Looking for Robust Properties in the Growth of an Academic Network: The Case of the Uruguayan Biological Research Community

Álvaro Cabana · Eduardo Mizraji · Andrés Pomi ·
Juan Carlos Valle-Lisboa

Received: 30 October 2007 / Accepted: 29 July 2008 /
Published online: 12 September 2008
© Springer Science + Business Media B.V. 2008

Abstract Graph-theoretical methods have recently been used to analyze certain properties of natural and social networks. In this work, we have investigated the early stages in the growth of a Uruguayan academic network, the Biology Area of the Programme for the Development of Basic Science (PEDECIBA). This transparent social network is a territory for the exploration of the reliability of clustering methods that can potentially be used when we are confronted with opaque natural systems that provide us with a limited spectrum of observables (happens in research on the relations between brain, thought and language). From our social net, we constructed two different graph representations based on the relationships among researchers revealed by their co-participation in Master's thesis committees. We studied these networks at different times and found that they achieve connectedness early in their evolution and exhibit the small-world property (i.e. high clustering with short path lengths). The data seem compatible with power law distributions of connectivity, clustering coefficients and betweenness centrality. Evidence of preferential attachment of new nodes and of new links between old nodes was also found in both representations. These results suggest that there are topological properties observed throughout the growth of the network that do not depend on the representations we have chosen but reflect intrinsic properties of the academic collective under study. Researchers in PEDECIBA are classified according to their specialties. We analysed the community structure detected by a standard algorithm in both representations. We found that much of the pre-specified structure is recovered and part of the mismatches can be attributed to convergent interests between scientists from different sub-disciplines. This result shows the potentiality of some clustering methods for the analysis of partially known natural systems.

Keywords Complex evolving network · Growing network · Early evolution of networks · Academic network

Á. Cabana · E. Mizraji (✉) · A. Pomi · J. C. Valle-Lisboa
Group of Cognitive Systems Modeling Sección Biofísica, Facultad de Ciencias,
Universidad de la República, Igua 4225, Montevideo 11400, Uruguay
e-mail: mizraj@fcien.edu.uy

1 Introduction

Complex networks embody many of the problems encountered when trying to understand the working of natural systems. In our group, we are mainly concerned with the study of cognitive phenomena, in particular those involved in the relationship between language, memory, thought processes and the brain. We are thus interested in exploring tools to extract hidden cognitive structures from the behavioural and linguistic productions of subjects, and the recent developments in complex networks create an interesting opportunity. In recent years, it has become clear that many of the properties of complex networks depend on their early evolution.

The most complete description of growing complex networks is available for social networks and in particular for scientific collaboration networks. As a consequence, we envision the study of social networks as a good intermediate research objective to understand complex network dynamics. We are going to assume that an evolving social system, in addition to its intrinsic interest, is an excellent system in which to explore these topics and to essay the robustness of some recently developed clustering methods under minimal pre-processing conditions.

The development of scientific communities requires the establishment of professional and social links. In the past few years, some aspects of the structure of scientific communities have been studied using complex-network approaches [1, 2]. Newman applied current network ideas to the social network of scientific collaborations, defined by co-authorship between scientists [3, 4]. His results show that collaboration networks have the characteristics of small-world networks, and their degree distributions appear to follow a power law. Furthermore, Barabási et al. [5] studied collaboration networks as prototypes of complex evolving networks, focusing on the dynamics and evolution of the graph.

To study an evolving network, precise data concerning the time at which the nodes and links were added should be available from the start of the growth process. Yet, due to the fact that computer databases cover only the past several decades, these data are limited to a recent segment of the evolution of each net. Consequently, the initial steps of growing networks have not been extensively described, and they surely deserve to be further explored. Our study benefits from the availability of data on our young scientific community since its origins.

