Preliminary Results on Mixed Integer Programming for Searching Maximum Quasi-Bicliques and Large Dense Biclusters

Dmitry Ignatov[1,3]
(0000-0002-6584-8534) dignatov@hse.ru,
Polina Ivanova[1]
(0000-0001-6010-7991) ivanova.p.m@gmail.com,
Albina Zamaletdinova[1]
(0000-0002-4116-9633) aazamaletdinova_1@edu.hse.ru, and
Oleg Prokopyev[2,1]
(0000-0003-2888-8630) droleg@pitt.edu

[1] National Research University Higher School of Economics, Russia
[2] University of Pittsburgh, USA
[3] St. Petersburg Department of Steklov Mathematical Institute of Russian Academy of Sciences, Russia

Abstract. This short paper is related to the problem of finding maximum quasi-bicliques in a bipartite graph (bigraph). A quasi-biclique in a bigraph is its "almost" complete subgraph; here, we assume that the subgraph is a quasi-biclique if it lacks $\gamma \cdot 100\%$ of the edges to become a biclique. The problem of finding the maximal quasi-biclique(s) consists of finding subset(s) of vertices of an input bigraph such that the induced by these subsets subgraph is a quasi-biclique and its size is maximal.
A model based on mixed integer programming (MIP) to search for a quasi-biclique is proposed and tested. Another its variant is tested that simultaneously maximizes both the size of the quasi-biclique and its density, using the least-square criterion similar to the one exploited by TriBox method for tricluster generation. Therefore, the output patterns can be called large dense biclusters as well.

Keywords: quasi-biclique, maximal quasi-biclique, mixed integer programming, biclustering, triclustering

1 Introduction

There are many data sources that can be represented as bipartite graphs. For example, social network data, where the underlying binary relation represents interactions between people and communities, advertisement data with a set of manufacturers and corresponding set of products etc.

In this study, we are interested in analysis of such bipartite data and search for largest dense communities, where almost all elements are connected. A situation where all elements of a community are actually involved can be described by a biclique (complete subgraph) of a bipartite graph.

Unfortunately, the community completeness requirement excludes almost complete communities frequently met in real-world data. Due to this reason we allow some edges to be absent and introduce the concept of quasi-biclique. In order to bound the size of quasi-biclique we can use the subgraph minimal density or the maximum number of absent edges needed to complete a subgraph.

The problem of searching for maximal quasi-clique as well as the problem of searching for maximal clique is NP-hard. Many algorithms that solve the problem are being developed. For instance, Pattillo et al. [1] offered an integer programming model for searching for maximal quasi-clique but the case of bipartite graph with biclique was not studied in details. Note that quasi-bicliques and dense biclusters can be considered as relaxed versions of formal concept definition [2, 3].

The aim of this paper is to propose a mixed integer programming models for finding a maximum quasi-bicliques in a bipartite graph and compare those models with existing ones.

The remainder of the paper is organised as follows. Section 2 proposes two MIP models for quasi-biclique search. In Section 3, we provide the preliminary experimental results. Section 4 concludes the paper.

2 Quasi-biclique searching models

Definition 1. *A γ-quasi-biclique in a bipartite graph $G = (U, V, E)$ is its bipartite induced subgraph $G[V', U']$ with $U' \subseteq U$, $V' \subseteq V$ and its number of edges $|E'| \geq \gamma |V'| \cdot |U'|$, where $\gamma \in (0, 1]$.*

The problem of maximum quasi-biclique in a bipartite graph $G(U,\ V,\ E)$ with fixed γ: $0 < \gamma \leq 1$ is to find $U' \subseteq U$ and $V' \subseteq V$ such that vertex-induced subgraph $G[U', V']$ is a γ-quasi-biclique of size $|U'| + |V'|$, maximum for this graph. Lets denote a maximum γ – quasi-biclique in the graph G by $\omega_\gamma(G)$

