Ströele et al. Journal of the Brazilian Computer Society (2018) 24:6
https://doi.org/10.1186/s13173-018-0070-6

Journal of the
Brazilian Computer Society

RESEARCH Open Access

Rational Erdös number and maximum flow as measurement models for scientific social network analysis

Victor Ströele[1*], Renato Crivano[2], Geraldo Zimbrão[2], Jano M. Souza[2], Fernanda Campos[1], José Maria N. David[1] and Regina Braga[1]

Abstract

In social network analysis, the detection of communities—composed of people with common interests—is a classical problem. Moreover, people can somehow influence any other in the community, i.e., they can spread information among them. In this paper, two models are proposed considering information diffusion strategies and the identification of communities in a scientific social network built through these two model concepts. The maximum flow-based and the Erdös number-based models are proposed as a measurement to weigh all the relationships between elements. A clustering algorithm (*k*-medoids) was used for the identification of communities of closely connected people in order to evaluate the proposed models in a scientific social network. Detailed analysis of the obtained scientific communities was conducted to compare the structure of formed groups and to demonstrate the feasibility of the solution. The results demonstrate the viability and effectiveness of the proposed solution, showing that information reaches elements that are not directly related to the element that produces it.

Keywords: Scientific social network analysis, Information diffusion, Measurement models, Clustering algorithm

Introduction

Web evolution and the increasing availability of data allow researchers to study the ways in which connections between people are established, and how they evolve over time. Social networks emerged to represent people and their relationships, and thereafter, many efforts have been made to analyze these networks, contributing to a better understanding of the social structures [45]. A relationship is defined as a specific contact or connection type between pairs of actors. Relationships may be direct, when an actor provides information and the other receives it directly, or indirect, when information reaches its destination through intermediate actors.

Since actors are directly or indirectly related, social networks play a key role in information diffusion, increasing the spread of new information and different points of view [2]. Facebook, for example, was very powerful in the Arab Spring in 2010 [20] and Twitter in the US presidential elections in 2008 [22]. More recently, these social networks were also powerful in the 2013 Brazilian protests [5, 46]. All these events are highly linked to social networking platforms, which contributed to disseminate information at specific moments.

Considering (i) information diffusion in social networks [17], (ii) some social network analysis studies ([16, 50] show that even two indirectly related people may be influenced by each other; and (iii) in previous studies from our research group in scientific social networks [41, 40], the authors proposed models for scientific social network analysis that include information diffusion strategies, throughout indirect relationships among actors [17, 27].

This paper aims to describe models capable of exploiting information diffusion in scientific social networks, i.e., how information propagates through people who have different potentials for spreading information (different tie weight). Our first objective is to build a scientific social network which allows developing studies to consider information diffusion in the network itself. It aims to analyze social network considering that all elements can somehow influence one another. *Influence* is related to the individual's ability

* Correspondence: victor.stroele@ice.ufjf.br
[1]Postgraduate Program in Computer Science/PPGCC, Federal University of Juiz de Fora (UFJF), Juiz de Fora, MG 36036-900, Brazil
Full list of author information is available at the end of the article

to affect other people in a social network community. In general, in a scientific context, influential researchers tend to create or strengthen ties, as they can propagate their knowledge, reaching a larger number of people. Our second objective, with the construction of a scientific social network, is to consider the proposed models and the identification of scientific communities using the clustering algorithm *k*-medoids.

As contributions, the authors can highlight the development of two evaluation measures to analyze how information diffusion occurs between pairs of researchers, namely *maximum flow* and *weighted resistance distance*. These measures consider that information travels throughout all possible paths between two researchers, like cascade analysis in information diffusion studies [27]. As a result, the influence that researchers have over each other is calculated.

Another contribution is the modeling of a scientific social network using those two evaluation measures and the use of a clustering algorithm to find scientific communities. The resulting groups are compared in order to validate the *weighted resistance distance* and to evaluate which measure produces the most homogeneous groups. In homogeneous groups, people of similar scientific interest are engaged. In this paper, maximum flow and weighted resistance distance are used as similarity measures.

The paper has five sections, besides its introduction. The "Scientific social networks and information diffusion" section presents the background. The "Measurement models to define tie weight" section discusses the proposed measurement models. The "Case study" section introduces a case study and analyzes the feasibility of the proposed models. The "Final remarks" section presents the conclusions and future work.

Scientific social networks and information diffusion

The growth of social networks is due to the Web evolution, and researchers are dealing with its challenges. Some works include mining newsgroups [50], predicting the popularity of links [35, 48, 52] and videos [8], and discovering useful information and patterns from data streams in sensor networks [25].

In addition, structure analysis of social networks helps in identifying critical regions and people [42]. Identifying them is a complex but essential task, as these elements are responsible for collaboration among all other peers in the network, and they are potential elements in information diffusion. Some studies have been conducted to gather information from social networks since appropriate aggregation of multiple social networks could offer a better opportunity for deep user understanding [38, 39].

Scientific social networks are specific types of social networks where two scientists are considered connected if they have co-authored a paper [23, 32]. In the real world, the nodes of a social network tend to be tightly connected, forming groups of people who work together, named *communities*. This effect also occurs in scientific social networks where researchers who have high concentration of common edges define *scientific communities* [33, 34]. Recovering these communities has been a challenging task, and many studies on social network analysis have been developed in order to identify them [48, 50].

In general, social network analysis involves the definition of a model to represent the relationship weights among researchers. According to [2], the higher the edge width/weight which defines the relationship between two individuals, the higher the influence that they exert on each other. Studies have shown that information spreads more easily between elements that have tighter relationships. However, looser relationships also play a key role in information diffusion [16, 24].

By means of a network topology, the behavior of its elements is used with several objectives. There are studies that attempted to model developers' networks to explore the coordination performance of open-source software (OSS) [19] or the behavior of developers in OSS communities [21]. Both apply information diffusion measures to help understand those complex networks. This kind of analysis has been carried out in the health scenario, for instance, for the identification of influential members [37] or communities [11].

This study differs from others in that it addresses (i) the identification of relationships and their quantification and its use by a clustering algorithm to identify *scientific communities*, (ii) considering *information diffusion* concepts, (iii) in relationship weight definition of *scientific social networks* in terms of direct and indirect influence, as well as tight and loose existing connections.

Measurement models to define tie weight

In some cases, social network analysis is carried out by a quantitative evaluation that indicates the link weight between related elements [16, 24, 44]. Based on such quantitative analysis, the weight of links between authors is defined, allowing not only the use of other metrics (degree, closeness, betweenness, PageRank, etc.) [43] but also the discovery of communities [50, 32, 14, 34, 49].

However, quantitative analysis alone is not enough to evaluate the effect that indirectly related elements produced on each other. Many studies are being developed in order to evaluate information diffusion [2, 17]. As aforementioned, in these studies, all the elements of the social network have some influence on information diffusion.

Previous studies have shown that there is influence between peer researchers even though they are not directly related [41]. This indirect influence can be calculated in

different ways. In this study, a comparison between the maximum flow-based model and the resistance distance-based model was made to evaluate the communication potential between nodes in a scientific social network.

Maximum flow-based model

The modeling process of the scientific social network was divided into three steps as follows: number of common relationships between researchers, age of the relationships that link these researchers, and loss of knowledge when the relationship between researchers is indirect [41].

Common relationships between researchers

To generalize this model, Eq. (1) considers that there may be different types of relationships between researchers, such as project participation, co-authored publications, advisory work, and technical productions. This study took into account only the co-authorship relationship since the database used has no other relationship type.

$$\mathrm{TR}(\alpha)^{a,b} = \sum_{i=1}^{t}\left[\alpha_i\left(\frac{\mathrm{CR}_i^{a,b}}{P_i^a + P_i^b}\right)\right], \quad (1)$$

where $\mathrm{TR}(\alpha)^{a,\,b}$ means the weight of relationships between researchers a and b weighted by the number of relationships of type i of these researchers, preventing relationships with the same frequency from having the same weight. In this equation, α_i is the weight of relationship type i, t is the total number of relationship types, $\mathrm{CR}_i^{a,b}$ is the number of common relationships of type i between researchers a and b, and P_i^a and P_i^b mean the total relationships of type i of researchers a and b, respectively. The result of Eq. (1) is normalized by the natural logarithm and min-max normalization (Eq. (2)), and a connected graph is obtained.

$$\mathrm{TR}(\alpha)_N^{a,b} = \frac{\ln\left(\mathrm{TR}(\alpha)^{a,b}\right) - \min(\ln(\mathrm{TR}(\alpha)))}{\max(\ln(\mathrm{TR}(\alpha))) - \min(\ln(\mathrm{TR}(\alpha)))} \quad (2)$$

where $\mathrm{TR}(\alpha)$ is a matrix with all relationship values of each researcher pair.

Relationship's age

Another factor considered in a relationship is its age. Relationship's age is useful to indicate whether the relationship reflects a link of present elements or if it is just a connection that existed in the past. In this sense, a penalty function was added to Eq. (1) considering the relationship in terms of years, resulting in Eq. (3).

$$\mathrm{TR}(\alpha,\rho)^{a,b} = \sum_{i=1}^{t}\sum_{j=0}^{d-1}\left[\rho(\mathrm{RY}+j)\alpha_i\left(\frac{\mathrm{CR}_i^{a,b}}{P_i^a + P_i^b}\right)\right], \quad (3)$$

where $\mathrm{TR}(\alpha,\rho)^{a,\,b}$ is the weight of the links between researchers a and b considering the base year of the relationship; d is the relationship duration in years. Exact relationships (e.g., co-authorship) have lasted 1 year, $d = 1$ and it contributes only once to the relationship weight between the researchers. On the other hand, ongoing relationships have equal duration considering the number of years in effect (e.g., projects), so if the authors worked, for example, in the same project for 3 years, then $d = 3$, and it will be considered more than once in the relationship weight. RY is the relationship year, i.e., the year in which the two researchers worked together or when they started to work (co-authored, participated in the same project, and others); t is the total number of link types, α_i is the relationship weight i, and ρ is as defined in Eq. (4).

$$\rho(y) = e^{1/(\mathrm{BY}-y+1)} \quad (4)$$

where BY is the base year, and y is the year to be considered in the penalty, as previously defined. Figure 1 represents a graph of the penalty function.