The immediate purpose of our work is to analyse the developmental process of a small scientific community, the Uruguayan basic sciences system. Science as a professional activity has a short but intense history in Uruguay and was born in the second half of the 1980s, in parallel with the democratic reconstruction of the country. Before the military dictatorial period that began in 1973, scientific activity in Uruguay had some very professional and successful contributions, restricted to a few areas and linked to singular personalities [6]. The Programme for the Development of Basic Sciences (PEDECIBA) was established in October 1986 by agreement between the Ministry of Education and Culture and the Universidad de la República, with strong support from the United Nations Development Programme, with the aim to establish a scientific platform and to support high-level academic training in the areas of biology, chemistry, informatics, mathematics and physics.

The present study focuses on the evolution of the biology area of PEDECIBA. To describe its development, we studied the relationships among scientists belonging to the Programme. We explored their social interconnection through their co-participation in MSc thesis committees. All researchers belonging to the Programme supervise graduate students

in areas of their specialties. Inside the Programme, each researcher is assigned to a sub-area according to his or her specialty. MSc thesis committees are appointed by the advisor of the student, who is a researcher of the Programme, and two or more other researchers belonging to the Programme designed by an academic committee according to the subject of the research work developed by the student. Hence, advisors and researchers are the same group of individuals.

We consider two networks arising from Master's degree thesis committees, which partly reflect the professional and social links between scientists of the area. In both networks, the nodes are the researchers of the biology area of the Programme. In the first network, links connect researchers, including the thesis advisor, whenever they coincide as members of a Master's thesis committee. In order to evaluate the dependence of the conclusions on the chosen graph representation, we explored an alternative graph to describe the underlying social and thematic relationships among biology researchers. In this network, only links that connect the advisor to the selected members of the committee are established. This is based on the idea that members of the thesis committees were chosen to cover different thematic aspects of the research undertaken by the student and are necessarily related to the work of the advisor but not necessarily to the thematic specialties of the other members of the committee. So, in this graph, we discarded the connectivity among the designated members of the committee.

The thesis committee networks expand by the addition of new nodes and the associated links, corresponding to researchers who take part of committees for the first time. The number of links is also increased by the co-participation of two unconnected researchers in a new committee. We also studied a weighted graph that takes into account the number of times each pair of researchers co-participates in a committee. Given that the results are not very different from the unweighted network, in what follows, we do not present these data.

Here, we study the growth and evolution of these thesis committee networks, analyzing aspects of their topology and dynamics. With this objective, we extract some descriptive empirical parameters of the network. We first present the two alternative graph representations used for the Master's thesis committees network (Section 2). Then, we characterize the topology and dynamics of the network using direct measures on the available data (Section 3).

During the preparation of this work, it has been noted in the literature that the classic, unweighted unipartite representation of social nets could give unreliable or incorrect results. The representation proposed as the natural, empirically standard case for social networks is a bipartite network [7]. In that framework, unipartite networks are envisioned as projections of these bipartite networks.

In spite of that, we consider it valuable as a first approximation, in order to describe the growth of our scientific community, to represent the relationships between researchers using classical methods. We think that in this case, the classical approach suffices, given that in contrast to networks obtained from large databases, we know exactly the modular or community structure of our social network, a structure that can be partially reconstructed using community-finding algorithms as we describe below.

The study of modularity in bipartite networks has been classically addressed constructing unipartite projections and then identifying modules. Methods to study modularity directly in bipartite networks were only recently proposed independently in the statistical physics community by Guimerà et al. [7] and by Barber [8]. However, previous work on co-clustering in bipartite networks has existed for a long time in graph theoretic and computer

sciences communities (see, for example, [9–11]) and in the neural network community [12]. These techniques are related to linear algebraic methods using bipartite representations as singular value decomposition analysis [13] and latent semantic analysis techniques [14, 15]. Using classical unipartite techniques, we analyze the community structure of the graphs and recover the natural classification of researchers (see Section 4). Finally, we discuss our findings and present our conclusions (Section 5).

2 Construction of Graph Representations

Each thesis committee is composed of three or more researchers of the Programme: the student's advisor and two, more, commonly three, and sometimes more, other scientists who accepted to participate in the committee. A straightforward graph representation of this situation consists of nodes corresponding to scientists and edges linking those who participate together in a given committee.

