Analyzing the formation of groups in a network adapting the modularity concept

Ana Carolina Wagner Gouveia de Barros
Fernanda Castello Branco Madeu
Moacyr Alvim Silva
Walter Wagner Carvalho Sande

SUMARY

- MOTIVATION
- CITATION NETWORKS
- STRATEGIC BEHAVIOR
- PROPOSED STUDY
- METHODOLOGY
- RESULTS
- CONCLUSIONS

MOTIVATION

- The problem of cutting a graph into “useful” subgraphs is classical in graph theory – relevant research field.
- Graphs representing data are usually directed.
- Different reasons and motivations for dividing graphs into smaller components:
  - they naturally arise as a consequence of simple interactions among people and do not require complicated mechanisms to be obtained and maintained (practical);
  - they have some useful properties, such as high internal connectivity, low path length among nodes and high robustness, which are of the most importance in real applications

(DUGUË and PEREZ, 2015), (GUIMERA and NUNES AMARAL, 2005), (NICOSIA et. al., 2009).

MOTIVATION

- A lot of methods to solve this – “clustering algorithms” to optimize a graph structure – guarantee certain desired features.

- Latest studies – algorithms not very useful for explaining partitioning patterns observed in social networks, such as the arising of “communities”, “groups” or “clubs”.

(NICOSIA et. al., 2009).

MOTIVATION

COMMUNITY (no precise definition):

"a community is a subgraph containing nodes which are more densely linked to each other than to the rest of the graph or, equivalently, a graph has a community structure if the number of links into any subgraph is higher than the number of links between those subgraphs"

(NEWMAN and GIRVAN, 2004)

MOTIVATION

- Real–life communities -> groups of strongly connected nodes (people in a football club, authors in a co–authorship paper, colleagues studying in the same school, journals that cites each other).

- Usually nodes in a community know each other – ↑ probability for two nodes to have a neighbor in common

(NEWMAN and GIRVAN, 2003), (NEWMAN, 2006),(NEWMAN and JUYONG, 2003),(NICOSIA et. al., 2009).

SOCIAL NETWORKS

- Growth dynamics – **preferential attachment** – more central nodes have a greater power of attraction for new connections;

- **Directional** character;

- Links depend on the connection degree of the nodes;

(SANDE, 2016)
(BARROS, 2016)

SOCIAL NETWORKS

- Nodes with higher degree have more centrality;
- Centrality gain is measured in degree of entrance – great challenge for nodes with less centrality (peripheral nodes);
- Center-periphery structure.

(SANDE, 2016)

STRATEGIC BEHAVIOR

- Creation of new links between peripheral nodes – contradicting preferential attachment.
- Peripheral nodes starts linking to each other to increase their centrality and consequently increasing their importance;
- These nodes continue to be seen by the rest of the network as ordinary nodes;
- Strategic group - not necessarily form a community; (inserted in the community as peripheral nodes);
- Modification of the methods for communities' identification so that it is possible to identify the emergence of strategic groups.

(SANDE, 2016)

MODULARITY

- G(*n*,*m*), *A*=adjacency matrix, k_v = degree of vertex *v;*
- *R* groups (communities);
- *S* is the matrix with elements S_{vr} = 1 if *v* belongs to group *r* and zero otherwise;
- 2 groups (strategic and non-strategic) in a DIRECTED network:

$$\boxed{B_{vw} = A_{vw} - \frac{k_v^{in} k_w^{out}}{m}}$$

and

$$\boxed{Q = \frac{1}{m} \sum_{vw} B_{vw} s_v s_w}$$

$$\boxed{Q = \frac{1}{m} S^T B S}$$

(NEWMAN, 2006)

STOCHASTIC BLOCK MODEL (SBM)

- Takes the following parameters:
  - the number *n* of vertices;
  - a partition of the vertex set $\{1,...,n\}$ into disjoint *R* subsets $\{C_1,..., C_R\}$, called communities;
  - a symmetric RxR matrix P of edge probabilities.
- The edge set is then sampled at random as follows: any two vertices are connected by an edge with probability P_{ij}

(ABBE, 2016)

PROPOSED STUDY

- To generate a random graph with two groups (strategic and non-strategic) through the SBM and calculate the "submodularity" to confirm the strategic behavior can not be identified by the modularity concept;

- Also, observe photographs of a social simulated network at different time intervals to verify the increase and/or drop of links within and between the groups, analyzing the communities in pairs, to verify the strategic behavior.

