolab.2019.05.011
Received 19 February 2019; Received in revised form 10 May 2019; Accepted 20 May 2019
Available online 25 May 2019

methodology, including chemometric techniques. JCA also provides a forum for the publication of original research and reviews on all aspects of separation science, comprising various chemometric methods. All the papers in ChemoLab and JChemom are chosen. And the chemometric papers in AC, ACA, JCA and Talanta are picked out. These datasets can establish a more comprehensive analysis in the chemometric field and the collaboration network among chemometricians [9].

We have organized this paper into the following sections. In Section 2, we provide a brief introduction to the network analysis methods. In Section 3, we illustrate the process of data collection and discuss the collaboration pattern analysis. We use the composite likelihood-Bayesian information criterion (CL-BIC) method to detect the significant communities and explain the components of chemometricians' collaboration network in Section 4. Our conclusions with some comments are included in Section 5, while the data pre-processing information are presented in the Appendix.

2. Network analysis methods

2.1. Basic features of the network

First, we present some notations. For an undirected network, $G = (V, E)$ stands for the network G with V nodes and E edges. The symmetric adjacency matrix $A_{i,j}$ is defined as follows: $a(i, j) = 1$ means that node i and node j are linked, and $a(i, j) = 0$ shows that they are not linked. If a pair of nodes has more than one link, then it is a weighted network; otherwise, it is an unweighted network. In this paper, we adopt the unweighted network.

One of the most concerned questions is to identify the important vertices in network analysis. Various centrality indexes for collaboration networks, such as clustering coefficients, shortest paths, betweenness, funneling, and average distances, have been studied in Newman's foundational work [10–12]. Those indexes are measured from different perspectives. For example, the betweenness centrality measures how often a vertex lies on short paths between other pairs of vertices [13]. The PageRank score considers not only the number of edges but also the influence of the linked nodes on certain nodes. And we will describe it in the next section.

Generally, the node with a larger number of links than the average number is called the hub node. When the degree of distribution follows a power law, then this is a scale-free network [12].

2.2. PageRank centrality

Google initially used the PageRank algorithm [14,15] to rank websites in its search engine and then expanded the algorithm to measure the centrality within a network. First, a state transition matrix k needs to be defined. As usual, the transition matrix is determined by the degree of nodes. The computational formula is as follows:

$$y_i = d\sum_{j\in V}\frac{1}{k_j}y_j + \frac{(1-d)}{N} \tag{1}$$

The first part of equation (1) sums up all centrality scores of the node j with a link to node i. The second part is the damping factor, which is the probability of random transitions between nodes. The default value of coefficient d in Equation (1) is 0.85 [14].

2.3. Community detection

Community detection is a major field in the network analysis. Considering a network $G = (V,E)$, the community detection problem can be shown as $V = V_1 \cup V_2 \cup \cdots V_K$, where V_i is the set of nodes that represents a community, and K denotes the number of communities. Besides chemists and chemical engineers, this research area has attracted a huge number of researchers from such fields as computer science, physics, and statistics. Therefore, various methods of community detection have been proposed from different perspectives, including modularity optimization [16], maximum likelihood [17–20], graph partition [21], and spectral clustering methods [22,23].

As a maximum likelihood method, the degree-corrected block model (DCBM) was proposed by Karrer and Newman in 2011 [17]. It is an extension of the stochastic block model [24]. In the DCBM, the adjacency probability between nodes considers not only a symmetric community-wise edge probability $P_{g_i g_j}$ but also the variations in the degree of nodes θ_i. The equation is defined as follows:

$$E(A_{i,j}) = \theta_i\theta_j P_{g_i g_j} \tag{2}$$

The links on the edges are independent Bernoulli random variables. Therefore, the formula for the likelihood function with the adjacency matrix A is as follows:

$$P(G|\theta,\omega,g) = \prod_{i<j}\theta_i\theta_j P_{g_i g_j}^{A_{i,j}}\left(1-\theta_i\theta_j P_{g_i g_j}\right)\left(1-A_{i,j}\right)^{} \tag{3}$$

Its log-likelihood form is given as follows:

$$cl(\widehat{\theta_c};A) = \log P(G|\theta,\omega,g) = 2\sum_i \theta_i log w_i + \sum_{rs}(m_{rs} log p_{rs} - p_{rs}) \tag{4}$$

where w_i and m_{rs} are defined as follows:

$$w_i = \frac{\theta_i}{\theta_{g_i}}\ ,\ m_{rs} = \sum\nolimits_{r<s} I(g_i = r, g_j = s)$$

Here, θ_{g_i} is the degree sum of group g_i. In recent years, the DCBM has been widely adopted in the new community detection methods [18–20].

2.4. Composite likelihood Bayesian information criterion

In the network analysis, it is a challenging task to determine the number of communities. Many algorithms assume the number of communities before detection [18–20,25]. Recently, some algorithms have been proposed to determine the number of communities, for instance, the cross-validation method by splitting the nodes [26] or edges [27], Markov Chain Monte Carlo [28,29], hypothesis testing [30,31], spectral estimation [32], among others.

In this paper, we introduce the $CL-BIC$ [33] to determine the optimal number of the communities. This criterion combines the block model likelihood with a penalization of the model complexity and is formulated as follows:

$$CL-BIC_k = -2cl(\widehat{\theta_c};A) + d_k^* log\left(\frac{N(N-1)}{2}\right) \tag{5}$$

The first part of Equation (5) can be a different log-likelihood model with the k communities, for example, the stochastic block model [24], the DCBM [17], or the mixed-membership stochastic block model [34]. In this paper, we choose the log-likelihood form of the DCBM because it is more suitable for the collaboration network. The second part of Equation (5) is the penalty term for the model complexity, where $d_k^* = trace(H_k^{-1}V_k)$, $H_k = E_\theta(-\nabla_\theta^2 cl(\theta;A))$ and $V_k = \mathrm{Var}_\theta(\nabla_\theta cl(\theta;A))$. The optimal K is determined by the minimal $CL-BIC$ value from a series of k' candidates.

