Brokerage Discovery in Social Networks

Luís Cavique,
LabMAg and Univ. Aberta, Portugal
lcavique@uab.pt

Abstract

In social networks two types of measures can be identified, the structural measures and community structure based on diameter and centrality. The community structure usually deals with network partition into communities. The key idea of this work is to explore the concept of strong and weak ties by finding brokers within communities. The strict partition problem is relaxed into a bi-objective set covering problem with k-cliques which allows over-covered and uncovered nodes. The information extracted from social networking goes beyond cohesive groups, allowing the finding of brokers that interact between groups.

Keywords: social networks, community partition, brokerage

1. Introduction

In social networks the set of vertices (or nodes) corresponds to the 'actors' (i.e. people, companies or social actors) and the set of edges corresponds to the 'ties' (i.e. relationships, associations or links), quite similar to the representation in graph theory.

In the late 1960s, while working on his Ph.D., Mark Granovetter interviewed people who had recently changed jobs, in order to come to a conclusion as to how they had discovered their new jobs. Surprisingly, he realized that the information about the new jobs had come from distant acquaintances instead of close friends. The concept of strong and weak ties (Granovetter 1973) introduced a novel principle in social networks. Weak ties are valuable because they are more likely the source of novel information, surprise and openness to new worlds. On the other hand, strong ties intensify group cohesion and the persistence of group identity. This resulted in the Triadic Closure property, which establishes that if the node has strong ties to two neighbors, these neighbors must have at least a weak tie between them.

Following this problematic, Burt (1992) developed a complementary approach coined Structural Holes, referring to the lack of links in a connected organization. He also introduced the concept of brokerage, signifying nodes that connect two dense groups.

Although there are different methods to find network partitions, the specific discovery of brokers between partitions is scarce or inexistent. In this work we are interested in finding, not only the communities, but also the elements that are among communities, the brokers, given their importance for the whole network.

In section 2 we present some related work about social networks and partition approaches. In section 3 we present the algorithm to discover brokerage in social networks. In section 4 the computational results are presented. Finally, in section 5 we draw some conclusions.

2. Related work

In this section some related work about social networks and networks partition concepts is presented.

2.1. Social Network Concepts

For many years, the centrality of the authors in networks has been an important issue in social network analysis. The central node (or hub) can be found using different measures: degree centrality, betweenness centrality, closeness centrality or eigenvector centrality.

A diverse subject is related to the bounds of each network component. The bridge and broker definitions can be stated as follows:

- bridge: is an edge whose removal increases the number of components in the network;
- broker or cut-vertex: is a vertex whose deletion increases the number of components in the network;
- local-bridge: is an edge whose removal increases significantly the distance between the components;
- local-broker: is a vertex whose deletion increases significantly the distance between components.

The brokers have some similarities with actors who score high in terms of centrality, the hubs. However, the "centrality" of the brokers lies between different communities instead of actors of the same group.

Given Figure 1, a graph with edges A-B and B-C, the Triadic Closure Property (Granovetter 1973) comes from the fact that the A-C edge has the effect of closing the third side of the triangle. The property is based on the fact that if the two people have a friend in common, it is probable that they will become friends in the future (Easley, Kleinberg 2010).

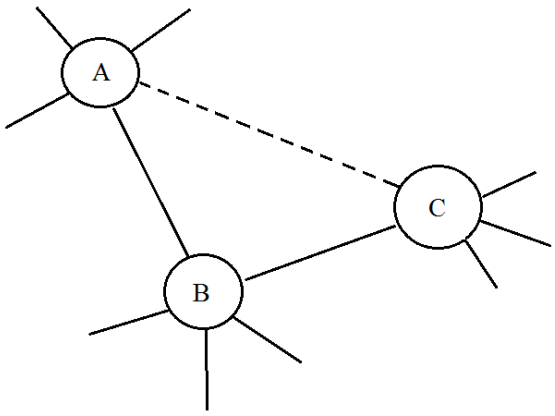

Fig 1. Triadic Closure

In graph theory, the k^{th} power of graph G returns a new graph G^k where each pair of vertices is adjacent when their distance in G is at most k. Graph G^2 is equivalent to the Triadic Closure, where a new A-C edge is inserted because the distance between A and

C is equal to 2. In this approach strong ties correspond to friends (solid lines), and weak ties correspond to acquaintances (broken line).

