Leveraging Multi-Criteria Integer Programming Optimization for Effective Team Formation

Pallavi Singh 1, Phat K. Huynh 2, Dang Nguyen 2, Trung Le 2, and Wilfrido Moreno 2

1University of South Florida
2Affiliation not available

December 7, 2023

Abstract

The research addresses a critical aspect of organizational settings – team formation – which is vital for effective team-based learning. In many organizational environments, successful teamwork depends on the composition of teams that consider various criteria, such as skill levels, background, and personality traits. Meanwhile, maintaining fairness and equity among teams is crucial to ensure equal opportunities for all team members. Our research framework lies in its ability to model a diverse range of factors, while simultaneously maximizing within-team diversity and minimizing conflict levels.

In this study, we introduce a novel application of multi-criteria integer programming (MCIP) that diverges from traditional and single-objective optimization methods to address the complexity of team formation with a focus on multiple criteria. This approach offers a comprehensive solution that accommodates multiple criteria simultaneously.

The contributions of this paper are:
1) An innovative optimization model for team formation that addresses the complex task of maximizing intra-group diversity while minimizing inter-group diversity.
2) Addressing the challenge of forming balanced and diverse student teams and providing educators with a practical tool for effective team formation.

Hosted file

Leveraging Multi-Criteria Integer Programming Optimization for Effective Team Formation

Pallavi Singh*, Phat K. Huynh*, Dang Nguyen, Trung Q. Le, and Wilfrido Moreno

Abstract—Team-based learning (TBL) plays a significant role in many organizational settings, necessitating a sophisticated and comprehensive approach to team formation. This study introduces an innovative application of multi-criteria integer programming, offering a solution that inherently accommodates multiple criteria simultaneously, thereby addressing the intricate task of team formation compared to traditional single-objective optimization methods. Our research designs a framework capable of modelling a broad range of diverse factors, including skill levels, background, and personality traits. The framework's objective function has been optimized to maximize within-team diversity, while minimizing conflict levels and the variance in diversity levels between teams. Our approach is further enhanced by the inclusion of explicit constraints such as potential interpersonal conflicts, a factor often overlooked in previous studies. Furthermore, we propose a two-stage optimization method that strategically divides the population into sub-populations using a weighted heterogeneous multivariate K-means algorithm and optimizes the team formation within these sub-populations through a surrogate optimization approach. This process effectively manages MCIP problems involving large populations. Results from this study demonstrate the robustness of our model across various simulation scenarios with different levels of data heterogeneity. Through rigorous validation, we offer a theoretically tested solution to the critical research gaps in the literature. In conclusion, our study provides a comprehensive, flexible, and well-validated approach to team formation. This advancement in the TBL field paves the way for future research to explore and enhance this model, providing more sophisticated and efficient team formation strategies.

Index Terms—Team formation problem, team-based learning, multi-objective integer programming, and optimization methods.

I. INTRODUCTION

Team formation problem (TFP) is a widely recognized challenge in various fields such as operation research [1], computer science [2], management [3], and education [4]. The central task in TFP is to organize a group with diverse characteristics into interconnected and effective teams [5-9]. Team-based learning (TBL) [10-13] is a pedagogical strategy that heavily relies on the effective resolution of the TFP in educational and professional settings. In educational settings, effective team formation can enhance students' learning experiences by fostering collaborative skills, and increasing engagement [10, 14]. Furthermore, diverse and balanced teams are known to enrich the learning experience, since members can benefit from different perspectives and experiences [15, 16]. In professional settings, the strategic assembly of teams plays a vital role in an organization's success. Teams are often formed to tackle specific projects, solve complex problems, or innovate new ideas. The diversity of a team's skills, expertise, and perspectives can greatly influence the quality of the team's outputs and overall effectiveness [17, 18]. However, forming optimal teams is a complex task, which requires the consideration of multiple factors, such as individual skills, compatibility, and workload balance.

The increasing class sizes in academic institutions and the task complexities in professional environments impose the development of automated and optimized methods for forming balanced and diverse teams. A well-structured team ideally represents a combination of various skills, backgrounds, and perspectives that can mutually enhance its problem-solving capacity and productivity [18, 19]. Conventional practices of forming teams, often rely on heuristic and subjective approaches, may be inadequate to handle the diversity and dynamics of contemporary classrooms and workplaces [14, 20]. Traditionally formed teams often fail to incorporate a comprehensive range of student attributes like academic background, technical skills, and interpersonal skills,
which profoundly influence the efficacy and satisfaction levels of a team [21, 22]. Furthermore, arbitrarily formed teams may lead to skill gaps and workload imbalances, hampering team performance and causing dissatisfaction and burnout [23, 24]. Therefore, the deployment of automated and optimized team formation strategies can provide a solution to these challenges by considering a broad spectrum of factors and forming teams that optimize specific objectives, such as skills diversity, equal workload distribution, and interpersonal compatibility [8, 25]. Using optimization algorithms in team formation can enable educators and managers to harness computational power and handle the task complexities, thereby creating teams that augment learning outcomes and enhance work productivity [26, 27].

This study aims to develop an optimization method for team formation, accounting for multiple criteria that impact team performance and satisfaction. Our goals are to address the complexities inherent in team formation, specifically the need for balance and diversity while minimizing conflict across various factors. Our approach aims to produce effective, balanced, and harmonious teams. The scope of this study primarily lies within the academic setting, focusing on team formation for collaborative learning initiatives such as TBL. We recognize the necessity for team diversity in these settings - a diversity that spans skills, experiences, and perspectives - as it fosters creativity, innovation, and improved problem-solving [17, 28]. Hence, one of the primary objectives of our optimization model is to maximize intra-group diversity, which encourages the formation of heterogeneous teams. Conversely, we aim to minimize inter-group diversity, ensuring equity and fairness across teams. It’s crucial in an academic setting to avoid disparities that might lead to uneven competition or exclusionary practices [29, 30]. Since diversity can lead to potential conflicts arising from differences in personal values, attitudes, or cultural beliefs [31, 32], our model also considers minimizing the conflict level to maintain smooth teamwork and collective productivity.