3. Collaboration pattern analysis

3.1. Data collection

We have collected the metadata of the papers published between 2001 and 2015 in the six journals, including papers' titles, authors, abstracts, keywords, and URLs. However, AC doesn't provide the keyword information in its webpages. Here, the keyword plus part of AC in the ISI system [35] is used. The original unfiltered dataset comprised a total of

62,042 papers and 207,452 authors. Notably, most of the papers published in AC, ACA, JCA and Talanta belong to the field of general analytical chemistry, and only a small part is relevant to chemometrics. The first step is that we select the related papers according to the subdirectories in ACA. The second step is that we use the keywords about chemometric methods to select the papers. We also have found that authors' names may have different variations, such as *Y.Z. Liang* or *Yizeng Liang* and *Bernard G.M. Vandeginste* or *BGM Vandeginste*. We use the R package *stringdist,* first name and last name to identify the different variations of authors' names.

After the process of paper selection and name disambiguation, only 3985 papers and 8389 authors have been left in the final dataset. And the 3985 papers include the original research and review papers, but exclude book reviews and so on. The Appendix presents more detailed information about the data pre-processing.

3.2. Publication and coauthor patterns

Fig. 1 shows the number of chemometric papers published each year between 2001 and 2015. Over the past 15 years, the number of papers published annually has steadily increased from 192 to 340. However, some minor fluctuations are observed in certain years, owing to conferences or special events. For example, in 2005, those papers presented at the 9th International Conference on Chemometrics in Analytical Chemistry are published in ACA, but in 2006, it did not publish any special events or conference papers. As such, the number of its published papers decreased sharply in 2006.

Fig. 2(a) shows the number of authors contributing chemometric papers each year, increasing from 515 to 1184 over the past 15 years. The growing number of the authors has exceeded that of the published papers, which indicates an increasing competition among chemometricians. In contrast, the average number of the chemometric papers written by each author between 2001 and 2015 has decreased from 0.37 to 0.28 based on Figs. 1 and 2.

Fig. 2(b) shows the average of each author's partner over the past 15 years. Overall, the average number has continuously risen from 3.5 to 4.6. Therefore, on average, the authors have increasingly collaborated in conducting chemometric research.

3.3. Imbalance in authors' contribution

In general, the individual authors' productivity shows a skewed distribution and only a few authors have contributed a high number of papers. As presented in Fig. 2(a), the papers written by each author range from 1 to 71. The empirical proportion among all the authors, publishing more than a given number of papers, is also shown in Fig. 3(a). The power-law model is often used to fit the distribution, and its formula is as follows:

$$p(z) \approx z^{-\gamma} \tag{6}$$

where z denotes the number of papers. The estimated power-law exponent γ for our data is 1.832. The R^2 value is 0.9975, and the p-value is smaller than 0.001, which suggests that our dataset fits the power-law distribution well and can help us infer the proportion of the authors with different productivity levels. For example, over 50% of the authors have published only one paper, and nearly 1% have published more than 40 papers.

Similar to the distribution of productivity, the number of coauthors has also substantially varied. The top five authors who have more collaborators are Yu (122), Liang (121), Heyden (120), Bro (111), and Massart (105). In contrast, as shown in Fig. 3(b), most of the chemometricians in the collaboration network have 10 or fewer coauthors. In Newman's fundamental research about the scale-free network, the following modified power-law distribution [36] is used to fit the distribution of the number of coauthors:

$$p(z) \approx z^{-\gamma} e^{-z/z_c} \tag{7}$$

For the dataset, the estimated γ is 0.815, and z_c is 12.22. The goodness-of-fit R^2 is 0.91, and the p-values of the parameter estimation are both smaller than 0.001, indicating a good fit for the modified power-law distribution. We can infer that near 8% authors have more than 10 coauthors and over 40% authors have one coauthor.

Above all, Fig. 3(a) clearly shows that the productivity and the number of coauthors among the authors have a huge variance. Here, we define each author's contribution as $\frac{1}{K}$ when one paper has K coauthors. Therefore, each author's contribution depends on the number of papers and coauthors. In economics, the Lorenz curve and Gini coefficient can measure the gap between the rich and the poor. In this paper, we use them to investigate the imbalance in the authors' contributions. In Fig. 4, the x-axis represents the fraction of the authors with the most contributions, and the y-axis denotes the cumulative fraction of the contributions. In economics, a 0.4 Gini coefficient is often regarded as the international warning for a dangerous level of inequality. In Fig. 4, the Gini coefficient is 0.5171, and the top 20% productive authors dominate 51.7% of the

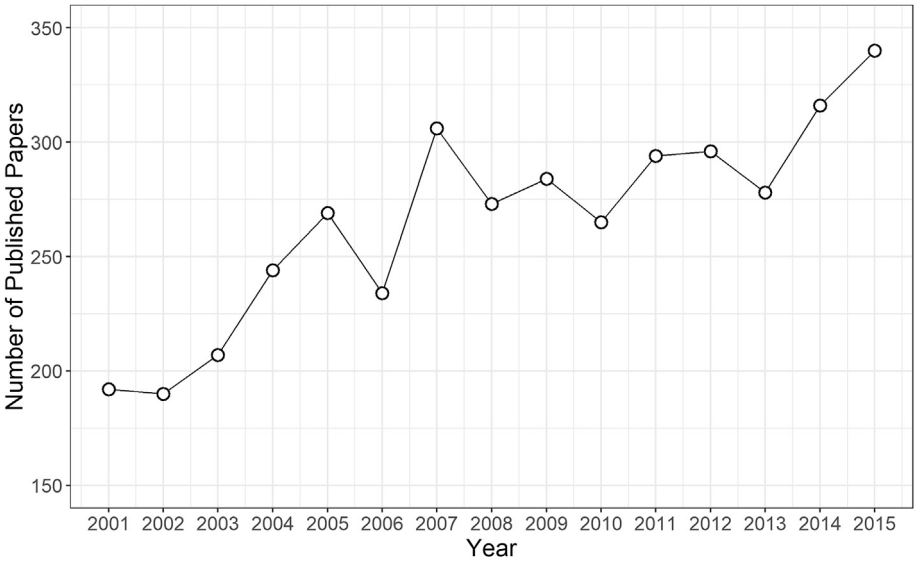

Fig. 1. The number of chemometrics papers published each year.

