**manuscript No.**
(will be inserted by the editor)

# A new mixed integer linear programming formulation for one problem of exploration of online social networks

**Aleksandra Petrović**

Received: date / Accepted: date

**Abstract** Enormous global popularity of online social network sites has initiated numerous studies and methods investigating different aspects of their use, so some concepts from network-based studies in optimization theory can be used for research into online networks. In Gajić (2014) are given a several new mixed integer linear programming formulations for first and second problem of exploration of online social networks introduced in Stanimirović and Mišković (2013). This paper introduces a new mixed integer linear programming formulation for remaining (third) problem. Numerical experiments show the performance of a commercial exact solver when applied to the proposed model. The new model is also compared with a model proposed in the literature using instances from single source capacitated facility location problem and randomly generated instances. The obtained results indicate that the proposed formulation clearly outperforms the existing formulation. Moreover, new formulation has small number of decision variables, so it is capable to find, by exact solver, upper and lower bounds on large-scale problem instances. The paper also noted and tried to correct some wrong considerations presented in the previous work.

**Keywords** Mixed integer linear programming · Location problems · Social networks · Combinatorial optimization

**Mathematics Subject Classification (2000)** 90C11 · 91D30 · 90B10

Faculty of Technical Sciences
University of Novi Sad
Trg Dositeja Obradovića 6
21000 Novi Sad, Serbia
E-mail: ap5309311@gmail.com

## 1 Introduction

In the modern era, social networking is gaining importance. Social networks are usually defined in the literature as structures made of actors (generally humans) linked by relations [4]. For example, these relations may be phone calls or kinship. Consequently, the social networks are mostly represented as graphs, therefore social network optimization is closely related to graph optimization techniques. Therefore, some concepts from these techniques may be used for research of online networks.

Stanimirović and Mišković [2] were introduced a mixed integer linear programming (MILP) formulations for three problems (authors wrongly named them as models) for efficient analysis of an online social network. In order to solve these problems of larger dimensions, they also proposed three evolutionary metaheuristic approaches. As it can be reported in [1]: "Unfortunately, their paper is full of errors and deceptions. For example, all three evolutionary approaches were used crossover probabilities which were *larger than 1* (Stanimirović & Mišković, 2013, p. 241). Since the authors performed experiments on the problem instances that were generated randomly by themselves, and are unavailable to other researchers, the only valuable scientic content of that paper seems to be MILP formulations". Unfortunately, some errors in [2] were unnoticed previously:

- P. 236. The proof of Proposition 3.1 is incomplete (wrong), since it is proved only in one direction! Therefore, Algorithm 4 on p. 240, which describes objective function for Problem 3 in all three metaheuristic approaches, is not mathematically correct, although it work correctly;
- P. 237. Algorithm 2 which describes objective function for Problem 1 could not work correctly in case of ties between costs, i.e. cost from one user node to two or more potential control devices can be equal;
- P. 238. Algorithm 3 which describes objective function for Problem 2 also could not work correctly in case of ties between costs;
- P. 241. state "However, in the hybrid EA-LS and EA-TS methods, the crossover rate parameter is *decreased* ...", while it is obvious that it is increased;
- P. 250. "... showed that the CPLEX 12.1 solver provided optimal solutions only for small problem instances with up to 100 user nodes, while larger instances were out of its reach of CPLEX." As it can be seen from definition, problem hardness obviously grows linearly with number of user nodes $m$. Therefore, each method could routinely solve to optimality instances with small number of potential locations $n$ and large number of user nodes $m$. Contrary, instances with small $m$ and large $n$ are very hard!
- P. 262. "The proposed mathematical models and hybrid metaheuristic methods may be used as additional tools for further research in this field". This sentence is only partially true, about models, but metaheuristic methods cannot be used at all, since they are wrongly and deceptively presented, with obvious false parameters (for example, crossover probability),

instances are unavailable to other researchers and even for small example has reported wrong optimal value.

Gajić [1] was proposed a several MILP formulations for first and second problem for efficient analysis of an online social network, incorporating several known closest assignment constraints from the literature. Experimental results contain the direct comparison between all formulations, performed on standard single source capacitated facility location problem instances. As it can be seen, the best formulation was based on *WF* closest assignment constraints proposed by Wagner and Falkson [3].

