Expert Systems With Applications 43 (2016) 261–270

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Social networks and genetic algorithms to choose committees with independent members

Eduardo Zamudio*, Luis S. Berdún, Analía A. Amandi

ISISTAN Research Institute (UNICEN/CONICET), Campus Universitario, Paraje Arroyo Seco, Tandil, Buenos Aires B7001BBO, Argentina

ARTICLE INFO

Keywords:
Committee
Group selection
Independence
Social network
Genetic algorithm

ABSTRACT

Choosing committees with independent members in social networks can be regarded as a group selection problem where independence, as the main selection criterion, can be measured by the social distance between group members. Although there are many solutions for the group selection problem in social networks, such as target set selection or community detection, none of them have proposed an approach to select committee members based on independence as group performance measure. In this work, we propose a novel approach for independent node group selection in social networks. This approach defines an independence group function and a genetic algorithm in order to optimize it. We present a case study where we build a real social network with on-line available data extracted from a Research and Development (R&D) public agency, and then we compare selected groups with existing committees of the same agency. Results show that the proposed approach can generate committees that improve group independence compared with existing committees.

1. Introduction

Organizations need representative individuals to make decisions about particular concerns. These representative individuals are appointed in committees, and we expect from his members to make decisions based on the benefit of the whole community they are representing, avoiding bias that could arise from closeness between them. In this context, the best committees are those which show the greatest independence between his members. How to choose these members based on objective criteria could be a difficult task, either because of the definition of the criteria or because of the analysis of the community from where members are chosen. Therefore, a committee in which some of its members are closely related is an unbalanced committee.

Fig. 1 shows a graphical example of difference between balanced and unbalanced committees that allow us to appreciate the distribution of selected nodes within a graph. A balanced distribution is essential to improve desirable features, such as independence. For instance, a committee to discuss about budget allocation must avoid biased decisions by ensuring that committee members are not closely related.

As mentioned before, Fig. 1 shows a simple example of individuals and their dispersal. Fig. 2 shows a graphical representation of the community used to evaluate this approach. This graph allows us to understand the problem complexity and critical importance of choosing the best committee members to maximize independence.

Initially, the committee member selection problem can be solved by a mathematical combination, but the computational cost associated to this approach could be very high. For instance, given a community with n members, the maximal number of groups is given by $2^n - 1$, and complexity is $O(2^n)$. In addition, if committees are r size groups, the number of possible solutions is given by applying binomial coefficient ${}_nC_r$ and complexity is $O(n!)$.

If there is no polynomial function to solve the problem, an alternative could be to adopt a non deterministic approach to approximate optimal solutions. For instance, a stochastic approach could produce random solutions, and then apply an independence function to rank these solutions. This approach is subjective because of the probability in selecting random committee members, and because of the joint probability of the committee.

However, the problem can be addressed by implementing some optimization strategy to approximate optimal solutions, such as genetic algorithms. A genetic algorithm could be implemented to search for the greatest independence between committee members, but not necessarily to guarantee the best solution. In other words, could be enough to approximate an optimal solution. For committee selection problem, the best solutions will be determined by the maximal independence between his members.

* Corresponding author. Tel.: +54 249 4439882/3764206786.
E-mail addresses: eduardo.zamudio@isistan.unicen.edu.ar, eduardozamudio@gmail.com (E. Zamudio), lberdun@exa.unicen.edu.ar (L.S. Berdún), amandi@exa.unicen.edu.ar (A.A. Amandi).

http://dx.doi.org/10.1016/j.eswa.2015.07.045
0957-4174/

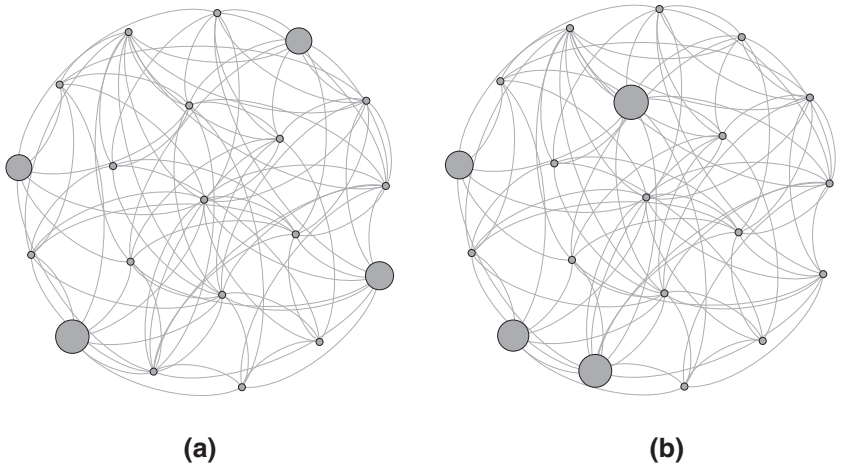

Fig. 1. Difference between balanced (a) and unbalanced (b) committees, where selected members are the largest 4.

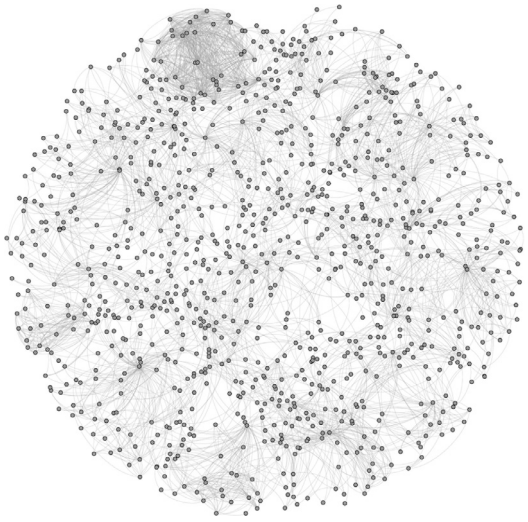

Fig. 2. Graphical representation of the community used to evaluate the approach.

If we consider committee candidates as individuals connected with each other through ties, it is possible to determine which of these ties could be relevant to analyze independence. These individuals and their relationship represent the basic elements of a social network; therefore, we can apply social network analysis to select committee members with the greatest independence. However, a social network approach requires a social network, and data to represent its elements, such as actors, ties, kind of network, and analysis object.

The current social network analysis techniques aim to identify the value or number of relations, roles or prominence of nodes, and to discover hidden groups or cohesive groups. The aim of this work is to present an alternative to the committee selection problem by choosing a set with maximal independence between members. To do this, we build a social network and then we define an independence group performance function and a genetic algorithm, to obtain n member committees with the greatest independence between members.

