N-body decomposition of bipartite author networks

R. Lambiotte* and M. Ausloos†
SUPRATECS, Université de Liège, B5 Sart-Tilman, B-4000 Liège, Belgium
(Received 20 July 2005; revised manuscript received 6 October 2005; published 15 December 2005)

In this paper, we present a method to project co-authorship networks, that accounts in detail for the geometrical structure of scientists' collaborations. By restricting the scope to three-body interactions, we focus on the number of triangles in the system, and show the importance of multi-scientist (more than two) collaborations in the social network. This motivates the introduction of generalized networks, where basic connections are not binary, but involve arbitrary number of components. We focus on the three-body case and study numerically the percolation transition.

DOI: 10.1103/PhysRevE.72.066117 PACS number(s): 89.75.Fb, 89.75.Hc, 87.23.Ge

I. INTRODUCTION

It is well known in statistical physics that *N*-body correlations have to be carefully described in order to characterize statistical properties of complex systems. For instance, in the case of the Liouville equation for Hamiltonian dynamics, this problem is at the heart of the derivation of the reduced BBGKY hierarchy, thereby leading to the Boltzmann and Enskog theories for fluids [1]. In this line of thought, it is primordial to discriminate *N*-body correlations that are intrinsic *N*-body interactions from those that merely develop from lower order interactions. This issue is directly related to a well-known problem in complex network theory, i.e., the "*projection*" of bipartite networks, i.e., composed of two kinds of nodes, onto unipartite networks, i.e., composed of one kind of node. As a paradigm for such systems, people usually consider co-authorship networks [2–4], namely networks whose nodes are scientists and articles, with links running between scientists and the papers they wrote. In that case, the usual projection method [5] consists in focusing, e.g., on the scientist nodes and in drawing a link between them if they co-authored a common paper (see Fig. 1). As a result, the projected system is a unipartite network of scientists that characterizes the community structure of science collaborations. Such studies have been very active recently, due to their complex social structure [6], to the ubiquity of such bipartite networks in complex systems [7–9], and to the large databases available.

A standard quantity of interest in order to characterize the structure of the projected network is the clustering coefficient [10], which measures network "transitivity," namely the probability that two co-authors of a scientist have themselves co-authored a paper. In topological terms, it is a measure of the density of triangles in a network, a triangle being formed every time two of one's collaborators collaborate with each other. This coefficient is usually very high in systems where sociological cliques develop [11]. However, part of the clustering in the co-authorship network is due to papers with three or more co-authors. Such papers introduce trivial triangles of collaborating authors (see Fig. 1), thereby increasing the clustering coefficient. This problem, that was raised by Newman *et al.* [5], was circumvented by studying directly the bipartite network, in order to infer the authors' community structure. Newman *et al.* showed on some examples that these high-order interactions may account for one half of the clustering coefficient. One should note, however, that if this approach offers a well-defined theoretical framework for bipartite networks, it suffers a lack of transparency as compared to the original projection method, i.e., it does not allow a clear visualization of the unipartite structure.

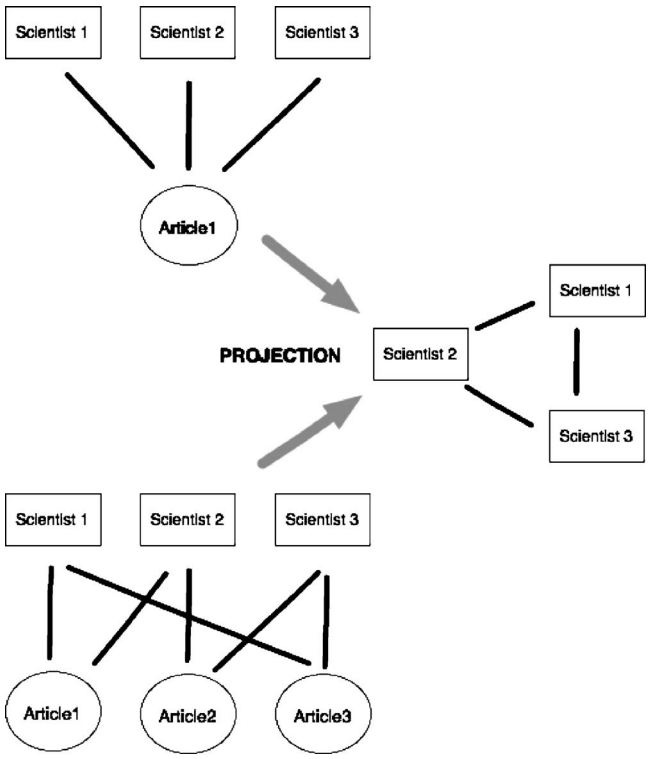


FIG. 1. Usual projection method of the bipartite graph on a unipartite scientists graph. Nonbijectivity of the application simplifies the structure of the network, thereby hiding, amongst others, the signification of triangles in the unipartite scientists graph.

