CHEMICAL ENGINEERING TRANSACTIONS Volume 24, 2011
Editor Sauro Pierucci
Copyright © 2011, AIDIC Servizi S.r.l., ISBN 978-88-95608-15-0 ISSN 1974-9791
DOI: 10.3303/CET1124054

A Comparative Study of Solving the Problem of Module Identification in a Complex Network

Toni Lastusilta[1], Lazaros G Papageorgiou[2] and Tapio Westerlund[*]

[1,*] Process Design and Systems Engineering Laboratory, Department of Chemical Engineering, Åbo Akademi University
[2] Centre for Process System Engineering, Department of Chemical Engineering, University College of London
[1,*] Åbo Akademi University, Biskopsgatan 8, FIN-20500 ÅBO, Finland,
Toni.Lastusilta@abo.fi

In this paper the problem to identify modularity in a complex network is studied. The ability to identify modularity can be vital for a clear understanding of how a complex network is constructed and the interaction in the network. The aim in this study is to compare different solving strategies and find the most appropriate ones to solve this type of problem. The strategies consist of using different problem formulations and solvers. Five network identification problems, where the networks are of different size, are studied with seven different solvers in the General Algebraic Modeling System (GAMS). Two different mathematical formulations are used: a convex Mixed-Integer Quadratic Programming (MIQP) formulation and a compact bilinear MIQP formulation. With two of the solvers the bilinear MIQP formulation is solved by starting the solution procedure from random network configurations. Furthermore, the impact of using symmetry breaking constraints, presented in (Xu et al., 2007), is evaluated.

1. Introduction

The problem of module identification in a complex network is a combinatorial problem. Combinatorial problems in optimization are formulated by using binary or/and integer variables. A linear increase in the number of these variables results in an exponential increase in combinatorial complexity. Some examples of complex networks are World Wide Web related networks (Flake et al., 2002; Eckmann and Moses, 2002), social networks (Girvan and Newman, 2002; Guimerà et al., 2003) and biochemical networks (Guimerà and Amaral, 2005).

The studied problem is a nonlinear problem. Nonlinear problems can in general be difficult to solve to global optimality, however, a local optimal solution is substantially easier to find. If the nonlinear problem is convex then a local optimal solution is also a global optimal solution. Typically the nonlinear solvers of today solve a problem only to a local optimal solution and therefore a convex nonlinear formulation is appealing. If the objective function and all constraints in a problem are convex inequality constraints then the problem is convex. A binary or integer variable makes a problem non-convex but if the integer relaxed problem is convex then global optimality can be proven, for example, with the branch and bound method. A bilinear term makes a problem non-

Please cite this article as: Lastusilta T., Papageorgiou L. and Westerlund T., 2011, A comparative study of solving the problem of module identification in a complex network, Chemical Engineering Transactions, 24, 319-324
DOI: 10.3303/CET1124054

convex, but it can be made convex by reformulating it, which is done for the studied problem. The down side of reformulating is that it typically increases the problem size, i.e. the number of variables and/or constraints. Furthermore, note that a nonlinear solver that guarantees global optimality is of little use if the solution time becomes unreasonable, which can be the case for a large scale problem.

The choice of solution method is equally important when the solving time becomes an issue. Different solution methods are provided by different solvers, which are typically included in modeling platforms. Some examples of modeling platforms are: AIMMS (Bisschop, 2006), AMPL (Fourer, 1990) and GAMS (Rosenthal, 2008). For solver descriptions the reader is advised to the respective platform documentations.

In this study we focus on the problem of module identification in a complex network by comparing different formulations and solvers on five test problems. The problem has earlier been studied, for example, in (Aloise et al., 2010).

2. Problem Description

The studied network type consist of nodes and links, undirected and unweighted, which can be grouped into modules. For example, a node can represent a person, a link between two nodes can represent friendship between those persons and a module can represent a group of friends. The goal is to divide the nodes into as few compact modules as possible. A module is compact if there are very few links outside from the module, but many links within the module. Each node can only belong to one module. If the number of nodes and links in a network is large we have a complex network.