After Eq. (3), an $M \times M$ symmetric weighted matrix is obtained representing the relationship weight between each pair of researchers, where M is the number of researchers. As relationship weight represents the similarity between researchers, this matrix was named similarity matrix, as represented in Eq. (5).

$$\mathrm{SM} = \begin{cases} \mathrm{TR}(\alpha,\rho)^{a,b} & \text{if researcher 'a' links to researcher 'b'} \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Content loss in long relationships

Communication potential (relation weight) between two nodes is calculated using the maximum flow algorithm, which evaluates all possible paths between two researchers and consolidates the largest flow of knowledge that can be transmitted. This strategy is interesting since it considers all communication possibilities; however, it disregards the number of intermediate elements to define the relationship weight. Figure 2 illustrates this situation, in which A and B represent two elements of distinct groups, where it is necessary to bind element C to one of these two groups. This simple graph shows that there are two ways from C to A and B. Element C is directly linked to A, and C is linked to B through D and E. Using the maximum flow (MF) calculation, obtained as $\mathrm{MF}_{\mathrm{CA}} = \mathrm{MF}_{\mathrm{CB}} = 4$, i.e., 2 from the edge between A and C, and from the other path allows passing only 2 through the edge that links

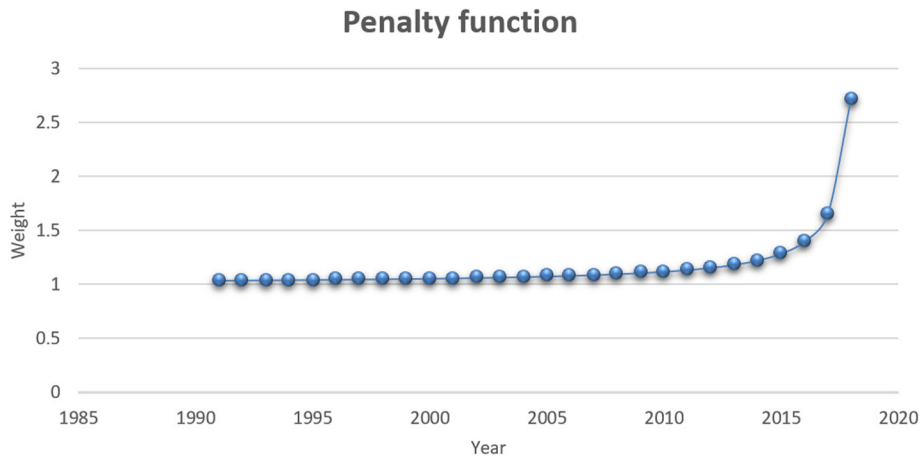

Fig. 1 Graph of Eq. 4

E and B. So, maximum flow between C and A is equal to the maximum flow between C and B. However, C is directly related to A, indicating a better communication between them.

Formally, the network maximum flow problem can be interpreted as a graph flow problem. The social graph with flow is represented by G = (*X*, *U*, **f**), in which **f** is a vector of dimension $m + 1$ and can be written in the form $\mathbf{f} = (f_0, f_1, \ldots, f_m)$.

Vector *f* is the flow in graph G, and each of the components indicates the value of the flow between the elements of G. The social graph is represented in this study by a non-oriented graph, so the maximum flow coming out of x_i to x_j is equal to the maximum flow of x_j to x_i, for all $x_i, x_j \in X$.

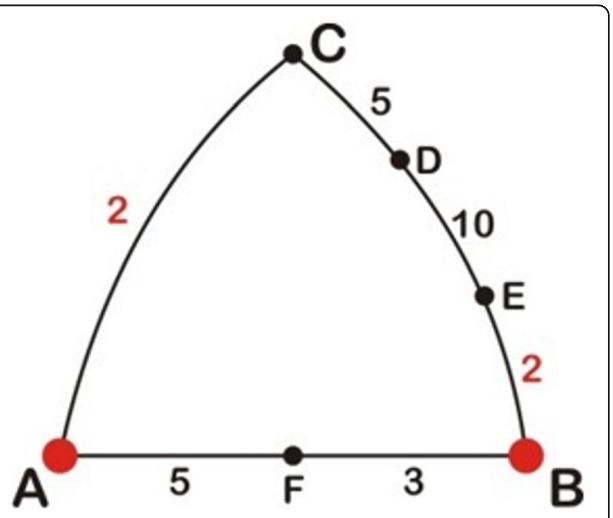

Fig. 2 Maximum flow analysis

Consider G as the social graph represented by SM defined in Eq. (5), *X* is the set of researchers, *U* is the set of relationships between these researchers, and *f* is the set of maximum flows between each pair of researchers. Thus, for any $x_i, x_j \in X$, the maximum flow between these two researchers will be equal to $f_w = \sum_{v:(i,v)\in E} \overline{f}_{iv}$, where *i* is the source node, $v \in X\backslash\{i, j\}$, $\overline{f}_{iv}$ represents the amount of flow passing from source *i* to node *v* and $0 \le w \le m$.

The maximum flow was calculated using the Edmonds-Karp algorithm [13]. This algorithm is a variation of the Ford-Fulkerson maximum flow algorithm [15]. The main difference between these two approaches is that the Edmonds-Karp algorithm targets the maximum flow between two elements considering the shortest paths. Thus, it is guaranteed that the algorithm will converge on a finite number of iterations, even for non-oriented graphs.

As the Edmonds-Karp algorithm considers the shortest paths, an adaptation was made in the algorithm so that the maximum flow was penalized with a percentage according to the path length.

In this context, it was considered that information is received by the researcher with a knowledge loss of *N*%, where *N* is the number of intermediate nodes between source and receiver. Thus, assuming that the maximum knowledge flow between A and B is $\text{MaxFlow}^{a,b}$, the new relationship weight between them is given by Eq. (6).

$$\text{TR}(\alpha, \rho, N) = \text{MaxFlow}^{a,b} \times (1 - N/100) \tag{6}$$

At the end of the multi-relational scientific social network modeling process, the *maximum flow matrix* with resistance is achieved, which will be used by the clustering algorithm. Algorithm 1 shows the step-by-step of the maximum flow-based model.

Algorithm 1 Maximum Flow-based Model

Input: X: Set of authors; α: weight of relationships set; n: number of intermediate nodes
Output: MF: Max-Flow Matrix returned by Algorithm

```
0:  Begin procedure
1:  MF ← 0;
2:  SM ← 0;                                          // Similarity Matrix
3:  pubDate ← null;                                  //Publication date
4:  ρ ← 0;                                           // Penalty value
5:      for each integer i in X do
6:          P ← getPublicationsOf(Xi);               // Return a set of publications of researcher Xi
7:          for each integer j in P do
8:              Y ← getAuthorsOf(Pj);                // Return a set of authors of publication Pi
9:              pubDate ← getDateOf(Pj);
10:             ρ ← e^(1/(CURRENT_DATE − pubDate+1)) ;
11:             for each integer k in Y do
12:                 a ← getIndexOf(Xi);              // Get researcher Xi corresponding index in the SM
13:                 b ← getIndexOf(Yi);              // Get researcher Yi corresponding index in the SM
14:                 SM(a, b) ← SM(a, b) + TR(α, ρ)^(Xi,Yj) ;
15:             end for
16:         end for
17:     end for
18: MF ← EdmonsKarpAdapted(SM, n);
19: end procedure
```

Erdös number-based model

Michael Barr proposed the rational Erdös number model (RENM) as a distance measure (Michael [31]). The idea was to consider a social network of researchers who have published together, as being akin to an electric circuit of resistances. To each person, a rational number is assigned representing the total resistance from that node to the center of the network—in this case, the famous mathematician Paul Erdös. If two researchers coauthored one paper, there would be a 1-Ω resistor between them. If they had two coauthored papers, then there would be two 1-Ω resistors connected in parallel, which, by the laws of electricity, are equivalent to a 0.5-Ω resistance. In Michael Barr's proposal, a paper that is written by more than two authors should be represented by a new node in the graph, and $N/4$ Ω resistors should be placed connecting that node to each one of the N authors involved.

The RENM was intended as a distance measure to Paul Erdös, who coauthored many articles and is somehow related to most of the mathematical community. But the idea of representing a social network as an electric circuit of resistances can be applied in other realities, and it is possible to calculate not only distances to a center but also distances between each pair of people. If the distance between each node of a graph is to be calculated, then an algorithm can be used to identify groups of tightly connected nodes and to identify elements that have the shortest distance to all other elements of their groups, i.e., identify the medoids.

Weighted resistance distance

Most of the complexity in this model is related to the calculation of the total resistance between any two arbitrary nodes of an electric circuit of resistors. If our problem is simpler, and all resistors could have a nominal value of 1 Ω (as in Fig. 3), then a better-studied situation will be dealing with the calculation of resistance distance in the graphs.

To solve the resistance distance, there is a method named the determinantal formula, which was proposed by Bapat [4].

