Constraint Community Detection: modelling approaches with applications

Oksana Pichugina[a], Lyudmyla Kirichenko[c], Yurii Skob[a] and Olha Matsiy[c]

[a] *National Aerospace University "Kharkiv Aviation Institute", 17 Chkalova Street, Kharkiv, 61070 Ukraine*

[b] *Kharkiv National University of Radio Electronics, 14 Nauki Avenue, Kharkiv, 61166 Ukraine*

[c] *V. N. Karazin Kharkiv National University, 4 Svobody Sq., Kharkiv, 61022, Ukraine*

Abstract

Community Detection (CD) is a fundamental issue in Network Analysis, focusing on identifying densely connected node groups within a network. Its broader interpretation, Constrained Community Detection (CCD), emerges when supplementary constraints are applied, expanding the scope of the problem. CD has been extensively explored in Network Analysis, boasting numerous developed exact and approximate methods. Conversely, CCD encompasses a more extensive array of real-world issues and applications within Network Analysis. There is a significant need to broaden the spectrum of CD problem variants by establishing rigorous mathematical models. These models would serve as the foundation for developing new exact and heuristic algorithms to solve these problems. This paper investigates various approaches to CCD problem (CCDP) modeling. Specifically, we introduce a novel method for problem modeling that encompasses a broader range of constraints and establish its correlation with the conventional CCDP modeling approach. Additionally, we demonstrate its distinct advantages. The integration of these approaches presents opportunities for extending the class of formalized CCDP as polynomial optimization problems. Consequently, these problems can be efficiently addressed using contemporary nonlinear solvers and can also be transformed into solvable QUBO models applicable to both quantum and digital annealing.

Keywords

Community Detection, Modularity, Constraint Binary Optimization, Integer Programming, Polynomial Optimization, Network, Node partition

Introduction

Network analysis (NA) aims to solve various problems and challenges related to understanding and extracting meaningful insights from network data [1, 2]. The main problems encountered in NA are Community Detection, Edge Clustering, Link Prediction, Centrality Analysis, Dynamic NA, Temporal NA, Heterogeneous NA, Large-Scale NA, Network Alignment, Network Visualization, Network Privacy and Security, Network Robustness, Diffusion and Information Spread, Sampling and Sampling Bias, Graph Isomorphism and Matching, etc. [1, 2]. This list demonstrates the breadth and complexity of NA challenges. Therefore, researchers in NA actively work on developing novel algorithms, techniques, and tools to address these issues and extract meaningful insights from network data of various types.

ProfIT AI 2023: 3rd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2023), November 20–22, 2023, Waterloo, Canada

o.pichugina@khai.edu (O. Pichugina); lyudmyla.kirichenko@nure.ua (L. Kirichenko); y.skob@khai.edu (Y. Skob); olga.matsiy@gmail.com (O. Matsiy)

0000-0002-7099-8967 (O. Pichugina); 0000-0002-2780-7993 (L. Kirichenko); 0000-0003-3224-1709 (Y. Skob); 0000-0002-1350-9418 (O. Matsiy)

CEUR Workshop Proceedings (CEUR-WS.org)

Community Detection and Edge Clustering are two main techniques of network analysis (NA) aiming to uncover structures and patterns within networks [3, 4, 5, 6, 7, 8, 9, 10, 11, 12].

Community detection (CD, graph partitioning or network clustering) is the process of identifying groups of nodes within a network called clusters that are more densely connected than nodes in other clusters.

CD aims to find natural subdivisions into dense clusters within the network called communities. Community detection is a fundamental aspect of network analysis with wide-ranging implications across various domains. Identifying communities within networks provides valuable insights into complex systems' structure, behavior, and function. Here are some reasons demonstrating the importance of community detection: it provides an understanding of complex systems, allows performing Social Network Analysis in particular, to explore collaboration networks and detect criminal networks, performs Cultural and Societal Analysis; enables solving many research problems in Biology and Bioinformatics, e.g., in Epidemiology and Disease Spread; helps to improve management of Recommendation Systems, Urban Planning and Transportation Systems, made proper Market Segmentation and study Customer Behavior; boost financial and economic systems by Fraud Detection and enhancement Cybersecurity issues; rise performance of real networks by improving their Robustness and Resilience; solve numerous problems of Natural Language processing by exploring Semantic Web and Content Organization. Community detection has a wide range of applications across various fields, where the goal is to uncover hidden structures, patterns, and relationships within networks. To understand the numerous applications in the listed research domain, we outline some applications [1, 2, 8, 11, 13, 14, 15, 16]:

1. Identifying groups of friends or communities in social networks, analyzing information flow and influence propagation, detecting online communities in forums and social media platforms.
2. Discovering research communities and collaborations in academic citation networks, identifying influential researchers or papers within specific fields.
3. Investigating criminal networks and identifying key actors and their associations, analyzing patterns of criminal activity and connections.
4. Identifying protein complexes and functional modules in protein-protein interaction networks, analyzing genetic regulatory networks and identifying co-regulated gene groups.
5. Tracking the spread of diseases through contact networks and identifying potential hotspots, analyzing transmission patterns and identifying groups at higher risk.
6. Enhancing recommendation algorithms by considering communities of users with similar interests, grouping web pages with similar content for better search results and content organization.
7. Optimizing routing and resource allocation in communication networks, Identifying clusters of devices in network traffic analysis.
8. Analyzing transportation networks to identify hubs, sub-communities, and traffic patterns, designing efficient public transportation routes based on community structure.
9. Segmenting customers based on purchasing behavior and preferences, analyzing social interactions to understand consumer trends.
10. Identifying groups of users engaging in coordinated fraudulent activities, detecting anomalies and security threats by analyzing network behavior.
11. Organizing and categorizing web content based on thematic communities, enhancing search results by considering community relevance.

Real-world CD problems are often accompanied by additional constraints on nodes, edges and communities, complicating their modeling and significantly affecting solution methods. This paper studies the issue of modeling CD problems with additional constraints (Constraint Community Detection). In particular, we propose a new approach to modeling the problems as Boolean constraint optimization problems.

1. Prerequisites

The application domain of CD is far from limited to the above list. Conducting CD is valuable whenever understanding network structure and relationships is important for making informed decisions or gaining insights into complex systems.

CD is conducted on networks, and it is necessary to distinguish networks and graphs. The terms "network" and "graph" are closely related concepts in the NA and Graph Theory field, but they are used in slightly different ways and contexts.

A graph is a mathematical object that consists of a set of nodes (vertices) and a set of edges that connect pairs of nodes. It is denoted as $G = (V, E)$, where $V = \{1, ..., n\}$ is a node-set (vertex set), E is an edge set. Graphs represent relationships or connections between different entities of various types. Graphs can be directed or undirected, weighted or unweighted.

A network is a collection of interconnected elements. Networks can represent a wide range of real-world systems with relationships or interactions between entities. A network can be represented by nodes representing entities and edges representing interactions of the entities, i.e. any network is representable by a certain graph.

The community detection problem (CDP) in a network G is formulated as an optimization problem, where the goal is to find a partition of nodes into communities that maximizes a certain objective function. Different objective functions capture different aspects of community structure. Modularity [11] and conductance [12] are the two most common.

1.1. Modularity optimization CDP

Modularity Q quantifies the difference between the observed number of edges within communities and the expected number of edges in a random graph. In this paper, modularity is chosen as a criterion of optimization.

The popular formalization (CDP statement) is

- **Input**: a network G by its weighted adjacency matrix $W = [W_{uv}]_{u,v \in V}$; the number k of desired communities (or an upper bound on the number of communities).
- **Output**: a partition $\mathbf{c}$ of the nodes V into communities.
- **Objective function**: maximization of modularity Q.

The modularity function $Q(\mathbf{c}|G)$ assesses the extent to which a partition of network nodes corresponds to the densely-connected node subsets in the network G and is defined as

$$Q(\mathbf{c}|G) = \frac{1}{2m} \sum_{u,v \in V} (W_{uv} - \frac{d_u d_v}{2m}) \mathbf{1}\{c_u = c_v\}, \tag{1}$$

where

- $d_u = \sum_{v \in V} W_{uv}$ is the weighted degree of the node u, $2m = \sum_{u,v \in V} W_{uv}$ is the total weight of the network (in particular, m is the number of edges in G if the graph is unweighted),

- $\mathbf{1}\{c_u = c_v\}$ is an indicator function equal to one if $c_u = c_v$, otherwise, it is annulled; $c_u = c(u) \in [1, k]$ is the community assignment for the node $u \in V$.

