Comment on "Evolutionary method for finding communities in bipartite networks"

Alberto Costa*
LIX, École Polytechnique, F-91128 Palaiseau, France

Pierre Hansen†
GERAD, HEC Montréal, 3000 Chemin de la Côte-Sainte-Catherine, Montréal, Canada, H3T 2A7 and LIX, École Polytechnique, F-91128 Palaiseau, France

(Dated: November 4, 2011)

In a recent paper, Zhan, Zhang, Guan, and Zhou [Phys. Rev. E **83**, 066120 (2011)] presented a modified adaptive genetic algorithm (MAGA) tailored to the discovery of maximum modularity partitions of the node set into communities in unipartite, bipartite, and directed networks. The authors claim that "detection of communities in unipartite networks or in directed networks can be transformed into the same task in bipartite networks." Actually, some tests show that it is not the case for the proposed transformations, and why. Experimental results of MAGA for modularity maximization of untransformed unipartite or bipartite networks are also discussed.

PACS numbers: 89.75.Hc, 02.10.Ox, 02.70.-c

Networks, or graphs, are increasingly used for modeling and optimization of complex systems in many fields [1]. A network $G = (V, E)$ consists of a set V of nodes, represented by points, and a set E of edges, represented by lines joining pairs of points. A simple, unipartite network has no multiple edges or loops. A bipartite network $G = (V_1, V_2, E)$ has two subsets of nodes V_1 and V_2 and all its edges join pairs of nodes in different subsets. A network is directed if its edges have an orientation, i.e., go from an initial node to a terminal one.

A basic problem is to find communities, or modules, in such networks, i.e., subsets of nodes that are more likely to be joined pairwise by an edge than nodes in different modules. Various authors have given mathematical expressions for this problem. In particular, Newman and Girvan [2] have proposed an attractive objective function called modularity, and defined it as

$Q =$(fraction of edges within communities)
- (expected fraction of such edges).

Modularity maximization has been extensively studied, first in unipartite networks, and more recently, in bipartite networks and other generalizations. Scores of heuristics have been proposed as well as a few exact algorithms. Heuristics rely on a large variety of approaches. In a recent paper [3], Zhan, Zhang, Guan, and Zhou present a modified adaptive genetic algorithm (MAGA) and apply it to modularity maximization in unipartite and bipartite networks. Moreover, these authors claim that "detection of communities in unipartite networks or in directed networks can be transformed into the same task in bipartite networks." In order to discuss this claim, we first recall the definition of modularity [2]:

$$Q = \frac{1}{2M} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[A_{i,j} - \frac{k_i k_j}{2M} \right] \delta(g_i, g_j), \qquad (1)$$

where M is the number of edges of the network, N is the number of nodes, $A_{i,j}$ is an element of the adjacency matrix equal to 1 if nodes i and j are joined by an edge and 0 otherwise, and k_i and k_j are, respectively, the degrees of nodes i and j, that is, the number of edges incident with i and with j. Finally, g_i and g_j are the communities to which belong nodes i and j, and δ is the Kronecker symbol equal to 1 if g_i and g_j are the same and 0 otherwise. The values of Q range from -1/2 to 1 ([4], Lemma 1). Modularity maximization in unipartite networks is NP-complete in the strong sense ([4], Theorem 3). To the best of our knowledge, the complexity status of modularity maximization in bipartite networks is an open problem.

Consider now a bipartite network. According to Barber [5] and Leicht and Newman [6], modularity becomes

$$Q_b = \frac{1}{M} \sum_{i=1}^{p} \sum_{j=p+1}^{N} \left[\tilde{A}_{i,j} - \frac{k_i k_j}{M} \right] \delta(g_i, g_j), \qquad (2)$$

where $V_1 = \{1, \ldots, p\}$, $V_2 = \{p+1, \ldots, N\}$, and the adjacency matrix A_b is

$$A_b = \begin{bmatrix} 0_{p \times p} & \tilde{A}_{p \times q} \\ (\tilde{A}^{\mathrm{T}})_{q \times p} & 0_{q \times q} \end{bmatrix}.$$

Optimization problems with or without constraints can either be solved directly or transformed into another problem and then solved. Such transformations must be justified in every case. They have two advantages which would be, for modularity maximization, to unify somewhat the field and to bring to bear heuristics for bipartite