*Electronic address: renaud.lambiotte@ulg.ac.be
†Electronic address: marcel.ausloos@ulg.ac.be

In this article, we propose an alternative approach that is based on a more refined unipartite projection and follows statistical mechanics usual expansion methods. To do so, we focus on a data set, retrieved from the arXiv database and composed of articles dedicated to complex network theory. This choice is motivated by their relatively few co-authors per article, a property typical to theoretical physics papers [12]. Our method consists in discriminating the different kinds of scientist collaborations, based upon the number of co-authors per article. This discrimination leads to a diagram representation [13,14] of co-authorship (see also [15] for the applicability of Feynman diagrams in complex networks). The resulting N-body projection reconciles the visual features of the usual projection and the exact description of Newman's theoretical approach. Empirical results confirm the importance of high-order collaborations in the network structure. Therefore, we introduce in the last section a simple network model, which is based on random triangular connections between the nodes. We study numerically percolation features in the model.

II. N-BODY PROJECTION METHOD

The data set contains all articles from arXiv in the time interval [1995:2005] that contain the word "network" in their abstract and are classified as "*cond-mat*"). In order to discriminate the authors and avoid spurious data, we checked the names and the first names of the authors. Moreover, in order to avoid multiple ways for an author to co-sign a paper, we also took into account the initial notation of the prenames. For instance, *Marcel Ausloos* and *M. Ausloos* are the same person, while *Marcel Ausloos* and *Mike Ausloos* are considered to be different. Let us stress that this method may lead to ambiguities if an initial refers to two different first names, e.g., *M. Ausloos* might be *Marcel Ausloos* or *Mike Ausloos*. Nonetheless, we have verified that this case occurs only once in the data set (*Hawoong Jeong, Hyeong-Chai Jeong, and H. Jeong*), so that its effects are negligible. In that sole case, we attributed the papers of *H. Jeong* to the most prolific *Jeong*, i.e., *Hawoong Jeong* in the data set. Given this identification method, we find $n_P=2533$ persons and $n_A=1611$ articles. By using the projection method of Fig. 1, the author network is made of a large connected island composed of 567 scientists and by a multitude of small disconnected clusters (Fig. 2). The size s distribution of the clusters (Fig. 3) shows a power law decrease $\sim s^{-2}$, compatible with the observations of [16]. Let us also stress that the distribution of the number of co-authors per article (Fig. 4) shows clearly a rapid exponential decrease, associated to a clear predominance of small collaborations.

Formally, the bipartite structure authors-papers may be mapped exactly on the vector of matrices $\mathcal{M}$ defined by

$$\mathcal{M}=[\mathbf{M}^{(1)},\mathbf{M}^{(2)},\ldots,\mathbf{M}^{(j)},\ldots,\mathbf{M}^{(n_P)}] \tag{1}$$

where $\mathbf{M}^{(j)}$ is a square n_P^j matrix that accounts for all articles co-authored by j scientists. By definition, the element $M^{(j)}_{a_1\cdots a_j}$ is equal to the number of articles co-authored by the j authors $a_1,\ldots,a_j$. In the following, we assume that co-

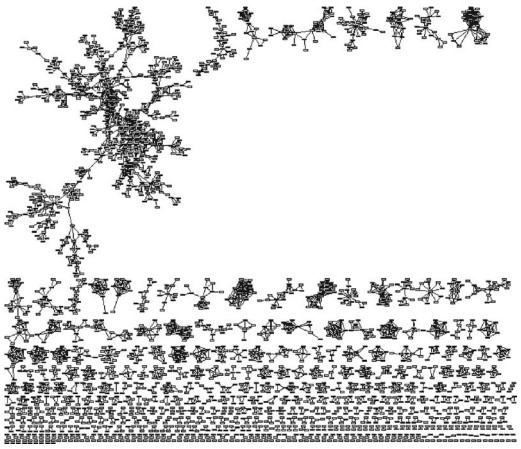


FIG. 2. Network of scientists having written a "network" article in the time interval [1995:2005] (see text for data acquisition). The main island of co-authors is composed by 567 authors and 1325 links.

authorship is not a directed relation, thereby neglecting the position of the authors in the collaboration, e.g., whether or not the author is the first author. This implies that the matrices are symmetric under permutations of indices. Moreover, as people cannot collaborate with themselves, the diagonal elements $M^{(j)}_{aa\cdots a}$ vanish by construction. For example, $M^{(1)}_{a_1}$ and $M^{(2)}_{a_1a_2}$ represent respectively the total number of papers written by a_1 alone, and the total number of papers written by the pair (a_1,a_2).

A way to visualize $\mathcal{M}$ consists in a network whose nodes are the scientists and whose links are discriminated by their shape. The intrinsic co-authorship interactions form loops (order 1), lines (order 2), triangles (order 3) (see Fig. 5),... . To represent the intensity of the multiplet interaction, the width of the lines is taken to be proportional to the number of collaborations of this multiplet. Altogether, these rules lead to a graphical representation of $\mathcal{M}$ that is much more refined than the usual projection method (Fig. 6).