The problem (further referred to as **Problem 1**) is to find the community assignment $\mathbf{c}^*$ maximizing the modularity function (1):

$$\text{(Problem 1): } Q(\mathbf{c}^*|G) = \max_{\mathbf{c} \in \mathcal{C}_G} Q(\mathbf{c}|G), \tag{2}$$

where $\mathcal{C}_G$ is a set of community node assignment over G.

The vector of integer variables $\mathbf{c} = [c_u] \in \mathbb{Z}^n_{>0}$ is the network-wide node community assignment satisfying the following constraints.

1. **Labels of communities** are in the range $[1, k]$:

$$1 \leq c_u \leq k, u \in V. \tag{3}$$

2. **Partition constraints.** Each node must belong to exactly one community. In other words, $\mathbf{c}$ induces a partition of V.
   To formalize this condition, let $C_1, ..., C_k$ be a set of communities induced by $\mathbf{c}$. They form a partition of the node set V if

$$\begin{aligned} &|C_1| + ...|C_k| = n, \\ &C_j \cap C_{j'} = \emptyset, j < j', j, j' \in [1, k]. \end{aligned} \tag{4}$$

   In terms of the introduced integer variables, now, the communities can be represented as

$$C_j = \{u \in V \,:\, c_u = j\}, j \in [1, k].$$

3. **Symmetry constraints.**

$$\forall \{u, v\} \subseteq V \quad c_u = c_v \Leftrightarrow c_v = c_u. \tag{5}$$

4. **Transitivity constraints**:
   - For every triple $\{u, v, w\} \subseteq V$, if u, v are in the same community and u, w are in the same community, then v, w are in the same community, i.e.

$$\forall \{u, v, w\} \subseteq V \quad c_u = c_v \text{ and } c_u = c_w \Rightarrow c_v = c_w. \tag{6}$$

   - For every triple $\{u, v, w\} \subseteq V$, if u, v are in the same community and u, w are in different communities, then v, w are also in different communities, i.e.

$$\forall \{u, v, w\} \subseteq V \quad c_u = c_v \text{ and } c_u \neq c_w \Rightarrow c_v \neq c_w. \tag{7}$$

Also, if k is an exact number of communities, the following constraints are used for mathematical formalization: $\forall \kappa \in [1, k] \; \exists u \in V \,:\, c_u = \kappa$ or

$$|C_j| \geq 1, j \in [1, k]. \tag{8}$$

1.2. Approaches to CD

Since CDP is NP-hard, various heuristic and approximation community detection algorithms (CDAs) are used to find approximate solutions effectively [3, 4, 5, 6, 7, 8, 10, 11, 12, 17, 18, 19, 20, 21]. The most common algorithms are:

- Louvain Method [17], which is a greedy optimization algorithm iteratively improving modularity by moving nodes between communities;
- Label Propagation [12], where nodes iteratively adopt the labels of their neighbors until stable labeling is achieved;
- Spectral Clustering [11] involving computing the eigenvectors of specific matrices derived from the graph to find clusters;
- The Girvan and Newman algorithm [22] is a hierarchical community detection method that divides communities by eliminating edges with higher betweenness;
- The Clauset community detection algorithm [13] identifies communities by optimizing the modularity of network partitions;
- The Brandes et al. Community Detection Algorithm is a greedy agglomerative method that utilizes Linear Integer Programming to optimize modularity for community detection;
- The Spin Glass Algorithm [23] is a hierarchical agglomerative approach that minimizes the Hamiltonian of the Potts-like spin model, with spin states symbolizing communities;
- The Walk Trap Algorithm developed by Pons and Latapy [24] is a hierarchical agglomerative method rooted in random walks, initiating from individual clusters.

Most of these methods are applied directly to Problem 1, where the constraints (3)-(7) are present in-explicitly. In addition, CDP can also be easily reformulated as a binary optimization problem (see below) and, respectively, be solved by Integer Programming methods [25] and other Nonlinear Programming techniques [26].