The remainder of this paper is structured as follows: Section II reviews existing TBL methods and identifies gaps in current team formation models. In Section III, we elaborate on the multi-criteria integer programming for TBL with the detailed mathematical formulation, the optimization strategy, and the approach to data simulation and validation. Section IV presents the results, including data simulation, proposed model validation, and a comparative performance analysis with existing models. Section V discusses the findings and their implications, while Section VI concludes the paper by summarizing the key findings and outlining future research directions.

II. LITERATURE REVIEW

Team formation within academic settings has its own unique challenges and requirements. The shift towards active learning methodologies, such as TBL, has amplified the importance of proper team formation in educational contexts. Factors such as student learning styles, academic performance, interpersonal skills, time availability, and even cultural backgrounds should be carefully considered [14, 33, 34]. Traditionally, instructors have often taken the reins in forming teams in academic environments, relying on their knowledge and judgment about the students’ behaviors, skills, personalities, and academic performance [35, 36]. This instructor-guided approach, although personalized and adaptable, is qualitative, subjective, and tends to be time-consuming in larger classes. Moreover, these methods may overlook some intricate team dynamics, like the potential conflicts or synergies between students, or fail to balance the team in terms of diverse skills and backgrounds [37, 38]. Furthermore, the complexity of the team formation process increases with the rise in class sizes and the diverse range of student attributes that need to be considered, highlighting the need for more sophisticated and scalable team formation methods in educational settings [39].

To enhance team formation for team-based learning, researchers have proposed various computational methods to overcome the limitations of traditional approaches. Early investigations in this field focused on enhancing team performance based on individual skills or expertise, utilizing models like linear and integer programming with single objective [40, 41]. In addition, matching algorithms like the Stable Marriage [42], Gale-Shapley [43], and Hungarian algorithms [44] have been employed, using mathematical models to balance team members’ compatibility based on preferences, skills, and characteristics. Clustering algorithms like K-means clustering [45], hierarchical clustering [46], and density-based clustering [47] aim to create teams by grouping individuals based on similarities in skills, interests, or other relevant criteria. Notably, more advanced methods such as genetic algorithms (GAs) have been utilized to optimize multiple team attributes like skills, preferences, and demographics [48-52]. Similarly, machine learning techniques have been used to suggest optimal team configurations based on individual attributes [53, 54], and expert systems [55, 56] have employed rules and heuristics based on domain knowledge to assist in team formation. However, these methods still face challenges, especially in scalability, robustness, and accommodating the complex nature of the problem, including factors like individual skillsets, potential for collaboration, and various constraints.
Following these computational strategies, a distinct approach known as the maximum diversity grouping problem (MDGP) has been explored to further enhance team formation in the context of team-based learning [57, 58]. This method promotes team diversity by considering a multitude of factors. Regarded as an extension of the MDGP, the MDGP's objective is to distribute students into unique, non-overlapping groups, thereby maximizing the sum of differences between each pair of individuals within the same group [59]. As this problem has attracted significant attention, there have been extensive research initiatives and algorithmic solutions proposed to tackle the MDGP formulation [58, 60, 61]. However, the key limitations of the MDGP approach include challenges in scalability as the computational complexity grows exponentially with the number of objects and groups [62], and sensitivity to the quality of input data, which influences the quality of obtained solutions [62]. Thus, while the MDGP presents a novel perspective on team formation, its limitations necessitate a more robust and scalable method that can comprehensively address the complex nature of TBL problem. Understanding the need for an evolved approach, we propose a new methodology utilizing multi-criteria integer programming (MCIP) to address these research gaps and overcome the limitations inherent in current team formation studies.

Our innovative application of MCIP diverges from traditional and single-objective optimization methods by inherently considering multiple criteria simultaneously, providing a comprehensive solution to the intricate task of team formation. In this study, we present several novel contributions to the field of TBL. First, our framework models a broad set of diverse factors including skill levels, background, and personality traits, addressing a significant research gap left by studies that only consider a few aspects of diversity. In addition, the framework's objective function is specifically designed to maximize within-team diversity while minimizing the conflict level and the variance in diversity levels between teams, offering a sophisticated approach to team formation. Our methodology also includes modeling explicit constraints such as potential interpersonal conflicts, which previous studies have overlooked. Moreover, we propose a two-stage methodology that strategically divides the student population into sub-populations by using a weighted heterogeneous multivariate K-means algorithm and optimizes the team formation for these sub-populations through a surrogate optimization approach. This can handle this MCIP problem effectively with large student populations, demonstrating an improvement over conventional method. Lastly, our framework includes rigorous model validation, demonstrating its efficacy with real-world scenarios with different levels of data heterogeneity, thus addressing the critical gap in the current literature that lacks robust validation. Hence, our proposed framework offers a comprehensive, flexible, and empirically validated approach to team formation.

III. PROPOSED MULTI-CRITERIA INTEGER PROGRAMMING TEAM FORMATION FRAMEWORK

A. Overview

The core proposition of our paper lies in the introduction and examination of a novel MCIP-based framework aimed at addressing the complexities inherent to team formation. The primary aim of our approach is to maximize the levels of intra-group diversity while minimizing the inter-group diversity and conflict levels within teams. The overall framework is visualized in Fig. 1 below:
**B. Objectives**

The primary objectives of our proposed MCIP-based model are:

1) **Maximizing Intra-group Diversity**: Our primary objective is to enhance the diversity within each team, which leads to more heterogeneous perspectives, skills, and ideas in problem-solving.

2) **Minimizing Inter-group Diversity**: The model aims to ensure that the variance between different groups is minimized, thus maintaining equity across teams.

3) **Minimizing Conflict Level**: Recognizing that diversity can also give rise to potential conflicts due to differences in personal values, attitudes, cultural beliefs, etc., the model strives to minimize the level of conflict within each team.