It is important to point out that the vector of matrices $\mathcal{M}$ describes without approximation the bipartite network, and

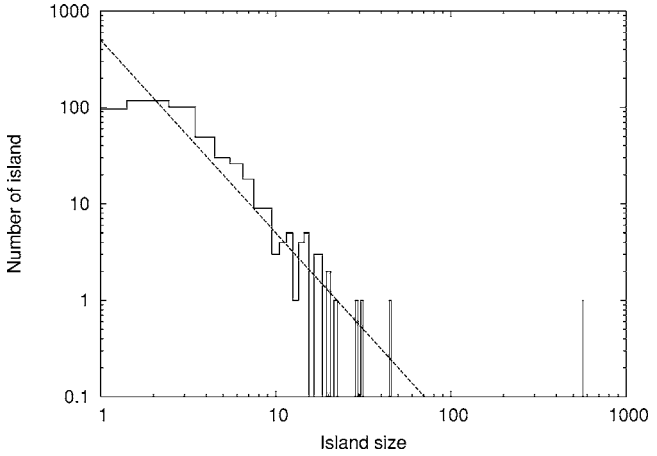


FIG. 3. Histogram of the size s of the disconnected islands of Fig. 2. The dashed line $\sim s^{-2}$ is a guide for the eye. The extreme event at $s=567$ corresponds to the main island.

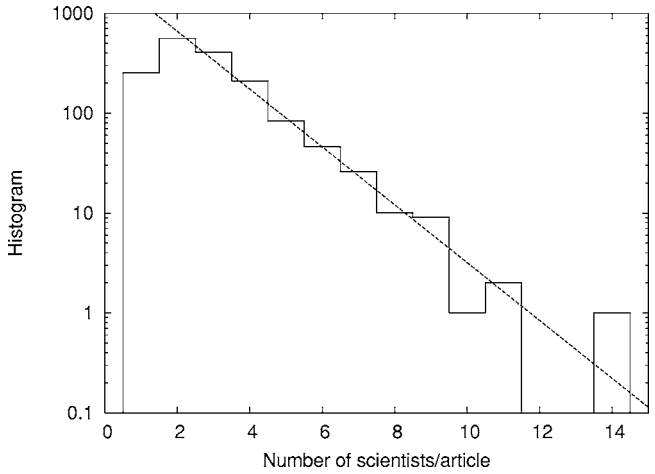


FIG. 4. Histogram of the number of scientists/articles, n, for the same data as in Fig. 2. The dashed line corresponds to the fit $e^{-n/1.5}$.

that it reminds the Liouville distribution in phase space of a Hamiltonian system. Accordingly, a relevant macroscopic description of the system relies on a coarse-grained reduction of its internal variables. The simplest reduced matrix is the one-scientist matrix $\mathbf{R}^{(1)}$ that is obtained by summing over the N-body connections, $N \geqslant 2$. $\mathbf{R}^{(1)}$ is a n_P vector for which the element a_1 is

$$R^{(1)}_{a_1} = M^{(1)}_{a1} + \sum_{a_2} M^{(2)}_{a_1 a_2} + \sum_{a_2} \sum_{a_3 < a_2} M^{(3)}_{a_1 a_2 a_3} + \cdots + \sum_{a_2} \cdots \sum_{a_j < a_{j-1}} M^{(j)}_{a_1 \cdots a_j} + \cdots . \quad (2)$$

It is straightforward to show that the elements $R^{(1)}_{a_j}$ denote the total number of articles written by the scientist a_j. The second order $(n_P \times n_P)$ matrix is

$$R^{(2)}_{a_1 a_2} = M^{(2)}_{a_1 a_2} + \sum_{a_3} M^{(3)}_{a_1 \cdots a_3} + \cdots + \sum_{a_3} \cdots \sum_{a_j < a_{j-1}} M^{(j)}_{a_1 \cdots a_j} + \cdots . \quad (3)$$

Its elements represent the total number of articles written by

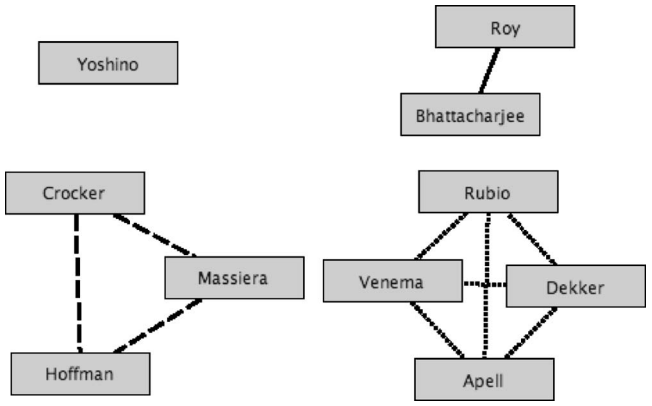


FIG. 5. Graphical representation of the four most basic authors' interactions, namely, 1, 2, 3, 4 co-authorships.

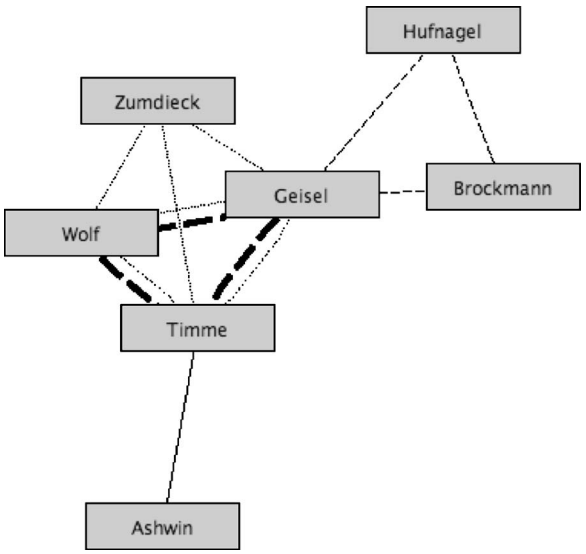


FIG. 6. Graphical representation of the co-authorship network. This small subnetwork accounts for one two-author collaboration (*Timme, Ashwin*); four three-author collaborations, three times the triplet (*Timme, Wolf, Geisel*) depicted by stronger links and once (*Geisel, Hufnagel, Brockmann*); one four-author collaboration (*Timme, Wolf, Geisel, Zumdieck*).