The main contributions of this work are summarized as follows. (1) We propose an approach for the committee selection problem with independent members as a group selection problem in social networks. (2) We define a novel group independence performance function to assess group fitness in social networks. Then, such a measure was optimized by means of a genetic algorithm. (3) We build a social network from a Research and Development (R&D) public agency with on-line available data. (4) We use such a social network to evaluate the proposed approach. Then, we compare results with current committees of the same public agency.

This document is organized as follows. Section 2 describes the construction process of the social network. Section 3 describes the implementation of the genetic algorithm and the function to evaluate group independence. Section 4 describes a case study and the configurations of the genetic algorithm, along with a discussion of the experiment results. Section 5 presents a discussion of the current literature. Finally, Section 6 presents conclusions as well as future work.

2. Social network construction

In order to choose committee members, we propose to build a social network to calculate distances between candidates, and then apply a genetic algorithm to get potential committees with the greatest distances between their members.

A social network is a set of individuals (actors) and relations (ties) between them; the social network analysis is used to study structures created by these relations and individuals.

We are particularly interested in the construction of a social network for its ability to represent analysis criteria based on ties. To clarify this concept, we built a network of researchers related through coauthorship and workplace. In this network, actors are the researchers, and ties are the criteria for calculating distance between each pair of researchers.

As mentioned above, relations between actors define what can be analyzed in the network. The aim of this analysis is to calculate distances between a set of actors. In order to do this, we built a consolidated graph. This graph contains every kind of relation proposed as analysis criterion. Fig. 3 shows a unified graph from two kinds of relations (*coauthor* and *same workplace*) of five researchers (A, B, C, D, and E) where relations are binary (relation is present or not), undirected (direction is meaningless), and irreflexive (a researcher does not publish with himself or does not work with himself).

Our proposal is to establish the greatest independence between committee members based on their distances. Thus, we need to calculate distances between committee members, for which we use the *shortest path* and *geodesic distance* (length of the shortest path) (Freeman, 1977) over the unified graph.

The graph must be connected to apply this metrics, which means that every actor must be reachable from every other actor in the network. This can be determined through a reachability matrix, which can be obtained through matrix multiplication (Wasserman & Faust, 1994).

Distances between each pair of actors is represented by a proximity matrix, obtained by applying power to the matrix

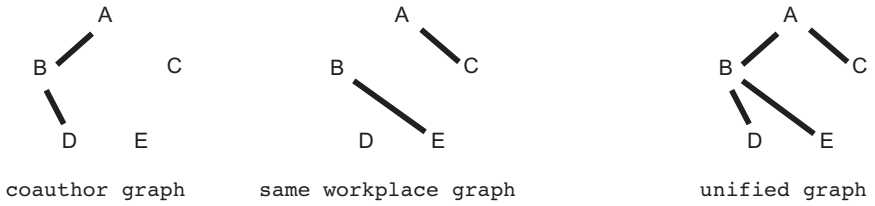

Fig. 3. Unified graph representing two kinds of relationships (coauthor and same workplace).

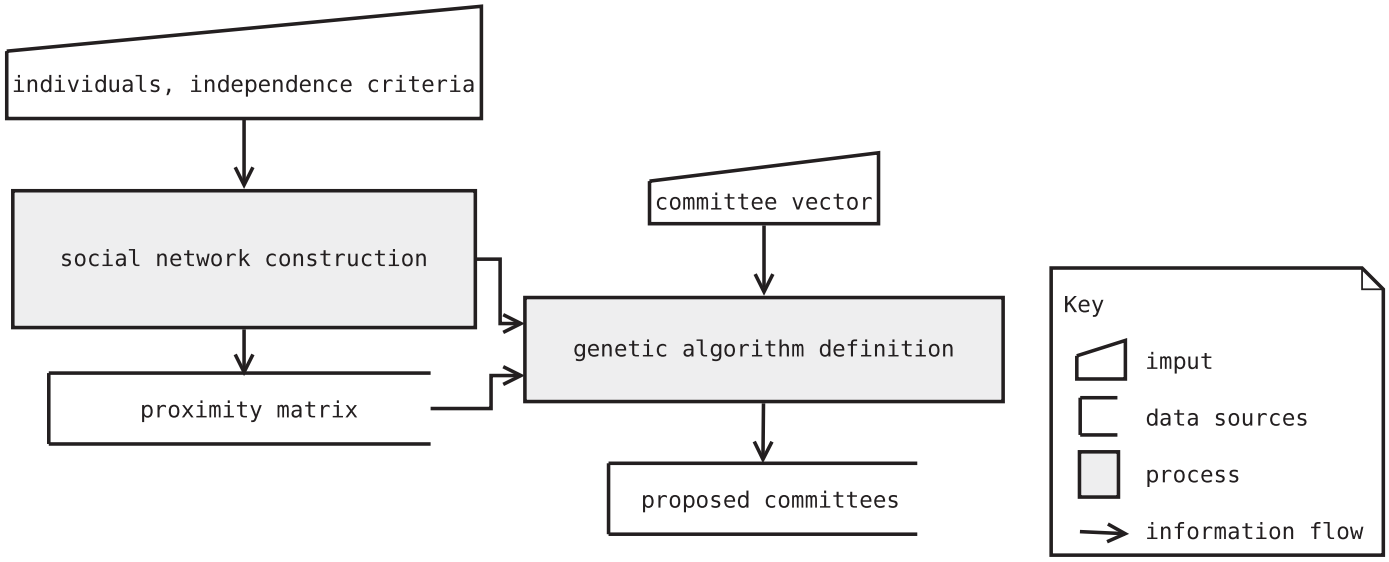

Fig. 4. Flow chart of the proposed approach showing inputs, datasources production, and processes related to the social network construction and the genetic algorithm definition.

representation of the unified graph. Proximity matrix contains input data for the algorithm whose aim is to choose a committee (a group of actors) with the greatest independence between its members. In this case, we work with a genetic algorithm which defines a function to optimize this distance to the largest one.

Fig. 4 shows the proposed approach in a flow chart, in which individuals and independence criteria are the inputs. Then, we generate the unified graph to determine relations between actors, and thus to build the social network. Next, we build the proximity matrix by calculating geodesic distances; then, the proximity matrix and the network data are put together into the genetic algorithm to produce optimized solutions.

3. Genetic algorithm definition

A genetic algorithm is a type of evolutionary algorithm that can be considered as a function optimization method (Smith & Eiben, 2008). Even though there is no definitive genetic algorithm, it is possible to adapt one using representations and operators considered suitable to the modeled problem. As an analogy of the biological model, chromosomes are the elements used in genetics algorithms to represent configurations, which contain genetic information represented by location and value of their genes. These chromosomes stand for solutions to the modeled problem.