Another way to interpret the absence of relationship A-C is named Structural Hole (Burt 1992). This approach replaces the concept of *closure* by *brokerage*, and the broker interactions between different groups are highlighted.

2.2. Partition Problem

For large networks the visualization is still an ongoing project, so the quantitative measures are very useful. Roughly, we can differentiate two types of measures, the structural measures and the community structure. The structural measures include the centrality measures, the measures based on the vertex degree and the measures based on the diameter, that is, the maximum distance between two vertices. The community structure usually deals with the network partition into communities or similarity groups.

Community or group can be defined as a set of nodes with similarity. A partition is a sub-division of a graph into groups of vertices such that each vertex is assigned to one group. The mathematic formulation of the Partition problem can be stated as follows, where matrix $[a_{i,j}]$ keeps the information about different communities and for each variable *x*, a cost can be associated by using a vector c_j:

$$\text{minimize } f=\sum c_j.x_j$$

$$\text{subject to } \sum a_{i,j}.x_j=1$$

$$\text{and } x_j \in \{0,1\} \qquad j=1,\ldots,n$$

As stated, the constraint with the equality is very restricted. So, this can lead to many problems, for instance when a node is shared by two or more communities.

One of the first studies is given by the Kernighan, Lin (1970) algorithm, which finds a partition of the nodes by dividing into two disjoint subsets A and B of equal size, such that the sum of the weights of the edges between nodes in A and B is minimized.

Recent studies, based on physics, introduced the concept of clique percolation (Derenyi, Palla, Vicsek 2005), where the network is viewed as a union of cliques.

The Girvan-Newman (2002) method has been applied in recent years to social networks in particular (Easley, Kleinberg 2010). This method successively deletes edges of high betweenness, and then recalculates all betweenness, breaking each component into smaller components.

A more relaxed problem that allows a node to share two components is the Set Covering problem. In the mathematic formulation the signal of the constraint is replaced for equal or greater instead of equal, allowing the existence of over-covered nodes.

In the following section, the Set Covering problem will be developed rather than the Partition problem.

3. The Algorithm

There are several partition algorithms but few studies about the linkage between them, especially issues related to the brokerage. In this paper we present a new approach that takes into account the common elements between partitions (over-covered) and elements that do not belong to any partition (uncovered), formulated as a new bi-objective set covering problem. Firstly, the set covering with k-cliques will be presented (Cavique, Mendes, Santos, 2009). Secondly, the problematic of the over-covered and uncovered nodes is highlighted (Cavique, Mendes, Santos, 2013) and finally a new bi-objective set covering problem with k-cliques is shown.

3.1. Set covering problem with k-clique

Given an undirected graph G=(V, E), where V denotes the set of vertices (or nodes) and E, the set of edges (or arcs), graph $G_1= (V_1, E_1)$ is called a sub-graph of G, if $V_1 \subseteq V$, $E_1 \subseteq E$ and for every edge $(v_i, v_j) \in E_1$, the vertices $v_i, v_j \in V_1$. A sub-graph G_1 is said to be complete, if there is an edge for each pair of vertices. Since the clique structure is very constrained to represent social networks, Luce (1950) introduced the distance base cohesion groups called a k-clique, where k is the maximum path length between each pair of vertices. To find all the maximal k-cliques in the graph, we use the k^{th} power of graph G in such a way that we can reuse an already well-known algorithm, the maximum clique algorithm. The transformation process adds edges to reach length *k* between every pair of nodes.

The Maximum Clique is a NP-hard problem that aims to find the largest complete sub-graph in a given graph. In this approach, we intend to find a lower bound for the maximization problem, based on the heuristics proposed by Johnson (1974) and in the meta-heuristic that uses Tabu Search developed by Soriano and Gendreau (1996).