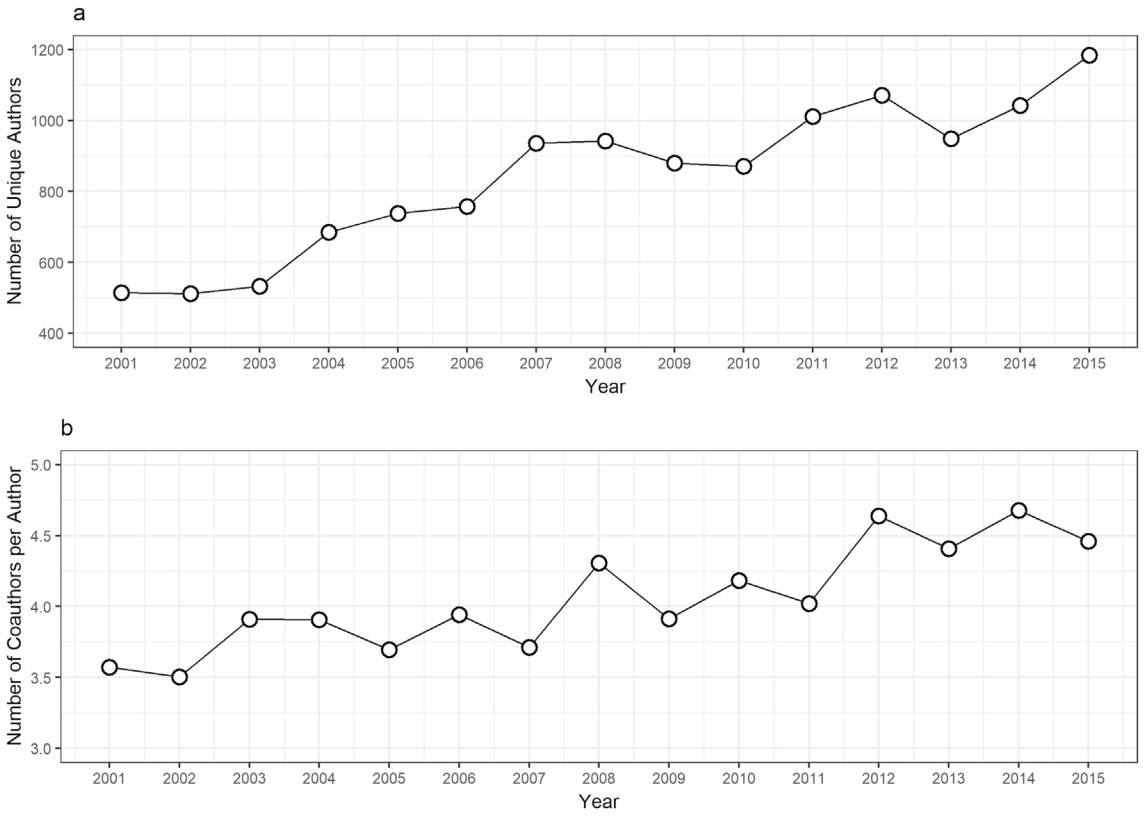

Fig. 2. (a). The number of authors each year. (b) The number of coauthors per author each year.

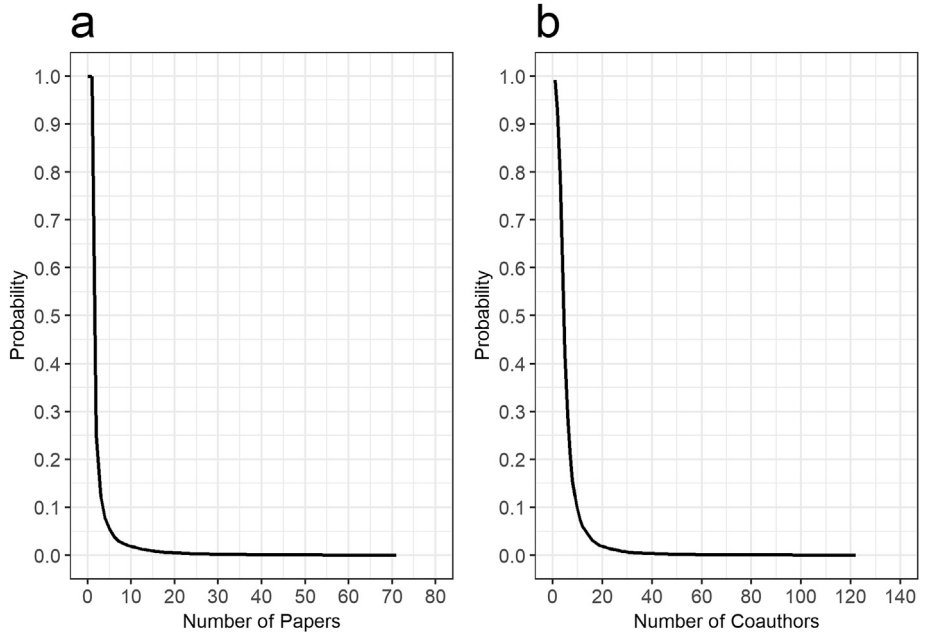

Fig. 3. (a). The distribution about the number of papers. (b). The distribution about the number of coauthors.