the pair of scientists (a_1, a_2). Remarkably, this matrix reproduces the usual projection method (Fig. 1) and obviously simplifies the structure of the bipartite structure by hiding the effect of high-order connections. The three-scientist matrix reads similarly

$$R^{(3)}_{a_1 a_2 a_3} = M^{(3)}_{a_1 a_2 a_3} + \sum_{a_4} M^{(4)}_{a_1 \cdots a_4} + \cdots + \sum_{a_4} \cdots \sum_{a_j < a_{j-1}} M^{(j)}_{a_1 \cdots a_j} + \cdots . \quad (4)$$

This new matrix counts the number of papers co-written by the triplet (a_1, a_2, a_3) and may be represented by a network whose links are triangles relating three authors. The generalization to higher order matrices $\mathbf{R}^{(j)}$ is straightforward, but, as in the case of the BBGKY hierarchy, a truncature of the vector $\mathcal{M}$ must be fixed at some level in order to usefully and compactly describe the system. It is therefore important to point out that the knowledge of $\mathbf{M}^{(2)}$ together with $\mathbf{R}^{(3)}$ is completely sufficient in order to characterize the triangular structure of $\mathcal{M}$. Consequently, in this paper, we stop the reduction procedure at the three-body level and define the triangular projection of $\mathcal{M}$ by the application

$$[M^{(1)}_{a1}, M^{(2)}_{a_1 a_2}, M^{(3)}_{a_1 a_2 a_3}, \ldots, M^{(n_P)}_{a_1 \ldots a_{n_P}}] \rightarrow [M^{(1)}_{a1}, M^{(2)}_{a_1 a_2}, R^{(3)}_{a_1 a_2 a_3}]. \quad (5)$$

The triangular projection is depicted in Fig. 7 and is compared to the usual projection method.

In order to test the relevance of this description, we have measured in the data set the total number of triangles