## 2 Existing mathematical formulation

In this section, Problem 3, that considers exploration of online social networks, is presented as it is defined in [2]. Let $I = \{1, 2, ..., m\}$ be the set of user nodes in the network, and $J = \{1, 2, ..., n\}$ a set of potential locations for establishing control devices. The non-negative matrix $c_{ij} | i \in I, j \in J$ represents the cost of searching through the data from a user node $i$ by a control device $j$. These costs may depend on the amount of time needed to explore one unit of data originating from $i$ by device $j$, the distance and speed of Internet connection between them, etc. Exactly $p$ potential locations must be established, and each user could be assigned to one or more open control devices in order to minimize the overall search process of online social network.

More precisely, in the Problem 3 (contrary from other two problems), a user node may split its out-coming data flow to several established control devices, i.e. user requirements can be (partially) served by some different established devices but out-coming flow of the considered user node must be equal to sum of these flow parts. The objective of the Problem 3 is to establish (exactly $p$) control devices and to divide user requirements through open control devices in order to minimize the load balance of control devices.

The Problem 3 can be mathematically formulated in the following way. Let $S \subset J$ denote the set of established control devices and $x_{ij}$ is the fraction of information flow originating from a user node $i$ and explored by a control device $j$. Then, objective function value $obj(S) = \max\limits_{i \in I, j \in S} c_{ij} \cdot x_{ij}$ subject to $\sum\limits_{j \in S} x_{ij} = 1$ for each $i \in I$. Note that for all $j \notin S$ hold $x_{ij} = 0$. The optimal objective value of Problem 3 can be defined as $opt_{P3} = \min\limits_{S \subset J, |S| = p} \text{obj(S)}$.

The mixed integer linear programming formulation of Problem 3 in [2] have the binary decision variables $y_j, j = 1, ..., n$, real variables $x_{ij}$ and real variable $z$. Using the previous notation, MILP was formulated as follows:

$$\min z \tag{1}$$

subject to:

$$\sum_{j\in J} y_j = p \tag{2}$$

$$x_{ij} \leq y_j, \quad for\ all\ i \in I, j \in J \tag{3}$$

$$\sum_{j\in J} x_{ij} = 1, \quad for\ all\ i \in I \tag{4}$$

$$c_{ij} \cdot x_{ij} \leq z, \quad for\ all\ i \in I, j \in J \tag{5}$$

$$x_{ij} \in [0,1], y_j \in \{0,1\}\,,\ z \geq 0 \quad for\ all\ i \in I, j \in J \tag{6}$$

The objective function (1) minimizes the load among control devices. Constraint (2) indicates that exactly $p$ resource nodes are established. By constraints (3) and (4) it is ensured that each user node is assigned to exactly one established resource node. Constraints (5) impose lower bounds on the value of objective value $z$, while (6) reflect the nature of decision variables $x_{ij}, y_j$ and $z$.

### 2.1 Illustrative example

*Example 1* [2] Let us consider a small network with $m = 10$ user nodes and $n = 6$ potential locations for establishing control devices. Suppose that exactly $p = 3$ control devices are to be established with cost matrix

$$C = \begin{bmatrix} 10163 & 14 & 73 & 489 & 14588 & 125 \\ 113 & 234 & 29 & 12365 & 12657 & 265 \\ 12050 & 12955 & 132 & 73 & 368 & 0 \\ 12114 & 12765 & 114 & 221 & 42 & 143 \\ 192 & 14245 & 122 & 13123 & 169 & 33 \\ 10533 & 12446 & 195 & 294 & 325 & 133 \\ 25 & 171 & 393 & 385 & 11333 & 10765 \\ 370 & 14645 & 116 & 292 & 14748 & 449 \\ 286 & 13273 & 245 & 14095 & 497 & 82 \\ 476 & 11263 & 187 & 124 & 14359 & 275 \end{bmatrix}$$

Optimal objective function value of Problem 3, on this example, is equal to 90.213645, with established control devices 1, 3 and 6. Note that, in [2] is erroneously presented optimal solution for this problem as 90.124 with correct set of established control devices. □

## 3 An improved mathematical formulation

This section presents new mixed integer linear programming model for Problem 3. The binary decision variables $y_j, j = 1, ..., n$ have the same meaning as previous, and it can be introduced one continuous decision variable $v$, which define objective function. Using the previous notation, MILP was formulated as follows:

$$\max v \tag{7}$$

subject to (2) and

$$\sum_{j\in J} \frac{1}{c_{ij}} \cdot y_j \geq v, \quad for\ all\ i \in I \tag{8}$$

$$y_j \in \{0, 1\}\,,\ v \geq 0 \quad for\ allj \in J \tag{9}$$

The optimal objective value of Problem 3 is defined as $opt_{P3} = \frac{1}{obj_{MILP}(y,v)}$, while $obj_{MILP}(y, v)$ is optimal objective function (7) subject to constraints (2), (8) and (9). New formulation have only $n$ binary and 1 continuous decision variables, instead of $n$ binary and $m \cdot n + 1$ continuous decision variables in *MS* model. Therefore, new MILP formulation have much smaller number of decision variables than previous one, so it can be useful even for large-scale problem instances.