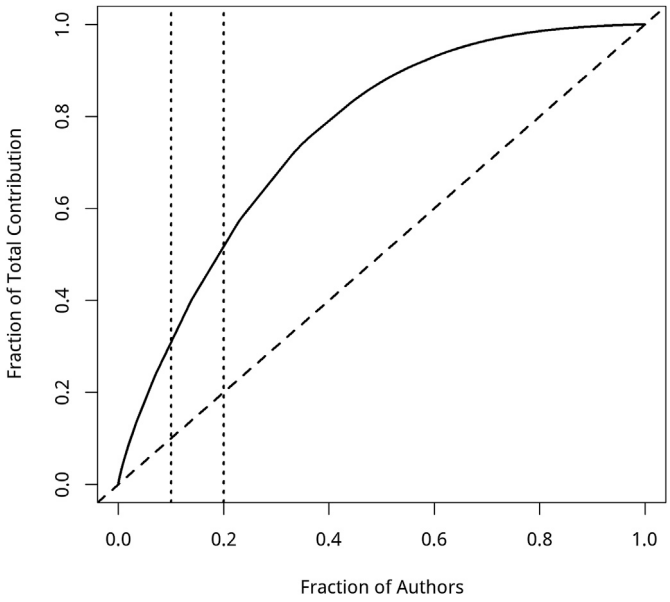

Fig. 4. The Lorenz curve.

total contributions. The reason can be partly attributed to the fact that a large proportion of the papers is produced by certain large research groups (or communities). We describe some detected communities in the following section.

3.4. Word cloud

A word cloud, which resembles different words and put together in the shape of a cloud, could be used to intuitively visualize the keywords listed in the selected papers and reveal the hot research topics. The size of each keyword represents its frequency of occurrence in the datasets. And different keywords with similar meanings are integrated into the same words. For example, PLS and partial least squares regression both represent the PLS algorithm. The frequencies of all keywords in the datasets are ranked, and the top 60 words are selected to make the word cloud, as shown in Fig. 5. It illustrates several hot topics, including multivariate calibration, PARAFAC (parallel factor analysis), MCR (multivariate curve resolution), classification, process monitoring, pattern recognition, optimization, and experimental design. Clearly, a number of the significant words focus on statistical methodology, for instance, PLS, PCA, variable selection, and QSAR (quantitative structure–activity relationship). The machine learning methods, such as support vector machine, neural networks, cluster analysis, and linear discriminant analysis, are also commonly used. These statistical methods have been extensively applied or expanded in chemometrics [37]. Other words indicate commonly used types of chemical data, such as near-infrared spectroscopy, gas (liquid) chromatography, raman spectroscopy, and mass spectroscopy.

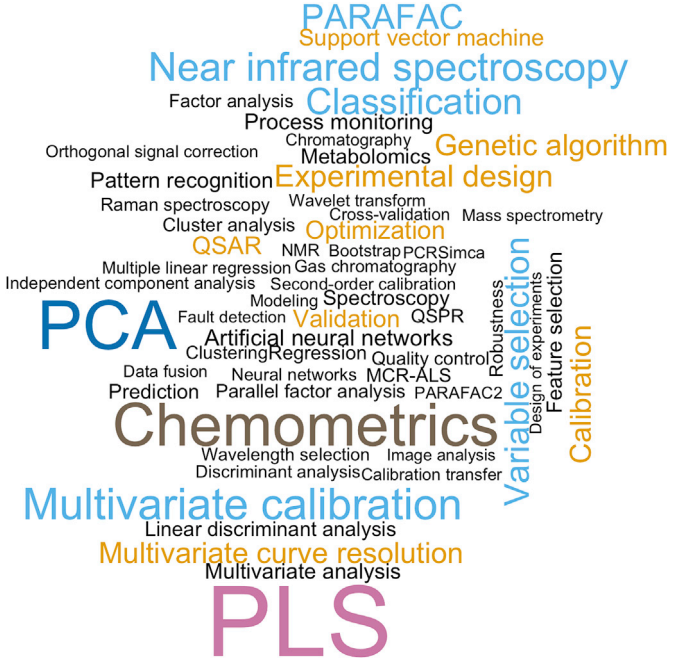

Fig. 5. Word cloud for the chemometric topics.

3.5. Centrality analysis

One of the major issues of concern is how to identify the most central nodes in the collaboration network. The number of papers, the number of collaborating authors, the betweenness centrality, and the PageRank score are chosen as the central indexes. These central measures indicate some differences from the detailed rankings. However, it can be observed from Fig. 6 that the most central authors are concordantly identified, such as Bro, Buydens, Heyden, Leardi, Liang, Massart, Oliverri, Rutledge, Smilde, Tauler, Walczak, Wu, Xu, and Yu (in alphabetical order). Compared with the other three centrality measures, the PageRank score is more convincing because it considers not only the number of edges but also the influence of the linked nodes. In other words, an author who collaborated with the more-central authors in the network should be ranked higher than an author who collaborated with the less-central authors. The top 100 authors based on these four indexes are provided in the Supplemental Material.

4. Collaboration network analysis

It is important to conduct an in-depth investigation of the more central nodes and the more persistent collaboration in a network. Considering the distribution of the PageRank scores, the threshold of the score is 0.000169. Next, 800 chemometricians with the top PageRank scores are chosen to construct the collaboration network. When two authors (nodes) collaborated in three or more papers, there would be a connection (edge) between them. Finally, all the authors are clustered into 405 components according to the links. Among these components, 297 authors have no connections, which are not analyzed in this paper. All the component analyses are completed by using the Gephi software (version 0.9.1). Limited to the size of the figure, we only label those nodes (authors) with top 230 PageRank scores in Figs. 7–10.

4.1. Largest component

The largest component of the network consists of 136 authors, including the most central authors identified in Fig. 6(d), namely Liang, Yu, Bro, Massart, Heyden, Tauler, Wu, Smilde, and Xu, and other central authors, such as Buydens, Walczak, Juan, Berg, and so on. These authors come from different countries and research institutes. As far as we know, there are some divergences in their research fields. For example, Bro puts more weight on food science and chemometric methods. Liang has conducted much research on traditional Chinese medicines, statistical methods and algorithms. Therefore, it is likely that the component can be further clustered into different communities. The CL-BIC procedure described in Section 2.3 is executed up to 100 iterations, and the optimal number of communities is 2. This two communities, which are colored red and green, respectively, are shown in Fig. 7.