To understand the determinantal formula, first, it is necessary to understand how the Laplacian matrix, denoted by $L = L(G)$, of a graph is formed. For each of the nodes i and j, $L_{ij} = -1$ if i and j are connected or 0 otherwise. If $i = j$, L_{ij} is the number of first neighbors of i. For the circuit above, L would be constructed as follows:

$$L(G) = \begin{Bmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 4 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 & 3 \end{Bmatrix}$$

The determinantal formula states that the resistance distance between two nodes can be obtained by (i)

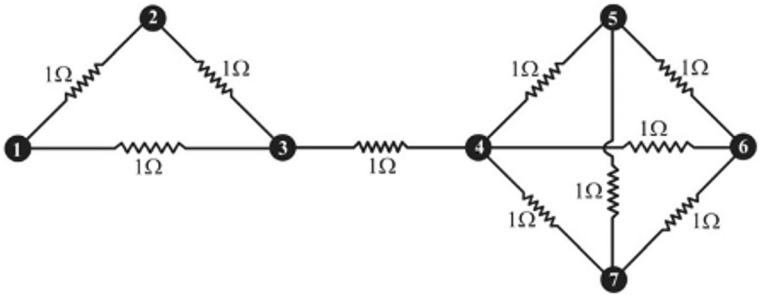

Fig. 3 A simple graph where all edges represent 1 Ω resistors

building the Laplacian matrix of the graph, (ii) removing both the lines and the columns related to these two nodes, (iii) calculating the determinant, and (iv) dividing it by the number of spanning trees (Eq. (7)).

$$r_{ij} = \frac{\det(L(i,j))}{t(G)} \tag{7}$$

Since the number of spanning trees $t(G)$ can be obtained by omitting the i^{th} line and the i^{th} column of L and calculating the determinant, another view of the determinantal formula can be seen in Eq. (8).

$$r_{ij} = \frac{\det(L(i,j))}{\det(L(i))} \tag{8}$$

Unfortunately, our preliminary studies showed that if all resistances are equal, no good groups are then identified. Thus, it is necessary to find an appropriate mathematical method to solve the circuit and obtain the effective resistance between each two authors.

The standard method to calculate the resistance between two points in an electrical network is to solve all equations provided by the first and second laws of Kirchhoff, and also by the Ohm's law [30]. Although being feasible, it would lead to very complex and tedious calculations. The formula for the determinant of the resistance matrix was derived by Bapat [3] and is shown to reduce to the formula obtained by Xiao and Gutman in the unweighted case [47]. Consequently, it was possible to calculate the effective resistance even if the circuit had different valued resistors.

Multiple terminals delta-star transformation

According to Michael Barr's original proposal, new nodes should be created to represent publications involving more than two people. Even though it is possible, the creation of more nodes would imply a bigger matrix, increasing the processing time of the algorithm that calculates the resistance distances. Therefore, this study worked with an alternative approach, which produces the same result, but does not raise the number of nodes.

The approach consists of a generalization of the well-known delta-star transformation on resistive circuits [28], and it was only possible because in this case, all resistances between the center point (publication) and the terminals (authors) are equal. So, when more than two people coauthored a publication, the star-like array of $N/4$ Ω resistors was replaced by a network of same valued resistors connecting each author to every other author, just like in a complete graph.

To calculate the precise value of the resistors that should be used, the determinantal formula for resistance was applied together with Cayley' formula [7, 9] for the number of spanning trees in complete graphs: $t(K_n) = n^{n-2}$, resulting in Eq. (9), where n is the number of researchers.

$$r_n = \frac{n}{2} \times \frac{n^{n-2}}{\det(L(i,j))} \tag{9}$$

The model in action

A scenario consisting of three publications was considered to validate the proposal. The first one was coauthored by three researchers named 1, 2, and 3; the second one by researchers 3 and 4; and the last one by researchers 5, 6, 7, and 8. The application of multiple terminals delta-star transformation formula, described above, leads to the electric circuit of Fig. 4.

The weighted Laplacian matrix of conductance for this RENM network can be written as follows:

$$L(G) = \begin{Bmatrix} 0.89 & -0.44 & -0.44 & 0 & 0 & 0 & 0 \\ -0.44 & 0.89 & -0.44 & 0 & 0 & 0 & 0 \\ -0.44 & -0.44 & 1.89 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1.75 & -0.25 & -0.25 & -0.25 \\ 0 & 0 & 0 & -0.25 & 0.75 & -0.25 & -0.25 \\ 0 & 0 & 0 & -0.25 & -0.25 & 0.75 & -0.25 \\ 0 & 0 & 0 & -0.25 & -0.25 & -0.25 & 0.75 \end{Bmatrix}$$

and by applying the determinantal formula to compute the resistances between each two authors, the following values can be found:

$$r(G) = \begin{Bmatrix} 0 & 1.5 & 1.5 & 2.5 & 4.5 & 4.5 & 4.5 \\ & 0 & 1.5 & 2.5 & 4.5 & 4.5 & 4.5 \\ & & 0 & 1 & 3 & 3 & 3 \\ & & & 0 & 2 & 2 & 2 \\ & & & & 0 & 2 & 2 \\ & & & & & 0 & 2 \\ & & & & & & 0 \end{Bmatrix}$$

After obtaining the resistances between all pairs of researchers, we can run the k-medoids algorithm [18, 26]

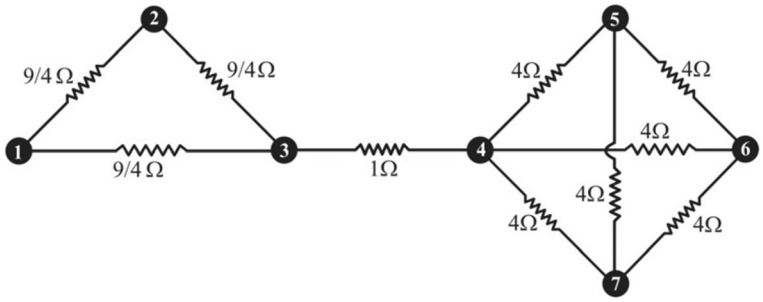

Fig. 4 A social network compliant with the RENM paradigm

using the resistances produced by the previous calculations as distance measures between each author in the social network. For the above example, and considering $K = 2$, the first group was formed by elements 1, 2, 3, and 4 and the second one by 5, 6, and 7, showing that the model is, in some way, feasible. Algorithm 2 shows the step-by-step of the Erdös number-based model.

Case study

The case study was conducted in order to evaluate the measurement models proposed to define tie weight. The scope of the evaluation, based on the GQM method [29], was described as follows: "*To analyze* the scientific communities generated by clustering techniques and/or information diffusion in social networks *for the purpose*

Algorithm 2 Erdös Number-based Model

Input: P: Set of Publications
Output: RM: Resistances Matrix

Begin procedure
RM ← 0; *resistors* ← 0;
circuit ← 0; *// Social network compliant with the RENM paradigm*
localL ← 0; *// Laplacian Matrix of complete graph (star-like)*
L ← 0; *// Weighted Laplacian matrix of conductance for the RENM network*
for each *integer i* **in** *P* **do**
    authors ← getCoauthorsOf(P_i) *// Returns a set of authors of publication P_i*
    if |*authors*| > 2 **then**
        n ← |*authors*|;
        localL ← getLaplacianMatrix(*n*); *// Builds Laplacian Matrix for a complete graph with n nodes*
        $resistors \leftarrow \frac{n}{2} \times \frac{n^{n-2}}{\det\left(L(i,j)\right)}$;
    else
        resistors ← 1;
    end if
    addResistors(*authors*, *resistors*, *circuit*); *// Adds the authors and the resistors among then in the circuit (RENM network)*
end for
L ← buildWeightedLaplacian(*circuit*); *// Builds the weighted Laplacian matrix of conductance for the RENM network*
nAuthors ← |L(i)|;
for each *integer* $0 \le i < nAuthors$ **do**
    for each *integer* $i < j < nAuthors$ **do**
        $RM(i,j) \leftarrow \frac{\det\left(L(i,j)\right)}{\det\left(L(i)\right)}$;
    end for
end for
end procedure

of evaluating scientific communities homogeneity and compare the two proposed approaches *in relation to* information diffusion *from the point of view* of the researchers *in the context of* scientific communities obtained based on researchers' information diffusion potential in a scientific social network."

Based on the scope of the case study, the main research question and three secondary ones were defined:

- How are scientific communities organized considering individual influences and measurement models that quantify information diffusion among researchers from a scientific social network?
- RQ1: Which information diffusion approach produces more homogeneous scientific communities?
- RQ2: Does the use of cluster analysis retrieve real scientific communities considering the activities developed by the researchers?
- RQ3: Are researchers from the same scientific community connected through direct or indirect relationships?

In view of the above research questions, the case study was a suitable choice as a research method, considering that a contemporary phenomenon was evaluated, in its "real-world context," according to Yin [51].

Dataset and case study process

Data for the construction of scientific social network were selected from DBLP,[1] one of the databases commonly used in scientific studies on social networks [10, 50]. For this case study, the data of five out of eight high-quality Brazilian institutions were extracted (COPPE/UFRJ, PUC-RJ, UFPE, UFRGS, and UFMG). Altogether, 169 researchers were analyzed from the area of Computer Science.

Figure 5 shows the DBLP data used in this study in the form of a social network. The nodes were highlighted according to its degree. As can be seen, there are many relationships between the pairs of researchers, indicating that these researchers have co-authored more than once.

In summary, the network contains the following: number of relationships, 1401; clustering coeficient, 0.325; and network density, 0.035. In order to explore some complex network metrics, we plotted graphs. Figure 6 shows the betweenness centrality distribution of the scientific network. The betweenness centrality of a node reflects the amount of control that this node exerts over the interactions of other nodes in the network [45]. This measure favors nodes that join communities, rather than nodes that lie inside a community. As can be seen in the network of Fig. 5, only three nodes have more control over other nodes.

Figure 7 shows the closeness centrality distribution of the scientific network. Closeness centrality is the measure

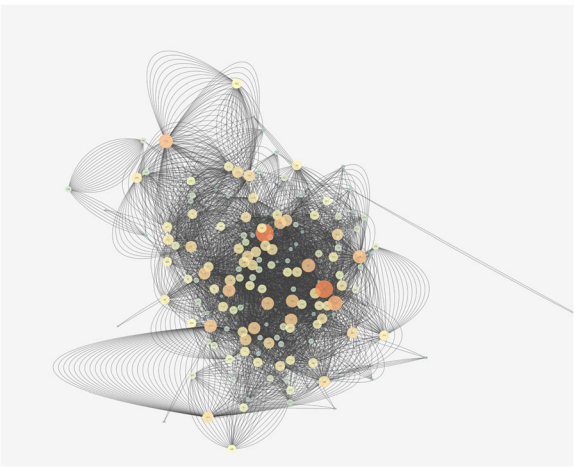

Fig. 5 DBLP scientific social network

of how fast information spreads from a given node to other reachable nodes in the network [45]. We can observe that, excepted by the two nodes with high degree, there is no other node that stands out from the others considering this metric.