**C. Mathematical Formulation and Objective Functions**

We propose a MCIP-based model, which optimizes these objectives to assign totally \( N \) individuals to \( K \) groups while ensuring that a set of constraints are met to maintain the feasibility of team formation. We denote the binary decision variables \( z_{ig} = 1 \), \( \forall i = 1, ..., N, g = 1, ..., G \), if the individual \( i \) is assigned to the group \( k \), and \( z_{ig} = 0 \) otherwise. Therefore, the decision variables are represented by a binary matrix \( Z \in \{0,1\}^{N \times G} \). We denote \( S_g = \{s_1, ..., s_{n_g}\} \), where \( n_g \) is the total number of individuals in group \( g \) and \( S_g \subset \{1, ..., N\} \), as the set of the individuals that belong to group \( g \). Based on Section III.B, the objective function of the proposed framework is formulated as a weighted sum of three key components — intra-group diversity level, inter-group diversity level, and intra-group conflict level. The objective function is mathematically defined as:

\[
\max_{s_1, ..., s_G} \sum_{g=1}^{G} \left[ \lambda_1 DL(S_g) - \lambda_2 CL(S_g) \right] - \lambda_3 \frac{2}{G(G-1)} \sum_{i=1}^{G} \sum_{j=i+1}^{G} \left[ DL(S_i) - DL(S_j) \right]^2
\]

where \( DL(S_g) \) and \( CL(S_g) \) denote the diversity level and of the group \( g \) respectively, and \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) are the normalized weights \( \lambda_1 + \lambda_2 + \lambda_3 = 1 \) associated with the intra-group diversity level, the conflict level, and the inter-group diversity level. Notably, this MCIP-based model can be applied to any team formation scenario that requires the balance of these components. Although the paper later provides a focused application in student team formation as an illustrative example, the
methods still hold potential for broader applications. The intra-group diversity level $DL(S_g)$ is formulated as:

$$DL(S_g) := \frac{2}{|S_g|(|S_g| - 1)} \sum_{i,j \in S_g, i < j} D_{ij} \quad (2)$$

where $|S_g|$ is the cardinality of the set $S_k$, and $D_{ij}$ is the diversity level of the individual pair $i$ and $j$. This diversity level $D_{ij}$ is computed by considering the weighted Euclidean distance $E(X_{c,i}, X_{c,j})$ [63] for continuous variables $X_c$ normalized to the range $[0, 1]$, the weighted Hamming distance $H(X_{d,i}, X_{d,j})$ [64] for discrete variables $X_d$, and the weighted Jaccard distance $J(X_{m,i}, X_{m,j})$ [65] for multi-valued variables $X_m$. The notations $X_{c,i}$, $X_{c,j}$, and $X_{m,i}$ represent the values of individual $i$'s continuous, discrete, and multi-valued attributes, respectively. Those distance functions are:

$$E(X_{c,i}, X_{c,j}) := \sum_{x_{c} \in X_{c}} w_{x_c} (x_{c,i} - x_{c,j})^2 \quad (3)$$

$$H(X_{d,i}, X_{d,j}) := \sum_{x_{d} \in X_{d}} w_{x_d} 1_{\{x_{d,i} \neq x_{d,j}\}} \quad (4)$$

$$J(X_{m,i}, X_{m,j}) := \sum_{x_{m} \in X_{m}} w_{x_m} \left(1 - \frac{|x_{m,i} \cap x_{m,j}|}{|x_{m,i} \cup x_{m,j}|}\right) \quad (5)$$

Here, $1\{x_{d,i} \neq x_{d,j}\}$ is an indicator function that equals 1 if the discrete values of individuals $i$ and $j$ differ, and 0 otherwise. Next, $D_{ij}$ is calculated as the average of these weighted distances:

$$D_{ij} := \frac{1}{3} \left[ E(\cdot) + H(\cdot) + J(\cdot) \right] \quad (6)$$

The normalized weights $w_{x_c}$, $w_{x_d}$, and $w_{x_m}$ represent the importance of each variable in the sets $X_c$, $X_d$, and $X_m$, in which $\sum_{x_c \in X_c} w_{x_c} = 1$, $\sum_{x_d \in X_d} w_{x_d} = 1$, and $\sum_{x_m \in X_m} w_{x_m} = 1$. The conflict level of a group $CL(S_g)$ represents the potential for disagreement or discord within a group due to differences in attributes such as personal values, attitudes, and cultural beliefs. Similar to $DL(S_g)$, it is measured at a group level by aggregating pair-wise conflict levels among the individuals, denoted by $C_{ij}$ in the group:

$$CL(S_g) := \frac{2}{|S_g|(|S_g| - 1)} \sum_{i,j \in S_g, i < j} C_{ij} \quad (7)$$

$$C_{ij} := \frac{1}{3} \left[ E(\cdot) + H(\cdot) + J(\cdot) \right] \quad (8)$$

where $Y_c$, $Y_d$, and $Y_m$ represent the sets of continuous, discrete, and multi-valued conflict attributes respectively, $w_{y_c}$, $w_{y_d}$, and $w_{y_m}$ are the weights associated with each variable in the sets $Y_c$, $Y_d$, and $Y_m$, reflecting the conflict potential of the respective variables. The normalization constraints for weights are:

$$\sum_{y_c \in Y_c} w_{y_c} = 1, \quad \sum_{y_d \in Y_d} w_{y_d} = 1, \quad \text{and} \quad \sum_{y_m \in Y_m} w_{y_m} = 1.$$

C. Constraints

The constraints of the proposed IP-based framework are intended to ensure the feasibility of the team formation and to align the teams with specific requirements and preferences. The constraints are as follows:

1) Minimum intra-group diversity level: Each team should have a diversity level of at least $\varepsilon_{\min}$. Mathematically, this can be represented as:

$$DL(S_g) \geq \varepsilon_{\min}, \forall g \in \{1, \ldots, G\} \quad (9)$$

2) Maximum inter-group diversity level: The squared difference in diversity levels between any two groups should be at most $\theta_{\max}$. This maintains a balance in the diversity levels across all teams, thus ensuring fairness and equality in team composition. This is expressed as:

$$[DL(S_i) - DL(S_j)]^2 \leq \theta_{\max} \forall i, j \in \{1, \ldots, G\} \quad (10)$$