PROPOSED STUDY

- Two values of “submodularity” are proposed here: Q_1 and Q_2
  - Where Q_1 refers to links inside the strategic group, and
  - Q_2 refers to the links between the two groups (directed from group 2 to group 1).

$$\boxed{Q_1 = \frac{1}{m} s_1^T B s_1}$$

$$\boxed{Q_2 = \frac{1}{m} s_1^T B s_2}$$

PROPOSED STUDY

$$s_{1Xn} = [1\ 1\ 0\ 0 \ldots 1]$$

- s_1 is define as:
  - s_{1v} = 1, if node *v* is starategic (belongs to group 1)
  - s_{1v} = 0, if node *v* is non-strategic
- s_2 is define as:
  - s_{2v} = 1, if node *v* is non-strategic (belongs to group 2)
  - s_{2v} = 0, if node *v* is strategic

$$Q_1 = \frac{1}{m} s_1^T B s_1$$

$$Q_2 = \frac{1}{m} s_1^T B s_2$$

PROPOSED STUDY

- Simulations with: 400 and 2000 nodes;
- 2 groups: strategic and non-strategic;
- Inputs:
  - Matrix of probabilities (P);
    - Probabilities of nodes to connect inside a group and between groups (2x2)
  - Partition vector (c);
    - Indicates if a node is strategic or non strategic (1xn)
- Output: Adjacency Matrix (A).

PROPOSED STUDY

- Generate a network through SBM;
- Based on A, calculate the values of Q_1 and Q_2;
- Calculate other 3 matrices A: random, strategic group and normal community;
- Calculate the following reasons:

PROPOSED STUDY

$$\frac{c_{21}(t)}{c_{11}(t)} > \frac{c_{21}(t+\Delta t)}{c_{11}(t+\Delta t)}$$

$$\frac{c_{12}(t)}{c_{22}(t)} \sim \frac{c_{12}(t+\Delta t)}{c_{22}(t+\Delta t)}$$

$$\alpha = \frac{c_{21}(t)}{c_{11}(t)}$$

$$\alpha^{+} = \frac{c_{21}(t+\Delta t)}{c_{11}(t+\Delta t)}$$

$$\beta = \frac{c_{12}(t)}{c_{22}(t)}$$

$$\beta^{+} = \frac{c_{12}(t+\Delta t)}{c_{22}(t+\Delta t)}$$

$$R_1 = \frac{\alpha}{\alpha^{+}}$$

$$R_2 = \frac{\beta}{\beta^{+}}$$

Where:

c_{ij} = number of links from group j to i. (i, j = 1,2)

PROPOSED STUDY

Expected results:

Strategic group: $R_1 > 1$ and $R_2 \sim 1$

Community: $R_1 > 1$ and $R_2 > 1$

METHODOLOGY

1. Network generated with SBM adapted for induction of strategic behavior.
   – Calculate the values of Q_1 and Q_2.

2. Network comparison: evolution analysis of a network from aleatory configuration to different situations:
   – SBM generating two normal communities;
   – SBM adapted to generate two groups – one of them with strategic behavior.
      • Calculate the reasons R_1 and R_2.

3. Application of step (2) on a network generated by simulation (e.g. citation network) with strategic behavior[1].

[1] (SANDE, 2016)

METHODOLOGY I

- Matlab code for SBM to generate A;
- Network with n= 400 and n=2000 nodes;
- Results were similar for both networks;

First step:

- Generate partition vector C:
  - Inputs: *a* = percentage of the nodes that are strategic (*a*=0:0.1:1);
  - C_{1xn} – the first a% of the nodes belongs to strategic group 1 and the others to non-strategic group 2.