In order to choose a subset of actors from a social network, we have defined an ad-hoc function to calculate distances between committee members. Consequently, we have defined a genetic algorithm to approximate solutions to an optimum by maximizing this function.

The development of a genetic algorithm requires defining representation, fitness function, parent selection and survivor selection mechanisms as well as mating and mutation operators. Next, we present selected configurations to the modeled problem.

3.1. Representation

The problem requires defining a representation of the chromosome. In this work, we do permutations of a vector of integers (chromosome), where each element references to only one node (gene). In this vector, every node in the network under study is included. Thus, a chromosome has as many genes as a community has individuals. Also, the participation in the committee is given by a vector with the same size as the vector of nodes, the vector of committee members, in which every location is binary valued. Therefore, if value $= 1$, then the node with same position in the vector of nodes must be included in the committee, and if value $= 0$, then the node is excluded from the committee. With this representation, a member appears only once in a given committee. It is important to note that in the modeled problem the order of members is not relevant. Fig. 5 shows a graphical representation of these vectors.

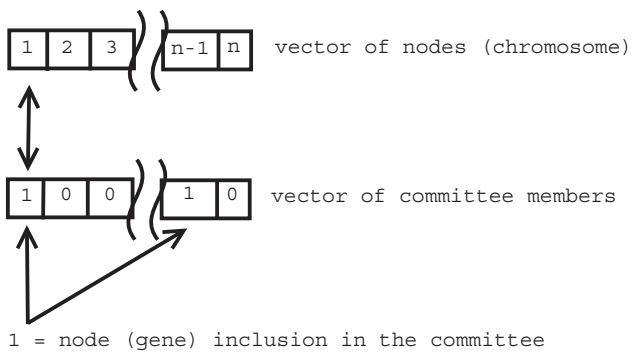

Fig. 5. Representation of the genetic algorithm through a vector of nodes which contains every node in the network under study, and a vector of committee members which indicates the elements of the vector of nodes to be included in the committee.

3.2. Fitness function

The aim of the fitness function is to calculate the solution value. In this work, we developed an ad-hoc fitness function to maximize distances, represented by the cumulative sum of distances between each pair of committee members. In order to get relative values, we consider the size of the committee and the network diameter. To improve results, we set a parameter to maximize minimum distances of the committees, defined as follows:

$$f = \frac{\left[\left(\sum_{i,j=0}^{k} d(i,j)\right) \Big/ k\right] + m}{2 * D}$$

Where d is the distance function between two members, $\forall i, j / i \neq j$ and $i, j \in S$, S represents the whole nodes set, k is the number of committee members, m is the minimum distance between each pair of members in the committee, and D is the network diameter. As previously established, it is necessary for the network to be connected.

3.3. Parent selection

The genetic information is obtained from the parents, which are chromosomes (solutions) of the previous generation. To this end, we need to define a strategy of parents selection by adopting one of the mechanisms suitable to the modeled problem. In this work, the mechanisms selected include *Stochastic Universal Sampling* (SUS) since we need to choose several parents from a community; and *Tournament*, since in both cases global fitness is unknown.

3.4. Crossover

Genetic information of new generations is determined by their parents. This process called genetic recombination is produced through crossover mechanisms. For instance, having two chromosomes representing distinct solutions, crossover implies that the new generation inherit genetic information from both parents.

To keep a valid permutation we have chosen recombination operators *Partially Mapped Crossover* (PMX) and *Order Crossover* (OX). Since the former is an algorithm designed for adjacency problems it is suitable to the modeled problem, and even though the latter is designed for order problems, the order in the second parent could be beneficial in new chromosome production.

3.5. Mutation

The other mechanism used in this work for genetic recombination is mutation, which implies to alter the genes within a chromosome. In permutations, mutation alters location of the values in the solution vector of the new generation.

We have selected *Swap Mutation* and *Insert Mutation*, since both operators are accepted to keep a valid permutation.

3.6. Survivor selection

Once a new generation is produced, the survivors must be selected in order to keep the number of solutions in every generation.

We have selected *Steady-state* and *Generational* mechanisms to keep solutions with the best fitness in the succeeding generations.

4. Case study

To evaluate the proposed approach, we decided to build a social network based on public information about researchers published by the National Council of Technical and Scientific Research (CONICET). This organization establishes committees for specific areas with different responsibilities. For instance, in the Informatics and Communications area there are 3 committees to evaluate *Admissions, Reports*, and *Fellowship awards*.

The prospective committee members are chosen from a set of experts in the field that could be internal or external to the organization.

We calculated fitness for distinct configurations of the genetic algorithm to propose committees based on the greatest distances. With the same criteria, we calculated fitness for existing committees.

4.1. Dataset

The dataset used here to produce the social network based on researchers (actors) information was built by applying *web crawling*, which consists in gathering information from web pages. In this case, we used basic information to characterize actors and their information about contributions and workplaces in order to discover ties between those actors. This process required disambiguation of actors and ties, since most of the information presented for every researcher is produced by themselves, particularly contribution data.

In addition, not every actor in the network is considered as candidate. For the Informatics and Communication area there is a list of qualified specialists that fulfill some requirements (i.e., to have a hierarchical degree), which means that only a limited set of actors qualify as committee members.

Thus, the social network in the case study is composed by 1293 nodes and 4322 ties, which produces 74 components (subgroups of actors disconnected from the rest of the network). From those components, the bigger one has 1058 (≈82%) actors (75 of them are qualified specialist), and 3878 (≈90%) ties.

4.2. Configuration

Having established the social network, we set up the genetic algorithm to evaluate groups of actors with the largest distances between them, which we assume as an independence criterion. This configuration has the following parameters:

- Community size: The number of solutions in every moment was given by P/n, where P is the set of all researchers, and n the size of the committees.
- Crossover probability: A generational parameter, selected from range [0.6; 0.9].
- Mutation probability: A mutation operator parameter, selected from range [0.01; 0.15].
- Stop condition: A generational parameter, set in 25 generations.
- Configurations: Sixteen different configurations emerged from the combination of the selected mechanisms in this approach (selection, mutation, and crossover). In addition, we use *Steady-state* and *Generational* as selection mechanisms. Table 1 shows these configurations.
- Runs: 40 runs produced by 5 runs per configuration. Average values and standard deviation (σ) were calculated.

4.3. Results

Here we show a fitness evaluation and social network centrality metric values for current committees of the Informatics and Communications area, and then we show results of the genetic algorithm runs.

4.3.1. Fitness of current committees

The current committees of the Informatics and Communications area had 6 members in 2014. In order to evaluate committee fitness we initially decided to apply the fitness function to committee members. This approach was modified since some members of the current

Table 1
16 proposed configurations for the genetic algorithm describing operators and selection mechanisms.

