Modeling Student Task Group Preferences Using Graph Theory and Spectral Clustering

Hafiz Zulkhairi[1*]

[1] Department of Informatics, Universitas Satya Terra Bhinneka, Medan, Indonesia

ABSTRACT

This study aims to model student preferences in forming task groups using graph theory and spectral clustering. The research involved 2024 cohort students of the IF A Siang class at Universitas Satya Terra Bhinneka, with preference data collected through a questionnaire consisting of 30 Likert-scale indicators. The data were transformed into numerical representations and normalized to ensure balanced feature contributions. Graph theory was applied to model relationships among students based on preference similarity, while spectral clustering was used to form optimal student groups. The results indicate that the optimal clustering consisted of two clusters, achieving a Silhouette Score of 0.185 and a Davies–Bouldin Index of 1.372, which reflect acceptable clustering quality for subjective preference data. Network visualization further shows stronger connectivity within clusters than between clusters, confirming the effectiveness of the proposed approach. Overall, this method provides an objective and data-driven alternative to conventional group formation methods and can support lecturers in forming balanced and effective task groups.

ARTICLE INFO

Keywords:
Student Preferences, Graph Theory,
Spectral Clustering,
Data Mining,
Task Grouping

Histori:
Received: 17 January 2026
Accepted: 1 February 2026
Published: 10 February 2026

* Corresponding Author at Department of Informatics, Universitas Satya Terra Bhinneka, Medan, Indonesia.
E-mail address: hafizzkr0403@gmail.com (author#1)

1. INTRODUCTION

The formation of task groups is an important component in the learning process at universities because it affects the effectiveness of collaboration and the achievement of learning objectives. However, grouping that is done without considering students' characteristics and preferences often results in role imbalances and lowers the quality of interaction within the group (Mueller et al., 2025; Müller et al., 2022; Poort et al., 2022).

The main problem faced in forming task groups is the process of selecting group members, which is still subjective and not data-driven. Mismatched preferences among students in the group can lead to low participation, increased internal conflicts, and decreased student satisfaction with the group-based learning process (Theobald et al., 2017).

In teaching practice, lecturers generally face limitations in time and information to thoroughly analyze the compatibility among students. As a result, group formation is often done randomly or based on simple considerations, without systematically taking into account the patterns of relationships and student preferences. This situation highlights the need for an approach that can process student preference data objectively and represent relationships between individuals in a more structured way (Putra & Margono, 2025).

Various approaches have been developed to analyze interactions and relationship patterns among individuals within a group. However, in the context of forming task groups at universities, the use of computational approaches that can model students' preference relationships relationally is still relatively limited. This indicates the need for more objective and systematic methods in the group formation process (Lee & Sharma, 2024).

Based on these issues, this study proposes a solution in the form of modeling student preferences in forming assignment groups using a graph theory approach and spectral clustering algorithm. This approach is expected to produce group divisions that are more objective, balanced, and in line with student preferences. Thus, the results of this study can serve as a decision support alternative for lecturers in organizing effective and data-driven assignment groups.

2. LITERATURE REVIEW

The formation of groups in the context of education has been extensively studied using mathematical and computational approaches to understand the patterns of relationships between individuals. One approach that is often used is graph theory, where individuals are represented as nodes and the relationships between individuals as edges. This approach allows for the analysis of structured and measurable relationship structures. Several studies have shown that graph theory is effective in analyzing group dynamics and social interactions in cooperative learning (Irwan, 2020; Samosir et al., 2024; Sirait et al., 2024).

In addition to modeling relationships using graph theory, clustering techniques in data mining are widely used to group objects based on certain degrees of similarity. The goal of clustering methods is to maximize similarity within a cluster and minimize similarity between clusters. However, conventional clustering methods such as K-Means have limitations in handling data with complex structures and non-linear relationships (Sari et al., 2025).