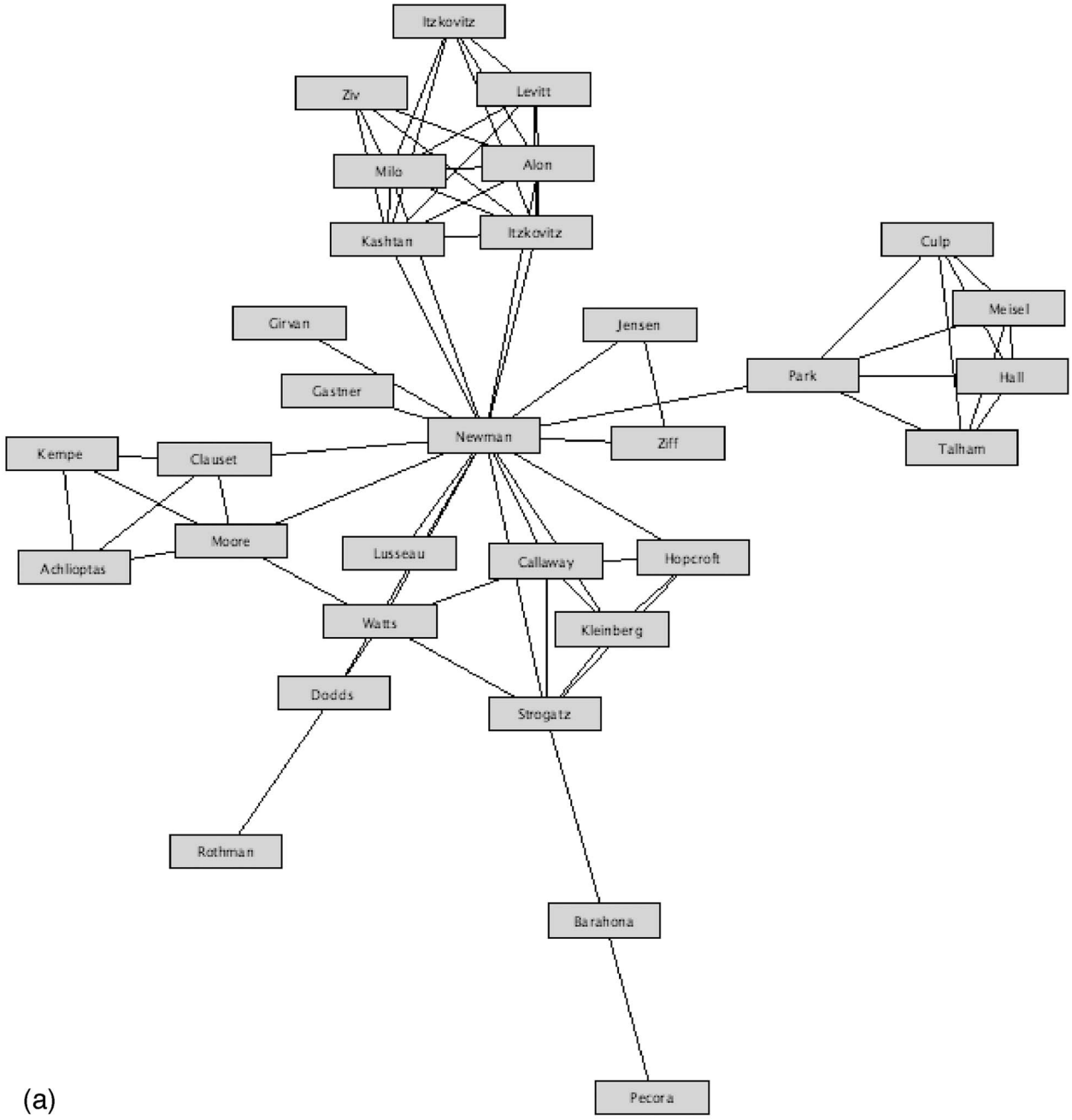


FIG. 7. Three-body projection of the bipartite network. For the sake of clarity, we focus on a small subcluster, centered around the collaborations of Newman. The upper figure is the usual projection method of Fig. 1. The lower figure is the triangular projection (4) of the same bipartite network.

generated by edges. We discriminate two kinds of triangles: those which arise from **one** three-body interaction of $\mathbf{R}^{(3)}$ and those which arise **only** from an interplay of different interactions. There are respectively 5550 and 30 such triangles, namely 99.5% of triangles are of the first kind. This observation by itself therefore justifies the detailed projection method introduced in this section and shows the importance of co-authorship links geometry in the characterization of network structures, precisely the clustering coefficient in the present case.

III. TRIANGULAR ERDÖS-RÉNYI NETWORKS

The empirical results of the previous section have shown the significance of N-body connections in social networks. A more complete framework for networks is therefore required in order to describe correctly the system complexity. In this article, we focus on the most simple generalization, namely a network whose links relate triplets of nodes. To do so, we base our modeling on the Erdös-Rényi uncorrelated random graph [17], i.e., the usual prototype to be compared with more complex random graphs. The usual Erdös-Rényi net-