Configuration	Crossover		Mutation		Parent selection		Survivor selection	
	PMX	OX	Swap	Insert	SUS	Tournament	Steady-state	Generational
1	X		X		X		X	
2		X	X		X		X	
3	X			X	X		X	
4		X		X	X		X	
5	X		X			X	X	
6		X	X			X	X	
7	X			X		X	X	
8		X		X		X	X	
9	X		X		X			X
10		X	X		X			X
11	X			X	X			X
12		X		X	X			X
13	X		X			X		X
14		X	X			X		X
15	X			X		X		X
16		X		X		X		X

committees were not present in the dataset. This situation occurs because of the low number of specialists in the area belonging to CONICET (actually there are 87 specialists in the Informatics and Communications area), which means that committees usually incorporate external researchers from other areas. Therefore, we have identified the current committees members present in the largest component of the proposed social network. In the *Admissions* committee, only 3/6 members are present in the social network; in the *Reports* committee, only 4/6 members are present in the social network; and in the *Fellowship awards* committee, only 5/6 members are present in the social network. Since names of the committee members are not relevant in this study, we enumerated members from 1 to 6 for each committee.

The *Admissions* committee of the Informatics and Communication area has fitness $=0.65152$ for members A1–A3, since A4 is present in another component and A5 and A6 are not classified as specialists. The other 2 committees are in similar situation. The *Reports* committee has fitness $=0.36364$ for members R1–R4, since the other members of the committee do not belong to CONICET (R5) or are not classified as specialists in the area (R6). And the *Fellowship awards* has fitness $=0.38636$ for members F1–F5, since F6 does not belong to CONICET. Table 2 shows current committee members with centrality metric values for those members present in the largest component of the social network.

Table 2
Current *Admissions, Reports, and Fellowship awards*committees with each degree, betweenness and closeness (last two metrics expressed in relative values).

Committee	Node	Degree	Betweenness	Closeness
Admissions	A1	49	0.05293	0.22404
(fitness = 0.65152)	A2	21	0.02593	0.17283
	A3	5	0.00001	0.15606
	[3]A4	–	–	–
	[2]A5	–	–	–
	[2]A6	–	–	–
Reports	R1	35	0.11858	0.20596
(fitness = 0.36364)	R2	51	0.11909	0.25101
	R3	37	0.03512	0.19495
	R4	34	0.14864	0.20989
	[1]R5	–	–	–
	[2]R6	–	–	–
Fellowship awards	F1	22	0.01246	0.15696
(fitness = 0.38636)	F2	6	0.00001	0.16307
	F3	19	0.00595	0.19317
	F4	42	0.07272	0.22751
	F5	46	0.06920	0.23731
	[1]F6	–	–	–

[1] Does not belong to CONICET.
[2] Not marked as specialist.
[3] Present in another component.

4.3.2. Social network metrics

The social network metrics for current committees shown in Table 2 can be compared with metrics of the whole component, which average degree $=7.316$, network diameter $=11$, and average path length $=5.76$. This indicates that almost every member (except for F2) of current committees has degree over the average component degree, but far away from the highest degree (80) in the component. Some committee members (A3 and F2) show very low betweenness, but their closeness is more balanced between each other.

4.3.3. Genetic algorithm runs

In order to compare the fitness of current committees with the fitness of the members proposed by the genetic algorithm, we decided to modify the genetic algorithm to generate committees of 3, 4, and 5 members.

For the *Admissions* committee, we set up the genetic algorithm in order to produce committees with 3 members. Table 3 shows results where maximal average fitness ≈ 0.72727 and minimal $\sigma = 0$ for configurations 9 and 11. Maximal fitness ≈ 0.72727 was reached by configurations 9, 11, 12, and 13, from which we infer that a local optimum is reached in these cases.

Table 3
Fitness of proposed configurations with average fitness, standard deviation, and maximal fitness for 3-member committees in 5 runs (best values in **bold**).

Configuration	Average fitness (runs = 5)	σ	Maximal fitness (with the shortest time in seconds)	
1	0.58788	0.01134	0.59091	1.548 s.
2	0.57879	0.01134	0.59091	1.510 s.
3	0.60303	0.02607	0.65152	1.563 s.
4	0.60909	0.02938	0.66667	1.537 s.
5	0.61818	0.02938	0.65152	1.625 s.
6	0.62424	0.03090	0.66667	1.468 s.
7	0.62121	0.02710	0.65152	1.544 s.
8	0.62121	0.03711	0.66667	**1.504 s.**
9	**0.72727**	**0.00000**	**0.72727**	31.135 s.
10	0.64545	0.00742	0.65152	31.493 s.
11	**0.72727**	**0.00000**	**0.72727**	30.744 s.
12	0.67879	0.02607	**0.72727**	31.325 s.
13	0.70606	0.02642	**0.72727**	31.024 s.
14	0.63939	0.02607	0.66667	32.174 s.
15	0.67879	0.00606	0.68182	33.181 s.
16	0.61515	0.02642	0.65152	38.198 s.

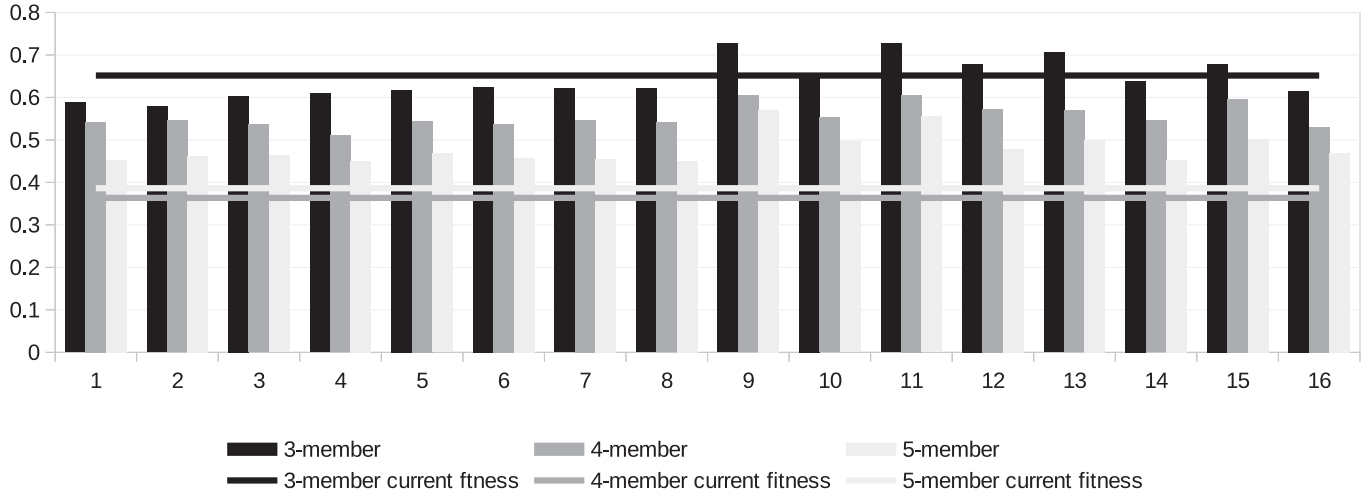

Fig. 6. Average fitnesses of 3-member, 4-member, and 5-member committees for the 16 configurations.