METHODOLOGY I

Second step:

- Generate A:
  - Define matrix P;
  - Enter with: C, P, directed(true);

Third step:

- Varies the configuration of the network changing the position of the elements in the partition vector C;
- Each new vector is calculated with a similarity degree *x* in relation to the first vector C proposed (*x*=0:0.1:1);
- For each new vectors C, calculate the values of Q_1 and Q_2;

METHODOLOGY

Fourth step:

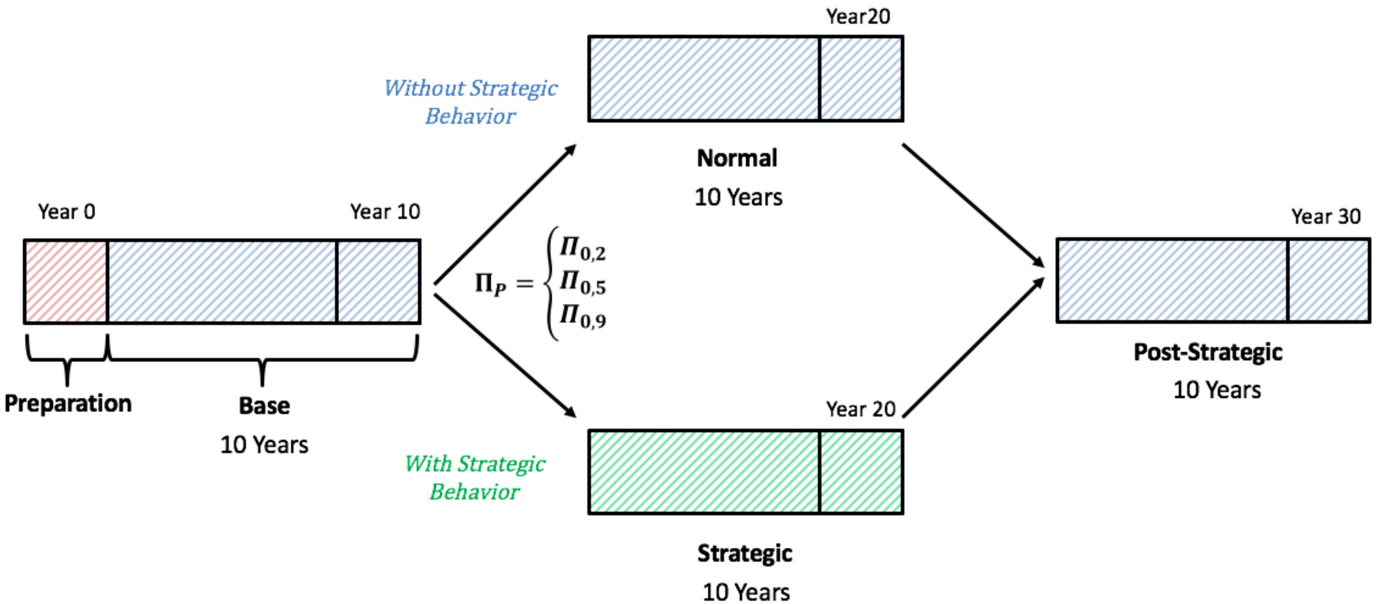
- Generate A aleatory, A with 2 communities and A with strategic group: A, A_{com} and A_{str}

Fifth step:

- Calculate the reasons R_1 and R_2 for the A_{com} and A_{str} .

METHODOLOGY

- Matlab code for a simulated citation network to generate A;
- Network with n = 400 journals (nodes);
- Strategic behavior: 20% of the nodes (journals).



(SANDE, 2016)
(BARROS, 2016)