Part of the work described in this section can also be found in Cavique, Rego and Themido (2002) and Cavique and Luz (2009). We define A(S) as the set of vertices that are adjacent to vertices of a current solution S. Let n=|S| be the cardinality of clique S and $A^k(S)$ the subset of vertices with k arcs incident in S. A(S) can be divided into subgroups $A(S) = \cup A^k(S)$, k=1,..,n. The cardinality of the vertex set |V| is equal to the sum of the adjacent vertices A(S) and the non-adjacent ones $A^0(S)$, plus |S|, resulting in $|V|= \Sigma|A^k(S)|+n$, k= 0,.., n. For a given solution S, we define a neighborhood N(S) if it generates a feasible solution S'. In this work we are going to use three neighbourhood structures. For the next flowchart consider the following notation:

$N^+(S) = \{S' : S' = S \cup \{v^i\}, v^i \in A^n(S)\}$
$N^-(S) = \{S' : S' = S \setminus \{v^i\}, v^i \in S\}$
$N^0(S) = \{S' : S' = S \cup \{v^i\} \setminus \{v^k\}, v^i \in A^{n-1}(S), v^k \in S\}$

where S is the current solution, S*, the highest cardinality maximal clique found so far, T, the tabu list and N(S), the neighborhood structures. Finding a maximal clique in graph G^k is the same as finding a maximal k-clique in a graph G. To generate a large set of maximal k-cliques, a multi-start algorithm is used, which calls the Tabu Heuristic for Maximum Clique Problem.

Algorithm 1 - The Tabu Heuristic for the Maximum Clique Problem
Input: graph G^k, complete sub-graph S
Output: clique S*
1. T=∅; S*=S;
2. while not end condition
2.1. if ($N^+(S)\backslash T \neq$ null) choose the maximum S'
2.2. else if ($N^0(S)\backslash T \neq$ null) choose the maximum S'; update T
2.3. else choose the maximum S' in $N^-(S)$; update T
2.4. update S=S'
2.5. if (|S|>|S*|) S*=S;
3. end while;
4. return S*;

Following the rule of three degrees of influence, i.e., our friends' friends' friends affect us, proposed by Christakis and Fowler (2013), we are going to use the 3-cliques, equivalent to a power graph G^3. In Figure 2, a pair of 3-cliques partially covers the given graph.

nodes	x_1	x_2	…	x_n
1	1			
2	1			
3	1			
4	1			
5	1			
6	1			
7	1	1		
8	1	1		
9		1		
10		1		
11		1		
12				
13		1		
14		1		
15		1		
16		1		
17		1		
18		1		

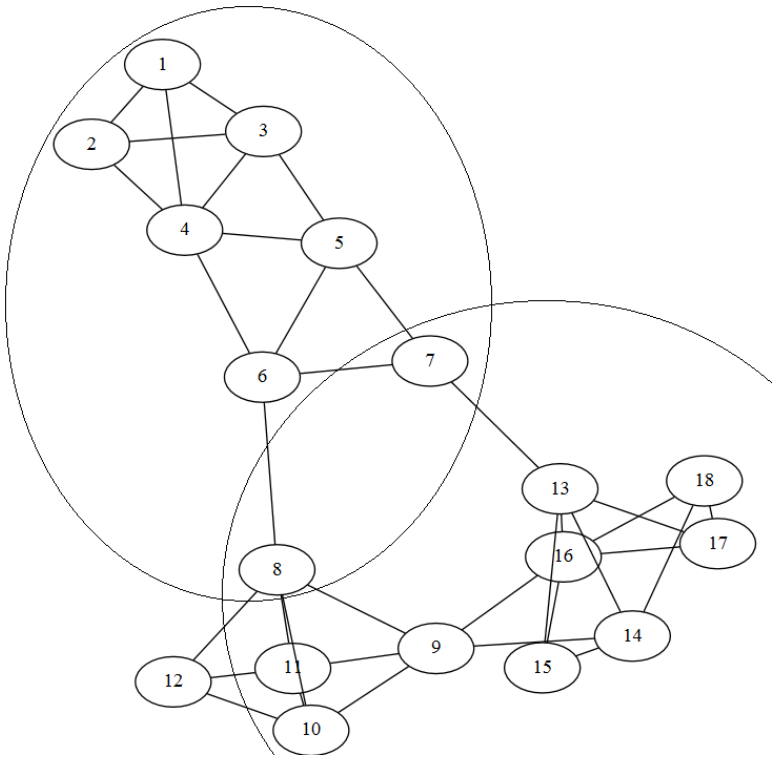

Fig 2. The pair of 3-cliques covers the graph partially

Matrix $[a_{i,j}]$ represents the *m* nodes which must be covered and *n* columns, where each column is a k-clique. Also in Figure 2 on the left, $[a_{i,j}]$ shows that nodes 7 and 8 are over-covered and node 12 is uncovered by any column. The partial solution of the example is {1, 2}.