The characteristics of the studied problems can be found in Table 1. Problem “Karate” and “Dolphin” are social network problems similar to the example given above. Problem “Miserables” is a co-appearance network of characters in the novel Les Miserables. Problem “Football” is a network of American football games between divisions and problem “Power” represents a network topology of the Western States Power Grid of the United States. The problems were found from Mark Newman’s web page: http://www-personal.umich.edu/~mejn/netdata/. In Table 1 the number of modules denotes the maximum number of allowed modules and it impacts heavily on the combinatorial complexity, i.e. the number of binary variables, of the problem. The maximum number of modules can in principle be as many as the number of nodes, but here we have chosen a reasonable upper limit. For the three smallest problems the number of modules found in (Xu et al., 2007) was used.

Table 1: Problem characteristics

Problem name	Short name	Nodes	Links	Modules	Binary variables	Links/Nodes ratio
Zachary's karate club	Karate	34	78	4	136	2.3
Dolphin social network	Dolphin	62	159	5	310	2.6
Les Miserables	Miserables	77	254	6	462	3.3
American College football	Football	115	613	10	1152	5.3
Power grid	Power	4941	6594	20	98820	1.3

3. Problem Formulations

The studied problem can, among others, be formulated as a compact bilinear non-convex MIQP problem or a convex MIQP problem. In this paper, the convex MIQP problem denotes an otherwise convex problem except for the integer requirement. The convex MIQP proposed by (Xu et al., 2007) was used and from that the bilinear MIQP can easily be formulated. Furthermore, the node data was reordered in descending connectivity order and symmetry breaking constraints proposed in (Xu et al., 2007) was applied for the formulations. The metric to estimate the module compactness was first proposed by (Newman and Girvan, 2004). In the following a full description of the used formulations is given.

The following notation is used:

Constants: M = the total number of nodes, L = the total number of links,

d_n = the number of links connected to node n

Indices: n,e = node, l =link, m =module

Variables: Q = modularity compactness measure and objective variable (continuous),

L_m = number of links among nodes within module m (positive and continuous),

D_m = degree of module m, i.e. the number of links connected to a node for each node in module m (positive and continuous),

$Y_{n,m}$ =1 if node n belongs to module m, otherwise 0 (binary),

$X_{l,m}$ =1 if link l belongs to module m, otherwise 0 (positive and continuous)

The following bilinear MIQP formulation, where Q is maximized, was used:

$$Q = \sum_{m}^{M} \left[\frac{L_m}{L} - \left(\frac{D_m}{2 \cdot L} \right)^2 \right] \tag{1}$$

$$\sum_{m}^{M} Y_{n,m} = 1 \qquad ,\forall n \tag{2}$$

$$D_m = \sum_{n}^{N} d_n \cdot Y_{n,m} \quad ,\forall m \tag{3}$$

$$L_m = \sum_{l | l \in \{n,e\}}^{L} Y_{n,m} \cdot Y_{e,m} \qquad ,\forall m \tag{4}$$

, where $l \in \{n,e\}$ denotes if there exist a link between the two nodes n and e.

Symmetry breaking was applied by replacing (2) with the following constraints:

$$\sum_{m | m \le n}^{M} Y_{n,m} = 1 \qquad ,\forall n | n \le M \tag{5}$$

$$Y_{n,m} = 0 \ \ ,\forall m,n | n < m \le M \tag{6}$$

$$\sum_{m}^{M} Y_{n,m} = 1 \qquad ,\forall n \middle| n > M \tag{7}$$

$$Y_{n,m} \leq \sum_{e \left| \substack{e<n \\ e \geq (m-1)} \right.}^{N} Y_{e,(m-1)} \qquad ,\forall m \middle| m \geq 3,\ n \middle| n \geq 3 \tag{8}$$

The convex MIQP was formulated with constraints (1) to (3) and the following additional constraints:

$$X_{l,m} \leq Y_{n,m} \qquad ,\forall m, l \middle| l \subseteq \{n,e\} \tag{9}$$

$$X_{l,m} \leq Y_{e,m} \qquad ,\forall m, l \middle| l \subseteq \{n,e\} \tag{10}$$

$$L_m = \sum_{l \left| l \subseteq \{n,e\} \right.}^{L} X_{l,m} \quad ,\forall m \tag{11}$$

The convex MIQP problem with symmetry breaking constraints was formulated by adding to (1), (2), (3), (5), (8) and (11) the following constraints:

$$X_{l,m} \leq Y_{n,m} \qquad ,\forall m, l \middle| l \subseteq \{n,e\},\ n \geq m \wedge e \geq m \tag{12}$$

$$X_{l,m} \leq Y_{e,m} \qquad ,\forall m, l \middle| l \subseteq \{n,e\},\ n \geq m \wedge e \geq m \tag{13}$$

$$X_{l,m} = 0 \qquad ,\forall m, l \middle| l \subseteq \{n,e\},\ n < m \vee e < m \tag{14}$$