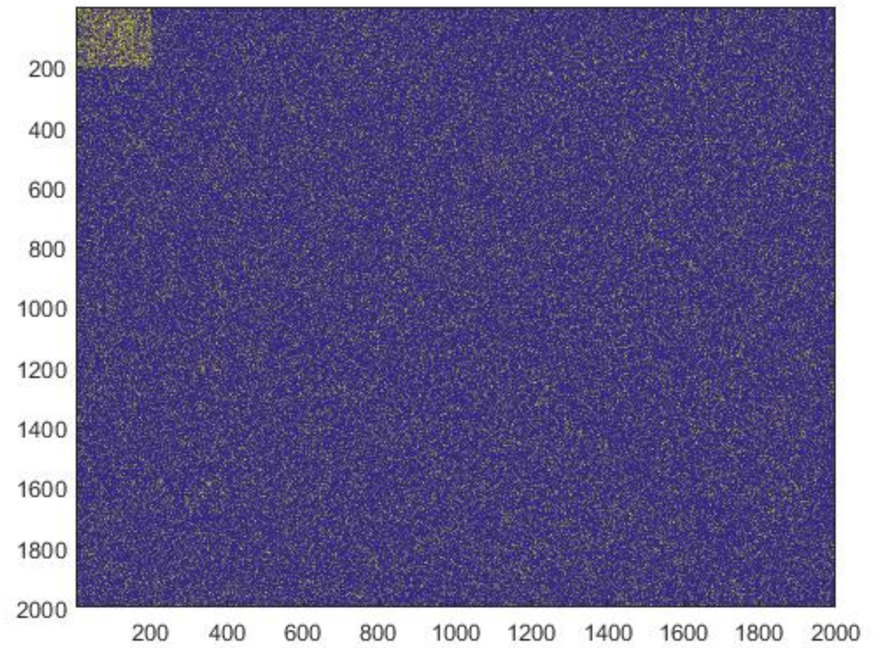
METHODOLOGY

- Plot the graphics of Q_1 and Q_2 for each value of *a* (varying the number of strategic nodes);
- Values of Q_1 and Q_2 in the same graphic;
- Plot the values of R_1 and R_2 in same graphic for a simulated citation network;

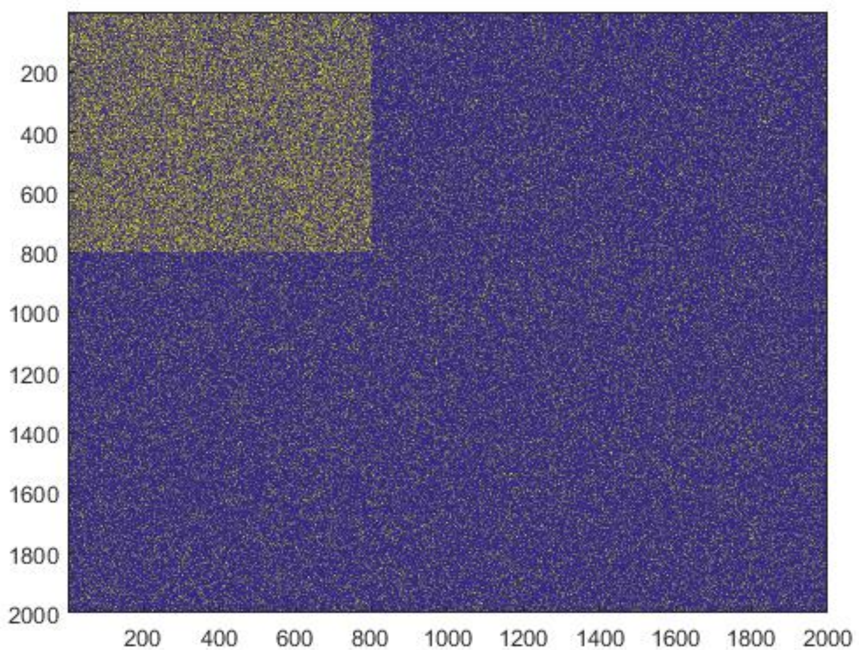
IT WAS FOUND THAT…

Matrix A for values of a=0.1, 0.4 and 0.9.