The optimization problem that finds the minimum number of columns that covers all the rows is the Set Covering problem. For each attribute *x*, a cost can be associated by using

a vector c_j, allowing a cost differentiation among attributes. The matrix and the cost vector are then used in the set covering problem, defined as:

$$\text{minimize } f = \sum c_j.x_j$$
$$\text{subject to } \sum a_{i,j}.x_j \geq 1$$
$$\text{and } x_j \in \{0,1\} \qquad j=1,\ldots,n$$

The Set Covering problem is a very well-studied problem in Combinatorial Optimization, with many computational resources which implement quasi-exact algorithms and heuristic approaches.

3.2. Bi-objective Set Covering problem with k-cliques

As the set covering problem is a relaxation of the partition problem, our goal is to create a more relaxed problem which also allows uncovered nodes in order to find solutions as shown in Figure 2. Table 1 shows some data extracted from figure 2:

- the cost of the solution, for unitary c_j, is equal to the number of columns, and is also equal to the number of communities in the social network;
- in the solution 17 covered nodes can be found;
- the over-covered nodes are 2 and they represent the brokers in the social network;
- node 12 is not covered, and represents outliers in the social network;
- the over-covered nodes and the uncovered nodes must be removed from the final solution and should be seen as errors, so the variable error coverage is the result of the over-covered plus the uncovered nodes.

Table 1. Information extracted from Figure 2

solution cost |x| = number communities	covered nodes	over-covered = brokers	uncovered	error_coverage = over_covered + uncovered
2	17	2	1	3

In this paper, we propose a trade-off between the minimization of the columns and the minimization of the error coverage.

Variable *x* represents the column set and has the same meaning as the previous Set Covering problem formulation. Variable *x* should be minimized in order to reduce the cost and to minimize the error coverage.

Our approach can be formulated as a Bi-objective Set Covering problem with k-cliques, such that:

(1) minimize $f_1 = \sum_{j=1}^{n} c_j \cdot x_j$

(2) minimize $f_2 = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{i,j} \cdot x_j - 1|$

(3) subject to $\sum_{j=1}^{n} a_{i,j} \cdot x_j \geq 0$

(4) $x_j \in \{0,1\} \quad j = 1, \ldots, n$

In a bi-objective formulation two objective functions must be stated, f_1 and f_2. The objective function f_1 (1) has the same meaning of the objective function *f* in the original Set Covering problem, which is to minimize the cost of the chosen columns.

In the formulation constraint (3) uses the sign "≥0" allowing uncovered nodes. To balance this relaxation the objective function f_2 (2) minimizes the sum of the uncovered nodes and the over-covered ones.

In multi-objective optimization the dominance concept is central. An objective vector $u=(u_1,\ldots,u_n)$ dominates $v=(v_1, \ldots, v_n)$, denoted $u > v$, if and only if, $u_i \geq v_i$: $\forall i$, and at least one component v is smaller, $u_i > v_i$: $\exists i$. A solution is non-dominated (or Pareto solution), if and only if, there is no solution that dominates it.

Figure 3 shows a bi-objective minimization problem. The set of all Pareto solutions is also called the Pareto Solution frontier or the Efficient Solutions frontier. The black circles of the Pareto solution dominate the solutions represented by triangles. The Efficient frontier is given by the line that includes all the Pareto solutions. In axis f_1 represents the sum of the costs of the chosen variables (objective 1) and f_2 shows the minimization of the error coverage (objective 2).

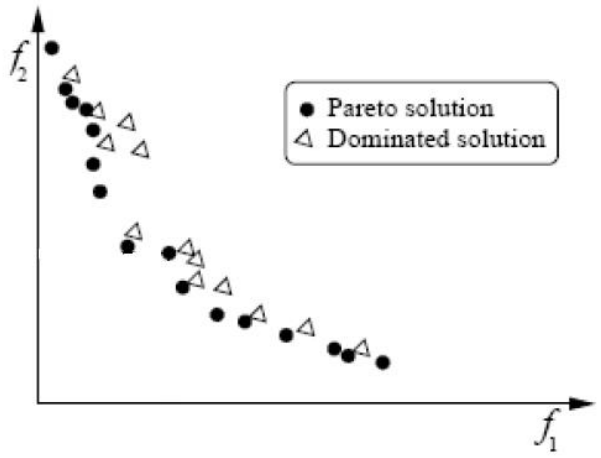

Fig.3. Bi-objective minimization problem

Given a dataset with social network G, the Heuristic for the Bi-objective Covering with k-cliques reuses a columns generation technique and can be specified as follows:

Algorithm 2 - The Heuristic for the Bi-objective Covering with k-cliques
Input: graph G
Output: the Pareto frontier $P=\{p^1,p^2,\ldots\}$
1. $P=\varnothing$
2. While not end condition
2.1. repeat column generation by adding k-cliques from $a_{i,j}$
2.2. find best (p)
2.3. if (p) is non-dominated then $P=P\cup p$
2.4. destructive phase by removing k-cliques from $a_{i,j}$
3. end while;
4. return P

In the constructive phase (Algorithm 2, 2.1) a large set of maximal k-cliques are generated, where a multi-start algorithm is used, which calls the Tabu Heuristic for Maximum Clique Problem. Then to find vector *x* a greedy heuristic is used to cover the nodes partially.

Meanwhile, in step (Algorithm 2, 2.3) the non-dominated solutions are added to the Pareto frontier. After the addition of a new solution, if the new solution dominates any other solution, the dominated solutions are removed from the Pareto frontier.

Finally, in the destructive phase (Algorithm 2, 2.4) some k-cliques are removed in order to find new solutions in the Pareto frontier.

3.3. The "best" solution

Given the set of solutions of the Pareto frontier, to choose a subset is a decision problem. The decision maker, or in this particular case the social network analyst, should decide based on his/her tacit or explicit knowledge.

Following the running example with the graph shown in this section, Table 2 presents three possible solutions. Figure 2 represents the second solution, of Table 2, with 2 over-covered nodes and 1 uncovered one.

Table 2. Solutions of the running example

|x|	covered	over-covered	uncovered	error coverage	ratio= errors / covered
1	8	0	10	10	125%
2	17	2	1	3	18%
3	18	8	0	8	dominated sol.

In Figure 4, the graphical representation of the three solutions shows clearly that the third solution is dominated by the second one. So, the Pareto frontier only includes the first and the second solutions.

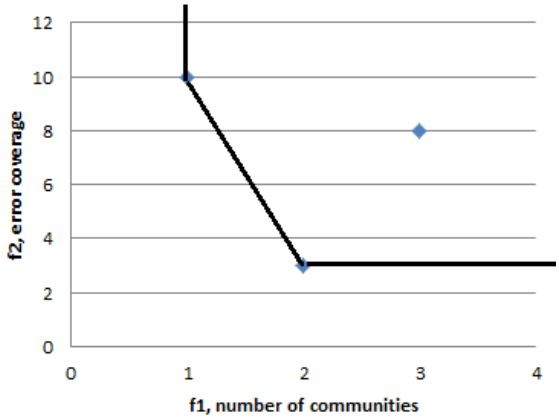

Fig. 4. Pareto frontier with two solutions

A way to solve this decision problem, using Value Analysis concepts, is finding the ratio of the error coverage by covered nodes. By identifying the smaller ratio a balanced solution is achieved. For the given example in Figure 2, which corresponds to the second solution, a ratio of 18% is found.

4. Computational Results

To validate the proposed method, two groups of datasets were used, the Erdös graphs and some clique DIMACS (1995) benchmark instances. In the Erdös graphs, each node corresponds to a researcher, and two nodes are adjacent if the researchers published together. The graphs are named "erdos-x-y", where "x" represents the last two digits of the year that the graphs were created, and "y", the maximum distance from Erdös to each vertex in the graph. The second group of graphs contains some clique instances from the second DIMACS challenge. These include the "brock" graphs, which contain cliques "hidden" within much smaller cliques, making it hard to discover cliques in these graphs. The "c-fat" graphs are a result of fault diagnosis data.

We select 4 brock datasets, 3 c-fat datasets and 3 erdos datasets. For each dataset we tested for k=1 to k=7, completing 70 runs. For large values of *k* only one community was found and there were no brokers. We extracted a sample of the runs to illustrate the computational results, in Table 3, where for each dataset, the number of communities, the total number of nodes, the nodes that only belong to one community (well-covered), the over-covered nodes, the uncovered nodes and other nodes with no links are presented.