1.3. Modularity optimization CCDP

Constrained Community Detection (Constrained CD, CCD) [9, 14] is a generalization of the standard community detection problem where additional constraints, such as (8), are introduced to guide the process of identifying communities within a network. These constraints can be in the form of prior knowledge, user preferences, or specific requirements related to the network's properties. The goal is to incorporate these constraints into the CD process while optimizing an objective function that captures the desired community structure, such as modularity and conductance.

Among known constraints in CCD are:

- **Community Size Constraints**: for preventing the formation of very small or very large communities, enforcing constraints on the sizes of communities are imposed;

  Let $\underline{n}_j, \overline{n}_j$ be a lower and an upper bound on the size of the community $C_j,\ j \in [1, k]$. In these notations, the constraint is

$$\underline{n}_j \leq |C_j| \leq \overline{n}_j, j \in [1, k]. \tag{9}$$

- **Balance Constraints**: balance requirement for certain community characteristics such as size of communities, distribution of node degrees and node/edge attributes.

  Suppose we are given an upper bound Δ on the difference of sizes of two communities in the community assignment $\mathbf{c}$. Mathematically, it can be expressed as

$$(|C_j| - |C_{j'}|)^2 \leq \Delta^2, j < j', j, j' \in [1, k]. \tag{10}$$

- **Seed Nodes or Labels**: the goal is to ensure that certain nodes are assigned to the specified communities.

  Let $\mathbf{I} \subseteq V$ and $J_u \subset [1, k]$ be a set of communities, where the node u can be assigned for $u \in \mathbf{I}$. This can be written as

$$c_u \in J_u \subseteq [1, k], u \in \mathbf{I}. \tag{11}$$

- **Conflicting constraints**: these constraints express the condition that some nodes must be assigned to different communities. For their formalization, we introduce a set $\mathbf{I}^c$, whose elements are collections of sets of conflicting nodes/ They are assigned to different communities if:

$$\mathbf{I}^c = \{I \subseteq V \, : \, \forall\{u, v\} \subseteq I \; c_u \neq c_v\}. \tag{12}$$

- **Forcing constraints**: these constraints require certain nodes to be assigned to the same community. Similar to conflict ones, we formalize them as follows. First, we introduce a set

$$\mathbf{I}^f = \{I \subseteq V \, : \, \forall\{u, v\} \subseteq I \; c_u = c_v\} \tag{13}$$

  consisting of the sets of forcing nodes, i.e. the ones that need to be assigned to the same community.

- **Hierarchical Constraints**: require hierarchical structuring of the detected communities, where the communities are nested within larger communities;

- **Similarity Constraints**: the goal is to ensure a certain level of similarity between nodes and edges, including their specific attributes.

To Problem 1 complemented by the constraints (9)-(13), we will refer to as **Problem 1.G**. It would be very desirable to reformulate Problem 1 and Problem 1.G as integer optimization problems with variables forming the vector $\mathbf{c}$, and for this, the approaches described in [27, 28] can be applied. However, this has not yet been achieved. That is why other modeling approaches are needed.

Additional constraints in CDP require implementing more complex optimization methods than the standard CD approaches because it is complicated to incorporate additional constraints in the standard community detection heuristics. These methods utilize mathematical models of CCD problems (CCDP) that are not uniquely determined. Accordingly, the effectiveness of using the CCD methods highly depends on these underlying mathematical models.

In this paper, we study approaches to CCDP modelling. In particular, we present a new approach to problem modeling that covers a larger number of constraints and establish its connection with the standard approach to CCDP modeling. Combining these two approaches opens up prospects for expanding the class of formalized CCD problems in the form of polynomial optimization problems. Respectively, these CCDPs can be solved effectively using contemporary nonlinear solvers [29] and also reduced to QUBO models [30] solvable by quantum and digital annealers.

2. Modelling CCD problems as binary programs

We will formalize additional constraints using binary variables. In order to accomplish this, first, we reformulate the integer programming problem (2), whose dimension is n, as a binary optimization problem of higher dimension.