Compared with current committee fitness ≈0.65152, maximal average fitness shows a fitness improvement of ≈8 points.

For the *Reports* committee, we set up the genetic algorithm in order to produce committees with 4 members. Results show maximal average fitness ≈0.60606 and minimal $\sigma = 0$ for configuration 11. Maximal fitness ≈0.60606 was reached by configurations 9, 11, and 15, from which we infer that a local optimum is reached in these cases.

Compared with the current committee fitness ≈0.36364, maximal average fitness shows a fitness improvement of ≈24 points.

For the *Fellowship awards* committee, we set up the genetic algorithm in order to produce committees with 5 members. Results show maximal average fitness ≈0.57091 for configuration 9, minimal σ ≈0.00530 for configuration 4, and maximal fitness ≈0.59091 for configurations 9 and 11.

Compared with current committee fitness ≈0.38636, maximal average fitness shows a fitness improvement of ≈20 points.

As shown in Fig. 6, *Generational* (configurations 9–16) selection mechanism produced better results than *Steady-state* (configurations 1–8), but Fig. 7 shows that *Generational* required more time than other configurations. For instance, in 5-member committees, the minimal time for *Steady-state* =4.73 s. (seconds) and for *Generational* =67.049 s. This situation is similar for 3-member and 4-member committees. In order to reach the time required by *Generational* configurations, we extended *Steady-state* stop condition to 25, 000 generations, resulting always in lower fitnesses than those obtained with *Generational* mechanism configurations.

For 3-member and 5-member committees, configuration 9 presents the fullest average fitness, and for all committees, configurations 9 and 11 show the highest maximal fitness values, from which we infer that in searching for optimal values in similar studies we should prefer the *Generational* selection mechanism and the PMX operator. In addition, in this case the mutation operator does not produce relevant differences. However, in bigger or more complex networks, computational cost improvement may be a requirement, in which cases we should prefer *Steady-state* selection mechanism instead of *Generational* selection mechanism. In addition, Fig. 8 shows that 3-member configurations 9 and 11 reached $\sigma = 0$, and that 3-member and 4-member configuration 9 reached $\sigma = 0$, from which we infer the stability of these configurations, at least for 3-member and 4-member committees.

Fig. 9 shows the social network built for the case study, in which current committee members are closer than the best fitness committee members obtained in experimentation. This representation shows a balance improvement of distances between the best fitness committee members compared to current committee members.

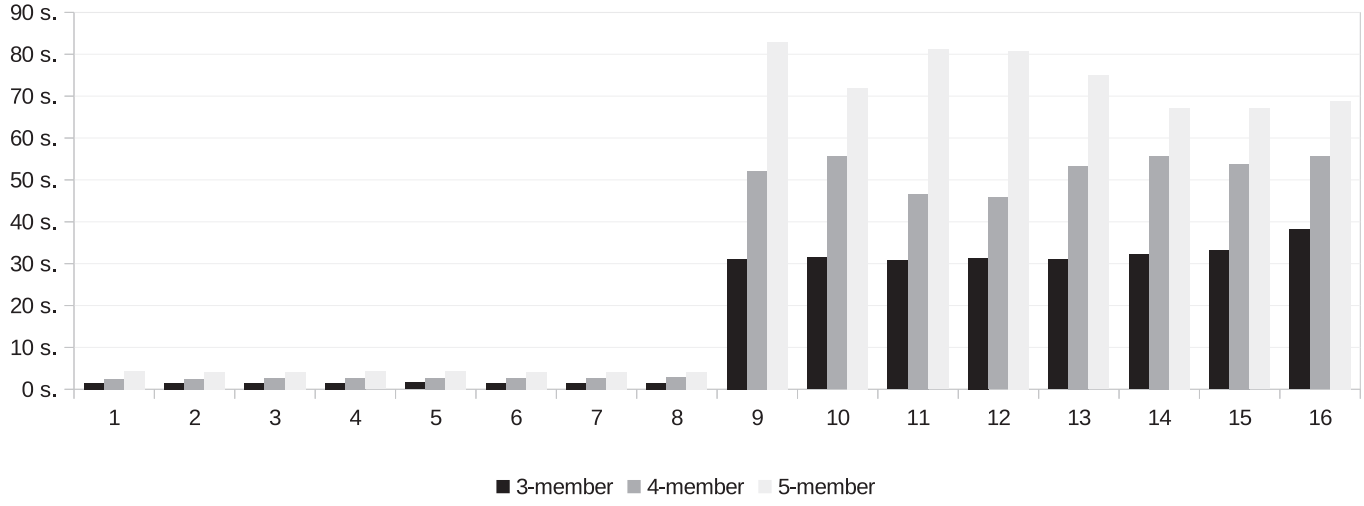

Fig. 7. Shortest times of 3-member, 4-member, and 5-member committees for the 16 configurations.

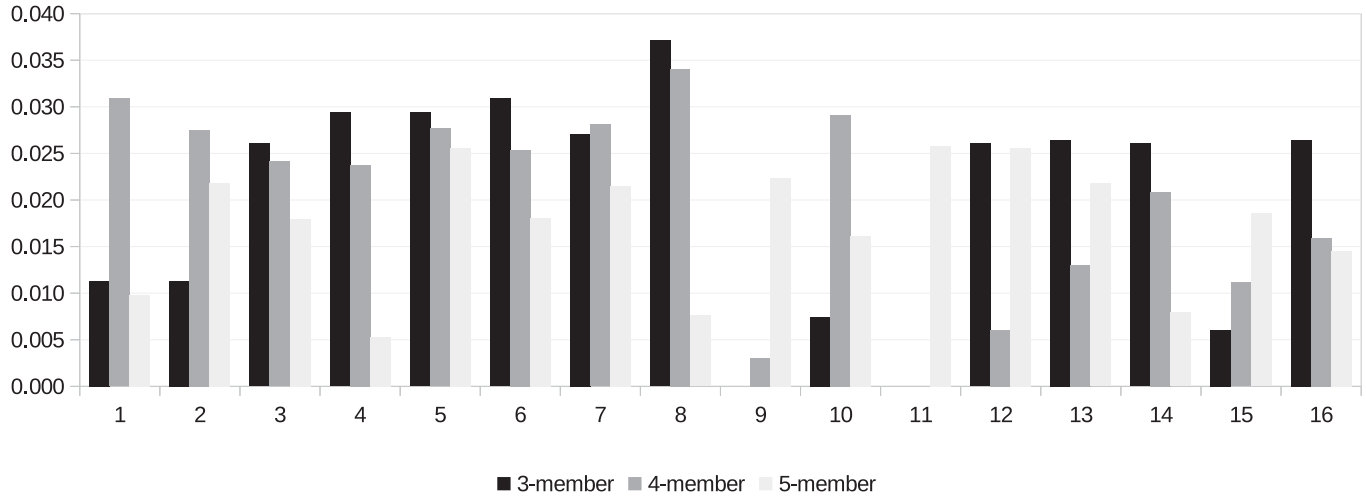

Fig. 8. Standard deviations of 3-member, 4-member, and 5-member committees for the 16 configurations.