## 4 Experimental results

In this section, the computational results of proposed methods and their comparison with existing methods are presented. All experiments were carried out on an Intel Core i5-4670K, 3.4 GHz with 4 GB RAM memory under Windows 7 Professional operating system. In order to compare efficiency of proposed MILP formulations, CPLEX 12.5.1 solver is used.

Computational experiments were first performed, as in [1], on a Single Source Capacitated Plant Location Problem (SSCPLP) instances `http://www-eio.upc.es/~elena/sscplp/index.html` from $m = 20, n = 10$ up to $m = 90, n = 30$ with $p = 5$. The results are presented in Table 1, with following meaning of columns:

- Instance name;
- Number of user nodes $m$;
- Number of potential locations $n$;
- Optimal solution (obtained using both formulations);
- Running time of CPLEX solver, in seconds, using the existing *SM* formulation ([2]);
- Running time of CPLEX solver, in seconds, using the new formulation.

**Table 1** Computational results on SSCFLP instances

| *Inst.* | $m$ | $n$ | *Opt* | *CPLEX time (sec.)* | |
|---|---|---|---|---|---|
| | | | | *SM MILP* | *New MILP* |
| p1 | 20 | 10 | 9.997334 | 0.109 | 0.031 |
| p2 | 20 | 10 | 8.768857 | 0.093 | 0.031 |
| p3 | 20 | 10 | 8.169854 | 0.109 | 0.015 |
| p4 | 20 | 10 | 10.500010 | 0.109 | 0.031 |
| p5 | 20 | 10 | 13.988354 | 0.078 | 0.015 |
| p6 | 20 | 10 | 9.986428 | 0.078 | 0.031 |
| p7 | 30 | 15 | 12.318358 | 0.437 | 0.031 |
| p8 | 30 | 15 | 8.239675 | 0.531 | 0.031 |
| p9 | 30 | 15 | 9.959379 | 0.484 | 0.062 |
| p10 | 30 | 15 | 8.309355 | 0.468 | 0.031 |
| p11 | 30 | 15 | 9.104891 | 0.468 | 0.078 |
| p12 | 30 | 15 | 9.729202 | 0.531 | 0.016 |
| p13 | 30 | 15 | 9.729202 | 0.500 | 0.031 |
| p14 | 30 | 15 | 10.461617 | 0.656 | 0.031 |
| p15 | 30 | 15 | 7.683923 | 0.437 | 0.031 |
| p16 | 30 | 15 | 7.683923 | 0.453 | 0.031 |
| p17 | 30 | 15 | 9.680279 | 0.453 | 0.031 |
| p18 | 40 | 20 | 9.939958 | 1.390 | 0.031 |
| p19 | 40 | 20 | 9.204059 | 2.390 | 0.109 |
| p20 | 40 | 20 | 9.324935 | 2.343 | 0.109 |
| p21 | 40 | 20 | 9.570781 | 1.734 | 0.109 |
| p22 | 40 | 20 | 9.139309 | 1.437 | 0.109 |
| p23 | 40 | 20 | 9.113619 | 1.953 | 0.078 |
| p24 | 40 | 20 | 9.271229 | 2.000 | 0.093 |
| p25 | 40 | 20 | 8.958497 | 1.875 | 0.093 |
| p26 | 50 | 20 | 10.284310 | 2.937 | 0.110 |
| p27 | 50 | 20 | 9.657404 | 2.656 | 0.109 |
| p28 | 50 | 20 | 9.218116 | 2.890 | 0.094 |
| p29 | 50 | 20 | 11.399892 | 2.406 | 0.109 |
| p30 | 50 | 20 | 10.466699 | 2.625 | 0.109 |
| p31 | 50 | 20 | 9.098856 | 1.812 | 0.093 |
| p32 | 50 | 20 | 8.811455 | 2.687 | 0.031 |
| p33 | 50 | 20 | 9.880332 | 3.875 | 0.140 |
| p34 | 60 | 30 | 10.236057 | 27.062 | 0.797 |
| p35 | 60 | 30 | 10.123260 | 33.562 | 0.718 |
| p36 | 60 | 30 | 9.465466 | 35.953 | 0.906 |
| p37 | 60 | 30 | 9.974811 | 30.953 | 0.515 |
| p38 | 60 | 30 | 10.160453 | 34.453 | 0.281 |
| p39 | 60 | 30 | 10.322895 | 45.953 | 0.734 |
| p40 | 60 | 30 | 10.617991 | 32.156 | 0.234 |
| p41 | 60 | 30 | 9.096601 | 22.718 | 0.234 |
| p42 | 75 | 30 | 10.541099 | 63.296 | 0.671 |
| p43 | 75 | 30 | 10.535353 | 52.234 | 0.343 |
| p44 | 75 | 30 | 9.414132 | 42.718 | 0.234 |
| p45 | 75 | 30 | 10.120970 | 46.187 | 0.328 |
| p46 | 75 | 30 | 10.568349 | 45.718 | 1.296 |
| p47 | 75 | 30 | 10.387008 | 53.313 | 0.531 |
| p48 | 75 | 30 | 11.359161 | 51.656 | 0.500 |
| p49 | 75 | 30 | 10.161895 | 32.312 | 0.641 |
| p50 | 90 | 30 | 10.499843 | 62.188 | 1.188 |
| p51 | 90 | 30 | 10.445646 | 78.078 | 0.375 |
| p52 | 90 | 30 | 10.485818 | 70.078 | 1.015 |
| p53 | 90 | 30 | 10.901124 | 73.421 | 1.078 |
| p54 | 90 | 30 | 10.942992 | 96.093 | 2.281 |
| p55 | 90 | 30 | 10.446344 | 74.125 | 0.750 |
| p56 | 90 | 30 | 10.857364 | 60.594 | 0.813 |
| p57 | 90 | 30 | 10.694566 | 53.296 | 0.312 |
| *Sum* | | | | 1261.121 | 18.859 |