The neighborhood connectivity distribution gives the average of the neighborhood connectivities of all nodes n with k neighbors for $k = 0,1,\ldots$. Figure 8 shows the neighborhood connectivity distribution for the network presented. As the neighborhood connectivity distribution is an increasing function in k, edges between highly connected nodes prevail in the network, i.e., nodes with more neighbors tend to have these neighbors more connected. On the other hand, nodes with few neighbors tend to have their neighbors less connected.

This case study follows the process illustrated in Fig. 9 with the following steps:

(1) Data extraction from DBLP and construction of the social graph where nodes represent researchers and edges represent their co-authoring relationships. This social graph was used by both proposed models and to compare them.
(2) After construction of this social graph, the relationship weight (tie weight) between each pair of researchers was defined. The weight was defined by the two measurement models, previously described.
(3) As a result of the application of these models, two matrices representing the communication potential between pairs of researchers were obtained.
(4) In order to identify the scientific communities, the k-medoids clustering algorithm was applied to each of these matrices.
(5) The obtained measurements were compared based on the quality of the clusters generated by each of them.

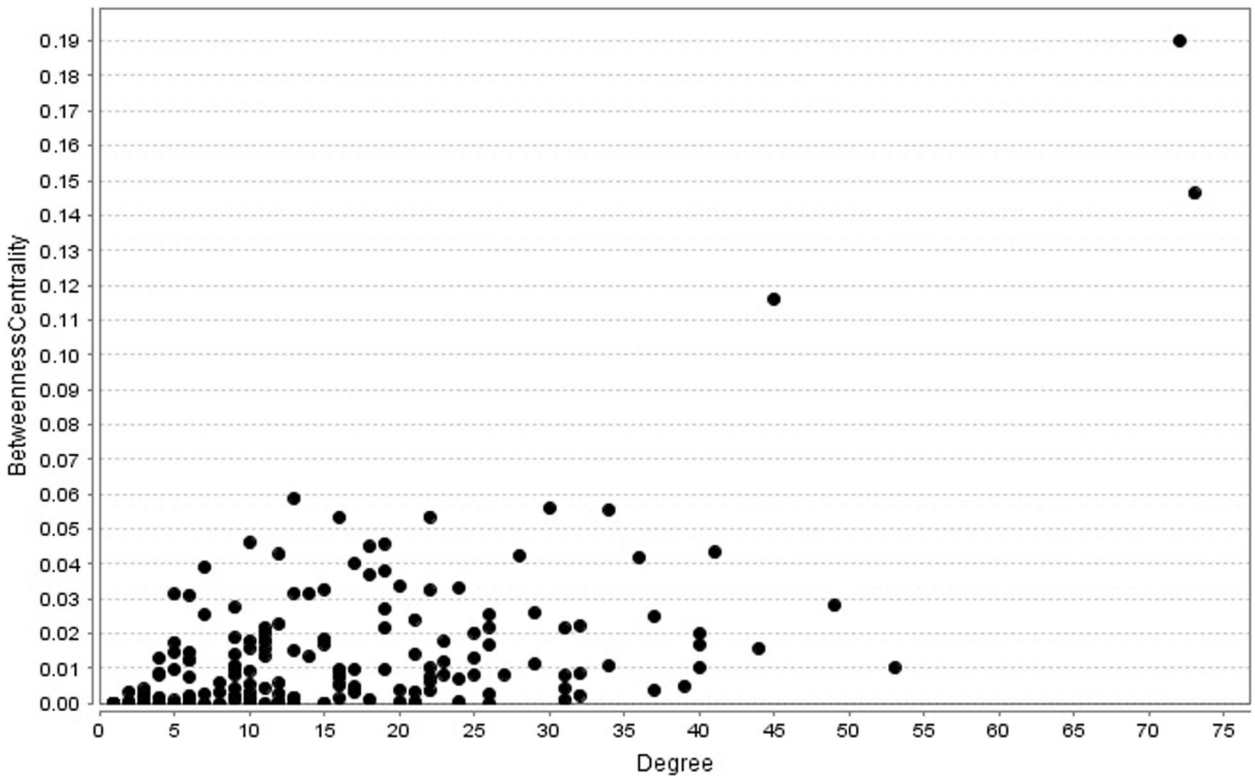

Fig. 6 Betweenness centrality of the network

k-medoids

The clustering algorithm developed aims to group researchers who have the greatest potential of communication with each other. To assist the development basis of the algorithm, the *k*-medoids algorithm was used [18].

As in the *k*-medoids algorithm, the proposed algorithm defined that *k* elements randomly represented the medoids. In the second step, each element of the data set is associated with the group (medoid) in which this element has the greatest potential for communication. In the third step, the medoids of each group are set once again, and the elements are regrouped.

The definition of medoids is based on the internal communication of each researcher group. The relationship weight of each researcher internal to the group is added, and the researcher who has the largest sum is considered the medoid of the group. After defining the new medoids, the algorithm comes back to the second step until there are no changes in the structure of the groups.

One of the biggest difficulties of some clustering techniques, including *k*-medoids, is in defining the ideal number of groups. Cluster analysis aims to identify homogeneous groups so that the sum of differences within the group (intragroup) is minimized, and the sum of differences among groups (intergroup) is maximized [1]. The groups are validated by evaluating which set of groups has the best grouping structure.

There are several techniques that can be used to assist in defining this number, such as PBM index, intragroup distance, and intergroup distance [6, 36].

The sum of the intragroup differences is a good measure to evaluate the homogeneity of the obtained groups. This measure assesses the position of objects in the variables space within their respective groups, so it will indicate whether the objects of the same group are close to the medoid. This sum is defined by Eq. (10), where k is the number of groups, m is the number of elements in the group i, x_j is an element of group i, and $\overline{x}_i$ is the medoids of group i. The value of k, which produces the smallest sum of intragroup distances, indicates the ideal number of groups to the problem that is being solved.

$$\text{Intragroup} = \sum_{i=0}^{k} \sum_{j=0}^{m} \left\| x_j - \overline{x}_i \right\|. \tag{10}$$

Another inherent difficulty in the *k*-medoids algorithm is the initial setting of medoids because the result of the clustering process depends on the initial selection of these elements. Therefore, to select the best group, it is necessary to define the number of groups and the best

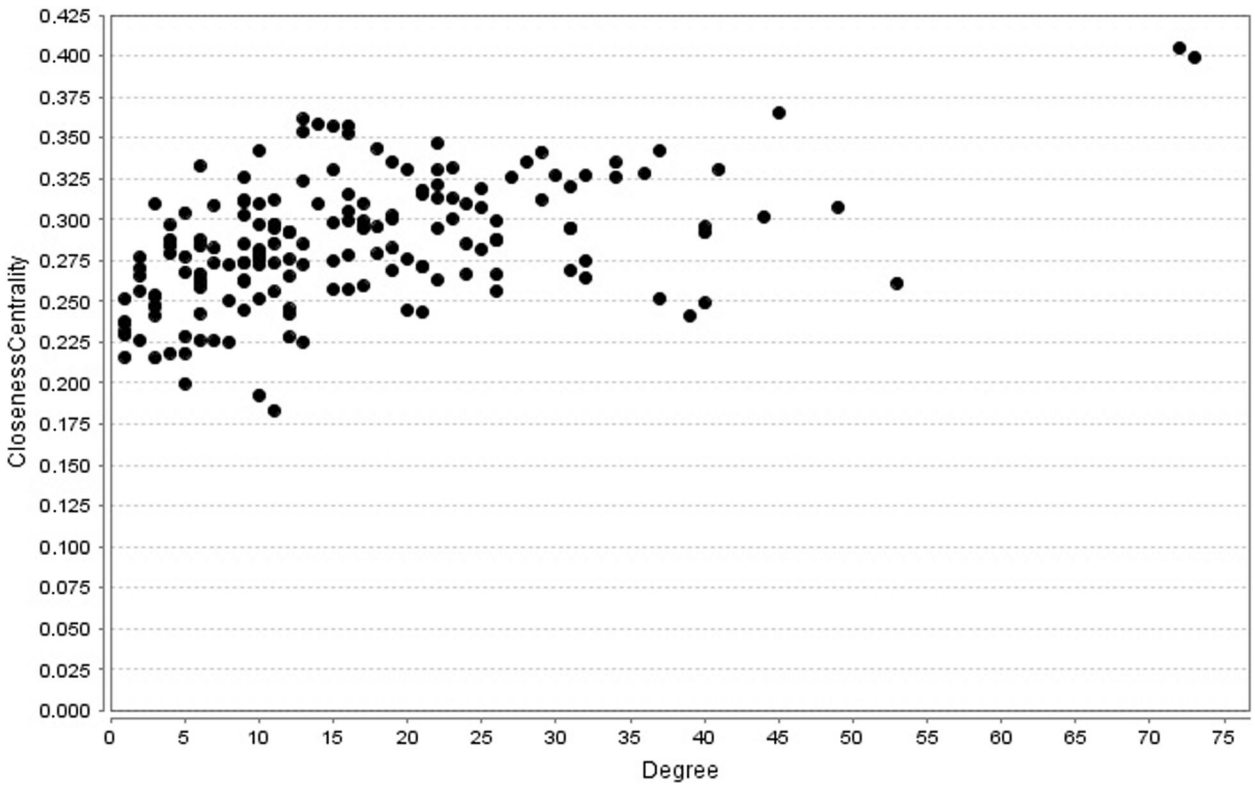

Fig. 7 Closeness centrality

set of elements that compose the initial medoids. In the next section, the details for setting these parameters and the results of clustering algorithm for the two proposed measurement models will be presented.