3) Unique assignment constraint: Every student must be assigned to exactly one team. This is represented as:

$$\sum_{g=1}^{G} z_{i,g} = 1, \forall i \in \{1, \ldots, N\} \quad (11)$$

4) Team size constraints: The number of students in each team should be within a specified range. This ensures that no team is too large or too small, which allows for effective collaboration and responsibility among team members. This is expressed as:

$$\omega_{\min} \leq \sum_{i=1}^{N} z_{i,g} = 1 \leq \omega_{\max}\forall g \in \{1, \ldots, G\} \quad (12)$$

5) Maximum conflict level: The conflict level within each team should not exceed $\xi_{\max}$, which is given as:

$$CL(S_g) \leq \xi_{\max}, \forall g \in \{1, \ldots, G\} \quad (13)$$

6) Integrality condition: The decision variables $z_{i,g}$ are binary variables, which implies:

$$z_{i,g} \in \{0,1\}, \forall i \in \{1, \ldots, N\}, \forall g \in \{1, \ldots, G\} \quad (14)$$

D. Two-stage Optimization Strategy for Large-Scale and Balanced Team Formation

We propose a two-stage methodology that strategically divides the student population into sub-populations by using a weighted heterogeneous multivariate $K$-means algorithm and optimizes the team formation for these sub-populations by a surrogate optimization method [66].

1) Student Sub-Population Division using Weighted Multivariate Heterogeneous $K$-Means

The first stage employs a division-and-conquer strategy to break down the large-scale team formation
problem into smaller, more tractable sub-problems (i.e., individual student sub-populations). This approach respects the goal of maximizing intra-group diversity while providing a more structured and less complex set of problems to solve. The basis of this strategy is the "weighted multivariate heterogeneous K-means (WMH K-means)" algorithm, an advanced version of the traditional K-means algorithm. The algorithm considers the weighted combination of different variable types, allowing us to appropriately consider and incorporate the diverse attributes of the students in the clustering process. Upon convergence, the algorithm forms $K$ clusters of students, where each student is assigned to the cluster that minimizes the diversity level of their attributes. Upon convergence, round-robin sampling [67] creates balanced $N_p$ sub-populations $\{\mathcal{P}_p\}_{p=1}^{N_p}$ with roughly equal sizes of $N_{sub}$. The resulting balanced sub-populations are optimal for the next stage of team formation optimization. The algorithm is elaborated in Supplementary Algorithm S1.

2) Surrogate Optimization Method for Sub-Population Team Formation

Following the division stage, each sub-population $\mathcal{P}_p$, $p \in \{1, ..., N_p\}$ is ready for the subsequent optimization process. This stage is implemented using a surrogate optimization (SO) method [66], which is effective when the objective function evaluation is time-consuming. The SO algorithm capitalizes on creating an approximation (surrogate) of the original problem and minimizing the surrogate within predefined bounds. The algorithm repeatedly generates trial points, constructs a surrogate model, finds an adaptive point, and updates the surrogate based on the obtained results, driving the search towards the global minimum of the problem. This iterative process continues until all trial points are within a specified minimum distance from the evaluated points, thereby providing a solution that is both efficient and effective for our team formation problem. Therefore, applying this method to our MCIP problem enables us to efficiently navigate the solution space and optimize the formation of diverse and balanced student teams.

To solve the optimization problem, we employ a mathematical optimization solver. The solver takes the MCIP-based model and applies numerical techniques to identify the optimal or near-optimal solution. There are several commercial and open-source solvers available such as Gurobi, CPLEX, SCIP, and MATLAB. MATLAB’s Optimization Toolbox provides functions for finding parameters that minimize objectives while satisfying constraints. The solver selection would largely depend on the size and complexity of the problem, and the available resources.

However, surrogate optimization is just one approach in a wide array of potential methods for solving this kind of problem. Alternative methodologies could be more suitable depending on the specifics of the problem and the resources. Metaheuristic techniques such as genetic algorithms (GA) [68], simulated annealing (SA) [69], or particle swarm optimization (PSO) [70] could provide better solutions when dealing with different kinds of constraints or objectives. These methods operate on different principles and may offer superior performance in different problem contexts.

E. Data Simulation and Validation

The effectiveness of the surrogate optimization approach for the MCIP-based team formation method is validated by running a thorough simulation and validation process.

1) Simulation setup

The setup of the simulation involves creating variables that are representative of a diverse set of scenarios, with each scenario having a unique heterogeneity level. Two key metrics are used to control these heterogeneity levels: the coefficient of variation (CV) [71] for continuous variables, and the Simpson’s diversity index (SDI) [72] for discrete and multi-valued variables. In our context, $CV_x$ measures the relative variability in the continuous attributes, which is calculated as the ratio of the standard deviation $\sigma_x$ to the mean $\mu_x$ of the student attribute $x$:

$$CV_x = \frac{\sigma_x}{\mu_x}, \quad \forall x \in X_c \cup Y_c$$

Conversely, $SDI_d$ quantifies the diversity in discrete and multi-valued variables $d$, which is defined as:

$$SDI_d = 1 - \sum_{c=1}^{C} p_{d,c}^2, \quad \forall d \in X_d \cup X_m \cup Y_d \cup Y_m$$

where $p_{d,c}$ represents the proportion of students whose variables taking the value $c$, and $C$ is the number of variable categories. In the next sections, we will explain the data generation process that aligns with these metrics and how the simulated data is used for validation.

2) Data simulation

To create a simulation scenario that closely mimics the reality of team formations, we generate data that satisfies the pre-specified CV and SDI values. For each continuous variable $x \in X_c \cup Y_c$, we assume the variable is non-negative and follows a Gamma distribution. With a pre-specified $CV_x$, we can determine the mean $\mu_x$ and standard deviation $\sigma_x$ of this distribution. If $\mu_x$ is a chosen constant, then the standard deviation $\sigma_x$ can be computed as: $\sigma_x = CV_x \mu_x$. To generate the Gamma-distributed random variable, the shape parameter $k_x$ and scale parameters $\theta_x$ are derived from $CV_x$ as follows:

$$k_x = \frac{1}{CV_x^2}, \quad \theta_x = \mu_x CV_x^2$$

(17)
Subsequently, we can use these calculated parameters $k_x$ and $\theta_x$ to generate our Gamma-distributed random variable, which ensures that our simulated data fits our prespecified $CV_x$.