2.1. Standard CCD modelling approach

First, we recall the standard approach [9, 14, 31, 32] that allows formalizing some of the abovementioned constraints.

Suppose variables form a square matrix of binary variables of the size n:

$$Y = [y_{uv}]_{u,v \in V} \in \mathbb{B}^{n \times n},$$

where $\mathbb{B} = \{0, 1\}$,

$$y_{uv} = \begin{cases} 1 \text{ if the nodes } u \text{ and } v \text{ are in the same community,} \\ 0, \text{ otherwise.} \end{cases} \tag{14}$$

Then the expressions (1) and (2) for the objective can be rewritten as

$$\text{(Problem 2):} \quad F(Y^*) = \max_{Y \in \mathbb{B}^{n \times n}} F(Y), \tag{15}$$

$$F(Y) = \frac{1}{2m} \sum_{u,v \in V} (W_{uv} - \frac{d_u d_v}{2m}) y_{uv}. \tag{16}$$

When the matrix Y^* is found, then the community assignment $\mathbf{c}^*$ is formed as follows: an arbitrary node $u \in V$ is selected and is assigned to the community C_1 along with all other nodes belonging to the same community as u. The process of community assignment continues iteratively for unassigned nodes and the communities $C_2, ..., C_k$.

Certain constraints must be added to (15).

1. CDP constraints
   a) The constraint (3) cannot be written in terms of the variables (14). However, there are CD algorithms where the number of communities can be specified in advance [12].
   b) The constraint (4) holds due to the above way of the node-to-community assignment.
   c) The symmetry constraint (5) is formalized as

$$\forall \{u, v\} \subseteq V \quad y_{uv} = y_{vu}. \tag{17}$$

   d) The constraint (6) takes the form of

$$\forall \{u, v, w\} \subseteq V \quad y_{vw} \geq y_{uv} + y_{uw} - 1. \tag{18}$$

   e) The constraint (7) becomes

$$\forall \{u, v, w\} \subseteq V \quad y_{vw} \leq y_{uv} + y_{uw}. \tag{19}$$

   f) Likewise (3), the constraint (8) cannot be written in terms of Y-entries.
2. CCDP constraints are only partially representable in terms of the Y-elements.
   a) (9), (10) and (11) are not formalized in terms of y_{uv}.

b) The condition (12) can be written as follows:

$$\forall I \in \mathbf{I}^c \ : \ \forall \{u, v\} \subseteq I \ y_{uv} = 0. \tag{20}$$

c) Similar to (20), the constraints (13) are rewritten as:

$$\forall I \in \mathbf{I}^f \ : \ \forall \{u, v\} \subseteq I \ y_{uv} = 1. \tag{21}$$

Thus, among the above CD and CCD constraints, five groups are not formalized in terms of y_{uv} and require other approaches for formalization.

The dimension of this CCDP given by (15)-(21) (further referred to as **Problem 2.G**) is n^2. It can be reduced to $\frac{n^2-n}{2}$ if the symmetry constraint is utilized for eliminating the variables y_{uv} satisfying the relation $u \geq v$.

The advantages of the formalized constraints in Problems 2, 2.G and objective functions are their linearity, while the disadvantage is the unknown, at the moment, the approach to formalizing the rest of the constraints in terms of entries of Y.

2.2. New CCD modelling approach

In this section, we present our approach to CCD modelling, also utilizing binary variables.

Let us introduce another matrix of binary variables

$$X = [x_{uj}]_{u \in V, j \in [1,k]} \in \mathbb{B}^{n \times k},$$

where

$$x_{uj} = \begin{cases} 1 \text{ if } c_u = j, \\ 0, \text{ otherwise.} \end{cases}$$