Experiments and results analysis

In this section, many experiments were conducted to analyze, compare, and evaluate the effectiveness of the two proposed models and answer the research questions.

Number of groups

To answer RQ1, it was necessary to set the number of groups and evaluate the best clustering, so the *k*-medoids algorithm was applied to the data generated by the two proposed measurement models. The value of *k* was evaluated in the range between 20 and 80 groups, and the values of seeds for random definition of medoids ranged from 1 to 199.

At this stage, the internal costs of the groups were analyzed for each seed. However, the sum of the internal maximum flow and the sum of the internal resistance (models proposed in the "Measurement models to define tie weight" section) produce values quantitatively and semantically different. Thus, it was necessary to define a criterion to calculate the internal distance of the groups so that the two methods could be compared.

As illustrated in Fig. 9, both models are based on a social graph where the relationship weight is defined by the scientific work done by them together. Since this graph is the same for both models, the intragroup distance (Eq. (10)) will be calculated based on it. It is worth noting that into the obtained groups may have elements that are not directly related to the medoid, i.e., there are elements that relate indirectly to the medoid. Thus, the distance between an element and the related medoid was calculated based on Dijkstra's algorithm [12], which calculates the shortest path between nodes in a graph.

Dijkstra's algorithm was used by the groups obtained for each seed in order to calculate the intragroup distance. As a result, for each seed, the internal average distances of the groups and the variation of this average were calculated, as can be seen in Fig. 10. The values in the graphs are ordered from smallest to largest average, i.e., from best to worst seed. As can be seen in Fig. 10a, the maximum flow-based model has the best result using seed 100, because this seed had the lowest intragroup average distances. On the other hand, seed 3 showed best results for the Erdös model (Fig. 10b). The seeds that produced more homogeneous groups were selected for a more detailed study of the intragroup distances variation of the groups.

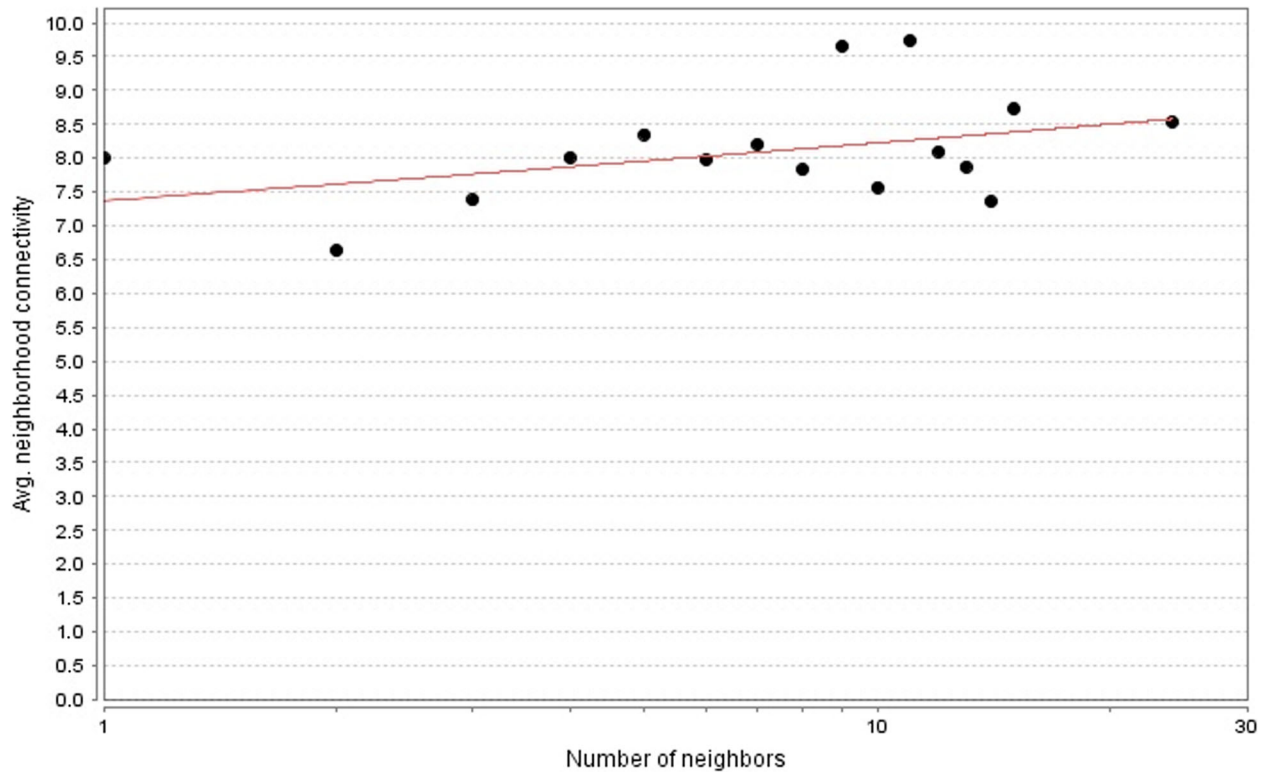

Fig. 8 Neighborhood connectivity distribution of the network, with a fitted line

With the aim of determining the best k value for each of the proposed models, a more detailed analysis was carried out considering seeds 100 and 3, as shown in Fig. 11. Considering it, $k = 43$ was chosen because there was a decrease in the intragroup distance value in this partitioning and the values decreased slowly after that. So, for the maximum flow-based model, it was determined that the optimal number of groups is 43 and the best seed is 100. The same analysis was conducted, as illustrated in Fig. 11, and $k = 47$ was chosen for the same reasons

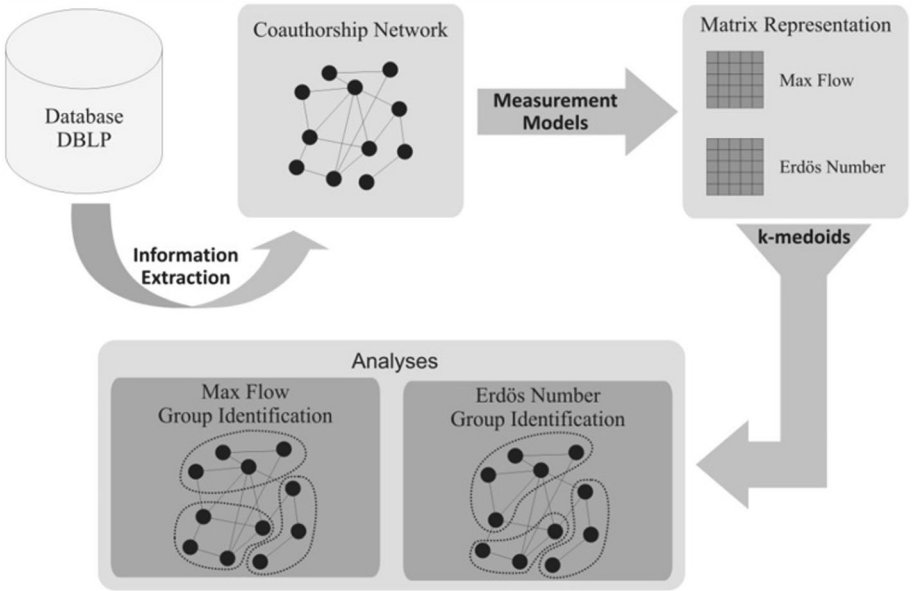

Fig. 9 Case study process steps

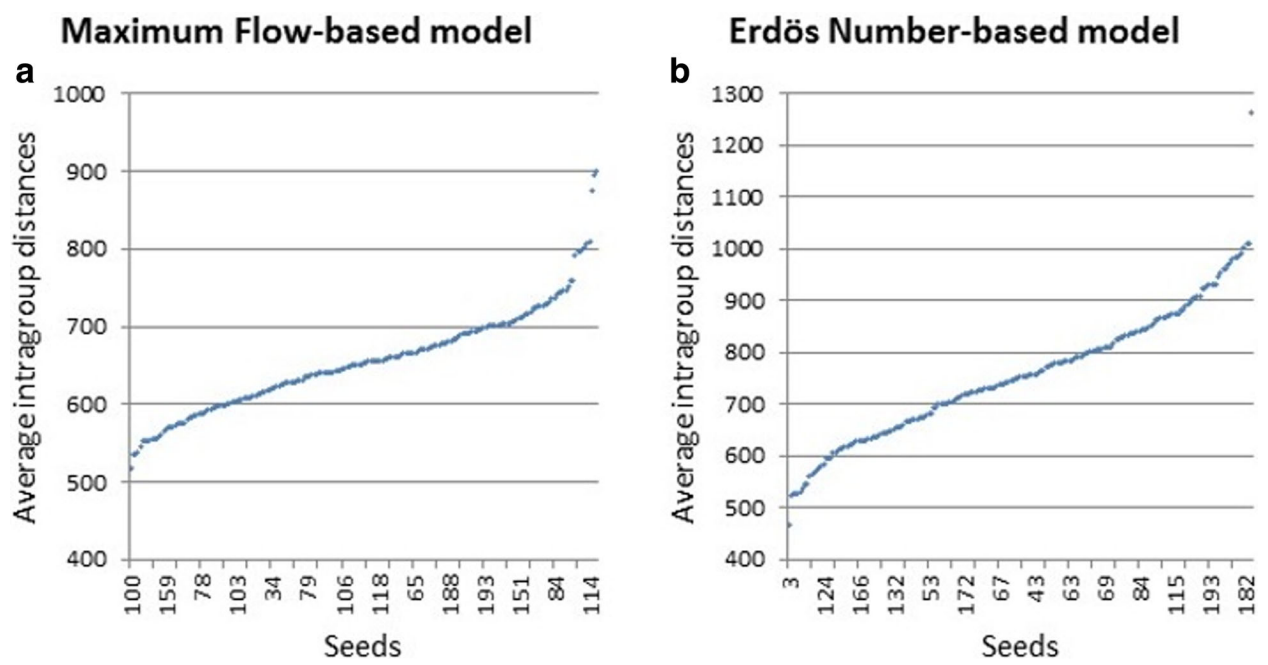

Fig. 10 Average variation of intragroup distances for each seed in (**a**) Maximum Flow and (**b**) Erdös Number models

described above. Therefore, for the Erdös number-based model, it was determined that the optimal number of groups is 47 and the best seed is 3.