For discrete and multi-valued variables, such as the first language and technical skills of the students, we simulate these data by using a modified stick-breaking process [73]. To apply this method to the multi-valued variables, they are binary encoded, which means creating additional binary variables for each possible value of the original variable. The simulation process is repeated $M$ times, each time generating a new set of discrete and multi-valued variables and computing the corresponding SDI$_d$. We select the dataset that yields an optimal SDI$_{opt}$ value closest to the pre-specified target SDI$_d$. The stick-breaking process is depicted in Supplementary Algorithm S2.

3) MCIP-based TFP model validation

We proceed to validate the proposed MCIP-based team formation problem (TFP) methods using these datasets. This step involves applying the proposed MCIP-based team formation methods to a set of problem instances generated from the simulated datasets. These problem instances are designed to reflect a wide range of realistic scenarios, governed by three main control parameters: SDI for discrete and multi-valued variables, CV for continuous variables, and the total number of students $N$.

To implement the methods, we input the simulated datasets into the MCIP-based team formation model. The model's constraints are set up to represent the unique characteristics of each problem instance. Subsequently, the MCIP-based algorithm is executed for each problem instance. The algorithm processes the defined attributes and constraints of each instance, including aspects such as team size and diversity level, and seeks to find an optimal solution. The output of this execution step is an optimal or near-optimal solution for each problem instance. These solutions represent the most effective distribution of students into teams, adhering to the constraints and objectives outlined in the problem instance. Finally, an evaluation process is undertaken to assess the performance of the MCIP-based method. This assessment focuses on the quality of the solutions (i.e., team formation outcomes) derived from the MCIP-based approach, which examines the qualities of the team compositions, how well the solutions satisfy the team formation criteria, and the desired diversity levels. In addition, the computational efficiency and robustness analysis are performed to investigate how well the MCIP-based approach handles variations in the scenarios. The proposed method is also benchmarked against other well-established team formation methods.

IV. RESULTS

In the section, we present our results in three distinct parts. First, we explored the data simulation process that generates scenarios mimicking realistic team formations. Second, we validated our MCIP-based team formation model using the simulated data by assessing the model's ability and the derived solution to facilitate the diverse skill sets and individual attributes while minimizing the potential conflicts. Third, the comparative performance analysis was performed to benchmark our model against other well-established team formation methods.

A. Data Simulation and Visualization

To ensure that our simulated data captures the intricacies and heterogeneity of real-world scenarios, we considered various attributes of students with continuous, discrete, and multi-valued variables. In our simulation setting, we considered 2 $N$ values: $N = 52$ and $N = 104$ to control to size of the problem. The summary of the variables and the parameter settings for data simulation were shown in Supplementary TABLES SI and SII. The simulated data for 3 representative variables — GPA (continuous), first language (discrete), and technical skills (multi-valued) — under varying degrees of heterogeneity and $N = 104$ were visualized in Fig. 2.
In Fig. 2, as CV increased, GPA distribution widened, indicating greater heterogeneity. Language diversity rose from being predominantly single-language (SDI = 0.2) to balanced (SDI = 0.5) and highly varied (SDI = 0.8). Technical skills also reflected a similar trend, moving from low diversity to a diverse distribution at SDI = 0.8.

**B. Validation of the MCIP Model Using Simulated Data**

In the validation phase, we utilized the MCIP-based team formation model to process the previously simulated data. We present the parameter settings of our MCIP-based model as shown in Supplementary TABLE SII for both scenarios: $N = 52$ and $N = 104$. We then proceeded to segment our simulated student populations into sub-populations, which were formed using a WMH K-means approach. The results of the sub-population division were visually represented in Fig. 3 below.
In Fig. 2, for $N = 52$, we achieved a successful division into 3 distinct sub-populations, as depicted in panels (a), (b), and (c). For the lower heterogeneity level of $CV = SDI = 0.2$ (panel a), the sub-populations appear closely grouped, indicating less diversity within sub-populations. As the heterogeneity level increases to 0.5 and then 0.8 (panels b and c), the sub-populations become increasingly scattered, illustrating a rise in internal diversity. For a larger student population size of $N = 104$, we managed to partition into 5 distinct sub-populations, represented in panels (d), (e), and (f). Again, the increasing spread of the sub-populations in these panels, correlating with the rise in heterogeneity levels ($CV = SDI = 0.2$, 0.5, and 0.8), confirms that the larger group has more diverse students, demanding more nuanced and careful sub-population divisions. Here, the success of the division is evident from the balanced and well-spread nature of the sub-populations across the heterogeneity levels. This successful division into high diversity sub-populations illustrates the efficiency of our WMH $K$-means approach in handling student data of different sizes and varying degrees of diversity, further setting the stage for applying the MCIP-based team formation model.

With the student populations effectively segregated into well-diversified sub-populations, the next step in our process was to facilitate the formation of effective teams within these sub-groups. This required us to deploy the two-stage optimization approach combining the WMH $K$-means method and the surrogate optimization method. The application of our two-stage optimization method on a student population of $N = 52$ was visualized in Fig. 4. The results were showcased for three heterogeneity levels: $CV = SDI = 0.2$, 0.5, and 0.8, represented by panels (a)-(c), (d)-(f), and (g)-(i) respectively.
WMH K-MEANS + SURROGATE OPTIMIZATION ($N = 52$)

Fig. 4. Implementation of the two-stage optimization method involving WMH $K$-means and surrogate optimization for a student population of $N = 52$. The figure presents outcomes for three different heterogeneity levels: CV = SDI = 0.2, 0.5, and 0.8, represented by panels (a)-(c), (d)-(f), and (g)-(i) respectively. Within each set, (a), (d), and (g) show the distribution of metrics across different groups, namely intra-group diversity level (DL), conflict level, and intra-group diversity level (DL) mean. (b), (e), and (h) present the solutions of 5 representative groups for continuous and discrete variables, namely GPA, language, and conflict management style. (c), (f), and (i) display the solutions for the multi-valued variable “technical skills” with 5 groups illustrated.