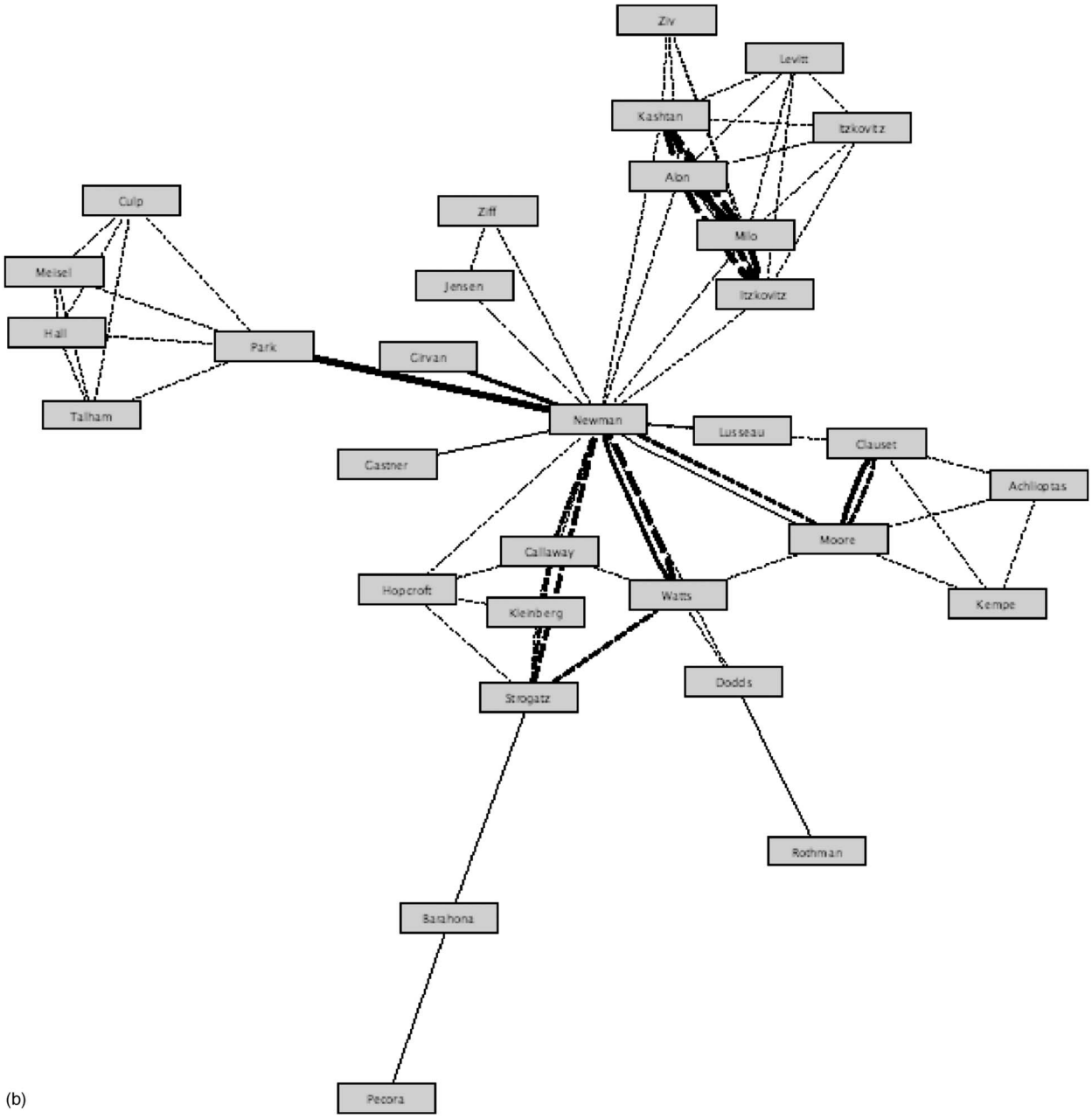
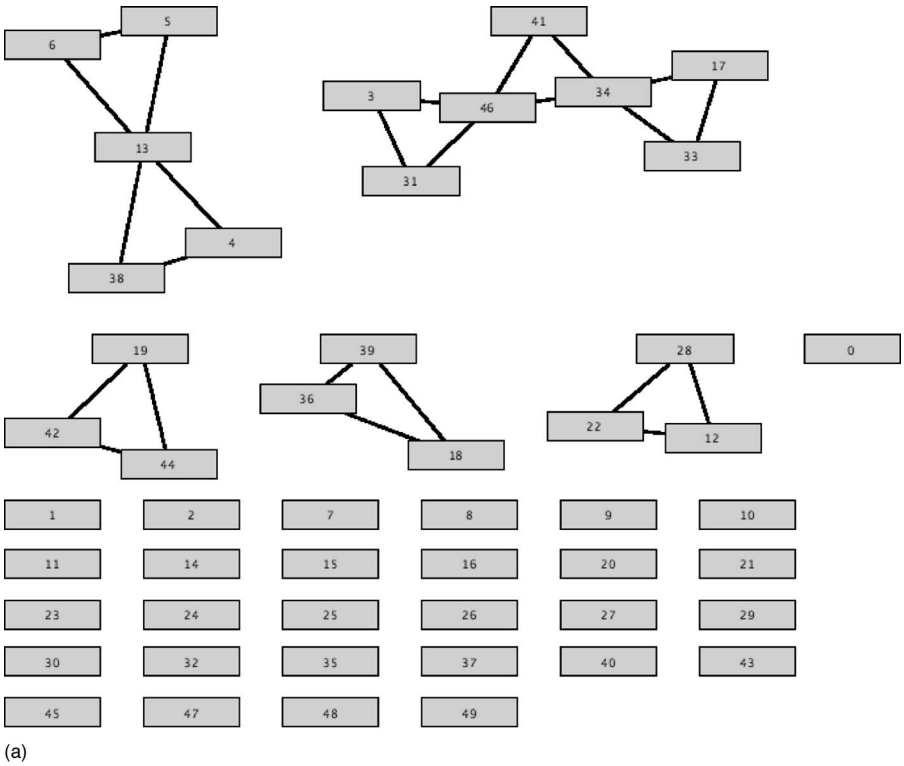


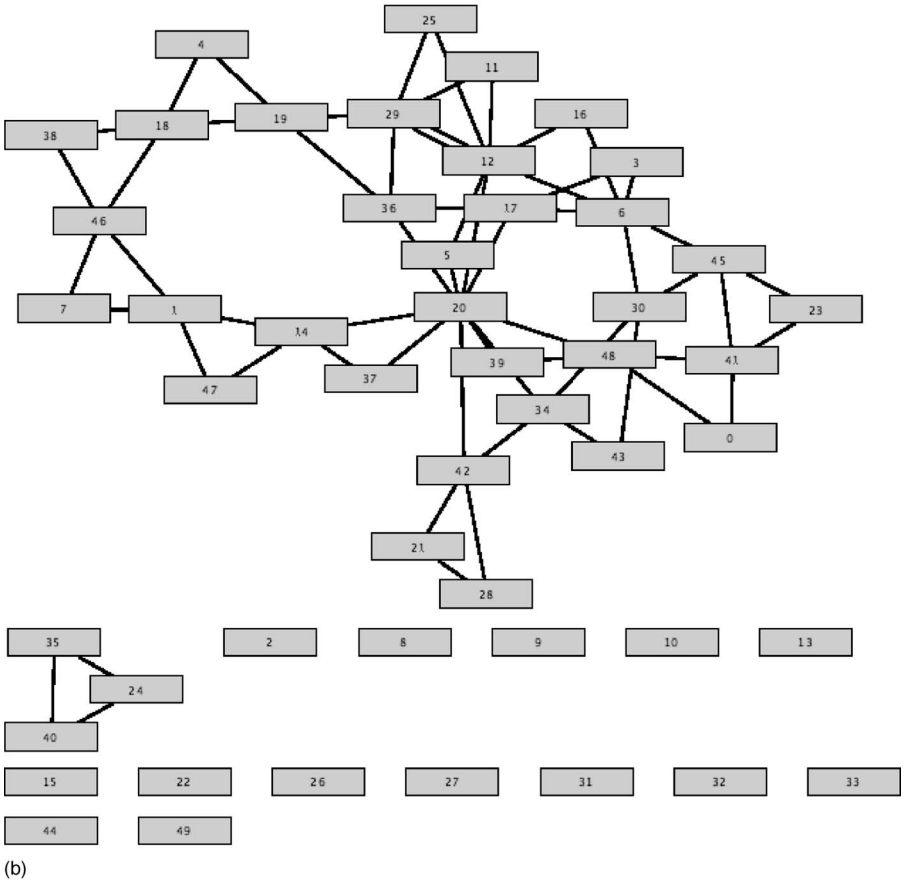
FIG. 7. (*Continued*).