In terms of these variables, the indicator function in (1) is representable as

$$\mathbf{1}\{c_u = c_v\} = f_1(X, u, v) = 1 - \frac{1}{2} \sum_{j=1}^{k} (x_{uj} - x_{vj})^2$$

that allows rewriting the modularity function (2) in terms of the introduced binary variables:

$$\begin{gathered} Q'(X) = \tfrac{1}{2m} \sum_{u,v \in V} (W_{uv} - \tfrac{d_u d_v}{2m}) f_1(X, u, v) = \\ A_0 + \sum_{u,v \in V} a_{uv} \sum_{j=1}^{k} (x_{uj} - x_{vj})^2, \end{gathered} \tag{22}$$

where

$$\begin{gathered} A_0 = 2 \sum_{u,v \in V} a_{uv}; \\ A = [a_{uv}]_{u,v \in V} \ : \ a_{uv} = -\tfrac{1}{2}(W_{uv} - \tfrac{d_u d_v}{2m}), u, v \in V. \end{gathered}$$

We came to the binary formulation of problem (2) (further referred to as **Problem 3**): find the binary matrix X^*, where the minimum of $Q'(X)$ is attained, i.e.,

$$\text{(Problem 3): } Q'(X^*) = \min_{X \in \mathbb{R}^{m \times k}} Q'(X), \tag{23}$$

where $Q'(X)$ is given by (22), and the one-hot constraints hold:

$$\sum_{j=1}^{k} x_{uj} = 1, u \in V. \tag{24}$$

Problem 3 is the constrained binary optimization problem with the quadratic objective (22) and linear equality constraints (24). Let us formalize in terms of X-entries the rest of the above CDP and CCDP constraints.

1. CDP constraints.
   a) The condition (3) holds automatically since the matrix X has k columns.
   b) Fulfillment of the condition (4) is ensured by the one-hot constraint (24).
   c) In order to write out the symmetry constraint (5) in terms of X, first, we establish connection between elements of the matrices X and Y:

   $$y_{uv} = \sum_{j\in[1,k]} x_{uj}x_{vj}, \{u, v\} \subseteq V. \tag{25}$$

   Making the substitution (25) into (17), we come to the quadratic equality constraint:

   $$\forall\{u, v\} \subseteq V \quad \sum_{j\in[1,k]} x_{uj}x_{vj} = \sum_{j\in[1,k]} x_{vj}x_{uj},$$

   which is, clearly, redundant.
   d) The constraints (6) and (7) also hold automatically.
   e) The condition (8) is easily written in terms of the new variables, taking into account that $|C_j| = \sum_{u\in V} x_{uj}, j \in [1, k]$:

   $$\textstyle\sum_{u\in V} x_{uj} \geq 1, j \in [1, k].$$

2. CCDP constraints:
   a) In terms of x_{uj}, the constraint (9) looks like:

   $$\underline{n}_j \leq \sum_{u\in V} x_{uj} \leq \overline{n}_j, j \in [1, k]. \tag{26}$$

   b) In terms of X-entries, the constraint (10) is:

   $$(\sum_{u\in V} x_{uj} - \sum_{u\in V} x_{uj'})^2 \leq \Delta^2, j < j', j, j' \in [1, k]. \tag{27}$$

   c) In terms of x_{uj}, the condition (11) is represented as

   $$\sum_{j\in J_u} x_{uj} = 1, u \in \mathbf{I}. \tag{28}$$

   d) The binary representation of (12) is

   $$\forall I \subseteq \mathbf{I}^c : \quad \sum_{\{u,v\}\subseteq I} \sum_{j=1}^{k} (x_{uj} - x_{vj})^2 = C^2_{|I|} \cdot 2 = |I|(|I| - 1). \tag{29}$$

e) The constraint takes the form of

$$\forall I \subseteq \mathbf{I}^f \; : \quad \sum_{\{u,v\}\subseteq I} (2 - \sum_{j=1}^{k} (x_{uj} - x_{vj})^2) = |I|(|I| - 1)$$

that can be simplified to

$$\forall I \subseteq \mathbf{I}^f \; : \; \textstyle\sum_{\{u,v\}\subseteq I} \sum_{j=1}^{k} (x_{uj} - x_{vj})^2 = 0. \tag{30}$$

The dimension of Problem 3 is $n \times k$.

It is seen that Problem 3 is a binary problem with a convex quadratic objective function and linear constraints, i.e. it is quadratic binary problem. Also, we came to its CCD generalization having the form of Problem 3 with the additional constraints (26)-(30), further referred to as **Problem 3.G**. These constraints can be present in any combination.