To find out how the properties attributed to a real network of scientific interactions depend on the graph representation chosen, we performed comparative analyses on two representations that differ slightly in their definitions and have different link densities. 'Committee graphs' were built linking all members of a given committee to each other, whereas in 'advisor graphs,' just the student's advisor is linked to the other members of the committee (Fig. 1).

The composition of the committees is available from the first thesis presented in 1990 until 2006. Graph representations were built using cumulative data from the beginnings of the Programme up to a given date. Given the smallness of our scientific community, we decided to carry out this study on a 2-year basis.

Fig. 1 Construction of 'committee' (a) and 'advisor' (b) graphs. Graphs generated with each procedure are shown for four time points (pictures were made with Pajek [30])

3 Topological Observables

3.1 Graph Growth

There were 303 Master's theses in biology presented within the framework of the Programme prior to May 2006, and 287 local and foreign scientists participated as referees and advisors. The time evolution of the number of theses, participants (as nodes in the graph) and links between them are shown in Fig. 2. It is clear that both representations differ, 'advisor' graphs being sparser than 'committee' graphs. In spite of this, the number of links increases at an approximately constant rate in both representations, albeit a different one in each case. The number of nodes and theses presented also increases linearly with time.

3.2 Small-world Observables

Small-world networks combine random graph-like low mean distance between nodes and a high degree of clustering reminiscent of ordered lattices [1]. In the case of social affiliation networks, where members are joined by common participation in a certain event or membership in some group, small-world observables are likely to arise [16]. Nevertheless, the extent to which these observables appear during the initial steps in the growth of such a network has not been addressed.

The clustering coefficient (CC_i) of a node i is defined as the number of links between its neighbors, divided by the number of links that could possibly exist between them. The clustering coefficient of a graph is just the mean clustering coefficient of all nodes $CC = \frac{1}{N} \times \sum CC_i$. Following Watts [17], characteristic path length was computed here as the median of mean distances from each node to the rest of the graph.

Figure 3 shows the clustering coefficients and characteristic path lengths for both graph representations as a function of time, together with what would be expected for a random graph of the same size and number of links. Both representations are small-world graphs, but the coefficients are not invariant across representations. Characteristic path length in both representations is nearly the same as what would be expected for random graphs and

Fig. 2 Number of nodes, committee and links as a function of time for both graph representations

Fig. 3 Clustering coefficient (**a**) and characteristic path length (**b**) for ‘committee’ and ‘advisor’ graphs. *Insets* show the ratio of observed quantities to what would be expected from a random graph of same size and number of links

rapidly approaches a stationary value of 3 for ‘committee’ and of around 4 for ‘advisor’ graphs (Fig. 3b). The clustering coefficient is greater than expected for random graphs and remains approximately constant. Despite this constancy, the ratio of observed clustering to that of a random graph of same size and link number grows from about 10 to 20 in ‘committee’ graphs and from about 5 to 10 for ‘advisor’ graphs, showing that clustering is relatively increasing. However, its absolute value is nearly 0.7 in ‘committee’ graphs and only about 0.1 in ‘advisor’ graphs, which is consistent with a lower link density in this type of representation. Notice that each thesis committee is only mapped onto a highly clustered connectivity pattern in ‘committee’ graph representations. In ‘advisor’ graphs, each committee corresponds to a star-like structure which adds no clustering. In this case, the clustering coefficient only grows when nodes participate in other committees with different advisors.

3.3 Degree Distribution

Small-world networks that have power law degree distributions have been called ‘scale-free networks’ and were found to be present in a wide range of natural and artificial networks [2, 18]. The degree distribution of both graph representations seems to be a power law with an exponent that varies between -1.41 and -1.53 for ‘committee’ graphs and between -1.16 and -1.41 for ‘advisor graphs’ (Fig. 4), as estimated by least-squares fitting. For both representations, the final value of the power law exponent is around -1.4 .