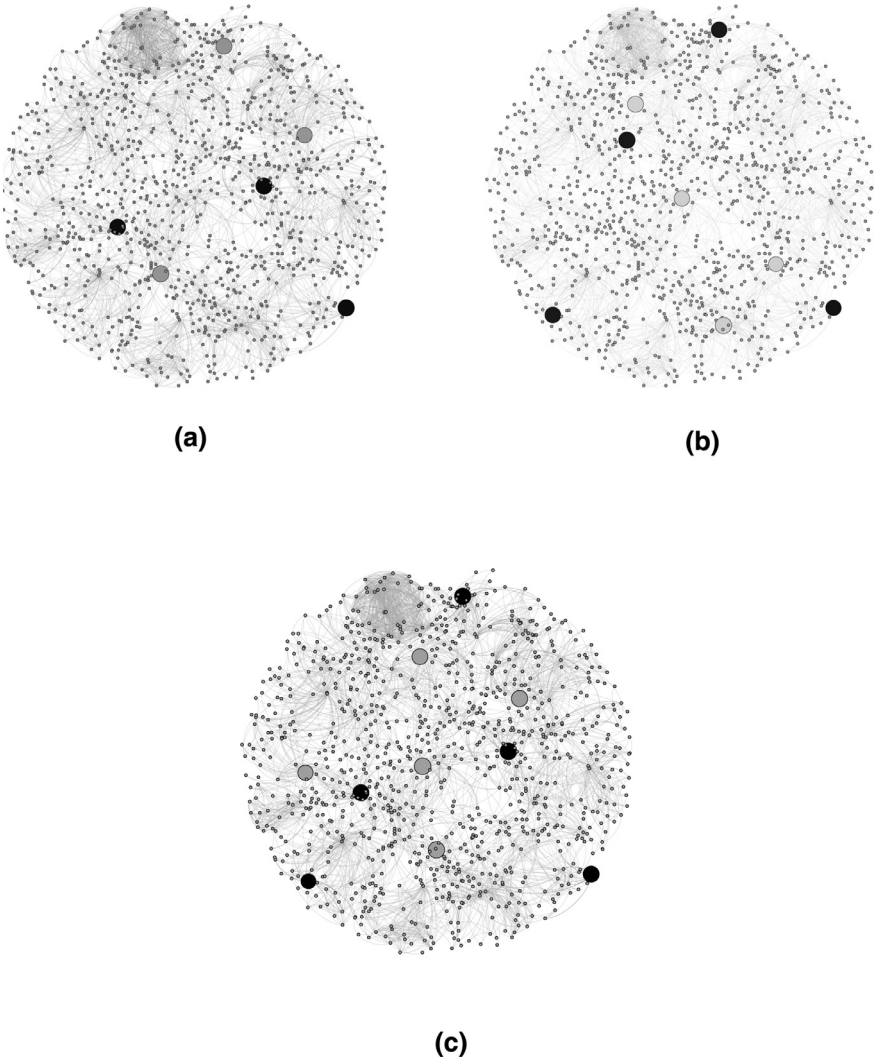

Fig. 9. Current committees members (big gray nodes) versus the best fitness committees members (big black nodes) for 3-member (a), 4-member (b), and 5-member (c) committees.

4.4. Discussion

The proposed social network is intentionally simple about tie complexity and node complexity. Here, ties are binary edges, and nodes do not have attributes considered in the committee setup. On real scenarios, other criteria could be taken into account, such as node prominence, related topic, or skill, in searching to fulfill certain requirements.

To test this approach, we used a new dataset based on public online available data. For simplicity, we built the social network starting from a set of specialists (those in the Informatics and Communication area), and then we created nodes and ties based on co-authorship and workplace information. To analyze other kind of specialists, social network should be built from all actors in the community or a new social network should be built starting from a new set of specialists from the area for which the committee is needed.

5. Related work

Previous work have contributed in the field of creating people committees applied to different areas, such as audit (Abbott & Parker, 2000), board directors (Shivdasani & Yermack, 1999; Westphal & Zajac, 1995), or public agencies (Loewenberg, Patterson, & Jewell, 1985). Some approaches have been focused on the diversity of the members (Aksela & Laaksonen, 2006; Hadjitodorov, Kuncheva, & Todorova, 2006; Kuncheva, 2005; Kuncheva & Whitaker, 2003; Shin & Sohn, 2005; Zouari, Heutte, & Lecourtier, 2005), while other approaches have been based different voting techniques (Bock, Day, & McMorris, 1998; Fishburn, 1981; Gehrlein, 1985). However, to the best of our knowledge there are no precedents in committee selections with independent members by using social networks.

Choosing committees with independent members in social networks can be regarded a group selection problem. Generally, this problem includes node group selection, structural consideration such as cohesion or centrality measures, and some optimization strategy since most of them are classified as NP problems.

Two well-known group selection problems in social networks are the target set selection problem and the community detection problem, however these problems present some differences with committee selection problem. The target set selection problem aims to select nodes that maximize influence in order to spread something in a network, such as information. Here, the focus is on the network, since the problem is determined by which set of nodes increase the influence. The community detection problem aims to discover node sets based on node relations or structural properties. Here, the focus is on the set and its internal structural properties, since the problem is determined by which nodes belong to a group or community.

However, committee member selection problem focuses on the group and the network, since the group considers relations between committee members and the group independence considers the whole network.

Current literature about target set selection problem shares some elements with this work. Wang, Deng, Zhou, and Jiang (2014) develop a set-based coding genetic algorithm (SGA) that converges in probability to the problem optimal solution. Here, the authors code chromosomes as sets, and choose operators based on the chromosome representation. However, SGA mainly differs with this work in the use of diffusion dynamics to measure performance. Cao, Wu, Wang, and Hu (2011) propose a transformation of the target selection problem into an optimal resource allocation problem. Here, the authors make use of the modular structural property of social networks, and propose a dynamic programming algorithm to solve the problem, which was proved to be NP-hard.