Due to the presence of the quadratic constraints (27)- (30), it belongs to the class of quadratically constrained binary optimization problems with the quadratic objective and constraints. Moreover, the constraints (29) and (30) are quadratic equality constraints, i.e. they are non-convex in contrast to the convex objective (22) and the convex inequality constraints (27). Thus, attacking Problem 3.G we deal with a non-convex binary optimization problem.

3. Discussion

Let us compare the models Problems 2 and 3 and their generalization, Problems 2.G and 3.G.

Comparing the dimensions of the models, the advantage remains with Problems 3 and 3.G since the upper bound k on the number of communities is normally much smaller than the number of nodes n, respectively, $n^2 \gg n \cdot k$.

In contrast to Problem 3 and Problem 3.G, Problem 2 and its CCD generalization, Problem 2.G, are linear binary problems but do not completely formalized.

Accordingly, only the models Problem 3 and Problem 3.G coped with formalizing the CDP and CCDP as a binary optimization problem. The obtained models have the quadratic objective representing modularity and linear or quadratic constraints. Therefore, the models can be directly solved by general nonlinear solvers, disregarding if the binary variables are supported.

The variables of the matrices X and Y have different meanings. Namely, the elements of X reflect the relationship between two nodes, while the elements of Y reflect the relationship between a node and a community. Supposedly, other constraints can be formalized by a combination of X, Y-entries, thus covering a much wider class of CCDPs.

4. Conclusion

This paper attacks a critical task in Network Analysis called Constrained Community Detection (CCD). Binary optimization was chosen as a modeling tool. A new approach to modeling these problems is presented, significantly expanding the set of formalized constraints. A comparison with the standard modeling approach is made, demonstrating the advantages of our approach. Both approaches can be combined, forming new CCD models in the form of polynomial binary optimization problems. Nonlinear, including polynomial, solvers can be used to solve them. The polynomial and binary nature of the models also allows their reduction to popular QUBO models, which are solved very effectively on quantum and digital annealers.

References

[1] J. Scott, Social Network Analysis, 4th ed., Sage Publications, 2017.

[2] M. E. V. Valkenburg, Network Analysis, 3rd ed., Pearson College Div, 1974.

[3] D. K. Sewell, Model-based edge clustering, Journal of Computational and Graphical Statistics 30 (2020) 390–405. doi:10.1080/10618600.2020.1811104.

[4] B. Farzad, O. Pichugina, L. Koliechkina, Multi - layer community detection, in: 2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), 2018, pp. 133–140. doi:10.1109/ICCAIRO.2018.00030.

[5] O. Pichugina, B. Farzad, A human communication network model, in: Proceedings of the 12th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer, volume 1614, CEUR, 2016, pp. 33–40. URL: https://ceur-ws.org/Vol-3403/paper21.pdf, issn 1613-0073.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (2008) P10008. doi:10.1088/1742-5468/2008/10/P10008.

[7] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Physical Review E 72 (2005) 027104. doi:10.1103/PhysRevE.72.027104.

[8] E. Eaton, R. Mansbach, A spin-glass model for semi-supervised community detection, Proceedings of the AAAI Conference on Artificial Intelligence 26 (2012) 900–906. doi:10.1609/aaai.v26i1.8320, number: 1.

[9] K. Eguchi, T. Murata, Constrained community detection in multiplex networks, in: G. L. Ciampaglia, A. Mashhadi, T. Yasseri (Eds.), Social Informatics, Lecture Notes in Computer Science, Springer International Publishing, 2017, pp. 75–87. doi:10.1007/978-3-319-67217-5_6.

[10] C. F. A. Negre, H. Ushijima-Mwesigwa, S. M. Mniszewski, Detecting multiple communities using quantum annealing on the d-wave system, PLOS ONE 15 (2020) e0227538. doi:10.1371/journal.pone.0227538, publisher: Public Library of Science.

[11] M. E. J. Newman, Modularity and community structure in networks, Proceedings of the National Academy of Sciences 103 (2006) 8577–8582. doi:10.1073/pnas.0601602103.

[12] P. Wagenseller III, F. Wang, Size matters: A comparative analysis of community detection algorithms, 2017. URL: http://arxiv.org/abs/1712.01690.