* costa@lix.polytechnique.fr
† pierre.hansen@gerad.ca

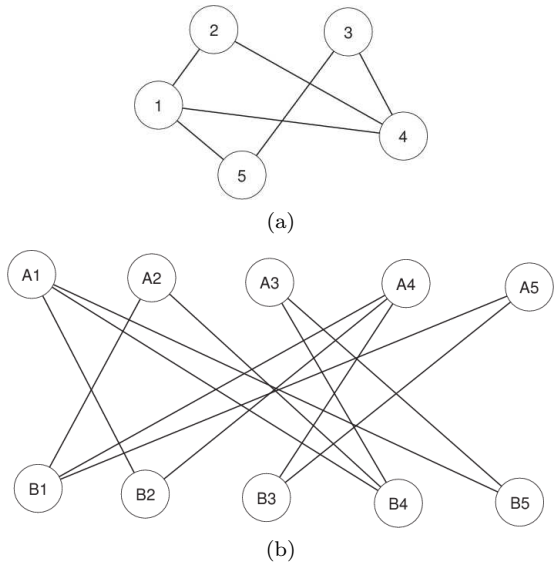

FIG. 1. Transformation of a simple unipartite network into a bipartite one. (a) A unipartite network with five nodes and six edges. (b) The bipartite network corresponding to (a).

maximization on the solution of modularity maximization problems in unipartite or directed networks.

Zhan *et al.* [3] propose the following transformation from a unipartite to a bipartite network: each node i of the original network is represented by two nodes A_i and B_i, and each edge $i-j$ is represented by two edges $A_i - B_j$ and $A_j - B_i$. This transforms a unipartite network with N nodes and M edges into a corresponding bipartite network with $2N$ nodes and $2M$ edges. An illustration is given in Fig. 1, borrowed from their paper.

In order to check for equivalence, the maximum modularity of the network in Fig. 1(a) has been computed with the clique-partitioning algorithm [7, 8]. This maximum modularity is equal to 0.111111, and corresponds to a partition in two modules $g_1 = \{1, 2, 4\}$ and $g_2 = \{3, 5\}$. To the best of our knowledge there is, as yet, no exact algorithm for modularity maximization in bipartite networks. Therefore we used a heuristic, i.e., LPAb' [9] to compute the maximum modularity of network 1(b). The near-optimal (or possibly optimal) partition obtained has a modularity of $Q_b = 0.347222$ corresponding to a partition in two modules $g_1 = \{A_1, A_3, B_2, B_4, B_5\}$ and $g_2 = \{A_2, A_4, A_5, B_1, B_3\}$ (this solution is not unique, another near-optimal one is $g_1 = \{A_1, A_2, A_3, B_4, B_5\}$ and $g_2 = \{A_4, A_5, B_1, B_2, B_3\}$). This example refutes the claim cited above.

Justification of the authors' claim is based on the three equations:

$$
\begin{aligned}
Q_b &= \frac{1}{2M}\sum_{i=1}^{N}\sum_{j=N+1}^{2N}\left[\tilde{A}_{i,j} - \frac{k_i k_j}{2M}\right]\delta(g_i, g_j) \\
&= \frac{1}{2M}\sum_{i=1}^{N}\sum_{j'=1}^{N}\left[\tilde{A}_{i,N+j'} - \frac{k_i k_{N+j'}}{2M}\right]\delta(g_i, g_{N+j'}) \quad (3) \\
&= \frac{1}{2M}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[A_{i,j} - \frac{k_i k_j}{2M}\right]\delta(g_i, g_j) = Q.
\end{aligned}
$$

The first line of (3) follows from the definition of Q_b, taking into account that the bipartite network (b) has $2N$ nodes and $2M$ edges. Going from the first equation to the second one, is a standard change of indices. The last equation is obtained by going back from A_b to A (or, in other words, focusing on the right upper square submatrix $\tilde{A}$ of A_b). It is implicitly assumed that the partition into communities does not change. This is expressed by the statement "where we have made use of the fact that the node A_i and B_i should be in an identical community," but as clearly shown by the example, this is not always true. A_i and B_i can be in different communities of the bipartite network. Indeed, it is likely to be so as, by construction, they are never joined by an edge.

One could then add explicit constraints specifying that each pair of nodes A_i and B_i must belong to the same community. Such constraints are usually easy to express, e.g., by identifying boolean variables for the assignment of entities to communities. However, if one adds some constraints, one gets into a different class of problems than modularity maximization in bipartite networks. This optimization problem with constraints will need a new or modified heuristic, as the heuristics for modularity maximization in bipartite networks of the literature do not apply anymore. Of course, one could use the constraints for each pair of nodes to merge them, but that brings one back to the original unipartite case.

It is hard to see what advantage there would be to transform a unipartite network into a larger bipartite one with constraints. Indeed, in the paper upon which we comment, when considering unipartite networks in their computational experiments the authors apply MAGA directly to these networks, without transforming them into bipartite ones.