In panel (a) with the lowest heterogeneity level ($CV = SDI = 0.2$), the intra-group DLs ranged from 0.07 to 0.28 with a mean of 0.1512, while the conflict level spanned from 0.06 to 0.27. As the heterogeneity level increased to 0.5 (panel d), the intra-group DL values ranged from 0.144 to 0.358 with a mean of 0.2099, and the conflict level varied between 0.167 and 0.304. At the highest heterogeneity level ($CV = SDI = 0.8$), panel (g) revealed an even wider spread in the intra-group DL, from 0.199 to 0.362, and a mean DL of 0.2742, along with a conflict level varying between 0.179 and 0.368. This increasing spread with higher heterogeneity levels indicates that the optimization approach successfully ensures a high diversity within groups, even under different diversity levels of the student population.

Panels (b), (e), and (h) highlighted the shift in group compositions as the heterogeneity levels changed from $CV = SDI = 0.2$, 0.5, to 0.8. At the lower heterogeneity level of $CV = SDI = 0.2$ (panel b), the grouping was relatively uniform, indicating that students with similar
GPAs, language proficiency, and conflict management styles were likely to be grouped together. However, as we moved to higher heterogeneity levels, the solutions for these variables displayed a greater degree of variation within groups, as shown in panels (e) and (h).

Panels (c), (f), and (i) provided a depiction of the solutions for the multi-valued variable of technical skills for five representative groups. The distribution of technical skills within each group varied significantly with the changes in heterogeneity level, a change clearly demonstrated as we moved from panel (c) to (i). For the lower heterogeneity level of $CV = SDI = 0.2$ (panel c), the distribution of technical skills within groups was more homogeneous, suggesting that students with similar skillsets were likely to be in the same group. However, at higher heterogeneity levels, the distribution of technical skills within each group was more diverse.

Following the results of the two-stage optimization method, the random search approach was implemented for the same student population of $N = 52$. The random search algorithm, outlined in Supplementary Algorithm S3. The outcomes of the random search algorithm were presented in Supplementary Fig. S1, which showcased the application of the best method producing the most effective solutions for benchmarking. The comparisons between the random search method and the optimization method (Fig. 4) across various heterogeneity levels ($CV = SDI = 0.2$, $0.5$, and $0.8$) indicated a slightly superior performance of the random search in terms of producing higher intra-group diversity and lower conflict levels.

C. Comparative Performance Analysis

This section intended to evaluate the performance of our two-stage optimization method against established team formation models. The assessment was carried out using key performance indicators such as intra-group diversity level (DL), conflict level, inter-group diversity level, computation time, and the optimality gap. The optimality gap indicates how closely the benchmarking models approximate the best possible team formation solution. Particularly, the best-known solution derived from the most effective algorithm (i.e., the random search) was used to estimate the optimality gap, which is the percentage difference between the best-known solution and the solution obtained from other algorithms. The performance comparison included 2 student population sizes, $N = 52$ and $N = 104$ as presented in TABLE I. In TABLE I, we presented a performance comparison of our two-stage optimization method (WHM $K$-means + SO) against three well-established team formation algorithms (random search, WHM $K$-means, WHM $K$-means + genetic algorithm (GA)) serving as the benchmark to evaluate the efficacy of our proposed model. The WHM $K$-means algorithm was clearly described in Supplementary Algorithm S4 as the extension of Algorithm S1. The third algorithm, WHM $K$-means + GA, replaces the surrogate optimization by the genetic algorithm, which has been known for its capability in solving complex optimization problems, in the 2nd stage of the two-stage optimization method.

### TABLE I

**COMPARATIVE PERFORMANCE ANALYSIS OF 3 TEAM FORMATION ALGORITHMS AND OUR PROPOSED ALGORITHM**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$N$</th>
<th>Intra-group DL</th>
<th>Conflict Level</th>
<th>Inter-group DL</th>
<th>Elapsed Time (sec)</th>
<th>Optimality Gap (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Search</td>
<td>5</td>
<td>0.216 ±</td>
<td>0.141 ±</td>
<td>0.007 ±</td>
<td>5028.737 ±</td>
<td>0 ± 0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.065</td>
<td>0.063</td>
<td>0.003</td>
<td>226.862 ±</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.204 ±</td>
<td>0.135 ±</td>
<td>0.004 ±</td>
<td>7412.777 ±</td>
<td>0 ± 0</td>
</tr>
<tr>
<td></td>
<td>04</td>
<td>0.067</td>
<td>0.064</td>
<td>0.001</td>
<td>4966.39 ±</td>
<td></td>
</tr>
<tr>
<td>WHM $K$-means</td>
<td>5</td>
<td>0.203 ±</td>
<td>0.203 ±</td>
<td>0.007 ±</td>
<td>801 ± 0.253 ±</td>
<td>43.730 ±</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.062</td>
<td>0.073</td>
<td>0.003</td>
<td>3.626 ±</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.199 ±</td>
<td>0.198 ±</td>
<td>0.003 ±</td>
<td>36.183 ±</td>
<td></td>
</tr>
<tr>
<td></td>
<td>04</td>
<td>0.071</td>
<td>0.074</td>
<td>0.001</td>
<td>5.36 ±</td>
<td></td>
</tr>
<tr>
<td>WHM $K$-means + GA</td>
<td>2</td>
<td>0.204 ±</td>
<td>0.206 ± 0.06</td>
<td>0.007 ±</td>
<td>960.307 ±</td>
<td>27.761 ±</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.194 ±</td>
<td>0.192 ±</td>
<td>0.004 ±</td>
<td>314.207 ±</td>
<td>2.97</td>
</tr>
<tr>
<td></td>
<td>04</td>
<td>0.073</td>
<td>0.069</td>
<td>0.001</td>
<td>1376.06 ± 435.77 ±</td>
<td>31.192 ±</td>
</tr>
<tr>
<td>WHM $K$-means + SO</td>
<td>2</td>
<td>0.203 ±</td>
<td>0.207 ±</td>
<td>0.007 ±</td>
<td>505.106 ± 66.749 ±</td>
<td>29.671 ±</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.192 ±</td>
<td>0.191 ±</td>
<td>0.004 ±</td>
<td>902.833 ± 29.532 ±</td>
<td>29.619 ±</td>
</tr>
<tr>
<td></td>
<td>04</td>
<td>0.064</td>
<td>0.062</td>
<td>0.001</td>
<td>5.079 ±</td>
<td></td>
</tr>
</tbody>
</table>