To overcome these limitations, the spectral clustering method was developed, which is a graph-based clustering technique that utilizes a similarity matrix and a Laplacian matrix. Spectral clustering has been shown to produce more optimal groupings on data with nonlinear structures compared to traditional clustering methods (Favati et al., 2020; Info, 2025; Pourkamali-Anaraki, 2020). Several studies indicate that this approach is effective in various fields, including regional segmentation, network analysis, and complex data clustering (Ding et al., 2024).

Although graph theory and spectral clustering have been widely applied, their application in modeling student preferences for task group formation is still relatively limited. Most group formations are still carried out subjectively and have not utilized computational analysis of preference relationships. Therefore, this study aims to fill this gap by combining graph theory and spectral clustering to model student preferences objectively.

3. METHOD, DATA, AND ANALYSIS

This research method uses a data mining approach with graph-based modeling and spectral clustering algorithms to form groups of students based on preference similarities. The research stages are systematically organized, starting from data collection, graph modeling, clustering process, to cluster quality evaluation.

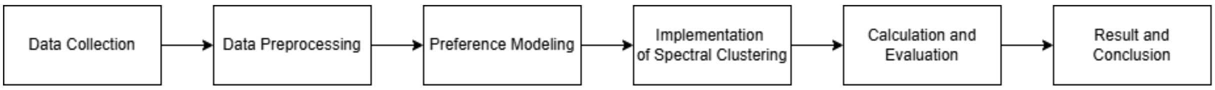

Figure 1. Research Flow

3.1. Data Collection

This research method uses a data mining approach with graph theory-based modeling and spectral clustering algorithms to form groups of students based on preference similarities. The research stages are systematically organized, starting from data collection, graph modeling, clustering process, to cluster quality evaluation.

3.2. Data Preprocessing

The data preprocessing stage includes checking data completeness, converting categorical data into numerical form, and normalizing values to prevent certain attributes from dominating. This process is important to ensure data quality before performing cluster analysis (Ding et al., 2024; Herdianti et al., 2025).

3.3. Preference Modeling Using Graph Theory

Graph theory is used to model relationships between students based on shared preferences. Each student is represented as a node, while the relationship between students is represented as an edge with a certain weight (Favati et al., 2020; Pourkamali-Anaraki, 2020; Putri et al., 2022).

The edge weight is determined based on the level of similarity in preferences among students, calculated from questionnaire data. This relationship is then represented in the form of a similarity matrix W, where the element w_{ij} indicates the closeness of preferences between student-*i* and student-*j* (Hidayati & Surono, 2021; José-García & Gómez-Flores, 2021; Sirait et al., 2024).

The graph theory approach allows for the visualization and analysis of relational structures among students in a mathematical and measurable way, enabling the connectivity patterns within the group to be analyzed objectively (Irwan, 2020; Khairunnisa et al., 2022).

3.4. Formation of Graph Laplacian Matrix

Based on the similarity matrix $W = [w_{ij}]$, this represents the relationships between students, where w_{ij} indicates the level of similarity between student-*i* and student-*j*.

The degree matrix D is defined as a diagonal matrix with elements (Herdianti et al., 2025):

$$D_{ii} = \sum_{j=1}^{n} w_{ij} \tag{1}$$

with:

- D_{ii} : diagonal element of the degree matrix
- w_{ij}: similarity wight between node-*i* and node-*j*
- n: number of nodes (students)

The graph Laplacian matrix is formulated as follows (Herdianti et al., 2025; Putri et al., 2022):

$$L = D - W \tag{2}$$

with:

- L: graph Laplacian matrix
- D: degree matrix
- W: similarity matrix

Next, the normalized Laplacian matrix in symmetric form is used (Hidayati & Surono, 2021; Putri et al., 2022):

$$L_{sym} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} \tag{3}$$

with:

- L_{sym}: normalized Laplacian matrix
- $D^{-\frac{1}{2}}$: inverse square root of the degree matrix

This Laplacian matrix serves as the main basis in spectral clustering algorithms because it can represent the global structure of the graph.