**Table 2** Computational results on randomly generated instances

| *Inst.* | $m$ | $n$ | $p$ | *Opt* | *SM MILP* | | *New MILP* | |
|---|---|---|---|---|---|---|---|---|
| | | | | | *Sol* | $t$ (sec) | *Sol* | $t$ (sec) |
| ins01 | 15 | 5 | 3 | 8.886831 | *opt* | 0.015 | *opt* | 0.015 |
| ins02 | 15 | 7 | 4 | 33.178690 | *opt* | 0.015 | *opt* | 0.015 |
| ins03 | 25 | 10 | 5 | 2.562945 | *opt* | 0.062 | *opt* | 0.016 |
| ins04 | 25 | 12 | 7 | 0.038890 | *opt* | 0.062 | *opt* | 0.031 |
| ins05 | 50 | 20 | 15 | 0.014998 | *opt* | 0.187 | *opt* | 0.046 |
| ins06 | 100 | 30 | 15 | 0.012646 | *opt* | 87.42 | *opt* | 0.610 |
| ins07 | 200 | 50 | 20 | 0.009827 | 0.01 | 5285 | *opt* | 129.5 |
| ins08 | 300 | 50 | 25 | 0.007676 | 0.0083 | 14357 | *opt* | 757.2 |

As it can be seen from Table 1, on SSCFLP instances new MILP formulation clearly outperform the *SM* one, since the running times on new formulation are about 2 orders of magnitude smaller than running times on *SM*.

In the paper [2] is performed experiments on the large-scale problem instances that were generated randomly. Unfortunately, these problem instances remained unavailable to the author of this article. Therefore, the author of this article is generated instances, on the same way as in [2], except the different random seed numbers. Generator of these instances is publicly available on link `https://docs.google.com/document/d/1INC7scPd0aTIZQzV97SDiGoRk7dKDsFXNZv69l0TLLw/pub`. This set of instances contains 40 instances with up to $m = 20000$ user nodes and up to $n = 500$ potential locations (same as in [2]), and it can be used as benchmark in future computational experiments on the presented problem.

The results of CPLEX solver using both formulations are given in Table 2, on similar way as in Table 1. Since the CPLEX solver using the *SM* formulation usually obtain "Out of memory" status in the initialization phase, without obtaining any integer solution, data about this formulation is omitted for large random instances (ins09-ins40). Although these instances are large, and CPLEX on new formulation also stopped its work with "Out of memory" status, integer solution (upper bound) and lower bound are reported.

As it can be seen from experimental results on random instances, new MILP formulation again clearly outperform the existing one. Moreover, it uses relatively small number of decision variables, so exact solvers can produce upper and lover bound on very large dimensions.