Model 1. Here, we adapt model F3 from [4] for searching for maximum quasi-bicliques. We introduce the following variables: $u_i = 1 \Leftrightarrow i \in U'$, $v_j = 1 \Leftrightarrow j \in V'$, $y_{ij} = 1 \Leftrightarrow \exists (i, j) \in E \cap (U' \times V')$, $z_k^{(1)} = 1 \Leftrightarrow |U'| = k$, $z_k^{(2)} = 1 \Leftrightarrow |V'| = k$, $\omega_l^{(1)}, \omega_u^{(1)}$ are the lower and upper bounds for vertex set U', $\omega_l^{(2)}, \omega_u^{(2)}$ are the lower and upper bounds for vertex set V'.

Then we build **Model 1**.

$$\omega_\gamma(G) = \max_{u,v,y,z} \left[\sum_{i \in U} u_i + \sum_{j \in V} v_j \right], \tag{1}$$

$$\text{under conditions } \sum_{(i,j) \in E} y_{ij} \geq \gamma \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} n \cdot m \cdot z_n^{(1)} \cdot z_m^{(2)}, \tag{2}$$

$$y_{ij} \leq u_i, y_{ij} \leq v_j, y_{ij} \geq v_i + v_j - 1, \forall i \in U, \forall j \in V, \tag{3}$$

$$\sum_{i \in U} u_i = \sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} n z_n^{(1)}, \sum_{j \in V} v_j = \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} m z_m^{(2)}, \tag{4}$$

$$\sum_{n=\omega_l^{(1)}}^{\omega_u^{(1)}} z_n^{(1)} = 1, \sum_{m=\omega_l^{(2)}}^{\omega_u^{(2)}} z_m^{(2)} = 1, \tag{5}$$

$$u_i \in \{0,1\}, v_j \in \{0,1\} \ \forall i \in U, \forall j \in V, y_{ij} \in \{0,1\} \ \forall i < j, (i,j) \in E, \tag{6}$$

$$z_n^{(1)} \geq 0 \ \forall n \in \{\omega_l^{(1)}, \ldots, \omega_u^{(1)}\}, z_m^{(2)} \geq 0 \ \forall m \in \{\omega_l^{(2)}, \ldots, \omega_u^{(2)}\}. \tag{7}$$

As in the model F3 we can bound $z_k^{(1)}$ and $z_k^{(2)}$ and recast them into continuous variables.

Remark. *In Model 1, condition 2 is not linear, so we linearize it further. In the worst-case scenario for dense graphs there are $|U|+|V|$ binary variables and $|E|+|U| \cdot |V|$ continuous variables to be optimized.*

Model 2. Let us have a look at a different maximizing criteria for our MIP model from [5] dedicated to triclustering generation. By narrowing this criteria to binary setting, it is possible to produce another maximising criteria for model GF3(f) from [4].

For a bipartite graph $G(U,V,E)$ and its subgraph $G[C_1,C_2]$ the function f is maximized over the density and the size of $G[C_1,C_2]$.

$$f(C_1,C_2) = \rho^2(G[C_1,C_2]) \cdot |C_1| \cdot |C_2| = \frac{(|\{(i,j) : i \in C_1, j \in C_2, (i,j) \in E\}|)^2}{|C_1| \cdot |C_2|}. \tag{8}$$

Using variables definitions from the previous model we can apply logarithm and rewrite f as follows:

$$f_{log}(C_1,C_2) = 2 \cdot \log \left(\sum_{(i,j) \in E} y_{ij} \right) - \log \left(\sum_{i \in U} u_i \right) - \log \left(\sum_{j \in V} v_j \right). \tag{9}$$

Without any extra bounds on vertex sets of a quasi-biclique and the minimum number of edges in it, the model has $2 \cdot (|U| + |V| + |E|)$ variables.