3.5. Spectral Clustering Algorithm

Spectral clustering algorithm works by utilizing the eigenvalues and eigenvectors of the graph Laplacian matrix. Several eigenvectors associated with the smallest eigenvalues are selected and arranged into a new matrix U (Hidayati & Surono, 2021; Putri et al., 2022):

$$U = [u_1, u_2, \dots, u_k] \tag{4}$$

with:

- u_k: the k eigenvector
- k: the desired number of clusters

Each row in the matrix U is then treated as a new representation of the student data in a lower-dimensional space. The subsequent clustering process is carried out on the matrix U using the k-means algorithm to obtain the final clusters (Info, 2025; Putri et al., 2022).

This method has been proven effective in handling data with nonlinear relationships and complex graph structures compared to conventional clustering methods (Info, 2025; Nurdiana et al., 2022).

3.6. Cluster Quality Evaluation

To assess the quality of clustering results, internal clustering evaluation methods are used, namely the Silhouette Coefficient and the Davies-Bouldin Index (DBI) (Clustering & Belajar, 2022; Sari et al., 2025).

3.6.1. Silhouette Coefficient

The silhouette value for each data point is calculated using the following equation (Sari et al., 2025):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

with:

- $a(i)$: is the average distance from a data point to all members in the same cluster
- $b(i)$: is the minimum average distance from a data point to another cluster

The silhouette value ranges from $[-1,1]$, where a value approaching 1 indicates good cluster quality.

3.6.2. Davies-Bouldin Index (DBI)

The Davies–Bouldin index is used to measure the ratio of intra-cluster similarity to inter-cluster distance. DBI is formulated as follows (Clustering & Belajar, 2022; José-García & Gómez-Flores, 2021; Vergara et al., 2020):

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}}\right) \tag{6}$$

with:

- S_i: is the average distance of cluster members to the centroid of the i cluster
- M_{ij}: is the distance between the centroids of the i and j clusters

- k: is the number of clusters

A smaller DBI value indicates better cluster quality.

4. RESULT AND DISCUSSION

4.1. Research Data

This study involved 52 students from the IF A Day class at Satya Terra Bhinneka University, class of 2024, as respondents. Data were obtained through a questionnaire on preferences in project group formation, consisting of 30 Likert scale statements with a score range of 1–5. One question regarding the frequency of group work experience was not included in the grouping process because it was descriptive in nature and did not represent a preference.

Before clustering is carried out, numerical data is normalized using the standardization method to ensure that each variable has a balanced contribution in distance calculations. The preprocessed data is then used as input for clustering algorithms and graph modeling.

4.2. Determination of the Optimal Number of Cluster

The determination of the optimal number of clusters was carried out using the Silhouette Score with variations in the number of clusters (k) from 2 to 8. This method is used to measure the closeness of objects within a cluster and their separation from other clusters.

The test results show that the highest Silhouette Score was obtained at k = 2 with a score of 0.185, so the optimal number of clusters to be used in the next stage is two clusters.

Table 1. Silhouette Scores for Various Numbers of Clusters

Number of Clusters k	Silhouette Score
2	0,185
3	0,145
4	0,134
5	0,139
6	0,147
7	0,136
8	0,144

A relatively low Silhouette value indicates that the students' preference data has complex and overlapping characteristics, but dividing it into two clusters still provides the most stable structure compared to other numbers of clusters.

4.3. Clustering Results Using Spectral Clustering

After the number of clusters is determined, the clustering process is carried out using the Spectral Clustering algorithm. This algorithm utilizes a similarity matrix and graph structure to group data based on non-linear relationships between objects. The results of Spectral Clustering show that the students are divided into two clusters with the following distribution.

Table 2. Distribution of the Number of Students in Each Cluster

Cluster	Number of Students
Cluster 0	40
Cluster 1	12

The distribution shows that the majority of students have relatively homogeneous preference patterns, while a small number of students form groups with different preference characteristics. This condition indicates the presence of variation in tendencies regarding how students choose members for task groups.

4.4. Evaluation of Clustering Result Quality

4.4.1. Silhouette Score

A Silhouette Score of 0.185 indicates that the quality of cluster separation is at a moderate level. Although it does not show very strong separation, this value is still acceptable considering that the data used is subjective and multidimensional, which tends to have overlaps between respondents' preferences.