4. Setup

To compare the problem formulations and their suitability for different solvers GAMS 23.5.2 was used. The following solvers were compared: CPLEX, AlphaECP, BARON, CoinBonmin, DICOPT, LINDOGlobal, SBB. The solvers were set to solve the five test problems so that each of the four formulation was used with two different time limits, see Table 2. Special attention was given to DICOPT and SBB when solving with the bilinear MIQP formulation without symmetry breaking constraints. These solvers terminated with a solution for each test problem within 2 minutes. With this motivation each test problem was solved 100 times with a random network configuration at startup, i.e. the discrete variable levels are selected randomly, see Table 3. The test computer was an Intel core i7 with 8 processors of 2,8GHz and 6GB of memory.

5. Computational Results and Conclusions

The result of the comparison can be found from Table 2 and 3. In Table 2 and 3 the following abbreviations are used: F=Formulation, B=Bilinear MIQP, C=Convex MIQP, S=Symmetry breaking applied, T=Time limit in hours, *=Terminated before time limit was reached. In Table 2 the last column in each row denotes how many solvers found for that formulation a modularity value above 0.1, i.e. a good or modest solution. The last row denotes the number of solved problems to a modularity value above 0.1 by the

Table 2: The objective value for the maximization problems

F	S	T	CPLEX	Alpha-ECP	BARON	Coin-Bonmin	DICOPT	LINDO-Global	SBB	
Karate										
B	No	1	0.42	0.00*	0.42*	0.41*	0.42*	0.42*	0.42*	6
B	No	2	0.42	0.00*	0.42*	0.41*	0.42*	0.42*	0.42*	6
B	Yes	1	0.29	0.29*	0.42*	0.42*	0.38*	0.42*	0.38*	7
B	Yes	2	0.29	0.29*	0.42*	0.42*	0.38*	0.42*	0.38*	7
C	No	1	0.42*	0.42*	0.42*	0.42*	0.00*	0.42	0.42*	6
C	No	2	0.42*	0.42*	0.42*	0.42*	0.00*	0.42	0.42*	6
C	Yes	1	0.42*	0.42*	0.42*	0.42*	0.35*	0.42*	0.42*	7
C	Yes	2	0.42*	0.42*	0.42*	0.42*	0.35*	0.42*	0.42*	7
Dolphin										
B	No	1	0.43	0.02*	0.53	0.50*	0.49*	0.53	0.49*	6
B	No	2	0.52	0.02*	0.53	0.50*	0.49*	0.53	0.49*	6
B	Yes	1	0.48	0.00*	0.53	0.50*	0.50*	0.52	0.50*	6
B	Yes	2	0.45	0.00*	0.53	0.50*	0.50*	0.53	0.50*	6
C	No	1	0.53	0.52	0.15	0.53	0.00*	0.51	0.46*	6
C	No	2	0.53	0.52	0.15	0.53	0.00*	0.51	0.46*	6
C	Yes	1	0.53*	0.53*	0.46	0.53	0.01*	0.47	0.53*	6
C	Yes	2	0.53*	0.53*	0.52	0.53*	0.01*	0.47	0.53*	6
Miserables										
B	No	1	0.45	0.17*	0.56	0.55*	0.55*	0.56	0.55*	7
B	No	2	0.49	0.17*	0.56	0.55*	0.55*	0.56	0.55*	7
B	Yes	1	0.24	0.12	0.56	0.56*	0.44*	0.56	0.45*	7
B	Yes	2	0.24	0.12	0.56	0.56*	0.44*	0.56	0.45*	7
C	No	1	0.56	0.56	0.12	0.56	0.00*	NA	0.54*	5
C	No	2	0.56	0.56	0.14	0.56	0.00*	NA	0.54*	5
C	Yes	1	0.56*	0.56*	0.15	0.56	0.25*	NA	0.56*	6
C	Yes	2	0.56*	0.56*	0.15	0.56	0.25*	NA	0.56*	6
Football										
B	No	1	0.48	0.16*	0.60	0.60*	0.48*	NA	0.48*	6
B	No	2	0.48	0.16*	0.60	0.60*	0.48*	0.52	0.48*	7
B	Yes	1	0.36	0.30*	0.54	0.57*	0.51*	NA	0.51*	6
B	Yes	2	0.39	0.30*	0.54	0.57*	0.51*	NA	0.51*	6
C	No	1	0.44	0.46	0.39	0.00	0.00*	NA	NA	3
C	No	2	0.49	0.46	0.39	0.00	0.00*	NA	NA	3
C	Yes	1	0.52	0.01	0.00	0.00	0.00*	NA	NA	1
C	Yes	2	0.52	0.01	0.00	0.00	0.00*	NA	NA	1
Power										
B	No	1	NA	0.07	NA	NA	0.66*	NA	0.70*	2
B	No	2	NA	0.07	NA	NA	0.66*	NA	0.70*	2
B	Yes	1	NA	NA	NA	NA	NA	NA	NA	0
B	Yes	2	NA	NA	NA	NA	NA	NA	NA	0
C	No	1	NA	0.00	NA	NA	NA	NA	0.81	1
C	No	2	NA	0.00	NA	NA	NA	NA	0.85	1
C	Yes	1	NA	NA	NA	NA	NA	NA	NA	0
C	Yes	2	NA	NA	NA	NA	NA	NA	NA	0
			32	24	30	28	22	21	32	