Average intragroup distances

To evaluate the behavior of the two proposed models, a detailed analysis was performed focusing on changes in the intragroup distances. Considering the boxplot graph, it is worth stressing the following three points: the first middle point (the median) and the middle points of the two halves. These three points divide the entire data set into quarters, named "quartiles" (Q1, Q2, and Q3). The distance between the first (Q1) and third (Q3) quartiles is a simple dispersion measure that represents the range containing the data on the average. This distance is named interquartile range (IQR), represented by the box on the chart. IQR can be used as a measure of how the values are spread out. If IQR is small, data dispersion is lower, i.e., the data are more homogeneous.

The maximum flow model has a lower interquartile range than the Erdös model, which suggests that the clusters produced from its data are more homogeneous. On the other hand, the Erdös model showed a larger interquartile range and with higher values. This issue

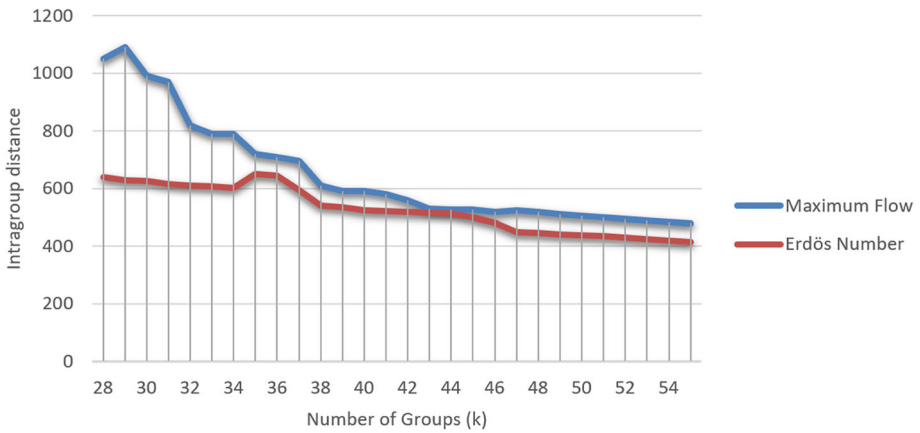

Fig. 11 Variation of intragroup distances for the maximum flow-based model (seed 100) and the Erdös number-based model (seed 3)

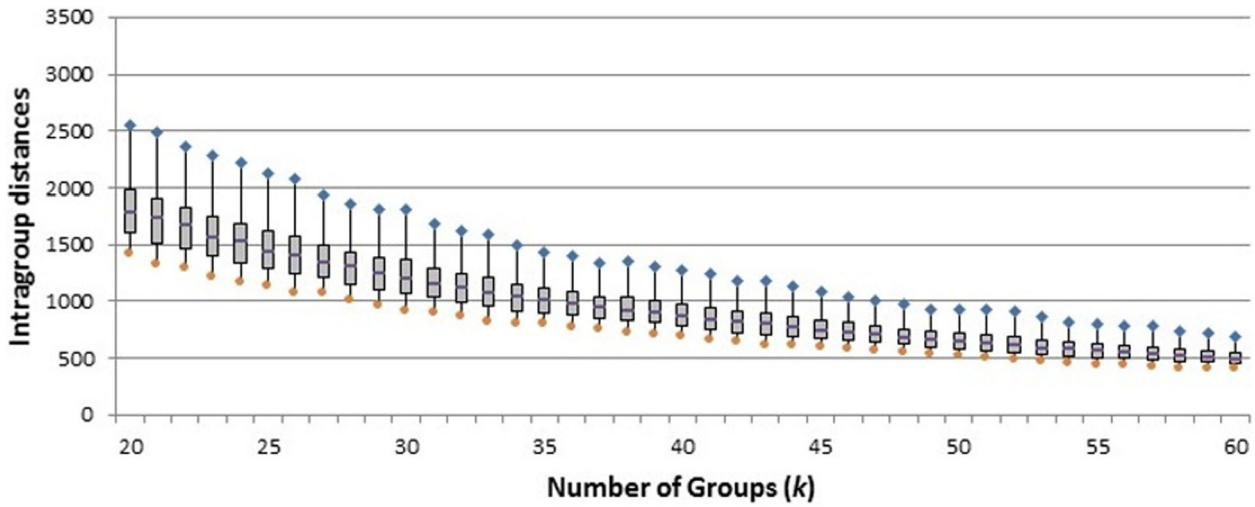

Fig. 12 Variation of intragroup distances per group (maximum flow model)

indicates that the clusters of this model are less homogeneous since there is a greater dispersion of the average intragroup distances and the average value is likewise higher. The graphs in Figs. 12 and 13 represent the variation of the intragroup distances for each group considering all seeds, i.e., for each value of *k*, the clusters were generated using all seeds (1–199), and the intragroup distance variation is in the *k*-group boxplot. As expected, a smaller number of groups produces a greater variation average of the distances. Thus, with an increase in the number of groups, there is a decrease in these mean values.

The maximum flow model has its interquartile range (variation around the median) always smaller than that of the Erdös model. This is another indication that the maximum flow model produces more homogeneous groups than the Erdös model does, regardless of the number of clusters.

Group elements

To answer RQ2, another important analysis related to the element distribution in the groups was conducted. The purpose of the analysis was to check which model could distribute more evenly the elements between groups, reducing the number of groups with a single element. The graph in Fig. 14 shows this distribution.

According to Fig. 14, there is a tendency that larger groups produced by the maximum flow model have

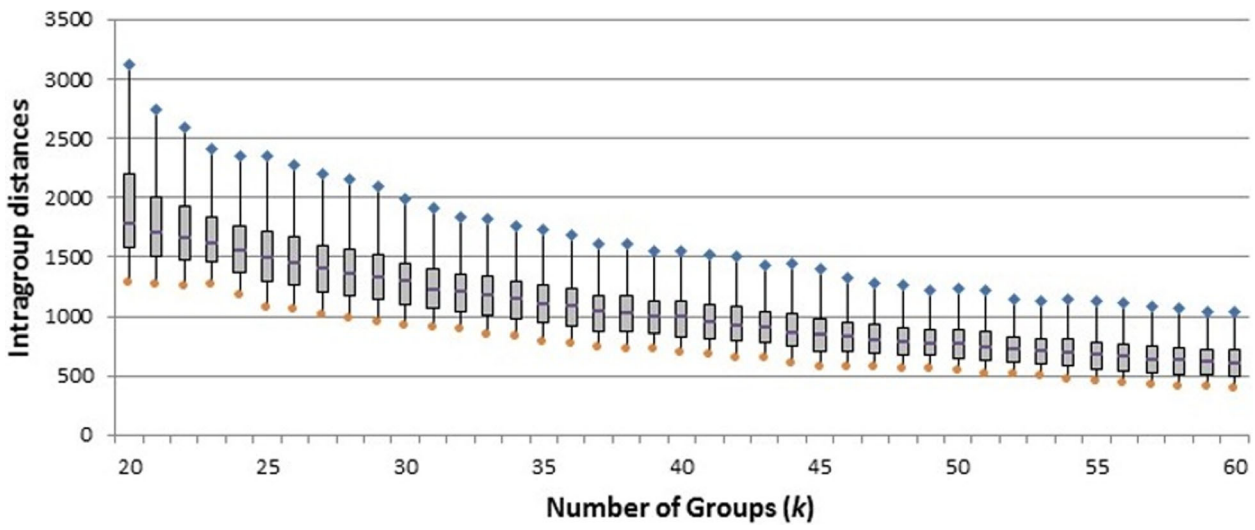

Fig. 13 Variation of intragroup distances per group (Erdös number model)

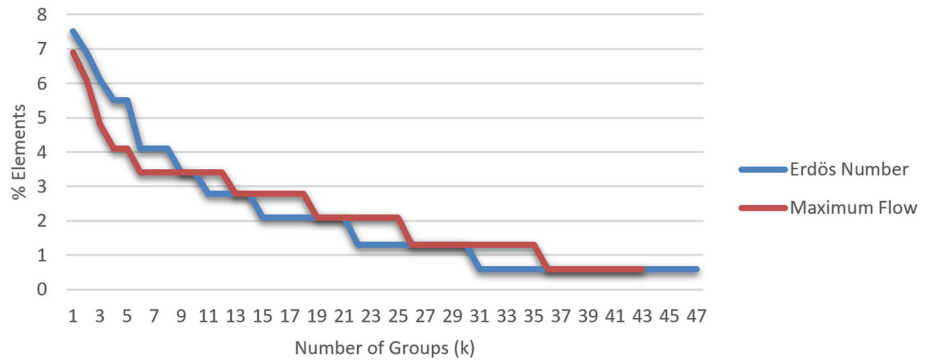

Fig. 14 Elements distribution in the groups

fewer elements than the larger groups produced by the Erdös model. Moreover, the maximum flow model tends to increase the smaller groups, while the Erdös model produces a greater number of groups with fewer elements. While the Erdös model has the lowest average intragroup distances, as shown in Fig. 10b, this model produces many groups with few elements. The maximum flow model tends to produce groups with a better distribution of elements.