We fit a power law to the log–log plot of the histogram. Of course, this is not a convenient procedure if sharp distinctions between models are required (see [19]). It has been remarked that in many cases, it is often better (see [18]) to fit a linear function to the log–log or the semi-log plot (for an exponential distribution) of the cumulative histogram of the data. As it has been recently noted, however, this procedure does not guarantee the best fit [19], and cannot be used to test goodness of fit. Given the smallness of our data set, we cannot affirm that these data follow a power law (i.e. we cannot rule out other reasonable possibilities), but if the heavy-tailed distributions are fit with a power law with a finite size cutoff using the methods in [19], the exponent lies in the range -1.2 to -1.7 .

Fig. 4 Degree distributions for the 2006 ‘committee’ (a) and ‘advisor’ (b) graph representations. *Insets* show the evolution of the estimated exponents. The exponents were obtained by fitting a straight line to the log–log plot of the histogram of the data by least squares. A better approach using the methods in [19] with a finite size cutoff gives slightly different results; in particular, the scaling parameter for the data of 2006 is -1.67 and with a minimal value of degree of 3 and a finite cutoff of 52. In this case, the p value calculated as in [19] gives 0.151

3.4 Preferential Attachment

The Barabási–Albert model for the growth of scale-free networks relies on the existence of linear ‘preferential attachment’ for the addition of new links [2], which implies that the probability $\Pi(k_j)$ of a new node i to link a given node j is proportional to that node’s degree k_j . To reduce noise impact in the measurements, usually, cumulative distributions of this probability are calculated; if linear preferential attachment is at work, cumulative distributions (usually denoted $\kappa(k_i)$) must be quadratic functions of degree.

We analyzed the growth pattern of the graph representations and estimated their cumulative distributions $\kappa(k_i)$ (Fig. 5a, b). In both representations, the cumulative distributions were found to be roughly quadratic functions of degree, supporting the existence of linear preferential attachment in the growth dynamics of both representations.

For the addition of new links between old nodes, preferential attachment has also been found to be at work [5]. In this case, the probability distribution $\kappa(k_i, k_j)$ is proposed to be proportional to the product of the connectivity of the nodes, $k_i \cdot k_j$. Figure 5 shows the cumulative distribution for this situation for ‘committee’ (c) and ‘advisor’ (d) graphs. Neither of these distributions shows the quadratic dependence predicted by linear preferential attachment; the former distribution is a curved one with slope greater than 2 at low $k_i \cdot k_j$ values and much smaller final slope. The latter distribution shows a straight line with slope less than 2. That is what would be expected from non-linear preferential attachment.

3.5 Clustering and Betweenness

The clustering coefficient and betweenness centrality are topological properties that have been found to display power law relationships with node connectivity. Some authors consider the former property as an indicator of modular or hierarchical topology of the network [20–22], and the latter property is a general relationship valid for large scale-free networks

Fig. 5 Cumulative distributions of the linkage probability of new nodes (**a**, **b**) and of existing nodes (**c**, **d**) for ‘committee’ and ‘advisor’ graphs, respectively, as functions of node degree

[23]. Betweenness is one of the proposed measures of a node’s centrality, i.e. how ‘central’ is the node in the sense of being part of shortest paths between nodes. The so called ‘shortest path betweenness centrality’ is measured as the sum of the fraction of shortest paths a node is part of, over all pairs of nodes [24].

Fig. 6 Power law exponents for clustering coefficient dependence on connectivity, $CC(k) \sim k^\delta$, for ‘committee’ (**a**) and ‘advisor’ (**b**) graphs. Exponents increase from nearly -1 to a final value around -0.79 . *Insets* show typical distributions with a power law fit to the mean clustering values for each k

While clustering depends on the local connectivity pattern, betweenness centrality measures the fraction of shortest paths that pass through a node, hence reflecting the topological structure of the whole graph.

In our case, the dependence of clustering coefficient on degree is well described by a power law in both graph representations. The value of the exponent for the clustering relationship changes monotonically from -1 to a final value of around -0.79 (Fig. 6).

If betweenness centrality holds a power law relationship with node degree (that is, $B(k) \sim k^\eta$), its cumulative distribution should also consist of a power law with exponent $\eta + 1$. Figure 7 shows the cumulative betweenness for ‘committee’ (a) and ‘advisor’ (c) graphs for three widely spaced epochs. At least for a certain range of degree values, the slope of the log–log plot is close to 3, implying that the value of the η exponent is near 2, which is a proposed maximum value of this exponent for large complex networks of small loop density [23]. Notice the constant dependence of betweenness on degree (shown in Fig. 7b, d), even when the number of nodes doubles and the number of links is three times greater in the interval 1996–2006.