Table 3: The best and average, in brackets, modularity value

Solver	F	S	Karate	Dolphin	Miserables	Football	Power
DICOPT	B	No	0.42 (0.39)	0.53 (0.49)	0.56 (0.51)	0.59 (0.54)	0.69 (0.67)
SBB	B	No	0.42 (0.40)	0.53 (0.50)	0.56 (0.52)	0.59 (0.54)	0.73 (0.71)

corresponding solver. In Table 3 the average value denotes the arithmetic mean value. From Table 2 and when considering the solving times (not shown) suggests that the solver CPLEX combined with the convex formulation that contains symmetry breaking constraints is to prefer if the problem is not too large. However, with the used time limits the formulation is not very suitable for the larger problems, i.e. the two largest problems in the comparison. In that case by using the more compact formulations, the remaining three, some solver could find the best solution. From Table 3 we can see that DICOPT and SBB were able to find good solutions when repeatedly solved starting from random network configurations. The solvers reported for each solver call a solution, except with DICOPT for problem “Power” when 29 times out of 100 a solution was reported. The total solving time for DICOPT was about 1 hour and for SBB about 8 hours, which suggests that it can be a good solving strategy especially for large networks. In the future a formulation with a reduced set of symmetry breaking constraints might be worth investigating.

References

Aloise D., Cafieri S., Caporossi G., Hansen P., Perron S. and Liberti L., 2010, Column generation algorithms for exact modularity maximization in networks, Phys. Rev E 82, 046112.

Bisschop J., 2006, AIMMS - Optimization Modeling, ISBN: 9781847539120, Lulu, North Carolina.

Eckmann J.P. and Moses E., 2002, Curvature of co-links uncovers hidden thematic layers in the World Wide Web, PNAS, vol. 99, no. 9, 5825-5829.

Flake G.W., Lawrence S., Giles C.L. and Coetzee F.M., 2002, Self-Organization and Identification of Web Communities, IEEE Comput. 35, 66-70.

Fourer R., Gay D.M. and Kernighan B.W., 2002, AMPL: A Modeling Language for Mathematical Programming, ISBN-13: 978-0534388096, Duxbury Pr., California.

Girvan M. and Newman M.E.J., 2002, Community structure in social and biological networks, Sci. U.S.A, vol. 99, no. 12, 7825-7826.

Guimerà R. and Amaral L.A.N., 2005, Functional cartography of complex metabolic networks, Nature, vol. 433, 895-900.

Guimerà R., Danon L., Díaz-Guilera A., Giralt F. and Arenas A., 2003, Self-similar community structure in a network of human interactions, Phys. Rev E 68, 065103.

Newman M.E.J. and Girvan M., 2004, Finding and evaluating community structure in networks, Rev. E 69, 026113.

Rosenthal R.E., 2010, GAMS -A User's Guide, GAMS Development Corporation, <http://www.gams.com/dd/docs/bigdocs/GAMSUsersGuide.pdf> accessed 3.3.2011

Xu G. , Tsoka S. and Papageorgiou L.G., 2007, Finding community structures in complex networks using mixed integer optimisation, Eur. Phys. J. B 60, 231-239.