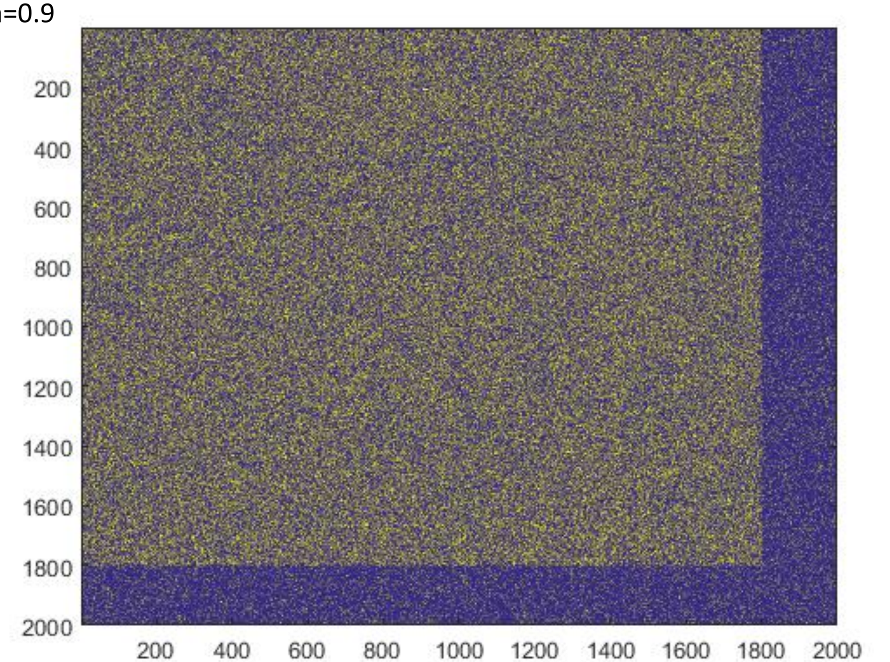
a=0.1

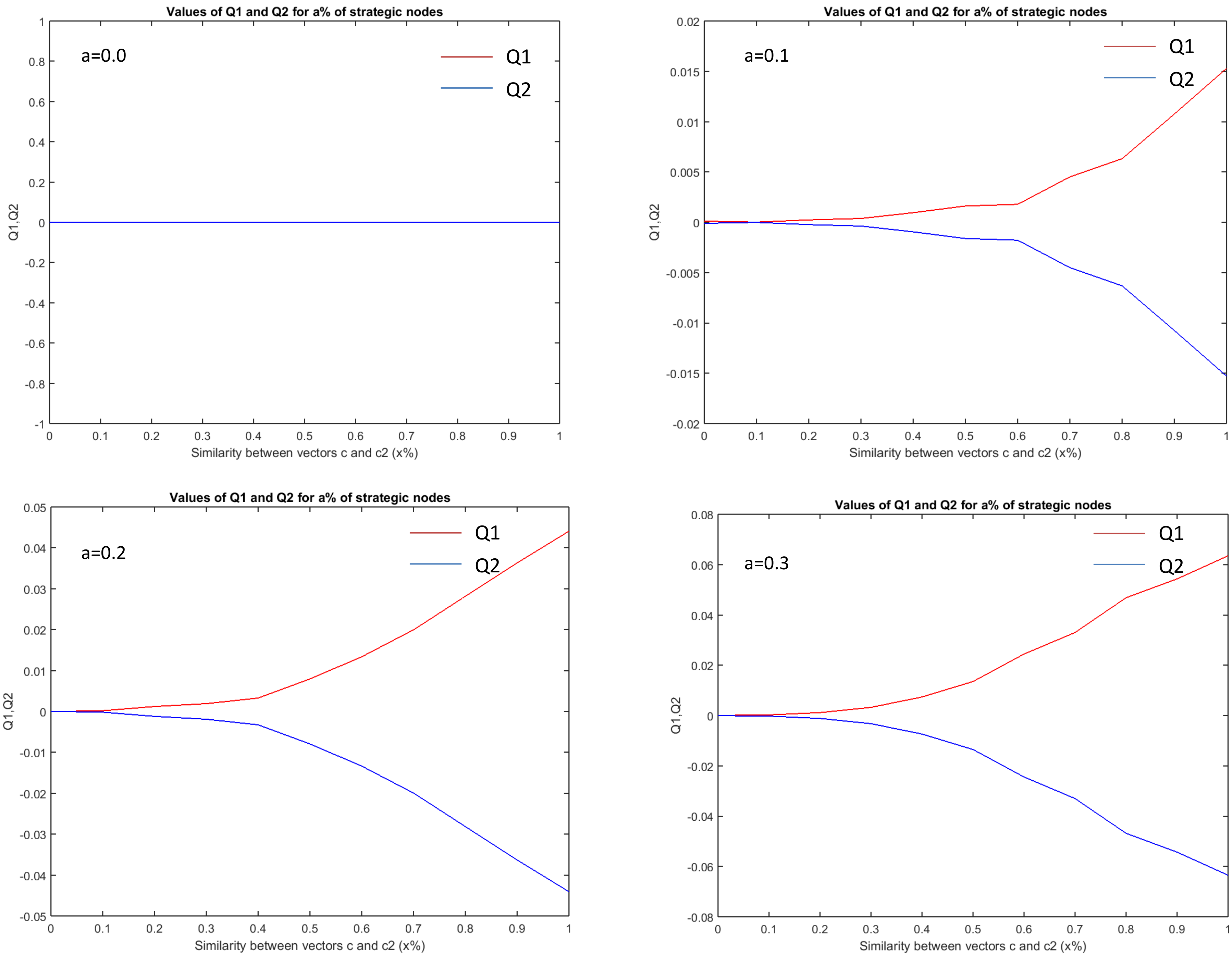


a=0.4



a=0.9




Values of Q1 and Q2 for a% of strategic nodes
a=0.0
Q1
Q2
Q1,Q2
Similarity between vectors c and c2 (x%)
Values of Q1 and Q2 for a% of strategic nodes
a=0.1
Q1
Q2
Q1,Q2
Similarity between vectors c and c2 (x%)
Values of Q1 and Q2 for a% of strategic nodes
a=0.2
Q1
Q2
Q1,Q2
Similarity between vectors c and c2 (x%)
Values of Q1 and Q2 for a% of strategic nodes
a=0.3
Q1
Q2
Q1,Q2
Similarity between vectors c and c2 (x%)

RESULTS

- Q_1 and Q_2 are symetric: $Q_1 = -Q_2$;
- To demonstrate:

$$Q_1 + Q_2 = \frac{1}{m}\left(S_1^T B S_1 + S_1^T B S_2 \right) =$$

$$= \frac{1}{m}\left(S_1^T B (S_1 + S_2)\right) = S_1^T \left(A - \frac{\overrightarrow{k_{in}}\,\overrightarrow{k_{out}}}{m} \right)\vec{u} =$$