3 Experimental Validation

The greedy algorithm of finding of maximal γ-quasi-bicliques in a bipartite graph was written in Python 2.7. The MIP models were implemented with the optimisation package *IBM Cplex*. All calculations were carried out on a laptop with the MacOs operating system, 2.7 GHz Intel Core i5 processor, RAM 8 GB 1867 MHz.

Datasets for testing the performance of the algorithms are taken from [6, 7]:

1. Divorse in US: 9×50 vertices, 225 edges;
2. Dutch Elite: 3810×937 vertices, 5221 edges;
3. Dutch Elite (TOP-200): 200×395 vertices, 877 edges;
4. Movie-Lens (ml-latest-small): 9125×20 vertices, 20340 edges.

The results of applying the algorithms are presented in the Table 1 for $\gamma = 0.6$. For each algorithm the main parameters are indicated: the algorithm running time (time)[4], the number of founded maximum quasi-bicliques (count) and the maximum size of the founded solution.

If one of the maximal quasi-biclique components has cardinality 1, this is marked in the table as (U', V'), where U' and V' are the sizes of the fractions.

Table 1: Results of maximum γ-quasi-biclique search. Parameters: $\gamma = 0.6$.

Data	**Model 1**			**Model 2**			**Greedy algorithm** [8]		
	time	count	size	time	count	size	time	count	size
Divorse in US	1.23 s	1	54	3.38 s	1	53	360ms	1	(37, 1)
DutchElite (top200)	7602 s	2	(26,1)	181 s	1	14	3 s	1	13
DutchElite	-	-	-	3265 s	1	22	1954 s	1	21
Movie-Lens (small)	28068	2	694	5042 s	1	712	2591 s	1	681

4 Results and conclusions

One can note, that MIP models work an order of magnitude slower than the [8] algorithm, but they find more quasi-cliques and generally each has larger size.

If we consider the results not in terms of speed, but in terms of quality, then Model 2 on the examples of data worked best. This model produced more unique and larger quasi-bicliques than other algorithms (however, only sizes of the maximum quasi-bicliques are mentioned in the tables).

[4] Dashes ("-") in the following tables mean that the algorithm worked too long and did not find a solution.

Acknowledgement. The work of Dmitry Ignatov shown in all the sections has been supported by the Russian Science Foundation grant no. 17-11-01276 and performed at St. Petersburg Department of Steklov Mathematical Institute of Russian Academy of Sciences, Russia.

References

1. Pattillo, J., Veremyev, A., Butenko, S., Boginski, V.: On the maximum quasi-clique problem. Discrete Applied Mathematics **161**(1) (2013) 244 – 257
2. Ignatov, D.I., Kuznetsov, S.O., Poelmans, J.: Concept-based biclustering for internet advertisement. In: 12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012. (2012) 123–130
3. Kaytoue, M., Kuznetsov, S.O., Macko, J., Napoli, A.: Biclustering meets triadic concept analysis. Ann. Math. Artif. Intell. **70**(1-2) (2014) 55–79
4. Veremyev, A.: Exact mip-based approaches for finding maximum quasi-cliques and dense subgraphs. Computational Optimization and Applications **64**(1) (May 2016) 177–214
5. Ignatov, D.I., Gnatyshak, D.V., Kuznetsov, S.O., Mirkin, B.G.: Triadic formal concept analysis and triclustering: searching for optimal patterns. Machine Learning **101**(1) (Oct 2015) 271–302
6. Borgatti, S.P., Everett, M.G., Freeman, L.C.: UCINET. In: Encyclopedia of Social Network Analysis and Mining. (2014) 2261–2267
7. Batagelj, V., Mrvar, A.: Pajek. In: Encyclopedia of Social Network Analysis and Mining. (2014) 1245–1256
8. Liu, X., Li, J., Wang, L.: Quasi-bicliques: Complexity and binding pairs. In Hu, X., Wang, J., eds.: Computing and Combinatorics, Berlin, Heidelberg, Springer Berlin Heidelberg (2008) 255–264