Table 3. Sample of the most interesting results of the 70 runs

Dataset	Number Communities	Total number nodes	Well covered nodes	Over covered nodes	Un-covered nodes	Others
example Fig.2, k3	2	18	15	2	1	0
brock22-k1	9	200	51	11	39	99
c-fat22-k3	5	200	178	22	0	0
erdos97-k1	4	472	63	19	35	355

In Figure 5, the nodes represent the communities and the edges the number of brokers between each pair of communities. The partition of graph c-fat-22-k3 shows poor connections among the nodes, the brokers are very well identified in two pairs of communities with 11 brokers each. In Table 3 the over-covered nodes of 22 correspond to the 2x11 brokers.

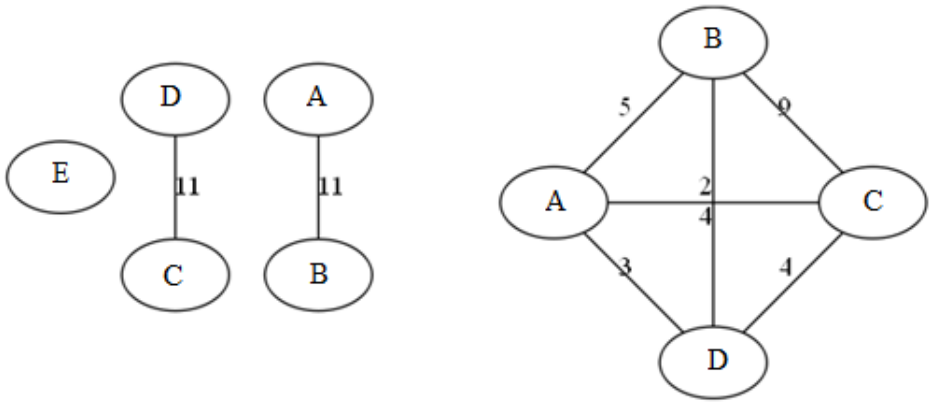

Fig. 5. Communities and respective number of brokers of the partitions of c-fatt22-k3 (on the left) and erdos-97-k1 (on the right)

Again in Figure 6, the partition of graph erdos-97-k1 shows a strong connectivity among communities, and an average of 4.5 brokers between each pair of communities. With this graph, the match with the values of Table 3 is more difficult to establish, because there are brokers that belong to three communities.

To overcome this drawback the graph must be seen with more detail. In Figure 7 on the left, 15 brokers can be identified (by numbers) which are linked to 2 communities. On the right of the figure, 4 brokers (301, 354, 405 and 441) are linked with 3 communities, using a different representation, where the brokers are identified and contained in boxes. So we have 15+4 brokers or over-covered nodes as mentioned in Table 3.

When the brokers link two communities the number of edges is equal to the number of brokers. On the other hand, when the brokers link three communities, the number of edges is times three in a graph representation. So, 15+4x3 is equal to the sum of the edges in Figure 6 on the right.

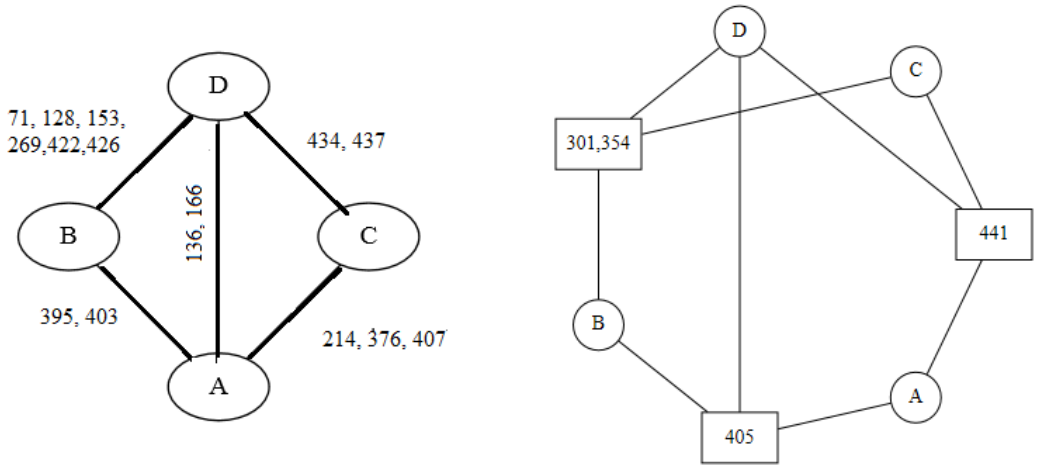

Fig. 6. Brokers of erdos-97-k1: brokers linked in 2 communities (on the left) and brokers linked with 3 communities (on the right)