Figure 15 shows the obtained result from the Erdös model, each rectangle indicates the researcher number, and the ones in parentheses indicate the group number to which they belong. The large areas behind researchers represent the institutions to which they are affiliated. It

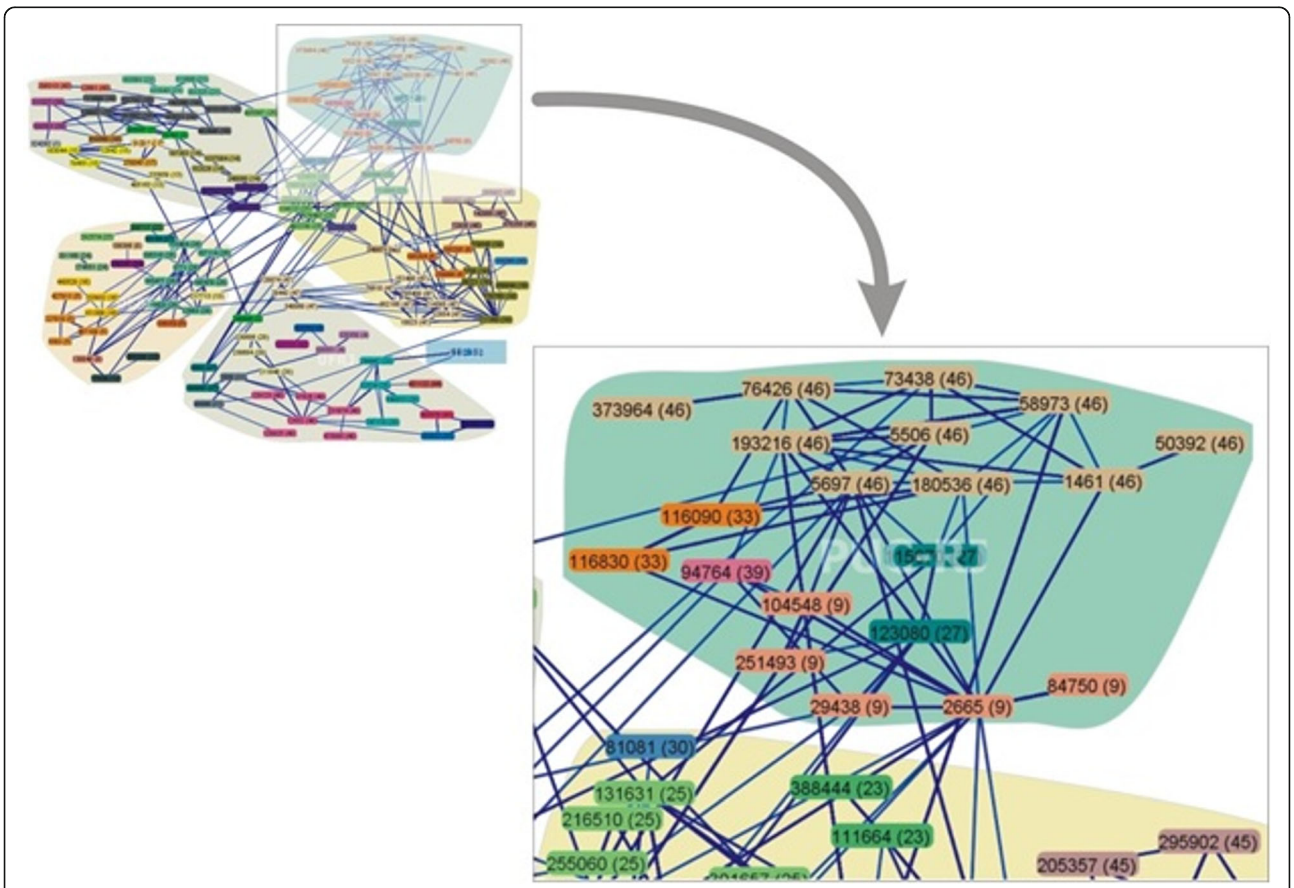

Fig. 15 General and detailed view of the generated groups for the Erdös number-based model

was possible to validate the cluster generated by the Erdös model by comparing it to that obtained by the maximum flow model, previously validated in Ströele et al. [41].

An important feature of the measurement models is the possibility of considering all the alternative paths between two elements to define their relationship weight. This feature answers RQ3, as can be seen in Fig. 15, where elements 50392 and 373964 are in group 46 due to their strong relationship with medoid 73438, and the relationship among them is defined by intermediary elements.

After the case study, it can be stated that both approaches produce good clusters, but the maximum flow-based approach produced more homogeneous groups with better distribution of the researchers among them. Moreover, the results showed that researchers participating in the same community could be indirectly related.

Final remarks

All elements in a social network have some influence on information diffusion. In addition, information reaches elements that are not directly related to the element that produces it. Therefore, this paper proposed two models (maximum flow-based model and Erdös number-based model) capable of measuring the influence that elements of a social network have on each other, even if they are not directly related.

The Erdös number-based model is a new approach for calculating resistance distance, which allows weights to be applied to resistance, considerably increasing the applicability of resistance distances in real-world applications. The maximum flow-based model is another way of calculating the relationship weight, which considers the maximum amount of information to be transmitted between two elements of the social network.

Both models were applied to scientific social networks built using the DBLP database, and a clustering algorithm (*k*-medoids) was used to identify scientific communities. These communities are composed of researchers who have great potential for communication among themselves. The obtained results allowed us to carry out a detailed analysis about the behavior of these models when identifying the scientific communities analyzed to assess whether the Erdös model also produces a good measure to set the relationship weight compared to the maximum flow model. This analysis shows that the results were satisfactory for both models.

By means of a case study, both the analyses and the obtained results confirmed the effectiveness of our approach using weighted resistance distance calculation in scientific social network analysis. However, additional experiments are needed so as to carry out a qualitative analysis.

As future work, the authors intend to improve the Erdös number-based model so that the elements connected by intermediate nodes can reduce their weight in the relationship, as proposed by maximum flow-based model. Thus, the distribution of elements in the groups can be enhanced, producing a smaller number of unit groups.

In this study, we considered only the Laplacian matrix for weighted graphs in the definition of the RENM. However, the spectral graph theory studies the structural properties derived from the matrices that represent graphs. The latter lead to the spectral properties of the representation matrices, which are the central element of the spectral theory of graphs. In this sense, as the spectral graph theory deals with matrices of weighted graphs similar to Laplacian, as future work, we intend to improve the RENM by considering other matrix representations.

Moreover, considering the complexity to set the parameters of *k*-medoids algorithm, the authors intend to explore other clustering algorithms that do not require these settings, such as density-based clustering algorithms.

Endnotes

[1] http://dblp.uni-trier.de/

Availability of data and materials

The dataset supporting the conclusions of this article is not available in an online repository, but all data used could be shared when asked via email from victor.stroele@ice.ufjf.br.

Authors' contributions

VS contributed to the data extraction, model development, results analysis, and paper writing. RC contributed to the model development, results analysis, and paper writing. GZ contributed to the students' guideline, model development, and results analysis. JMS contributed to the students' guideline and results analysis, FC, JMND, and RB contributed to the students' guideline, results analysis, and paper writing. All authors read and approved the final manuscript.

Authors' information

Victor Ströele: BS degree in Computer Science from Federal University of Juiz de Fora (2005), master's degree (2007) and Ph.D. (2012) in Systems Engineering and Computer Science Program from Federal University of Rio de Janeiro. He is currently an Associate Professor II at Federal University of Juiz de Fora. He has experience in Computer Science, with emphasis on Data Mining and Complex Network, working mainly on the following topics: Clustering Algorithm, Social Network Analysis, Recommender Systems, and Informatics in Education.

Renato Crivano: Founding partner and hands-on CTO; Renato designed most of the technological solutions developed by the company and recruited and led the team of programmers. With the other partners, he participated in all the strategic decisions and various investment fundraising processes. Always interested in cutting-edge computer technologies, he also has a master's degree in Database from the Systems Engineering and Computer Science Program/COPPE/UFRJ.

Geraldo Zimbrão: BS degree in Computer Science from Federal University of Rio de Janeiro (1993), master's degree in Applied Mathematics from Federal University of Rio de Janeiro (1993), Ph.D. in Systems Engineering and Computer Science from Federal University of Rio de Janeiro (1999), post-doctoral degree from FernUniversitat (2005). He is currently a professor at Federal University of Rio de Janeiro and has a fellowship from CEDERJ (Center for Distance Higher Education of Rio de Janeiro). He has an experience in the area of Computer Science, with emphasis on Computer Systems, focused mainly on the following topics: Spatial Join and Spatial Databases.

Jano M. Souza: BS degree in Mechanical Engineering from Federal University of Rio de Janeiro (1974), master's degree in Computer Science from COPPE-Federal University of Rio de Janeiro (1978), and Ph.D. in Information Systems from the University of East Anglia (1986). Sabbatical at CERN from 1989 to 1993 (3 months a year). Researches and teaches Computer Science, focusing on the following

topics: Databases, Knowledge Management, Social Networks, CSCW, Autonomic Computing, and Negotiation Support Systems.
Fernanda Campos: BS degree in Mathematics from Federal University de Juiz de Fora (1978). Master's degree in Systems Engineering and Computer Science from Federal University of Rio de Janeiro (1994). Ph.D. in Systems Engineering and Computer Science from Federal University of Rio de Janeiro (1999). She is currently a senior professor at the Federal University of Juiz de Fora, working at the Computer Science undergraduate program, and is a permanent member of the Masters Program in Computer Science. She has an experience in Computer Science, with emphasis on software engineering, specifically in the following topics: e-Science, e-Learning, Recommender Systems, and Ecosystems.
José Maria N. David: BS degree in Electrical Engineering from Military Institute of Engineering, IME, (1983), master's degree (M.Sc.) in Computer Science from Federal University of Rio de Janeiro (1991), and doctoral degree (D.Sc) in Computer Science from Federal University of Rio de Janeiro (2004). He is currently an associate professor at the Federal University of Juiz de Fora, working at the Computer Science undergraduate program, and is a permanent member of the Masters Program in Computer Science. He has an experience in Computer Science, focusing on Software Engineering, acting on the following topics: Groupware, CSCW, CSCL, Software Ecosystems, and Middleware.
Regina Braga: BS degree in Computer Science from Federal University de Juiz de Fora (1991). Master's degree in Systems Engineering and Computer Science from Federal University of Rio de Janeiro (1995). Ph.D. in Systems Engineering and Computer Science from Federal University of Rio de Janeiro (2000). She is currently an associate professor at the Federal University of Juiz de Fora, working at the Computer Science undergraduate program, and is a permanent member of the Masters Program in Computer Science. She has an experience in computer science, with emphasis on software engineering and databases, specifically on the following topics: Software Reuse, Ontologies, Data Integration, Ecosystems, and Scientific Workflows.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Postgraduate Program in Computer Science/PPGCC, Federal University of Juiz de Fora (UFJF), Juiz de Fora, MG 36036-900, Brazil. [2]Systems Engineering and Computation Program/COPPE, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, RJ 21941-972, Brazil.