Fig. 7 Betweenness centrality shows a power law relationship with node degree. Cumulative plots (a, c) show a slope of about 3, which implies an exponent of 2 for the betweenness power law (a, b ‘committee’ graphs; c, d ‘advisor’ graphs)

4 Thematic Clustering of Biology Researchers

We used a community-detecting algorithm [25] which is a close variant of one described by Girvan and Newman [26] to cluster the researchers in thematic groups or communities. As long as our graph representation captures the underlying social process, we expect the community structure derived from the graph to partially match the formal community structure imposed by the PEDECIBA. As we mentioned above, all biology researchers are assigned to a sub-area (standard sub-disciplines of biology) based on their main specialty within biology. Obviously, this does not preclude researchers from one area to work in other themes of research and to collaborate with scientists of other disciplines. Hence, it is expected that the communities detected by the algorithm do not perfectly match the classification of the researchers according to the biological sub-disciplines of the Programme, but rather match instead the actual thematic interest of their work. After applying the algorithm of Radicchi [25], we obtained the community structure depicted in Fig. 8, where we indicate the sub-area each researcher belongs to.

As can be seen, most researchers are grouped in accordance to their sub-area, indicating that the graph representation is at least good enough to capture the underlying academic relationships. According to what has just been noticed, some sub-areas are partially dispersed and that the members are closer to members of other sub-areas than to members of their sub-areas. In most of these cases, these researchers have thematic ties with researchers in other sub-areas, showing the effect of collaboration within the biology section of the Programme. For example, genetics researchers adopt two different approaches, one linked to zoology and the other linked to cell and molecular biology. The same happens with researchers in ecology who collaborate with zoologists and plant biologists or biophysicists who show links to neuroscientists or cellular biologists.

Fig. 8 Community structure produced by the algorithm of Radicchi et al. [25] in the case of ‘committee’ graphs (a) and advisor graphs (b), using the edge clustering coefficient for cycles of order 4 and the weak sense community definition. Each dendrogram is displayed in a circular layout and represents the hierarchical structure of the network communities. Only the nodes corresponding to researchers belonging to a given PEDECIBA sub-area are *marked* in each picture

In all cases, these partitions are apparent from the output of the community-detecting algorithm.

5 Conclusions

The objective of this article is to describe how the usual observables of complex networks change in the early steps of graph development. In most models describing complex networks, time is a variable that disappears in the limit, so the usual distributions could be termed stationary. We are interested in the transient steps, because we think that, as happens in many cases, much information is available in these transients. In order to describe the growth process, we created two different representations for the same network, aiming to find robust indicators of the underlying mechanisms.

Although the two representations differ in many aspects, there are noteworthy similarities. Both graph representations show a wide range of preferential attachment, mostly linear in the case of new incoming nodes but non-linear for the case of new links between existing nodes. Linear preferential attachment for the addition of new nodes results in the establishment of a power law degree distribution in the limit [2]. In any case, our data suggest a power law degree distribution, but better fitting techniques should be used to address this issue. If fitted with a power law, the time evolution of the exponents seems to be another robust property that emerges in both representations. There are differences though, since in the ‘advisor’ graphs, the exponent starts from values closer to -1 than what is observed in ‘committee’ graphs. However, the final values they achieve are almost indistinguishable.

Clustering coefficients are consistently different between the two representations, something that comes as no surprise given the different ways of setting up the graphs. In spite of this difference, however, the dependence of clustering on degree follows very similar power laws that change in the same way as time progresses (Fig. 6). Whether this signals a robust measure connected to the graph evolution, we do not know yet. But, we notice that if the exponents of these power laws reflect the relationships between the clustering of hubs and that of barely connected nodes, this difference should be comparable in both cases. Thus, the common evolution of this exponent probably means that, in both representations, the underlying social network becomes less partitioned, both by a decreasing of low-degree nodes’ clustering and an increase of high-degree nodes’ clustering. Both types of changes imply the establishment of links between different communities.