4.4.2. Davies-Bouldin Index (DBI)

In addition to the Silhouette Score, cluster quality evaluation was also conducted using the Davies–Bouldin Index (DBI). The calculation results showed a DBI value of 1.372. This value indicates that the distance between clusters is relatively larger compared to the variation within clusters, suggesting that the formed cluster structure can still be considered suitable for analyzing task group formation.

Tabel 3. Clustering Quality Evaluation Score

Evaluation Method	Value
Silhouette Score	0,185
Davies-Bouldin Index	1,372

4.5. Visualization of Student Preference Networks

To strengthen the interpretation of clustering results, network visualization is carried out using a graph theory approach. Each node represents a student, while the edges indicate the level of similarity in preferences based on cosine similarity. The color of the nodes represents the clustering results from the Spectral Clustering algorithm.

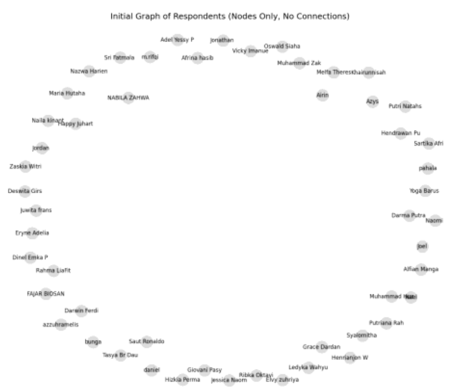

Figure 2. Graph Before Clustering

Based on Figure 2, it can be seen that the raw graph produced from the node formation process has not undergone spectral clustering, and therefore does not produce interactions to connect with each other.

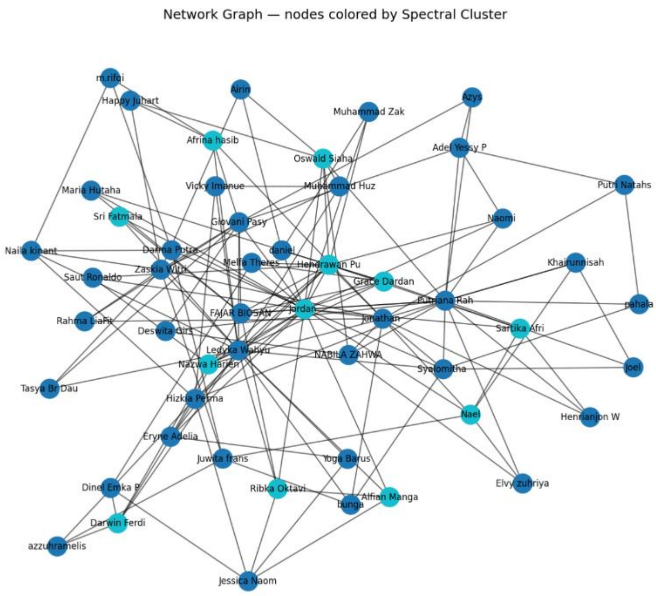

Figure 3. Student Preference Network Graph Based on Spectral Clustering

Based on Figure 3, it can be seen that students within the same cluster tend to form groups of nodes that are more closely connected. Conversely, connections between clusters are relatively rare, indicating a separation in the preference structure between groups. This visualization reinforces the results of the quantitative evaluation that Spectral Clustering is able to capture the pattern of student preference relationships more representatively.

4.6. Discussion

The research results show that an approach based on graph theory and Spectral Clustering can be used to model student preferences in forming task groups objectively. Compared to centroid-based methods like K-Means, Spectral Clustering is better able to capture non-linear relationships that arise from similarities in student preferences.

The Adjusted Rand Index (ARI) value of 0.239 between the results of K-Means and Spectral Clustering indicates a fairly significant difference in cluster structure. This confirms that a graph-based approach provides an additional perspective in understanding student preference patterns. Thus, the clustering results obtained can be used as a supporting basis for lecturers in organizing more balanced task groups that align with student preferences.