Similar to the target set selection problem is the key player problem (KPP) (Ballester, Calvó-Armengol, & Zenou, 2006; Borgatti, 2006; Everett & Borgatti, 2010). KPP identifies key player sets with two different approaches, KPP-Neg and KPP-Pos. KPP-Neg searches for key players sets that if removed will disrupt the network. KPP-Pos searches for key players sets optimally connected to all other nodes in a network. The main difference with this work is on the structural property, since KPP-Pos uses set cohesion and KPP-Neg uses closeness centrality. Also, the authors suggest some evolutionary strategies for function optimization.

An early effort on maximizing the impact in social networks is presented in Liberman and Wolf (1997) that proposes a strategy to increase impact of information flow on scientific communities. This work has historical value, but it shows that similar problems in social networks have had different names over time.

Current literature about community detection problem shows a growing interest in topics such as social circles, topic models, or complex networks. However, there still are community detection approaches mainly based on structural properties. Bhattacharyya and Bickel (2014) use graph distances to detect communities in graphs by using a block model approach. The authors use geodesic distances which have underlying problems, such as the impossibility to measure geodesic distances in unconnected graphs. The authors solve this constraint by replacing distances of disconnected pair of nodes with the largest geodesic distance in the graph.

About the use of genetic algorithms as an optimization strategy for community detection, Freeman (1993) presents a review of the group selection problem and recognizes the computation constraint of uncovering groups based on proximity matrix representation. He also recognizes the need for a search strategy, therefore he proposes a simple genetic algorithm. The main differences with our work are in the chromosome representation and in the fitness function, which uses the proximity matrix information and a binary node classification.

As a precedent on using a structural approach to select people groups, Burt (1978) proposes a process that uses sociometric measures for sampling firm representatives of interlocking directorates to overcome profit constraints of an industry.

We found other areas that use distance as social network structural property for group selection. For instance, in the recommendation area, Hwang, Wei, and Liao (2010) suggest articles based on a co-authorship network and different schemes to measure the closeness of author sets. Here, the social network graph representation includes directed and valued ties which affect closeness measure implementation. In the social network analysis homophily area, Preciado, Snijders, Burk, Stattin, and Kerr (2012) take geographical proximity as distance in order to analyze likelihood of friendship existence and dynamics within social networks. A related approach is presented by Morgan and Carley (2011, 2014) which uses social distance as part of an impact factor set to candidate selection for hiring processes.

As another group selection approach, Wi, Mun, Oh, and Jung (2009a, 2009b) use social network structural properties along with genetic algorithms. The authors propose a quantitative method for the team member selection problem based on knowledge and collaboration of candidates. This problem aims to select teams based on abilities of candidates to fulfill project requirements and to predict team performance. Network structural properties are used to measure familiarity between candidates which is translated in what they call knowledge competence. Also, they use structural properties to select project managers from teams.

A previous work that uses geodesic paths as structural property for group selection (Kolaczyk, Chua, & Barthélemy, 2009) proposes a metric called co-betweenness, which extends betweenness centrality to sets of nodes in order to measure the information flow of the set. Co-betweenness considers the geodesic paths that pass through all nodes in the set.

Out of the social network scope, some works in artificial intelligence use a committee based concept to select other kinds of groups, such as classification (Aksela, 2003; Argamon-Engelson & Dagan, 1999; Li, Zou, Hu, Wu, & Yu, 2013; Wang & Wang, 2006; Zheng, 1998) or clustering (Hadjitodorov et al., 2006; Tao, Ma, & Qiao, 2013).

6. Conclusions

A novel social network approach to the committee member selection problem has been proposed. This approach consists in a mechanism that models the problem as a social network group selection problem.

In this group selection problem for committee member selection, independence is the main selection criterion, for which a novel group independence function is defined. This group independence function uses geodesic distances to measure social distances between all node pairs in the social network. Also, a genetic algorithm is defined to generate committee candidates. Then, the group independence function is maximized to choose candidate groups with the best fitness.

A case study is presented where the proposed approach is applied to a real social network. The social network was built with on-line available data extracted from a public R&D funding agency. Further, results were compared with current committees of the same agency. Results show that the proposed approach can generate committees that improve group independence compared to the current committee performances.

Assisting committee selection processes may be the greatest competitive advantage offered by the proposed approach, since we have proved that the best performance groups can be selected within seconds for a real scenario. Also, alternative group selections can be preferred by experts in charge for committee appointments. Moreover, this work is built upon a simple infrastructure because there are many genetic algorithm implementations, and social network manipulation software, that allow the implementation and the execution of the approach in standard hardware and software configurations. As practical usage, this approach can be implemented in recommendation processes to propose alternative group selections, or even group member replacements in order to improve group performances. Also, this approach can be used in opinion polls where there is a need to select less related respondents, such as focus groups.

Although this approach is presented as a simple alternative to the committee selection problem, there still are some limitations. These limitations include an underlying problem, which implies that the geodesic distances must be calculated between every node pair in the network. Another limitation of the geodesic distance as underlying measure is that distance between nodes from different components cannot be determined. Also, despite the proposed genetic algorithm returns the best performance solutions, it is still an approximation strategy to the global optimum. Finally, the proposed approach is intentionally designed for simple social networks with undirected and unvalued ties, therefore its application in other scenarios, such as complex networks, may require some modifications.

Future works aim to test the proposed approach in other domains that require committee member selection. Despite this approach uses a simple network representation, more complex committee member selection processes may include criteria other than the group independence, therefore future works may include multiple criteria in group selection for the committee member selection problem. Further, other optimization strategies could be evaluated, particularly for scalability scenarios. Moreover, a complex social network representation will allow to include other kinds of network properties, such as directed ties or node attributes.

References

Abbott, L. J., & Parker, S. (2000). Auditor selection and audit committee characteristics. *AUDITING: A Journal of Practice & Theory, 19*(2), 47–66. doi:10.2308/aud.2000.19.2.47.

Aksela, M. (2003). Comparison of classifier selection methods for improving committee performance. In T. Windeatt, & F. Roli (Series Eds.), *Multiple Classifier Systems, Lecture Notes in Computer Science: Vol. 2709* (pp. 84–93). Berlin, Heidelberg: Springer.

Aksela, M., & Laaksonen, J. (2006). Using diversity of errors for selecting members of a committee classifier. *Pattern Recognition, 39*(4), 608–623. doi:10.1016/j.patcog.2005.08.017.