The biology section of PEDECIBA is subdivided in sub-disciplines, as zoology or cellular biology. We are currently investigating whether the reduction in the dependence of clustering on degree might mean more inter-disciplinary connections. We approached this investigation by the use of the methods developed to detect communities, like those of Newman and coworkers [27] or the method of Radicchi et al. [25]. We found a close correspondence between the ‘assumed’ classification and the one that the algorithm finds. There are some differences that are explained by convergent interests from researchers of different sub-disciplines. The fact of joining a committee is only partially determined by the thematic domain of the scientist within PEDECIBA. These committees are mainly established on research thematic affinities and/or technical complementarities, which are well captured by the algorithm of Radicchi and coworkers [25]. This can be verified by the analysis of the community structure and the thematic specialty of each researcher (Fig. 8). The complete graphs, including the names of the researchers, are available upon request from the authors.

Admittedly, there are some researchers that are misclassified by the method, and the correct assignment to a class requires the refinement of our techniques. There is a recent concern that graph representations of social networks should not be based on unipartite graphs. Social groups are usually clearly defined and allow unambiguous representations using bipartite graphs, but in other areas of science, this is not always possible or reasonable. For instance, when trying to understand the relationships between thought, language and the brain, the lexical production of an individual is an objective manifestation of the underlying brain activity. This lexical production can be classified a posteriori as a set of documents containing words. This classification allows researchers to define a bipartite network linking documents and vocabulary, and this network can be analysed using powerful matrix methods [14, 15, 28]. Yet, this a posteriori classification of lexical information represents one of many possible organizations of the data and as such has as much natural information as prejudice. In fact, when using textual data in the information retrieval community, the definition of a document is ambiguous enough to be left to the researcher [29]. If we want to characterize the development of cognitive categories using operational observables, like lexical production, it is important to continue exploring the topology of the relations directly generated by these observables, their imperfection notwithstanding. The graphs we used to describe the scientific community show that at least part of the information underlying the evolution of the network is captured in the unipartite graph, and we think that the tools are valuable in searching for the hidden structure in the data. We consider this a positive result that encourages us to use these methods as first approaches to attack hard problems as the previously mentioned relationships among the brain, thought and language, where, despite advances in neuroimaging, cellular and molecular biology and the neurosciences, the underlying structures are opaque and any definition is non-natural.

From the point of view of social network evolution, we are not yet in possession of a concrete model that can explain our data. The evolution of the topological and growth parameters shown in this work could be valuable to the construction of such a model.

Acknowledgements We thank the kind help of Graciela Bocage and Beatriz San Martín at PEDECIBA. This work was partly supported by PEDECIBA. A.C. and J.C.V.-L. were partly supported by FCE—S/C/IF/54/002.

References

1. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998). doi:[10.1038/30918](https://doi.org/10.1038/30918)
2. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. *Science* **286**, 509–512 (1999). doi:[10.1126/science.286.5439.509](https://doi.org/10.1126/science.286.5439.509)
3. Newman, M.E.J.: Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E* **64**, 016131 (2001). doi:[10.1103/PhysRevE.64.016131](https://doi.org/10.1103/PhysRevE.64.016131)
4. Newman, M.E.J.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 016132 (2001). doi:[10.1103/PhysRevE.64.016132](https://doi.org/10.1103/PhysRevE.64.016132)
5. Barabási, A.-L., Jeong, H., Nédá, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. *Physica A* **311**, 590–614 (2002). doi:[10.1016/S0378-4371\(02\)00736-7](https://doi.org/10.1016/S0378-4371(02)00736-7)
6. Mañé-Garzón, F., Mazzella, H.: Historia de la Fisiología en el Uruguay. Oficina del Libro AEM, Montevideo (2000)
7. Guimerà, R., Sales-Pardo, M., Amaral, L.A.N.: Module identification in bipartite and directed networks. *Phys. Rev. E* **76**, 036102 (2007). doi:[10.1103/PhysRevE.76.036102](https://doi.org/10.1103/PhysRevE.76.036102)
8. Barber, M.J.: Modularity and community detection in bipartite networks. *Phys. Rev. E* **76**, 066102 (2007). doi:[10.1103/PhysRevE.76.066102](https://doi.org/10.1103/PhysRevE.76.066102)

9. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, California, USA (2001)
10. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Washington, DC, USA (2003)
11. Zha, H., He, X., Ding, C., Simon, H., Gu, M.: Bipartite graph partitioning and data clustering. In: Proceedings of the Tenth ACM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA (2001)
12. Ding, C.: Document retrieval and clustering: from principal component analysis to self-aggregation networks. In: Bishop, C.M., Frey, B.J. (eds.) Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, FL (2003)
13. Golub, G., van Loan, C.: Matrix Computations, 2nd edn. The Johns Hopkins University Press, Baltimore, MD (1989)
14. Berry, M., Dumais, S., O'Brien, G.: Using linear algebra for intelligent information retrieval. *SIAM Rev.* **37**(4), 573–595 (1995). doi:[10.1137/1037127](https://doi.org/10.1137/1037127)
15. Valle-Lisboa, J.C., Mizraji, E.: The uncovering of hidden structures by latent semantic analysis. *Inf. Sci.* **177**, 4122–4147 (2007). doi:[10.1016/j.ins.2007.04.007](https://doi.org/10.1016/j.ins.2007.04.007)
16. Newman, M.E.J., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **99**, 2566–2572 (2002). doi:[10.1073/pnas.012582999](https://doi.org/10.1073/pnas.012582999)
17. Watts, D.J.: *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press (1999)
18. Amaral, L.A.N., Scala, A., Barthélemy, M., Stanley, H.E.: Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**(21), 11149–11152 (2000). doi:[10.1073/pnas.200327197](https://doi.org/10.1073/pnas.200327197)
19. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. arXiv: 0706.1062v1 [physics.data-an] (2007)
20. Barabási, A.-L., Ravasz, E., Vicsek, T.: Deterministic scale-free networks. *Physica A* **299**, 559–564 (2001). doi:[10.1016/S0378-4371\(01\)00369-7](https://doi.org/10.1016/S0378-4371(01)00369-7)
21. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.-L.: Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002). doi:[10.1126/science.1073374](https://doi.org/10.1126/science.1073374)
22. Dorogovtsev, S.N., Mendes, J.F.F.: The shortest path to complex networks. arXiv:cond-mat/0404593 (2004)
23. Barthélemy, M.: Betweenness centrality in large complex networks. *Eur. Phys. J. B* **38**, 163–168 (2004). doi:[10.1140/epjb/e2004-00111-4](https://doi.org/10.1140/epjb/e2004-00111-4)
24. Newman, M.E.J.: A measure of betweenness centrality based on random walks. *Soc. Netw.* **27**(1), 39–54 (2005). doi:[10.1016/j.socnet.2004.11.009](https://doi.org/10.1016/j.socnet.2004.11.009)
25. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **101**(9), 2658–2663 (2004). doi:[10.1073/pnas.0400054101](https://doi.org/10.1073/pnas.0400054101)
26. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**(12), 7821–7826 (2002). doi:[10.1073/pnas.122653799](https://doi.org/10.1073/pnas.122653799)
27. Newman, M.E.J.: Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**(23), 8577–8582 (2006). doi:[10.1073/pnas.0601602103](https://doi.org/10.1073/pnas.0601602103)
28. Mizraji, E.: Neural memories and search engines. *Int. J. Gen. Syst.* (2008, in press). doi:[10.1080/03081070802037738](https://doi.org/10.1080/03081070802037738)
29. Dumais, S.T.: LSA and information retrieval: getting back to basics. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) *LSA: A Road to Meaning*, pp. 293–321. Lawrence Erlbaum, Mahwah, NJ (2006)
30. Batagelj, V., Mrvar, A.: Pajek—program for large network analysis. <http://vlado.fmf.uni-lj.si/pub/networks/pajek> (2007)