Received: 23 October 2017 Accepted: 11 June 2018
Published online: 04 July 2018

References
1. Aldenderfer MS, Blashfield RK (1984) Cluster analysis. Sage, Beverly Hills. https://doi.org/10.4135/9781412983648,
2. Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: Proceedings of the 21st international conference on World Wide Web SE—WWW '12, pp 519–528
3. Bapat RB (2004) Resistance matrix of a weighted graph. Communications in Mathematical and in Computer Chemistry/MATCH 50:73–82
4. Bapat RB, Gutman I, Xiao W (2003) A simple method for computing resistance distance. J Phys Sci 58:494–498
5. Bastos MT, Recuero RDC, Zago GDS (2014) Taking tweets to the streets: a spatial analysis of the vinegar protests in Brazil. First Monday. https://doi.org/10.5210/fm.v19i3.5227
6. Bezdek JC, Pal NR (1998) Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics) 28:301–315. https://doi.org/10.1109/3477.678624
7. Cayley A: A theorem on trees, Quart. J. Pure Appl. Math. 23 (1889); 376–378.
8. Chen J, Song X, Nie L et al (2016) Micro tells macro. In: Proceedings of the 2016 ACM on Multimedia Conference—MM '16. ACM Press, New York, New York, USA, pp 898–907
9. Collected Mathematical Papers Vol. 13, Cambridge University Press 1897, 6–28.
10. Cumpsty NA (2009) Some Lessons Learned. In: Volume 7: Turbomachinery, parts A and B. ASME, Lyon, France, pp 785–794
11. Dias A, Chomutare T, Botsis T (2012) Exploring the community structure of a diabetes forum. Studies in Health Technology and Informatics 180:833–837. https://doi.org/10.3233/978-1-61499-101-4-833
12. Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1:269–271. https://doi.org/10.1007/BF01386390
13. Edmonds J, Karp RM (1972) Theoretical improvements in algorithmic efficiency for network flow problems. J ACM 19:248–264. https://doi.org/10.1145/321694.321699
14. Evans TS, Lambiotte R, Panzarasa P (2011) Community structure and patterns of scientific collaboration in business and management. Scientometrics 89:381–396. https://doi.org/10.1007/s11192-011-0439-1
15. Ford LR, Fulkerson DR (1956) Maximum flow through a network. Can J Math 8:399–404. https://doi.org/10.4153/CJM-1956-045-5
16. Grabowicz P, Ramasco J, Moro E et al (2012) Social features of online networks: the strength of intermediary ties in online social media. PLoS One 7:e29358. https://doi.org/10.1371/journal.pone.0029358
17. Guille A (2013) Information diffusion in online social networks. Proceedings of the 2013 Sigmod/PODS PhD symposium on PhD symposium 1:31–36. https://doi.org/10.1145/2483574.2483575
18. Han J, Kamber M (2006) Data mining: concepts and techniques. Morgan Kaufmann Publishers, USA
19. Hossain L, Zhu D (2009) Social networks and coordination performance of distributed software development teams. J High Technol Management Res 20:52–61. https://doi.org/10.1016/j.hitech.2009.02.007
20. Howard PN, Duffy A, Freelon D et al (2011) Opening closed regimes: what was the role of social media during the Arab Spring? Project on Information Technology and Political Islam:1–30. https://doi.org/10.1007/s13398-014-0173-7.2
21. Huang Y, He P, Li B (2012) IEEE Xplore—Applying centrality measures to the behavior analysis of developers in open source software Communit... In: cloud and green computing (CGC), 2012 second international conference on. IEEE pp 418–423
22. Hughes AL, Palen L (2009) Twitter adoption and use in mass convergence and emergency events. Int J Emerg Manag 6:248. https://doi.org/10.1504/IJEM.2009.031564
23. Ichise R, Takeda H, Ueyama K (2005) Community mining tool using bibliography data. In: Proceedings of the International Conference on Information Visualisation IEEE, pp 953–960
24. Jones JJ, Settle JE, Bond RM et al (2013) Inferring tie strength from online directed behavior. PLoS One 8:e52168. https://doi.org/10.1371/journal.pone.0052168
25. Jung JJ (2010) Integrating social networks for context fusion in mobile service platforms. J Universal Computer Sci 16:2099–2110
26. Kaufman L, Rousseeuw PJ (2005) Finding groups in data: an introduction to cluster analysis. Wiley-Interscience 33:368. https://doi.org/10.1007/s00134-006-0431-z
27. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining—KDD '03, p 137. https://doi.org/10.1145/956755.956769
28. Kennelly AE (1899) Equivalence of triangles and three-pointed stars in conducting networks. Electrical World and Engineer 34:413–414
29. Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering version 2.3. Engineering 45:1051. https://doi.org/10.1145/1134285.1134500
30. Kuphaldt TR (2009) Lessons In electric circuits: DC. In: Open Book Project. http://www.ibiblio.org/kuphaldt/electricCircuits/DC/, pp 1–526
31. Michael Barr (2001) Rational Erdős number. https://www.oakland.edu/Assets/upload/docs/Erdos-Number-Project/barr.pdf
32. Newman M (2004a) Who is the best connected scientist?A study of scientific coauthorship networks. In: Complex networks. SFI working paper 00-12-64, Santa Fe, pp 337–370
33. Newman MEJ (2004) Coauthorship networks and patterns of scientific collaboration. In: Proceedings of the National Academy of Sciences 101 (suppl 1) 5200-5205. https://doi.org/10.1073/pnas.0307545100

34. Newman M (2004c) Detecting community structure in networks. Eur Phys J B 38:321–330. https://doi.org/10.1140/epjb/e2004-00124-y
35. Nowell DL, Kleinberg J (2003) The link prediction problem for social networks. In: Proceedings of the twelfth international conference on information and knowledge management, vol 58, pp 556–559. https://doi.org/10.1145/956863.956972
36. Pakhira MK, Bandyopadhyay S, Maulik U (2004) Validity index for crisp and fuzzy clusters. Pattern Recogn 37:487–501. https://doi.org/10.1016/j.patcog.2003.06.005
37. Rice E, Tulbert E, Cederbaum J et al (2012) Mobilizing homeless youth for HIV prevention: a social network analysis of the acceptability of a face-to-face and online social networking intervention. Health Educ Res 27:226–236. https://doi.org/10.1093/her/cyr113
38. Song X, Nie L, Zhang L et al (2015a) Interest inference via structure-constrained multi-source multi-task learning. In: IJCAI International Joint Conference on artificial intelligence, pp 2371–2377
39. Song X, Nie L, Zhang L et al (2015b) Multiple social network learning and its application in volunteerism tendency prediction. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval—SIGIR '15. ACM Press, New York, New York, USA, pp 213–222
40. Strode V, Campos F, Pereira CK et al (2016) Information extraction to improve link prediction in scientific social networks. 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD):515–520. https://doi.org/10.1109/CSCWD.2016.7566043
41. Ströele V, Zimbrão G, Souza JM (2013) Group and link analysis of multi-relational scientific social networks. J Syst Softw 86:1819–1830. https://doi.org/10.1016/j.jss.2013.02.024
42. Trajanovski S, Kuipers FA, Ilic A et al (2015) Finding critical regions and region-disjoint paths in a network. IEEE/ACM Trans Networking 23:908–921. https://doi.org/10.1109/TNET.2014.2309253
43. Varlamis I, Eirinaki M, Louta M (2010) A study on social network metrics and their application in trust networks. In: Proceedings - 2010 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2010, pp 168–175
44. Wang D, Pedreschi D, Song C et al (2011) Human mobility, social ties, and link prediction. In: Acm. ACM, New York, NY, USA, pp 1100–1108
45. Wasserman S, Faust K (1994) Social network analysis: methods and applications. https://doi.org/10.1525/ae.1997.24.1.219
46. Watts J (2013) Brazil protests: president to hold emergency meeting. In: The Guardian. http://www.guardian.co.uk/world/2013/jun/21/brazil-protests-president-emergency-meeting. Accessed 27 Mar 2018
47. Xiao W, Gutman I (2003) Resistance distance and Laplacian spectrum. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 110:284–289. https://doi.org/10.1007/s00214-003-0460-4
48. Yan E, Guns R (2014) Predicting and recommending collaborations: an author-, institution-, and country-level analysis. Journal of Informetrics 8:295–309. https://doi.org/10.1016/j.joi.2014.01.008
49. Yang C, Ma J, Silva T et al (2014) A multilevel information mining approach for expert recommendation in online scientific communities. Comput J 58:1921–1936. https://doi.org/10.1093/comjnl/bxu033
50. Yang J, Leskovec J (2012) Defining and evaluating network communities based on ground-truth. ACM SIGKDD Work Min Data Semant 42:745–754
51. Yin RK (2009) Case study research: design and methods, fifth edn. Sage Publications, Beverly Hills
52. Yuan G, Murukannaiah PK, Zhang Z, Singh MP (2014) Exploiting sentiment homophily for link prediction. In: Proceedings of the 8th ACM Conference on Recommender systems—RecSys '14, pp 17–24