5. CONCLUSION AND SUGGESTION

Based on the research results, student preference data obtained through a questionnaire with 30 Likert scale indicators can be represented quantitatively and processed objectively through a preprocessing stage that includes feature selection and normalization. The application of graph theory has been proven effective in modeling the similarity relationships among students, where network visualizations show stronger connections within clusters compared to between clusters. Clustering using Spectral Clustering with two optimal clusters resulted in fairly good separation quality, as indicated by a Silhouette Score of 0.185 and a Davies–Bouldin Index of 1.372, which is still acceptable for subjective preference data. The difference in results between Spectral Clustering and K-Means, with an Adjusted Rand Index of 0.239, highlights the advantage of the graph-based approach in handling non-linear relationships. Overall, this approach has the potential to serve as a data-driven and objective decision support alternative in forming student task groups to enhance the quality of collaboration and the effectiveness of group-based learning.

As a recommendation for future research, it is suggested to increase the number of respondents, combine students' preferences with other indicators such as academic ability or learning styles, and compare Spectral Clustering with other graph-based clustering methods. With these developments, the task group formation model is expected to become more comprehensive and adaptive to various learning contexts.

ACKNOWLEDGEMENT

Thanks are extended to all parties who have supported the implementation of this research.

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used ChatGPT and Consensus AI to help develop ideas and find relevant reference sources. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the content of this publication.

REFERENCE

Clustering, A. K., & Belajar, G. (2022). *Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokkan Mahasiswa Berdasarkan Gaya Belajar*. *12*. https://doi.org/10.34010/jati.v12i1

Ding, L., Li, C., Jin, D., & Ding, S. (2024). Survey of spectral clustering based on graph theory. *Pattern Recognition*, *151*, 110366. https://doi.org/10.1016/j.patcog.2024.110366

Favati, P., Lotti, G., Menchi, O., & Romani, F. (2020). Construction of the similarity matrix for the spectral clustering method: Numerical experiments. *Journal of Computational and Applied Mathematics*, *375*, 112795. https://doi.org/10.1016/j.cam.2020.112795

Herdianti, R. A., Syaifullah, W., Saputra, J., & Prasetya, D. A. (2025). *Pemetaan Karakteristik Wilayah Miskin di Jawa Timur Menggunakan Spectral Clustering untuk Penentuan Target Kebijakan*. *10*(2), 362–375.

Hidayati, Q. R., & Surono, S. (2021). Implementasi Algoritma Spectral Clustering Untuk Analisis Sentimen. *Delta: Jurnal Ilmiah Pendidikan Matematika*, *9*(1), 27. https://doi.org/10.31941/delta.v9i1.1229

Info, A. (2025). *Segmentasi Wilayah Indonesia Berdasarkan Ihk*. *5*(1), 65–76.

Irwan, S. E. (2020). Aplikasi Pewarnaan Graf pada Penempatan Kamar Mahasiswa (Studi Kasus: Asrama Institut Teknologi Sumatera). *Sainmatika: Jurnal Ilmiah Matematika Dan Ilmu Pengetahuan Alam*, *17*(1), 17. https://doi.org/10.31851/sainmatika.v17i1.3137

José-García, A., & Gómez-Flores, W. (2021). A survey of cluster validity indices for automatic data clustering using differential evolution. *Proceedings of the Genetic and Evolutionary Computation Conference*, 314–322. https://doi.org/10.1145/3449639.3459341

Khairunnisa, A., Nazir, A., Novrianto, N., Yusra, Y., Oktavia, L., & Akhyar, A. (2022). Klastering Penyebaran Covid-19 Di Indonesia Tahun 2020 Menggunakan K-Medoids Clustering. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *5*(3), 391–398. https://doi.org/10.32672/jnkti.v5i3.4409

Lee, M., & Sharma, P. (2024). Leveraging the Louvain algorithm for enhanced group formation and collaboration in online learning environments. *International Journal of Educational Technology in Higher Education*, *21*(1), 65. https://doi.org/10.1186/s41239-024-00495-w