Zhan *et al.* [3] (see also [10]) propose a transformation analogous to the previous one of a directed network into a bipartite one. A node i is represented by two nodes A_i and B_i and a directed edge from i to j as an (undirected) edge between A_i and B_j. This transforms a directed network with N nodes and M directed edges into a corresponding bipartite network with $2N$ nodes and M undirected edges. An illustration is given in Fig. 2. The argument defending the claim that this transformation does not change the modularity value is similar to the case of the transformation from the unipartite to

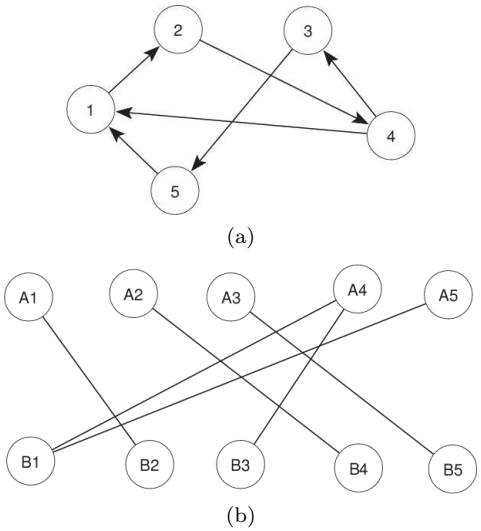

FIG. 2. Transformation of a simple directed network into a bipartite one. (a) A directed network with five nodes and six edges. (b) The bipartite network corresponding to (a).

the bipartite. It relies on the three equations (4):

$$\begin{aligned} Q_b &= \frac{1}{M}\sum_{i=1}^{N}\sum_{j=N+1}^{2N}\left[\tilde{A}_{i,j}-\frac{k_i k_j}{M}\right]\delta(g_i,g_j) \\ &= \frac{1}{M}\sum_{i=1}^{N}\sum_{j'=1}^{N}\left[\tilde{A}_{i,N+j'}-\frac{k_i k_{N+j'}}{M}\right]\delta(g_i,g_{N+j'}) \quad (4) \\ &= \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[A_{i,j}-\frac{k_i^{\text{out}}k_j^{\text{in}}}{M}\right]\delta(g_i,g_j)=Q, \end{aligned}$$

where the symbols are defined above, except for the half-degrees k_i^{out} and k_j^{in}, which are equal to the number of directed edges going out of i and into j, respectively. Once more, it is assumed that each pair of nodes A_i and B_i belongs to the same community.

Optimizing modularity of network 2(a), with the clique partitioning algorithm previously mentioned, gives an optimal partition into two communities $g_1 = \{3, 5\}$ and $g_2 = \{1, 2, 4\}$ with a modularity of $Q = 0.111111$. Computing the maximum modularity of the network 2(b) with LPAb' leads to a partition into four communities $g_1 = \{A_1, B_2\}$, $g_2 = \{A_2, B_4\}$, $g_3 = \{A_3, B_5\}$, and $g_4 = \{A_4, A_5, B_1, B_3\}$ with a modularity of $Q_b = 0.666667$

Zhan *et al.* [3] mention that MAGA can be applied directly to modularity maximizing in unipartite or bipartite networks without transforming them, and in their abstract they claim "Experimental results show that the MAGA outperforms existing methods in terms of modularity for both bipartite and unipartite networks." They report empirical results for both kinds of problems. In the bipartite case, they compare the results of MAGA with a standard genetic algorithm (SGA) and with a multiobjective genetic algorithm (MOGA) [11]. The superiority of MAGA and SGA over MOGA is very clear for artificial bipartite networks. Moreover, they studied two real-world networks, namely, the southern women [12] and Scotland corporate interlock [13] networks. For the first network MAGA gives the same modularity value as BRIM [5], while MAGA gives better results than BRIM, SGA, and MOGA in the second case. Recently, Liu and Murata [9],[14] obtained an almost as good partition than Zhan *et al.* [3] for the second case (the difference in value is in the fourth decimal place), with an improved version of the label propagation algorithm for bipartite networks (LPAb+), as shown in Table I.

Network	Nodes	Edges	Q LPAb+	Q MAGA
Southern women	18+14	89	**0.3455**	**0.3455**
Scotland interlock	86+131	348	0.7091	**0.7093**

TABLE I. Maximum values of modularity Q for real-world bipartite networks obtained by heuristics LPAb+ and MAGA.

Six well-known unipartite networks were also used (i.e., Zachary karate club [15], jazz musicians [16], *C. elegans* metabolic [17], e-mail [18], PGP [19], and Condmat [20] networks). Results are compared with those of the Girvan Newman heuristic [21], extremal optimization [22], spectral relaxation [23, 24], and simulated annealing [25]. MAGA obtained the best results in all cases, improving the records for the three largest ones (e-mail, PGP, and Condmat). However, other researchers did recently obtain equally good results for the first three problems and better ones for the three last ones [i.e., Liu and Murata's label propagation algorithm for unipartite networks (LPAm+) [26], and Noack and Rotta with the single-step multi-level algorithm (SS-ML) [27])], as shown in Table II.