work (ERN) is composed by N_n labeled nodes connected by $N_e^{(2)}$ edges, which are chosen randomly from the $N_n(N_n-1)/2$ possible edges. In this paper, we define the triangular ER network (ERN$^{(3)}$) to be composed by N_n labeled nodes, connected by $N_e^{(3)}$ triangles, which are chosen randomly from the $N_n(N_n-1)(N_n-2)/6$ possible triangles. As a result, connections in the system relate triplets of nodes (a_1,a_2,a_3), and the matrix vector $\mathcal{M}$ reduces to the matrix $\mathbf{M}^{(3)}$. Before going further, let us point out that the clustering coefficient of triangular ER networks is very high by construction, but, contrary to intuition, it is different from 1 in general. For instance, for the two triplets (a_1,a_2,a_3) and (a_1,a_4,a_5), the local clustering coefficient of a_1 is equal to $\frac{1}{3}$.

In this paper, we focus numerically on the percolation transition [18] in ERN$^{(3)}$, i.e., on the appearance of a giant component by increasing the number of links in the system (Fig. 8). This transition is usually associated with dramatic changes in the topological structure, which are crucial to ensure communicability between network nodes, e.g., the spreading of scientific knowledge in the case under study. In the following, we work at a fixed number of nodes and focus on the proportion of nodes in the main cluster as a function of the number of binary links in the system. Moreover, in order to compare results with the usual ERN, we do not count twice redundant links, i.e., couples of authors who interact in different triplets. For instance, the triplet (a_1,a_2,a_3) accounts for three binary links, but (a_1,a_2,a_3) and (a_1,a_2,a_4) account together for five links, so that



(a)



(b)

FIG. 8. Percolation transition in a triangular Erdös-Rényi network (see text for definition) made of 50 nodes, from a dilute phase with small disconnected islands (8 triangles) to a percolated phase with one giant cluster (20 triangles).

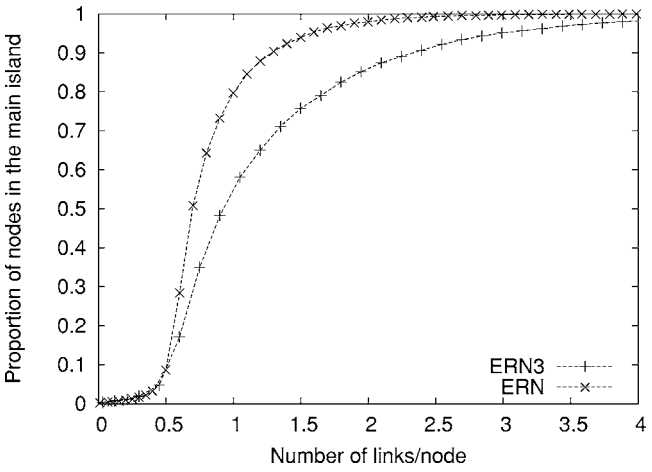


FIG. 9. Proportion of nodes in the main island, as a function of the number of links/node, in the ERN and the ERN$^{(3)}$ model. The networks are composed of 1000 nodes.

$N_e^{(2)} \neq 3N_e^{(3)}$ in general. Let us note, however, that this detailed counting has small effects on the location of the percolation transition. Numerical results are depicted in Fig. 9, where we consider networks with $N_n = 1000$ and where 50 realizations of the simulations have been performed for each value of the number of links/node in order to improve the statistics. Obviously, the triangular structure of interactions displaces the bifurcation point, by requiring more links in order to observe the percolation transition. This feature comes from the triangular structure of connections that restrains the network exploration as compared to random structures. Indeed, three links relate only three nodes in ERN$^{(3)}$, while three links relate *at least* three nodes in ERN (Fig. 10). Finally, let us stress that the same mechanism takes place in systems with high clustering coefficients [19,20].