## 5 Conclusions

This paper considers the problem of efficient exploration of online social networks and a new MILP model for the proposed problem is presented. A new model has small number of decision variables so its running time on each instance is about 2 orders of magnitude smaller than previous one, when obtaining optimal solution. Additionally, it can be used for obtaining upper and lover bounds of reasonable quality in solving large-scale instances. It is also noted some wrong sentences from the literature and tried to correct them.

**Table 3** Upper and lower bounds on large randomly generated instances

| *Inst.* | $m$ | $n$ | $p$ | *New MILP* | | |
|---|---|---|---|---|---|---|
| | | | | *UB* | *LB* | $t$ (sec) |
| ins09 | 500 | 75 | 25 | 0.008548 | 0.006226 | 650 |
| ins10 | 1000 | 100 | 20 | 0.015568 | 0.007452 | 760 |
| ins11 | 1000 | 100 | 50 | 0.003187 | 0.002692 | 566 |
| ins12 | 1000 | 150 | 35 | 0.005276 | 0.003240 | 1348 |
| ins13 | 2000 | 100 | 20 | 0.021249 | 0.007846 | 1577 |
| ins14 | 2000 | 100 | 35 | 0.007164 | 0.004684 | 1416 |
| ins15 | 2000 | 200 | 50 | 0.003387 | 0.002093 | 5679 |
| ins16 | 2000 | 200 | 75 | 0.001916 | 0.001436 | 6509 |
| ins17 | 5000 | 200 | 25 | 0.021089 | 0.004764 | 3390 |
| ins18 | 5000 | 200 | 50 | 0.004493 | 0.002362 | 13178 |
| ins19 | 5000 | 250 | 75 | 0.002242 | 0.001416 | 21780 |
| ins20 | 5000 | 250 | 100 | 0.001429 | 0.001078 | 10509 |
| ins21 | 6000 | 200 | 100 | 0.001475 | 0.001184 | 9685 |
| ins22 | 6000 | 250 | 100 | 0.001570 | 0.001068 | 6762 |
| ins23 | 7000 | 200 | 100 | 0.001530 | 0.001193 | 8056 |
| ins24 | 7000 | 250 | 100 | 0.001580 | 0.001089 | 1543 |
| ins25 | 8000 | 250 | 100 | 0.001721 | 0.001094 | 1350 |
| ins26 | 8000 | 300 | 100 | 0.001594 | 0.001003 | 3786 |
| ins27 | 9000 | 250 | 100 | 0.001681 | 0.001100 | 9781 |
| ins28 | 9000 | 300 | 100 | 0.002105 | 0.001046 | 546 |
| ins29 | 10000 | 300 | 100 | 0.002001 | 0.001018 | 462 |
| ins30 | 10000 | 300 | 150 | 0.001039 | 0.000690 | 532 |
| ins31 | 11000 | 300 | 150 | 0.001001 | 0.000699 | 406 |
| ins32 | 12000 | 300 | 150 | 0.001109 | 0.000696 | 410 |
| ins33 | 13000 | 300 | 150 | 0.000958 | 0.000702 | 499 |
| ins34 | 14000 | 350 | 150 | 0.001152 | 0.000664 | 643 |
| ins35 | 15000 | 350 | 150 | 0.001243 | 0.000685 | 713 |
| ins36 | 16000 | 400 | 200 | 0.000820 | 0.000486 | 643 |
| ins37 | 17000 | 400 | 200 | 0.000723 | 0.000489 | 765 |
| ins38 | 18000 | 450 | 200 | 0.000854 | 0.000472 | 744 |
| ins39 | 19000 | 450 | 250 | 0.000621 | 0.000378 | 835 |
| ins40 | 20000 | 500 | 250 | 0.000634 | 0.000371 | 923 |

The proposed method has the obvious potential to be applied to similar problems that arise from exploration of online social networks. Construction of some exact method based on this formulation is other possible direction of future work.

## References

1. Gajić, Z.: A several new mixed integer linear programming formulations for exploration of online social networks. Working papers, Faculty of Technical Sciences, University of Novi Sad (2014). URL `http://www.optimization-online.org/DB\_HTML/2014/04/4334.html`
2. Stanimirović, Z., Mišković, S.: Efficient metaheuristic approaches for exploration of online social networks. In: W. Hu, N. Kaabouch (eds.) Big Data Management, Technologies, and Applications, pp. 222–269. IGI Global, Hershey, PA (2013)

3. Wagner, J., Falkson, L.: The optimal nodal location of public facilities with price-sensitive demand. Geographical Analysis **7**, 69–83 (1975)
4. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK (1994)