According to TABLE I, the random search method yielded the most optimal solutions for both population sizes. However, this came at the cost of longer computation times, specifically $5028.737 ± 226.862$ (secs) and $7412.777 ± 4966.39$ (secs) for
N = 52 and N = 104 respectively, indicating a low computational efficiency as the student population sizes increased. Additionally, the random search approach resulted in intra-group DLs of 0.216 ± 0.065 for N = 52 and 0.204 ± 0.067 for N = 104, with corresponding conflict levels of 0.141 ± 0.063 and 0.135 ± 0.064, respectively. This method provided minimal inter-group DLs (0.007 ± 0.003 for N = 52 and 0.004 ± 0.001 for N = 104), suggesting it can effectively balance the team diversity across different groups. In contrast, the WHM K-means algorithm reduced computation times to only 0.801 ± 0.253 (secs) for N = 52 and 5.083 ± 2.513 (secs) for N = 104, indicating superior computational speed. However, this was at the expense of an increased optimality gap — 43.730 ± 3.626% for N = 52 and 36.183 ± 5.36% for N = 104, implying a trade-off between computational speed and solution quality. This method yielded intra-group DLs of 0.203 ± 0.062 for N = 52 and 0.199 ± 0.071 for N = 104, and conflict levels of 0.203 ± 0.073 and 0.198 ± 0.074, respectively, indicating a slight increase in conflicts compared to the Random Search approach. The inter-group DLs remained low (0.007 ± 0.003 for N = 52 and 0.003 ± 0.001 for N = 104), demonstrating a robust capability to ensure balanced team diversity across groups, which was similar to the random search method.

The hybrid model, WHM K-means + GA, yielded an encouraging balance between computation time and solution quality. The optimality gap was significantly reduced compared to WHM K-means (to 27.761 ± 2.97% for N = 52 and 31.192 ± 5.079% for N = 104), suggesting that the hybrid model achieved superior solution quality. However, the algorithm did increase computation time compared to WHM K-means, but still maintained significantly lower computation times than Random Search, with average times of 960.307 ± 314.207 (secs) for N = 52 and 1376.06 ± 435.77 (secs) for N = 104, reflecting a great balance between solution quality and computational time. Interestingly, the intra-group DLs were 0.204 ± 0.055 for N = 52 and 0.194 ± 0.073 for N = 104, while the conflict levels were 0.206 ± 0.06 for N = 52 and 0.192 ± 0.069 for N = 104, implying slight improvements in team homogeneity and conflict management compared to the standalone WHM K-means. Similar to previous models, it maintained low inter-group DLs (0.007 ± 0.003 for N = 52 and 0.004 ± 0.001 for N = 104), demonstrating its capability to foster balanced team diversity across different groups.

Finally, our proposed model, WHM K-means + SO, delivered a strong performance. It outperformed WHM K-means in terms of the optimality gap, reducing it to 29.619 ± 5.076% for N = 52 and 29.619 ± 3.27% for N = 104, thereby affirming the superior quality of its solutions. It also had significantly lower computation times than random search and WHM K-means + GA, clocking in at 505.106 ± 66.749 (secs) and 902.833 ± 29.532 (secs) for N = 52 and N = 104, respectively, underscoring its computational efficiency. This model reported intra-group DLs of 0.203 ± 0.055 for N = 52 and 0.192 ± 0.064 for N = 104, and conflict levels of 0.207 ± 0.057 and 0.191 ± 0.062, respectively. It maintained low inter-group DLs (0.007 ± 0.002 for N = 52 and 0.004 ± 0.001 for N = 104), supporting its capability to guarantee balanced diversity across teams.

The detailed evaluations of all scenarios with different CV and SDI values for N = 52 and N = 104 were reported in Supplementary TABLES III and IV. As shown in the tables, an increased CV led to a higher intra-group diversity level (DL) and conflict level. This implies that as the variability of continuous attributes within a team increases, the team becomes more diverse and tends to experience more internal conflicts. The impact of SDI was less consistent. Though an increased SDI led to a rise in conflict level, indicating a higher disagreement level as the diversity in variables increased, its influence on the intra-group DL was inconsistent, which suggested a more complex interaction between these factors. With regards to algorithm performance, the random search method, despite yielding optimal results, consistently had the longest elapsed times across all CV and SDI values. Our proposed methods, WHMKM + GA and WHMKM + SO, offered a more balanced performance in terms of time efficiency and result optimality.

V. DISCUSSION

This research utilized a novel two-stage optimization approach to address the complexities of team formation in heterogeneous environments. The findings revealed that our two-stage optimization method was not only effective at various levels of heterogeneity, but it also presented a superior balance of computational efficiency and solution quality compared to other established team formation models like the random search method, the WHM K-means, and the WHM K-means combined with GA. The results showed that the two-stage optimization model had the ability to form teams with high intra-group diversity level, low conflict level, and low inter-group DL in all instances. In addition to these, our model demonstrated an impressive performance in terms of computation time and optimality gap. These findings have validated our hypothesis that the two-stage optimization model can balance the need for solution quality and computational speed, while ensuring high diversity within teams, even under varying conditions of heterogeneity.