$$= S_1^T \left(A\vec{u} - \overrightarrow{k_{in}}\frac{\overrightarrow{k_{out}}}{m}\vec{u} \right) = S_1^T \left(\overrightarrow{k_{in}} - \overrightarrow{k_{in}} \right) = 0$$

$$\boxed{Q_1 + Q_2 = 0 \rightarrow \boldsymbol{Q_1 = -Q_2}}$$

$$\boxed{where\ (S_1 + S_2) = \vec{u} = unit\ vector\ nx1}$$

RESULTS

- Matrices P for each A:

$$P = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.1 \end{bmatrix}$$

$$P_{com} = \begin{bmatrix} 0.12 & 0.087 \\ 0.087 & 0.1088 \end{bmatrix}$$

$$P_{str} = \begin{bmatrix} 0.12 & 0.1 \\ 0.087 & 0.1 \end{bmatrix}$$

RESULTS

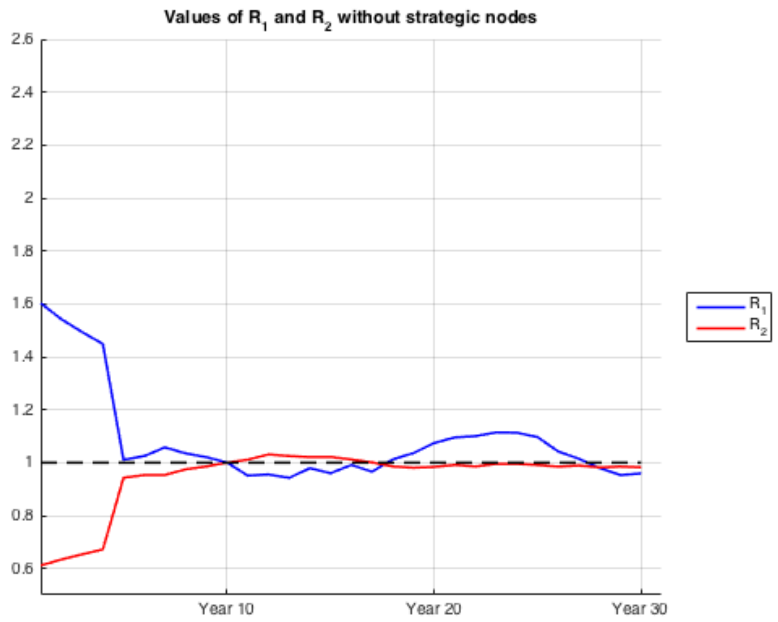
	Community	Strategic
R_1	1.3778	1.3799
R_2	1.2473	0.99136