[13] A. Clauset, M. E. J. Newman, C. Moore, Finding community structure in very large networks, Physical Review E 70 (2004) 066111. doi:10.1103/PhysRevE.70.066111.

[14] W. D. Viles, A. J. O'Malley, Constrained community detection in social networks, 2017. URL: http://arxiv.org/abs/1708.04354.

[15] A. Lancichinetti, S. Fortunato, Community detection algorithms: A comparative analysis, Physical Review E 80 (2009) 056117. doi:10.1103/PhysRevE.80.056117, publisher: American Physical Society.

[16] M. E. J. Newman, Communities, modules and large-scale structure in networks, Nature Physics 8 (2012) 25–31. doi:10.1038/nphys2162, number: 1 Publisher: Nature Publishing Group.

[17] V. A. Traag, L. Waltman, N. J. van Eck, From louvain to leiden: guaranteeing well-connected communities, Scientific Reports 9 (2019) 5233. doi:10.1038/s41598-019-41695-z, number: 1 Publisher: Nature Publishing Group.

[18] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences 99 (2002) 7821–7826. doi:10.1073/pnas.122653799, publisher: Proceedings of the National Academy of Sciences.

[19] S. Yakovlev, O. Kartashov, O. Pichugina, Optimization on combinatorial configurations using genetic algorithms, in: Proceedings of the Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2019), volume 2353, CEUR, 2019, pp. 28–40. URL: http://ceur-ws.org/Vol-2353/paper3.pdf, issn 1613-0073.

[20] S. Fortunato, Community detection in graphs, Physics Reports 486 (2010) 75–174. doi:10.1016/j.physrep.2009.11.002.

[21] S. E. Schaeffer, Graph clustering, Computer Science Review 1 (2007) 27–64. doi:10.1016/j.cosrev.2007.05.001.

[22] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E 69 (2004) 026113. doi:10.1103/PhysRevE.69.026113.

[23] J. Reichardt, S. Bornholdt, Statistical mechanics of community detection, Physical Review E 74 (2006) 016110. doi:10.1103/PhysRevE.74.016110, publisher: American Physical Society.

[24] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: p. Yolum, T. Güngör, F. Gürgen, C. Özturan (Eds.), Computer and Information Sciences - ISCIS 2005, Lecture Notes in Computer Science, Springer, 2005, pp. 284–293. doi:10.1007/11569596_31.

[25] L. A. Wolsey, Integer Programming, 2nd ed., Wiley, 2020.

[26] S. Yakovlev, O. Pichugina, On constrained optimization of polynomials on permutation set, in: Proceedings of the Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2019), volume 2353, CEUR, 2019, pp. 570–580. URL: http://ceur-ws.org/Vol-2353/paper45.pdf, issn 1613-0073.

[27] O. Pichugina, S. Yakovlev, Euclidean combinatorial configurations: Continuous representations and convex extensions, in: V. Lytvynenko, S. Babichev, W. Wójcik, O. Vynokurova, S. Vyshemyrskaya, S. Radetskaya (Eds.), Lecture Notes in Computational Intelligence and Decision Making, Advances in Intelligent Systems and Computing, Springer International Publishing, 2020, pp. 65–80. doi:10.1007/978-3-030-26474-1_5.

[28] O. Pichugina, S. Yakovlev, Euclidean combinatorial configurations: Typology and applications, in: 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2019, pp. 1065–1070. doi:10.1109/UKRCON.2019.8879912.

[29] A. Wächter, L. T. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming, Mathematical Programming 106 (2006) 25–57. doi:10.1007/s10107-004-0559-y.

[30] A. P. Punnen, The Quadratic Unconstrained Binary Optimization Problem: Theory, Algorithms, and Applications, 1st ed., Springer Nature, 2022.

[31] S. Aref, H. Chheda, M. Mostajabdaveh, The bayan algorithm: Detecting communities in networks through exact and approximate optimization of modularity, 2023. doi:10.48550/arXiv.2209.04562.

[32] G. Agarwal, D. Kempe, Modularity-maximizing graph communities via mathematical programming, The European Physical Journal B 66 (2008) 409–418. doi:10.1140/epjb/e2008-00425-1.