Our study demonstrates the robustness, versatility, and computational efficiency of the two-stage optimization model across varying heterogeneity levels. Despite the increased problem complexity with rising heterogeneity, our model adeptly ensures diversity while maintaining low conflict levels, thus optimizing team harmony. While the random search method offers optimal results, it lacks computational speed, rendering it impractical for real-time applications. The WHM K-means model, although efficient, struggled to balance diversity and conflict. Its hybrid version, WHM K-means + GA, has demonstrated improvements but still underperformed relative to our model in solution quality and computational efficiency. Therefore, our findings underscore the potential of our model in facilitating effective team formation without compromising on diversity or speed, underlining its utility in diverse, heterogeneous environments.

The practical implications of our study extend across several domains. Foremost, in an educational setting, our multi-objective optimization model can help educators form more balanced and diverse student teams [74]. This is an improvement over the methods used in studies like that of
Meulbroek et al. [43], who employed the Gale-Shapley algorithm for team formation, and Akbar et al. [45], who utilized a clustering approach for the same purpose. By considering multiple factors such as skills, personality traits, and learning styles, our model fosters better collaborative learning experiences, potentially leading to improved student outcomes as suggested by Weidmann et al. [40]. In the corporate environment, human resource managers can utilize our model for team formation during project assignments. The algorithm's ability to balance diverse skills and personalities could promote synergistic collaboration, as was also concluded in the study by Berktaş and Yaman [41]. This could enhance project success rates and improve overall organizational productivity, much like the methods suggested by Liu et al. [54] and Fatahi and Lorestaní [55]. Our model can also be applied in online collaborative platforms. By creating optimally balanced teams, it could contribute to increased user satisfaction, resulting in better user retention rates and effective collaboration. Moreover, the flexibility and adaptability of our model can accommodate diverse team formation requirements, thereby promoting inclusivity, and efficiency in team formation strategies, as supported by Krousa and Virvou's study [75].

Our study has several limitations which, in turn, open avenues for future research in TBL. Firstly, the MCIP-based algorithm's effectiveness is heavily reliant on the quality and comprehensiveness of input data. Although our model managed to utilize several diverse data points, the inclusion of more detailed and varied information, such as cultural backgrounds, language proficiency, or previous project experiences, could potentially enhance the team formation process even further. Secondly, while our proposed framework has proven successful under the conditions tested, its performance in real-world domains and applications, such as corporate environments, non-profit teams, or online collaboration platforms, needs to be extensively evaluated. Thirdly, despite our model aiming to create balanced and diverse teams, it does not consider the dynamics of team interactions after the team assignment. Understanding how these teams function and adapt over time is a crucial aspect of the team formation process [76, 77]. Lastly, our algorithm, while efficient, is computationally intensive, especially for large-scale applications. This presents an interesting area for future research, where efforts are directed towards improving computational efficiency while maintaining or even enhancing the algorithm's accuracy and robustness. This could involve the exploration of parallel processing or the use of more efficient data structures and algorithms [78, 79].

V. CONCLUSION

In conclusion, our study presents a novel method for forming balanced and diverse teams using a MCIP-based two-stage optimization strategy, which combines WHM K-means and surrogate optimization algorithms for TBL. The proposed two-stage strategy ensures high intra-group diversity, low intra-group conflict, and low inter-group diversity variance heterogeneity on key characteristics, thereby promoting a beneficial mixture of compatible yet diverse team members. Our numerical analysis using simulation data in academic setting demonstrated the effectiveness and robustness of our algorithm, achieving superior outcomes compared to traditional methods and recent approaches in terms of both performance and team member satisfaction. Furthermore, the results highlight the potential benefits of harnessing machine learning in the realm of team formation, paving the way for more sophisticated and dynamic team-building strategies. The algorithm's implications are profound, revolutionizing team formation in diverse fields such as education, project management, and human resources management. By ensuring balance and diversity in teams, the proposed method can foster an environment of constructive collaboration and innovation, while minimizing conflict and ensuring fair workload distribution. Nevertheless, we recognize the limitations of our work, which provide avenues for future research. These include the need to enrich the data used for team formation, the requirement for extensive real-world testing in various contexts, and the importance of considering team dynamics post-formation. As we move forward in an increasingly digital and interconnected world, tools like the one proposed in this study will be instrumental in fostering a collaborative and harmonious work culture. Future research is not only encouraged but essential for harnessing the full potential of diverse and balanced teams, ultimately enhancing productivity, innovation, and job satisfaction.

REFERENCES


Pallavi Singh is a Ph.D candidate and currently working toward her Ph.D degree in Electrical Engineering from University of South Florida. Her main area of research centers around Machine Learning applications for Engineering Education, with the goal of enhancing students' performance and learning. This includes Digital Engineering, Model-Based Systems Engineering (MBSE), and Educational Data Mining.

Phat K. Huynh is a Ph.D candidate of the Industrial and Management Systems Engineering Department at University of South Florida. His work mainly focuses on nonlinear complex systems, probabilistic models, and predictive analytics in healthcare. His research work has been featured in the high-impact peer-reviewed journals, such as Artificial Intelligence in Medicine, IIE Transactions on Healthcare Engineering, and PLOS ONE.

Dang Nguyen, BS, is currently working as a Clinical Research Coordinator at Massachusetts General Hospital. He graduated with Tau Beta Pi and Honors Distinction from the University of South Florida, holding a Bachelor's degree in Biomedical Engineering. Dang has authored over 20 peer-reviewed journal articles, and his work has been featured in many outlets such as The Lancet, PNAS, and The Journal of Physical Chemistry B.

Dr. Trung (Tim) Q. Le, Ph.D, is an assistant professor of the Industrial and Management Systems Engineering (IMSE) Department at University of South Florida. He is affiliated with the Medical Engineering Department at University of South Florida. His research focuses in 3 main directions: 1) data-driven and sensor-based modeling, 2) medical device manufacturing and bio-signal processing, and 3) predictive analytics for personalized healthcare.

Dr. Wilfrido Moreno, Ph.D, is an assistant professor of the Electrical Engineering Department at University of South Florida.His primary area of research are in Instructional Technologies Solutions for Engineering Education including Distance Learning and System integration for industrial applications in the areas of Industrial Controls & Instrumentation, Robotics, Digital Signal Processing, Reconfigurable Architectures, Analog-Digital Mixed Signal Design & Synthesis, Communications and Biomedical Engineering.