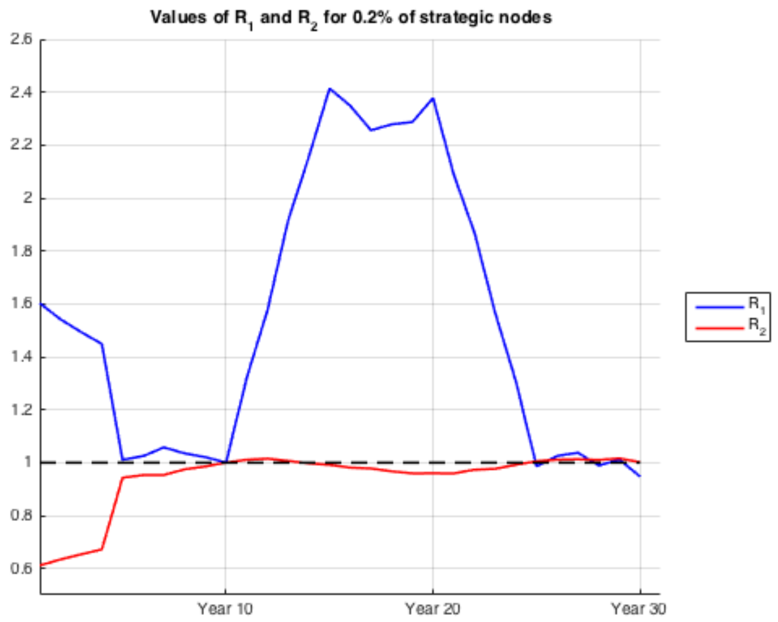
Results as expected!

Community: $R_1 > 1$ and $R_2 > 1$

Strategic group: $R_1 > 1$ and $R_2 \sim 1$

RESULTS


Values of R_1 and R_2 without strategic nodes
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
Year 10
Year 20
Year 30
R_1
R_2


Values of R_1 and R_2 for 0.2% of strategic nodes
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
Year 10
Year 20
Year 30
R_1
R_2

CONCLUSIONS

- The identification of the 2 groups depends only on one value ($Q_1 = -Q_2$);
- It is necessary to plot the behavior of the network at different time stamps;
- The method shows to be effective to identify 2 groups in a network, and specifies which one is strategic (if the behavior is known);
- **Application:** it was proposed an algorithm for generating a network with certain premises (such as strategic behavior);
- **Future studies:** apply this method in some real social networks to find if the strategic behavior occurs and if the method is efficient in identifying it (without previously knowing it).

REFERENCES

- ABBE, E. Community detection and the stochastic block model. Princeton University, February 20, 2016.
- BARROS, A. C. W. G. Dinâmica da reciprocidade periférica em uma rede de citações acadêmicas. Escola de Matemática Aplicada (FGV-EMAp), 2016.
- DUGUÉ, N. and PEREZ, A. Directed Louvain : maximizing modularity in directed networks. [Research Report] Université d'Orléans. 2015.
- NEWMAN, M. E. J. The Structure and Function of Complex Networks. SIAM Review, 45(2):167– 256, 2003.
- NEWMAN, M. E. J. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America. 103 (23): 8577–8696. 2006.
- NEWMAN, M. E. J. and GIRVAN M. Mixing patterns and community structure in networks, pages 66–87. Springer, Berlin, 2003.
- NEWMAN, M. E. J. and GIRVAN M. Finding and evaluating community structure in networks. Physical Review E, 69:026113, 2004.
- NEWMAN, M. E. J. and JUYONG, P. Why social networks are different from other types of networks. Physical Review E, 68:036122, 2003.
- NICOSIA, V., MANGIONI, G., CARCHIOLO, V. and MALGERI, M. Extending the definition of modularity to directed graphs with overlapping communities.
  J. Stat. Mech. P03024, 2009.
- SANDE, W. W. C. Reciprocidade periférica como estratégia para aumento de centralidade: estudo de rede de citações acadêmicas. PhD Thesis – Escola Brasileira de Administração Pública e de Empresas (FGV-EBAPE), 2016.