Argamon-Engelson, S., & Dagan, I. (1999). Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research, 11*, 335–360. doi:10.1613/jair.612.

Ballester, C., Calvó-Armengol, A., & Zenou, Y. (2006). Who's who in networks. wanted: the key player. *Econometrica, 74*(5), 1403–1417. doi:10.1111/j.1468-0262.2006.00709.x.

Bhattacharyya, S., & Bickel, P. J. (2014). *Community detection in networks using graph distance* arXiv:1401.3915 [cs, stat], Netherlands.

Bock, H.-H., Day, W. H., & McMorris, F. (1998). Consensus rules for committee elections. *Mathematical Social Sciences, 35*(3), 219–232. doi:10.1016/S0165-4896(97)00033-4.

Borgatti, S. P. (2006). Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory, 12*(1), 21–34. doi:10.1007/s10588-006-7084-x.

Burt, R. S. (1978). A structural theory of interlocking corporate directorates. *Social Networks, 1*(4), 415–435. doi:10.1016/0378-8733(78)90006-0.

Cao, T., Wu, X., Wang, S., & Hu, X. (2011). Maximizing influence spread in modular social networks by optimal resource allocation. *Expert Systems with Applications, 38*(10), 13128–13135. doi:10.1016/j.eswa.2011.04.119.

Everett, M. G., & Borgatti, S. P. (2010). Induced, endogenous and exogenous centrality. *Social Networks, 32*(4), 339–344. doi:10.1016/j.socnet.2010.06.004.

Fishburn, P. C. (1981). Majority committees. *Journal of Economic Theory, 25*(2), 255–268. doi:10.1016/0022-0531(81)90005-3.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*(1), 35–41. doi:10.2307/3033543.

Freeman, L. C. (1993). Finding groups with a simple genetic algorithm. *Journal of Mathematical Sociology, 17*(4), 227–241. doi:10.1080/0022250X.1993.9990109.

Gehrlein, W. V. (1985). The Condorcet criterion and committee selection. *Mathematical Social Sciences, 10*(3), 199–209. doi:10.1016/0165-4896(85)90043-5.

Hadjitodorov, S. T., Kuncheva, L. I., & Todorova, L. P. (2006). Moderate diversity for better cluster ensembles. *Information Fusion, 7*(3), 264–275. doi:10.1016/j.inffus.2005.01.008.

Hwang, S.-Y., Wei, C.-P., & Liao, Y.-F. (2010). Coauthorship networks and academic literature recommendation. *Electronic Commerce Research and Applications, 9*(4), 323–334. doi:10.1016/j.elerap.2010.01.001.

Kolaczyk, E. D., Chua, D. B., & Barthélemy, M. (2009). Group betweenness and co-betweenness: inter-related notions of coalition centrality. *Social Networks, 31*(3), 190–203. doi:10.1016/j.socnet.2009.02.003.

Kuncheva, L. I. (2005). Using diversity measures for generating error-correcting output codes in classifier ensembles. *Pattern Recognition Letters, 26*(1), 83–90. doi:10.1016/j.patrec.2004.08.019.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning, 51*(2), 181–207. doi:10.1023/A:1022859003006.

Li, L., Zou, B., Hu, Q., Wu, X., & Yu, D. (2013). Dynamic classifier ensemble using classification confidence. *Neurocomputing, 99*, 581–591. doi:10.1016/j.neucom.2012.07.026.

Liberman, S., & Wolf, K. B. (1997). The flow of knowledge: scientific contacts in formal meetings. *Social Networks, 19*(3), 271–283. doi:10.1016/S0378-8733(96)00303-6.

Loewenberg, G., Patterson, S. C., & Jewell, M. E. (1985). *Handbook of legislative research* (1st ed.). Cambridge, MA: Harvard University Press.

Morgan, G. P., & Carley, K. M. (2011). Exploring the impact of a stochastic hiring function in dynamic organizations. In *Proceedings of BRIMS* (pp. 106–113).

Morgan, G. P., & Carley, K. M. (2014). Comparing hiring strategies in a committee with similarity biases. *Computational and Mathematical Organization Theory, 20*(1), 1–19. doi:10.1007/s10588-012-9130-1.

Preciado, P., Snijders, T. A. B., Burk, W. J., Stattin, H., & Kerr, M. (2012). Does proximity matter? Distance dependence of adolescent friendships. *Social Networks, 34*(1), 18–31. doi:10.1016/j.socnet.2011.01.002.

Shin, H., & Sohn, S. (2005). Selected tree classifier combination based on both accuracy and error diversity. *Pattern Recognition, 38*(2), 191–197. doi:10.1016/j.patcog.2004.06.008.

Shivdasani, A., & Yermack, D. (1999). CEO involvement in the selection of new board members: an empirical analysis. *The Journal of Finance, 54*(5), 1829–1853. doi:10.1111/0022-1082.00168.

Smith, J. E., & Eiben, A. E. (2008). Introduction to evolutionary computing. Springer.

Tao, H., Ma, X.-p., & Qiao, M.-y. (2013). Subspace selective ensemble algorithm based on feature clustering. *Journal of Computers, 8*(2). doi:10.4304/jcp.8.2.509-516.

Wang, C., Deng, L., Zhou, G., & Jiang, M. (2014). A global optimization algorithm for target set selection problems. *Information Sciences, 267*, 101–118. doi:10.1016/j.ins.2013.09.033.

Wang, X., & Wang, H. (2006). Classification by evolutionary ensembles. *Pattern Recognition, 39*(4), 595–607. doi:10.1016/j.patcog.2005.09.016.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. Cambridge University Press.

Westphal, J. D., & Zajac, E. J. (1995). Who shall govern? CEO/board power, demographic similarity, and new director selection. *Administrative Science Quarterly, 40*(1), 60. doi:10.2307/2393700.

Wi, H., Mun, J., Oh, S., & Jung, M. (2009a). Modeling and analysis of project team formation factors in a project-oriented virtual organization (ProVO). *Expert Systems with Applications, 36*(3, Part 2), 5775–5783. doi:10.1016/j.eswa.2008.06.116.

Wi, H., Oh, S., Mun, J., & Jung, M. (2009b). A team formation model based on knowledge and collaboration. *Expert Systems with Applications, 36*(5), 9121–9134. doi:10.1016/j.eswa.2008.12.031.

Zheng, Z. (1998). Naive Bayesian classifier committees. In *Naive Bayesian classifier committees*. Springer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.3003&rep=rep1&type=pdf.

Zouari, H., Heutte, L., & Lecourtier, Y. (2005). Controlling the diversity in classifier ensembles through a measure of agreement. *Pattern Recognition, 38*(11), 2195–2199. doi:10.1016/j.patcog.2005.02.012.