			SS-ML	LPAm+		MAGA	
Network	Nodes	Edges	Q	Q	Time	Q	Time
Zachary	34	78	**0.420**	**0.420**	0.014 s	**0.420**	0.1 s
Jazz	198	2742	**0.445**	**0.445**	0.368 s	**0.445**	19 min
C. elegans	453	2025	0.446	**0.452**	1.247 s	**0.452**	12 min
e-mail	1133	5451	0.577	**0.582**	3.589 s	0.581	72 min
PGP	10680	24316	**0.884**	**0.884**	114.221 s	0.881	610 min
Condmat	27519	116181	**0.814**	0.755	461.599 s	0.802	3517 min

TABLE II. Maximum values of modularity Q for real-world unipartite networks obtained by heuristics SS-ML, LPAm+, and MAGA. Computing times for LPAm+ on a 2.53-GHz Intel Core 2 Duo CPU and MAGA on a PC with two 2.93-GHz Intel processors.

The comparison of results of MAGA and state-of-the-art heuristics for unipartite modularity maximization reported on in this Comment shows that MAGA gets a solution equal to the best previously known in half of the cases and worse in the other cases. The differences in value are twice in the third decimal place and once in the second. Moreover, the computing time of MAGA is quite large and increases rapidly. On comparable computers,

MAGA takes 7 to 3000 times more time than LPAm+ to solve the problems of Table II. Nevertheless, MAGA has some advantages: mainly, robustness. Indeed, as other evolutionary algorithms, it provides several near-optimal solutions instead of a single one, as done by most heuristics of other families.

ACKNOWLEDGMENTS

Financial support by Grants Digiteo 2009-14D "RMNCCO" and Digiteo 2009-55D "ARM" is gratefully acknowledged.

[1] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
[2] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[3] W. Zhan, Z. Zhang, J. Guan, and S. Zhou, Phys. Rev. E **83**, 066120 (2011).
[4] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE Transactions on Knowledge and Data Engineering **20**, 172 (2008).
[5] M. J. Barber, Phys. Rev. E **76**, 066102 (2007).
[6] E. A. Leicht and M. E. J. Newman, Phys. Rev. Lett. **100**, 118703 (2008).
[7] M. Grötschel and Y. Wakabayashi, Math. Program. **45**, 59 (1989).
[8] D. Aloise, S. Cafieri, G. Caporossi, P. Hansen, S. Perron, and L. Liberti, Phys. Rev. E **82**, 046112 (2010).
[9] X. Liu and T. Murata, Journal of Advanced Computational Intelligence and Intelligent Informatics **14**, 408 (2010).
[10] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **76**, 036102 (2007).
[11] N. L. Law and K. Y. Szeto, in *Proceedings of the 20th international joint conference on Artifical intelligence* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007) pp. 2330–2333.
[12] B. B. G. A. Davis and M. R. Gardner, *Deep South: A Social Anthropological Study of Caste and Class* (University of Chicago Press, Chicago, 1941).
[13] J. Scott and M. Hughes, *The Anatomy of Scottish Capital: Scottish Companies and Scottish Capital, 1900-1979* (CroomHelm, London, 1980).
[14] In [9], Liu and Murata report a best-known value of 0.7194 for the full Scotland interlock network. The value they found for the main connected component of that network, which was also studied by Zhan *et al.*, reported in this Comment, was kindly communicated by Mr. Liu on August 7, 2011.
[15] W. W. Zachary, Journal of Anthropological Research **33** (1977).
[16] P. Gleiser and L. Danon, Advances in Complex Systems **6**, 565 (2003).
[17] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, Nature **407**, 651 (2000).
[18] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103 (2003).
[19] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys. Rev. E **70**, 056122 (2004).
[20] M. E. J. Newman, Proceedings of the National Academy of Sciences of the United States of America **98**, 404 (2001).
[21] M. Girvan and M. E. J. Newman, Proceedings of the National Academy of Sciences **99**, 7821 (2002).
[22] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, Journal of Statistical Mechanics: Theory and Experiment **9**, 8 (2005).
[23] M. E. J. Newman, Proceedings of the National Academy of Sciences of the United States of America **103** (2006).
[24] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[25] R. Guimera and L. A. N. Amaral, Nature **433**, 895 (2005).
[26] X. Liu and T. Murata, Physica A: Statistical Mechanics and its Applications **389**, 1493 (2010).
[27] A. Noack and R. Rotta, in *Proceedings of the 8th International Symposium on Experimental Algorithms*, SEA '09 (Springer-Verlag, Berlin, Heidelberg, 2009) pp. 257–268.