5. Conclusions

The social networks analysts often referred the problematic of the strong and weak ties and the associated issue of the brokerage. There is a lack of automatic procedures to find, not only the communities but also the actors that play within communities.

The community partition can be relaxed for the Set Covering problem allowing brokerage (over-covered nodes). However, the large number of intersections is not compatible with a good visualization. So we created a model that allows not only over-covered nodes, but also uncovered nodes. With this purpose in mind we defined community as a k-clique and the community partition as a bi-objective Set Covering problem with k-cliques which allows over-covered and uncovered nodes. The uncovered nodes are called outliers and the over-covered nodes are called brokers.

The data extracted from social networking goes beyond the structure of communities, allowing the finding of the brokers that interact between groups. In this paper we show how to find the communities and identify their related brokers.

Acknowledges

The author would like to thank the FCT support in the Funding of Strategic Projects with Public Interest promoted by Associated Laboratories and RD Units, PEst-OE/EEI/UI0434/2011.

REFERENCES

Burt R.S. (1992), Structural Holes: The Social Structure of Competition, Harvard University Press.

Cavique L., A.B. Mendes, J.M.A. Santos (2009), An Algorithm to Discover the k-Clique Cover in Networks, in Progress in Artificial Intelligence, L. Seabra Lopes et al. (Eds.): EPIA 2009, LNAI 5816, Springer-Verlag Berlin Heidelberg, pp. 363–373.

Cavique L., A.B. Mendes, J.M.A. Santos (2013), Clique Communities in Social Networks, in Quantitative Modelling in Marketing and Management, World Scientific Publisher, edited by Luiz Moutinho and Kun-Huang Huarng.

Cavique L., C.J. Luz (2009), A heuristic for the stability number of a graph based on convex quadratic programming and tabu search, Journal of Mathematical Sciences, 161 (6), pp. 944-955.

Cavique, L., Rego, C.,Themido, I. (2002), A scatter search algorithm for the maximum clique problem, In Ribeiro, C. e Hansen, P. (Eds.) Essays and Surveys in Meta-heuristics. Kluwer Academic Pubs.: Dordrecht, The Netherlands, pp 227-244.

Christakis N., J. Fowler (2011), Connected: The surprising power of networks and how they shape our lives, Back Bay Books/Little, Brown and Company, Hachette Book Group.

DIMACS (1995), Maximum clique, graph coloring, and satisfiability, Second DIMACS implementation challenge, URL http://dimacs.rutgers.edu/ Challenges/.

Derenyi I., G. Palla, T. Vicsek, (2005), Clique Percolation in Random Networks, Physical Review Letters, vol. 94(16), pp. 160202.

Easley D., J. Kleinberg, 2010, Networks, Crowds and Markets: Reasoning About a Highly Connected World, Cambridge University Press.

Girvan M., M.E.J. Newman (2002), Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA, 99(12), pp. 7821–7826.

Granovetter M. (1973), The strength of weak ties, American Journal of Sociology, 78, pp.1360–1380.

Johnson D.S. (1974), Approximation algorithms for combinatorial problems, Journal of Computer and Systems Sciences, 9 (9), pp. 256-278.

Kernighan, B.W., Lin, Shen (1970), An efficient heuristic procedure for partitioning graphs, Bell Systems Technical Journal, 49, pp. 291–307.

Luce, R.D. (1950), Connectivity and generalized cliques in sociometric group structure, Psychometrika, 15 (15), pp. 159-190.

Soriano P., Gendreau M. (1996), Tabu search algorithms for the maximum clique, In: Johnson, D.S.; Trick, M.A. (Eds.). Clique, Coloring and Satisfiability, Second Implementation Challenge DIMACS, American Mathematical Society, pp. 221-242.