We label the larger community as “Big Europe”. This community consists of the researchers or the research groups from European countries. The researchers from Vrije Universiteit Brussel, led by Massart and Heyden, have focused on such areas as analytical chemistry,

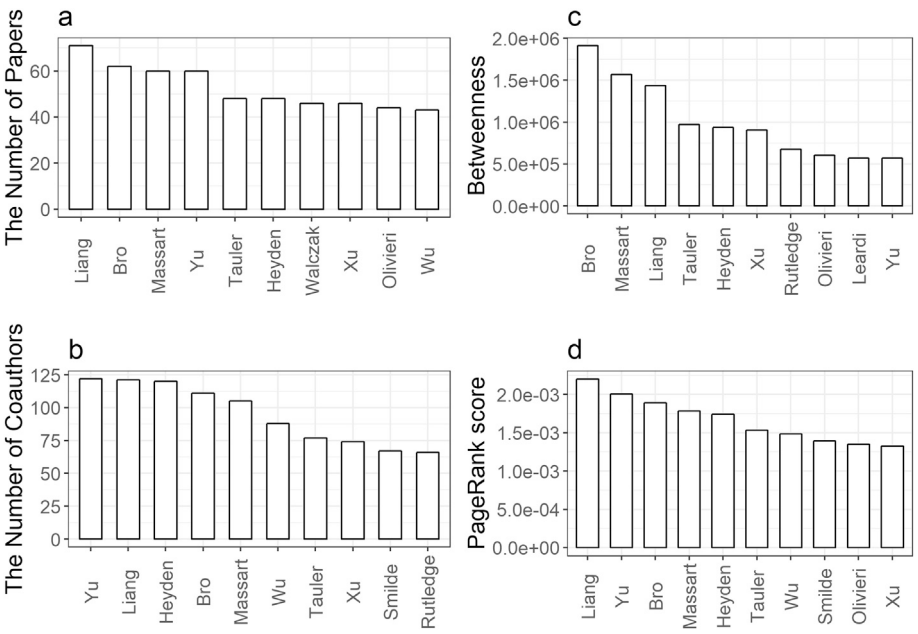

Fig. 6. The top 10 central authors ranked by four indexes. (a). The number of published papers, (b). The number of coauthors, (c) Betweenness centrality, (d) PageRank score.

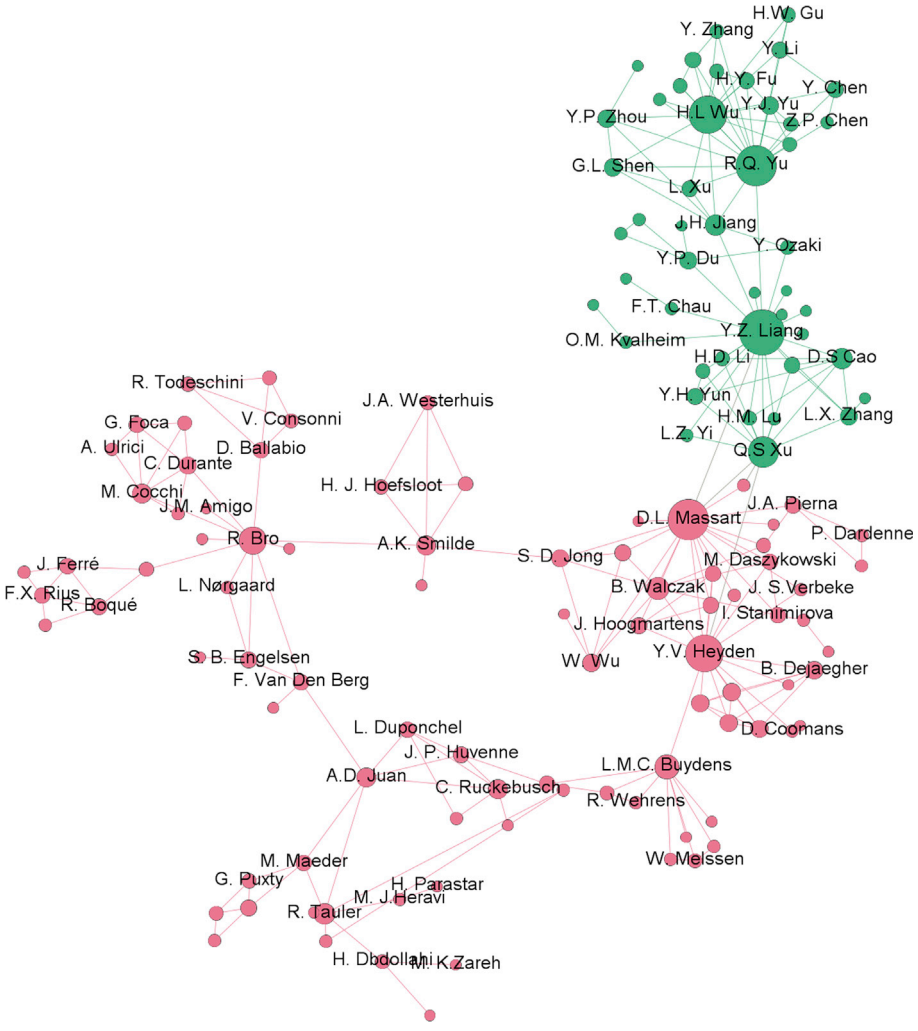

Fig. 7. Community detection results for the largest component.

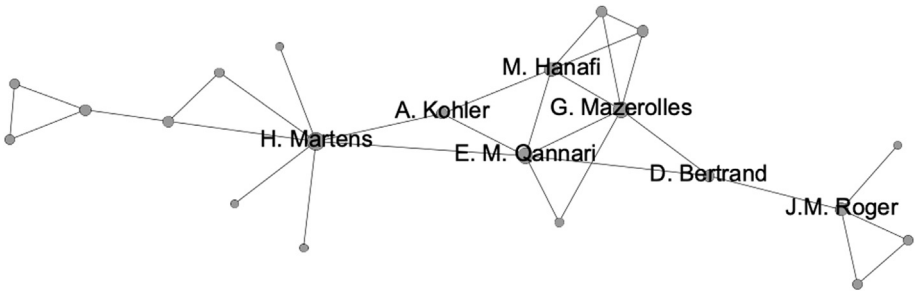

Fig. 8. The second largest component.