Mueller, A., Konert, J., Röpke, R., Genc, Ö., & Bellhäuser, H. (2025). Group formation based on extraversion and prior knowledge: a randomized controlled study in higher education online. *Journal of Computing in Higher Education*, *37*(3), 780–808. https://doi.org/10.1007/s12528-024-09406-4

Müller, A., Bellhäuser, H., Konert, J., & Röpke, R. (2022). Effects of Group Formation on Student Satisfaction and Performance: A Field Experiment. *Small Group Research*, *53*(2), 244–273. https://doi.org/10.1177/1046496420988592

Nurdiana, N., Nilogiri, A., & Rahman, M. (2022). Penerapan Algoritma Fuzzy C-Means dan Metode Elbow untuk Mengelompokkan Provinsi di Indonesia Berdasarkan Indeks Demokrasi Indonesia Application of The Fuzzy C-Means Algorithm and Elbow Method to Grouch Provinces in Indonesia Based on The Indonesian Democ. *Jurnal Smart Teknologi*, *3*(5), 544–551. https://jurnal.unmuhjember.ac.id/index.php/JST/article/view/7924

Poort, I., Jansen, E., & Hofman, A. (2022). Does the group matter? Effects of trust, cultural diversity, and group formation on engagement in group work in higher education. *Higher Education Research & Development*, *41*(2), 511–526. https://doi.org/10.1080/07294360.2020.1839024

Pourkamali-Anaraki, F. (2020). Scalable Spectral Clustering With Nyström Approximation: Practical and Theoretical Aspects. *IEEE Open Journal of Signal Processing*, *1*, 242–256. https://doi.org/10.1109/OJSP.2020.3039330

Putra, Y. A. N. S., & Margono, H. (2025). Simulation of Student Study Group Formation Design Using K-Means Clustering. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(2), 598–608. https://doi.org/10.57152/malcom.v5i2.1795

Putri, M. W., Nur, I. M., & Wasono, R. (2022). Implementasi Spectral Clustering Algorithm Untuk Pengelompokan Sasaran Vaksinasi Covid-19 Di Indonesia. *Jurnal Statistika Universitas Muhammadiyah Semarang*, *10*(1), 26. https://doi.org/10.26714/jsunimus.10.1.2022.26-31

Samosir, N. E., Sitindaon, M. C., Manik, R. M., Reslina, I., Simanullang, T. A., Sitorus, R. C., & Haris, D. (2024). *Pemodelan Teori Graph untuk Menganalisis Dinamika Kelompok dalam Pembelajaran Kooperatif Matematika di SMP Swasta Santa Maria Medan*. *8*, 45605–45615.

Sari, F., Kuntari, M., Yati, W., Khaulasari, H., & Hafiyusholeh, M. (2025). Implementasi K-Means Clustering Melalui Pemanfaatan Sampling Kombinasi Pada Pengelompokan Pola Kesehatan Mental Mahasiswa Sains dan Teknologi. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *11*(1), 9–16. https://doi.org/10.25077/teknosi.v11i01.2025.9-16

Sirait, L. T., Simanjuntak, H. S., Sinaga, M. C., Hijriyanti, P., Putra, O., Simamora, M., Putri, A., Dameria Sianturi, Y., Haris, D., Willem, J., Pasar, I., Baru, K., Tuan, S., & Serdang, D. (2024). Penerapan Teori

Graph untuk Meningkatkan Interaksi Siswa dalam Pembelajaran Matematika Berbasis Diskusi di SMP Negeri 27 Medan. *Journal on Education*, *07*(01), 7339–7349.

Theobald, E. J., Eddy, S. L., Grunspan, D. Z., Wiggins, B. L., & Crowe, A. J. (2017). Student perception of group dynamics predicts individual performance: Comfort and equity matter. *PLOS ONE*, *12*(7), e0181336. https://doi.org/10.1371/journal.pone.0181336

Vergara, V. M., Salman, M., Abrol, A., Espinoza, F. A., & Calhoun, V. D. (2020). Determining the number of states in dynamic functional connectivity using cluster validity indexes. *Journal of Neuroscience Methods*, *337*, 108651. https://doi.org/10.1016/j.jneumeth.2020.108651