IV. CONCLUSION

In this paper, we show the importance of *N*-body interactions in co-authorship networks. By focusing on data sets extracted from the arXiv database, we introduce a way to project bipartite networks onto unipartite networks. This approach generalizes usual projection methods by accounting for the complex geometrical figures connecting authors. To do so, we present a simple theoretical framework and define *N*-body reduced and projected networks. The graphical representation of these simplified networks rests on a "shape-based" discrimination of the different co-authorship interactions (for a "color-based" version, see the first author's website [21]) and allows a clear visualization of the different

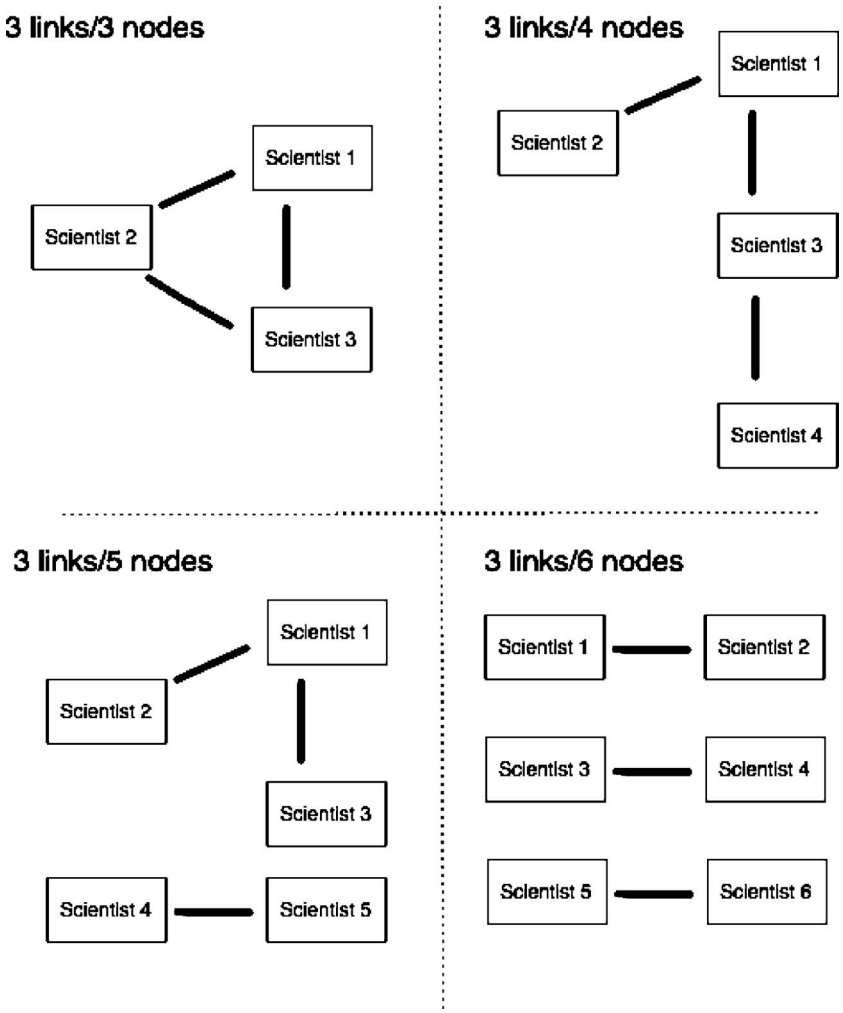


FIG. 10. Different ways to distribute three links in a network, thereby connecting from three to six nodes. Obviously, only the first case (three nodes) occurs in a triangular Erdös-Rényi network.

mechanisms occurring in the system. Finally, we apply the method to some arXiv data subset, thereby showing the importance of such "high-order corrections" in order to characterize the community structure of scientists. The empirical results motivate therefore a better study of networks with complex weighted geometrical links. In the last section, we focus on the simplest case by introducing a triangular random model, $ERN^{(3)}$, and restrict the scope by analyzing the effect of the three-body connection on percolation. A complete study of the topological of $ERN^{(3)}$ as well as its generalization to higher order connections is left for a forthcoming work.

ACKNOWLEDGMENTS

Figures 2 and 5–8 were plotted thanks to the *visone* graphical tools [22]. R.L. is supported by European Commission Project CREEN FP6-2003-NEST-Path-012864. We also thank the COST P10 Physics of Risk European Project.

[1] P. Résibois and M. De Leener, *Classical Kinetic Theory of Fluids* (Wiley, New York, 1977).
[2] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[3] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **31**, 590 (2002).
[4] R. Lambiotte and M. Ausloos, e-print physics/0508234.
[5] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[6] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, Proc. Natl. Acad. Sci. U.S.A. **99**, 2566 (2002).
[7] A. L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **311**, 590 (2002).
[8] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras, Phys. Rev. E **70**, 036106 (2004).
[9] R. Lambiotte and M. Ausloos, e-print physics/0508233.
[10] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[11] D. Fenn, O. Suleman, J. Efstathiou, and N. F. Johnson, e-print physics/0505071.
[12] J. W. Grossman, Congr. Numer. **158**, 202 (2002).
[13] R. D. Mattuck, *A Guide to Feynman Diagrams in the Many-Body Problem* (Dover, New York, 1992).
[14] J. E. Mayer and M. G. Mayer, *Statistical Mechanics* (Wiley, New York, 1940).
[15] J. Berg and M. Lässig, Phys. Rev. Lett. **89**, 228701 (2002).
[16] F. Pütsch, Adv. Complex Syst. **6**, 1 (2003).
[17] P. Erdös and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).
[18] I. Derenyi, G. Palla, and T. Vicsek, Phys. Rev. Lett. **94**, 160202 (2005).
[19] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
[20] R. Lambiotte and M. Ausloos, e-print physics/0508233.
[21] http://www.lambiotte.be/parallel/nBody/nBody.html
[22] http://www.visone.de/