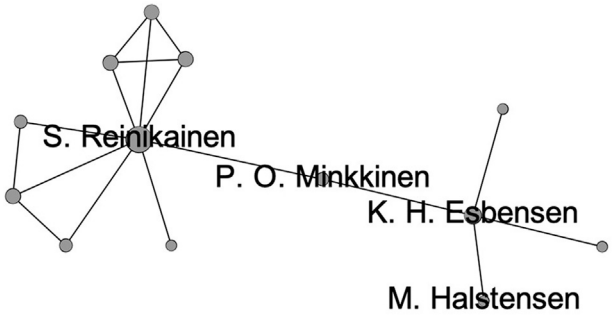

Fig. 9. The third largest component.

pharmaceutical analysis, and chemometric methods. Walczak (from the University of Silesia, Poland) is closely connected to the group because of her long-term cooperation with Massart and Heyden in data exploration and modeling methods. Buydens (Radboud University, Netherlands) and her group have carried out methodological research in molecular chemometrics and spectroscopic image analysis by using machine learning and statistical methods, such as support vector machine, neural networks, classification techniques, and global optimization. She graduated from Vrije Universiteit Brussel and has collaborated with Heyden in chemometric methods.

Led by Tauler and Juan, the research team from the Spanish National Research Council and the University of Barcelona has focused on chemometric methods for environmental omics (genomics and metabonomics). Tauler and Juan have also undertaken numerous works using

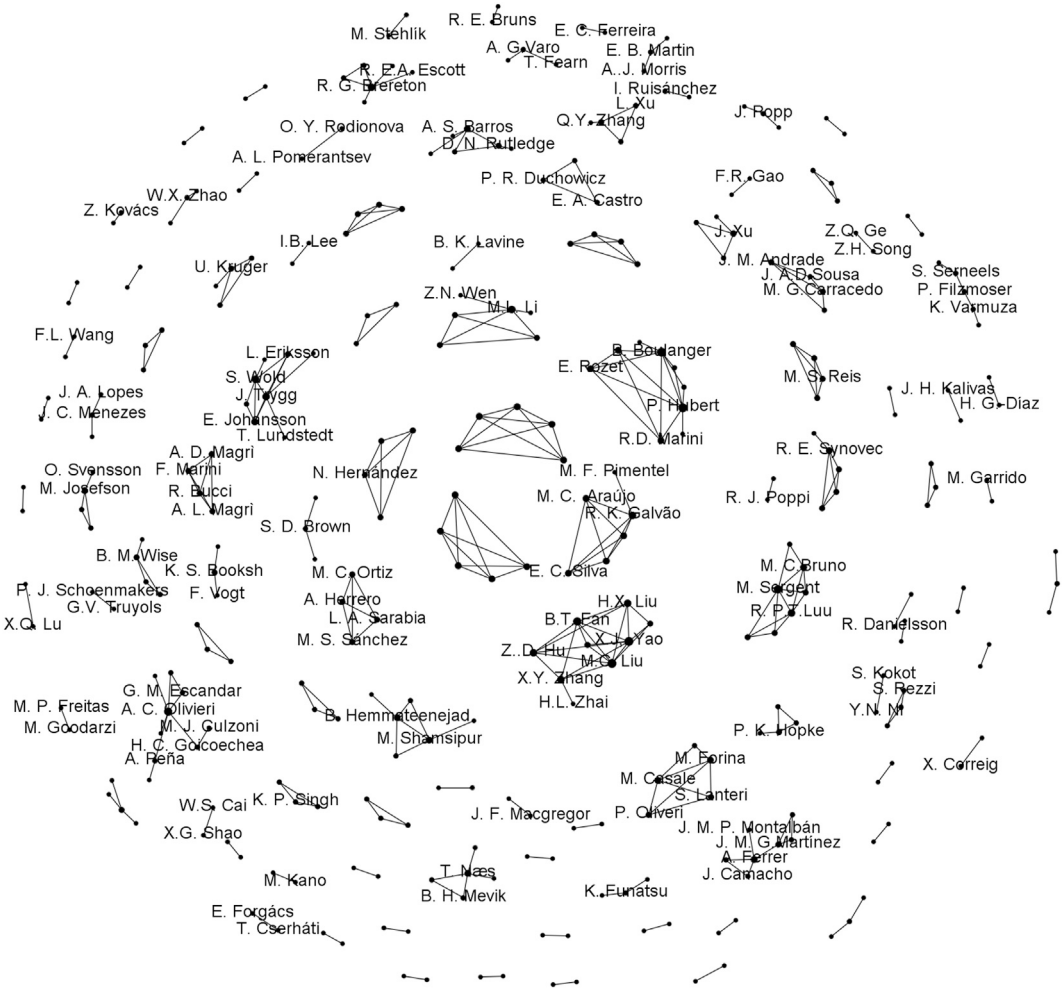

Fig. 10. The rest of the components.

the MCR methods. Fig. 7 shows a connection between Tauler's team and Buydens' team through Ruckebusch (Lille University of Science and Technology, France), Duponchel (Lille University of Science and Technology, France), Wehrens (Radboud University, Netherlands) and so on.

The team led by Bro, from the University of Copenhagen, has conducted numerous studies in many areas of chemometrics, such as experimental design, spectroscopy, metabonomics, process analytical technology, and in particular, multiway analysis methods. Bro's research links Tauler's team via Berg (University of Copenhagen, Denmark), who has cooperated with Juan in such aspects as the MCR method. Smilde and his team's (University of Amsterdam, Netherlands) research has focused on the chemometric method and application in analytical chemistry and metabonomics, among others. And Smilde has collaborated with Massart and Heyden's team via De Jong, as well as with Bro's team in chemometric methods. Rius' group (Rovira i Virgili University, Spain) has engaged in the development and the application of multivariate data analysis methods to explore measurements in chemistry.

The other community shown in Fig. 7 is named "China-Changsha". This community consists of two teams of researchers, one led by Yu and Wu (Hunan University, China) and the other headed by Liang and Xu (Central South University, China). Both teams are based in Changsha. Yu and Wu's team has carried out a large number of works on various aspects of chemometrics, especially second-order calibration, multiway calibration, and pattern recognition. Yu and Wu's team has a long-term cooperation with Liang and Xu's team in multivariate calibration and chemometric modeling. Liang and Xu's team has developed numerous chemometric methods, such as experimental design, curve resolution, PLS, classification, and variable selection. They have especially performed much work on the quality control of traditional Chinese medicines. Liang has connections with Chau (Hong Kong Polytechnic University, Hong Kong), Ozaki (Kwansei Gakuin University, Japan) and Kvalheim (University of Bergen, Norway). Liang has collaborated with Chau, Ozaki and Kvalheim in the traditional Chinese medicines, near-infrared spectroscopy and calibration methods, respectively. Fig. 7 also shows that the two communities are linked by Xu, who has cooperated a great deal with Massart in statistical methods and applications.

4.2. Other components

The second largest component of the network, with 21 nodes, is a loose group. It is represented by Martens (Norwegian University of Science and Technology, Norway). His research has focused on multivariate data modeling. The other important nodes in this component are Roger (National Research Institute of Science and Technology for Environment and Agriculture, France), Qannari (Nantes-Atlantic National College of Veterinary Medicine, France) and so on. Martens has collaborated with Qannari in multiblock analysis methods.

The third largest network has 13 nodes, represented by Esbensen (Aalborg University, Denmark) and Reinikainen (Lappeenranta University of Technology, Finland). Esbensen has studied the sampling theory, the process analytical technology, and so on. Reinikainen's main research interest is multivariate data analysis.

In Fig. 10, it can be seen that the rests of the components are smaller collaboration nets. Those top authors are in the rest components, maybe because they have less long-term partners. For example, Rutledge is in the twelfth component although he has high PageRank score shown in Fig. 6. Because he has only collaborated with three authors (nodes) on more than three papers in our datasets. The fourth to the tenth largest components are respectively represented by Olivieri (Rosario National University, Argentina), Hu (Lanzhou University, China), Hubert (University of Liège, Belgium), Wold and Trygg (Umeå University, Sweden), Ferrer (Polytechnic University of Valencia, Spain), Sergent (Aix-Marseille University, France), and Rutledge (AgroParisTech, France). And the more details about the fourth to the fifty largest components are also given in the Supplemental Material.

5. Conclusions

In this study, we have collected the chemometric papers published between 2001 and 2015 in the six journals to analyze the collaboration patterns and network of chemometric research. The trends of the publications and coauthors have revealed chemometrics as a fast-growing and competitive area. Centrality analysis and word cloud has respectively shown the most active chemometricians and the hot research topics. In the collaboration network, those most central authors are clustered into the "Big Europe" and the "China-Changsha" communities and the two communities are connected to be the largest component of the chemometric society. We have also analyzed the other major components in the network. The network analysis presents the scientific community structure of chemometricians in a much clearer manner.

It should be pointed out that our results are based on the collected data. The results may have some differences if we consider papers published in more journals and in a longer period of time. However, our work provides a new perspective for people to explore chemometrics and understand how chemometricians work together.

Acknowledgements

We thank the editor and the referee for constructive suggestions that substantially improved this work. We also thank Wentao Huang for collecting the related datasets. This work was supported by the National Natural Science Foundation of China [Grant No. 11801105 and 11561010] and the Innovation Program of Central South University [Grant No. 502221807].

Appendix. Data pre-processing

1. Data selection

From 2001 to 2015, AC, ACA, JCA and Talanta published 20214, 10156, 11686 and 17336 papers, respectively, but only a few of them relate with chemometrics [9]. ACA have classified the original research papers into different categories, such as featured article, atomic spectrometry, electrochemistry, molecular spectrometry, separation methods, chemometrics, and so on. However, ACA had no exact categories before 2005. AC, JCA and Talanta had no specific subdirectories named or related to chemometrics. Therefore, we should determine whether the papers in the four journals belong to the field of chemometrics.

First, we select the chemometric papers according to the keywords. Because the keywords are the most core words in a paper. And there is no keyword information in AC's webpages. Here, we use the keyword plus part of AC in the ISI system [35] to replace. What's more, the keywords indicating chemometric methods are expressed in a unified or a similar way, such as PLS, PCA, and PCR, so they are concise and easy to distinguish. These keywords about chemometric methods, for example, pattern recognition, curve resolution, or multivariable calibration, are found in review papers [8,38–42] and books [43,44]. Thus, we calculate the proportion that the terms about chemometric methods appear in keywords for every paper. If the proportion is more than 0.3, the paper is automatically regarded as a chemometric article.

From 2001 to 2005, ACA also included the papers published in special issues and presented in some conferences, such as the 7th International Conference on Chemometrics and Analytical Chemistry (CAC-2000), the 8th International Conference on Chemometrics and Analytical Chemistry (CAC), and so on. Thus, these papers are also taken into the datasets. Finally, we respectively choose 130, 734, 254 and 217 papers from AC, ACA, JCA and Talanta.

2. Name disambiguation

The authors' names in the academic journals are occasionally vague

and inconsistent. In some cases, the initial letters of an author's first and middle names are used, or the first name is spelled out, for example, *Y.Z. Liang* or *Yizeng Liang* and *Bernard G.M. Vandeginste* or *BGM Vandeginste*. Even worse, sometimes, we cannot identify the same author with several names, such as J. Chen, Jing Chen, and Jun Chen. We have encountered many similar situations in the pre-processing. It takes a lot of time to disambiguate the authors' names. The detailed workflow is described as follows. First, the R package *stringdist* can be used to measure the similarity of authors' names through the Jaccard method. These names with high similarities are checked carefully to ensure that they refer to a single person. Second, if a pair of names with low similarities have the same first name and last name, some extra information should be used, such as the email address and the affiliations. After the name disambiguation, the number of authors has been reduced from 9171 to 8389.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.05.011.

References

[1] S. Wold, Spline functions, a new tool in data-analysis, Kem. Tidskr. 3 (1972) 1–11.
[2] S. Wold, Chemometrics; what do we mean with it, and what do we want from it? Chemometr. Intell. Lab. Syst. 30 (1995) 109–115.
[3] B. Lavine, J. Workman, Chemometrics, Anal. Chem. 80 (2008) 4519–4531.
[4] R. Kiralj, M.M.C. Ferreira, The past, present, and future of chemometrics worldwide: some etymological, linguistic, and bibliometric investigations, J. Chemom. 20 (2006) 247–272.
[5] M.E.J. Newman, Coauthorship networks and patterns of scientific collaboration, Proc. Natl. Acad. Sci. 101 (2004) 5200–5205.
[6] T. Martin, B. Ball, B. Karrer, M.E.J. Newman, Coauthorship and citation patterns in the physical review, Phys. Rev. E. 88 (2013), 012814.
[7] P.S. Ji, J.S. Jin, Coauthorship and citation networks for statisticians, Ann. Appl. Stat. 10 (2016) 1779–1812.
[8] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools, Anal. Bioanal. Chem. 409 (2017) 5891–5899.
[9] R.G. Brereton, A short history of chemometrics: a personal view, J. Chemom. 28 (2014) 749–760.
[10] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. 98 (2000) 404–409.
[11] M.E.J. Newman, Scientific collaboration networks: I. Network construction and fundamental results, Phys. Rev. E. 64 (2001), 016131.
[12] M.E.J. Newman, Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, Phys. Rev. E. 64 (2001), 016132.
[13] L.C. Freemana, S.P. Borgattib, D.R. Whitea, Centrality in valued graphs: a measure of betweenness based on network flow, Soc. Network. 13 (1991) 141–154.
[14] L. Page, S. Brin, The anatomy of a large-scale hypertextual Web search engine, in: Proceedings of the Seventh International World-wide Web Conference, vol. 30, 1998, pp. 107–117.
[15] L. Shen, Q.S. Xu, D.S. Cao, X. Huang, The research of semi-supervised manifold learning algorithm based on Markov property, Acta Math. Sci. 45 (2015) 703–712.
[16] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E. 69 (2004), e026113.
[17] B. Karrer, M.E.J. Newman, Stochastic blockmodels and community structures in network, Phys. Rev. 83 (2011) 1436–1462.
[18] A.A. Amini, A.Y. Chen, P.J. Bickel, E. Levina, Pseudo-likelihood methods for community detection in large sparse networks, Ann. Stat. 41 (2013) 2097–2122.
[19] P.J. Bickel, A.Y. Chen, A nonparametric view of network models and Newman-Girvan and other modularities, Proc. Nat. Acad. Sci. 106 (2009) 21068–21073.
[20] M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (2006) 8577–8582.
[21] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech (2008) P10008.
[22] K. Rohe, S. Chatterjee, B. Yu, Spectral clustering and the high-dimensional stochastic blockmodel, Ann. Stat. 39 (2011) 1878–1915.
[23] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, Adv. Neural Inf. Process. Syst. 14 (2001) 849–856.
[24] P.W. Holland, K.B. Laskey, S. Leinhardt, Stochastic blockmodels: first steps, Soc. Netw. 5 (1983) 109–137.
[25] J.S. Jin, Fast community detection by score, Ann. Stat. 43 (2015) 57–89.
[26] K.H. Chen, J. Lei, Network cross-validation for determining the number of communities in network data, J. Am. Stat. Assoc. 113 (2018) 241–251.
[27] T.X. Li, E. Levina, J. Zhu, Network Cross-Validation by Edge Sampling, 2016. 1612.04717.
[28] M.E.J. Newman, G. Reinert, Estimating the number of communities in a network, Phys. Rev. Lett. 117 (2016), e078301.
[29] M.A. Riolo, G.T. Cantwell, G. Reinert, M.E.J. Newman, Efficient method for estimating the number of communities in a network, Phys. Rev. E. 96 (2017), e032310.
[30] P.J. Bickel, P. Sarkar, Hypothesis testing for automated community detection in networks, J. R. Stat. Soc. Ser. B 78 (2016) 253–273.
[31] J. Lei, A goodness of fit test for stochastic block models, Ann. Stat. 44 (2016) 401–424.
[32] C.M. Le, E. Levina, Estimating the Number of Communities in Networks by Spectral Methods, 2015. 1507.00827v1.
[33] D.F. Saldana, Y. Yu, Y. Feng, How many communities are there? J. Comput. Graph. Stat. 26 (2017) 171–181.
[34] E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing, Mixed membership stochastic blockmodels, J. Mach. Learn. Res. 9 (2008) 1981–2014.
[35] E. Garfield, KeyWords PIus, 1S1's breakthrough retrieval method. part 1. expanding your searching power on current contents on diskette, Curr. Comm. 13 (1990) 3–7.
[36] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Nat. Acad. Sci. 98 (2001) 404–409.
[37] L. Shen, Q.S. Xu, D.S. Cao, Y.Z. Liang, H. S Dai, The hybrid of semisupervised manifold learning and spectrum kernel for classification, J. Chemom. 32 (2018) e2955.
[38] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part II: modelling, validation, and application, Anal. Bioanal. Chem. 409 (2018) 6691–6704.
[39] B. Lavine, J. Workman, Chemometrics, Anal. Chem. 78 (2006) 4137–4145.
[40] B. Lavine, J. Workman, Chemometrics, Anal. Chem. 82 (2010) 4699–4711.
[41] B. Lavine, J. Workman, Chemometrics, Anal. Chem. 85 (2013) 705–714.
[42] S.D. Brown, R.S. Bear, T.B. Blank, Chemometrics. Anal. Chem. 64 (1992) 22–49.
[43] S.D. Brown, R. Tauler, B. Walczak, Comprehensive Chemometrics Chemical and Biochemical Data Analysis, Elsevier, Amsterdam, 2009.
[44] Y.Z. Liang, Q.S. Xu, H.D. Li, D.S. Cao, Support Vector Machines and Their Application in Chemistry and Biotechnology, CRC Press, New York, 2011.