Applying Community Detection Algorithms
To Examine Group Formation in Online Discussions

Abstract

Many group formation strategies have been used within the design of asynchronous online discussions to increase their effectiveness. In this research, we used a modularity optimization method (Blondel 2008) for community detection to explore group effectiveness by examining the network measures in global and local levels using asynchronous online discussion data. Our analysis showed some variations of network measures in groups that clearly underscore group effectiveness. Therefore, we believe this method is promising for analyzing the relations in groups to anticipate the group effectiveness and identify the implications of using this algorithm for designing and analyzing effective groups in online discussions.

Introduction

Group formation is key to group processes and outcomes (Kozlowski & Bell, 2003) and has been a central research agenda in Computer Supported Collaborative Learning(CSCL) (Ouyang & Scharber, 2017; Rodríguez et al., 2011; Weinberger et al.,2005). Learner grouping influences collaborative learning, and the outcomes often support heterogeneous groups, in contrast to learners' preference for homogeneous groups which may negatively affect learning (Bell, 2007). Heterogeneous groups aim for participants to bring unique attributes to the collaboration to achieve group goals, while homogeneous groups are based on similarity; individuals who share goals face lower levels of relationship conflict along with better performance (De Dreu & Weingart, 2003). Algorithmic group formation which is economically beneficial is another way to group students, but there is no general or evidence-based recommendation for group formation to feed the algorithms (Magpili & pazos 2018, as cited in Müller et al., 2022). In this research, we use the community detection algorithm based on randomization and weights to explore group formation in asynchronous discussion with the aim of better supporting instructors and designers in designing effective collaborative learning.

Community detection in networks

In most educational research, SNA (social network analysis) has been used to capture learner interaction and how their individual relationships affect their learning outcomes, so the purpose of using SNA is generally to assess whether and to what degree interaction and collaboration occur (Froehlich et al., 2020). However, few studies have focused on SNA to identify communities to examine learning. Jan & Vlachopoulos (2018) used the integrated methodological framework using SNA to structurally identify communities in online learning. They suggest that community-based learning and structural similarity between networks and communities make SNA a natural choice for deeper understanding of group interaction. The study substantiated their method as an effective framework for structural identification of a Community of Inquiry (CoI) and Community of practice (CoP). Yassine (2020) used the label propagation algorithm and Louvain algorithm for community detection and found those different community detection algorithms can be implemented on learning networks and detect communities.

In the community detection algorithms, communities are defined as nodes that have similar affiliations different from the rest of the network (Yang et al., 2010) or as network structures that have a cohesive characteristic with the possibility of separation (Newman 2018). Community can be separated when the internal links are larger than the external links, so a strong community has more internal links than external links (Wasserman &Faust, 1994). Though communities can be detected using a variety of methods, modularity-based algorithms are most frequently cited in relation to community detection.

The Girvan-Newman algorithm is based on betweenness and identifies communities by removing bridges between nodes and clusters. The algorithm iterates through the process of identifying and removing the largest link. This procedure ends when all links are removed, and nodes are isolated to optimize the modularity which is a function for the approximation of communities. Louvain method (Blondel et al., 2008) is the most popular modularity optimization method (Menczer et al., 2018). It is an agglomerative procedure where communities continue to be turned into supernodes. The procedure stops until no further grouping of the clusters in the partition increases the modularity to find the partition with the largest modularity. Specifically, clusters are iteratively changed to supernodes by two steps: first, every node is assigned to the neighboring community yielding the partition with the maximum possible modularity. Second, the network turns the cluster into a supernode, transforming each set of links between two distinct communities into a single weighted link between their respective supernodes, and each community's inner links into its own loop of its own supernode, transforming it into a smaller weighted graph (Blondel et al., 2008).

Methods

The Modularity function in Gephi, a software to draw social networks is based on the Louvain algorithm. It is an agglomerative procedure where communities continue to be turned into supernodes. The procedure stops until no further grouping of the clusters in the partition increases the modularity to find the partition with the largest modularity. We used the default setting of modularity resolution in Gephi. This tool helped us optimize visual representations of communities.

Purpose of the study

In this research, we used modularity optimization, the Louvain algorithm, as a community detection method to group students who participated in asynchronous discussion to explore the group effectiveness by examining the network measures in global and local levels. Our main research questions are:

- What are the network characteristics in courses at a global level?
- Do we clearly see defined groupings, or is the network highly connected in many directions?
- What are the characteristics in groups at a local level? How do centrality measures allow us to analyze interaction?

Data Collection

We used anonymized discussion data from an undergraduate business course, which matched our criteria for amount of interaction and course duration. The datasets were collected from a learning management system and are from three separate instances of the course. The total number of posts and students in each course are in table 1. Students were given different reflective questions each week and required to write their opinions as well as replies to others.

Result

We first explored the data at a global level where every single participant is captured as an entire network to understand the structures of the network with network measures and investigated one of the courses to understand the interaction by groups generated by the Louvain algorithm. Table 2 shows key metrics related to each of the course datasets (C4, C5, and C6).

Global level

The number of nodes represents the number of students participating in the discussion, and the edge is the number of connections students made during the discussion. Degree describes the number of links or neighbors, and in an undirected network, the average weighted degree describes the average undirected link-related weight, which means repetitive interactions in pairs were counted without directionality. The results show a slightly higher number of edges and average weighted degree in C6 than in C4 though the number of nodes is smaller than C4 indicating interaction in C6 was more active. The number of nodes plays an essential role in the degree, but the degree does not proportionally increase by the nodes number. Students in Course 5 had less active interaction relative to the other two courses.

The connectivity among neighbors is important in the structure because it represents how closely connected or clustered the network is. The clustering coefficient is the ratio between the actual number of triangles and the potential maximum number of triangles where the node could participate or the number of closed triangles relative to the potential triangles in the network, so when the clustering coefficient is 1 (maximum), possible triangles and actual triangles are equal. Generally, a high clustering coefficient can predict well knit social groups. In C6, about 37 percent of all possible graph triangles are complete, so we can assume this course has relatively dense networks because of this clustering coefficient relative to the C5, 28 percent.

As the diameter refers to the longest graph distance (maximum of the eccentricity) between any two nodes in the network, if it is 5, that represents the longest distance between two nodes in one network. Despite nodes being at a greater distance from the center of the graph, none should be farther than 5 nodes in C4 and 4 nodes in C5 and C6. The average path length (maximum eccentricity) was between 1.8 to 2. We do not have outliers at the fringes of the network who should travel more than five steps to cross the graph in all courses. The graph density is a measure of the level of connected edges within a network relative to the total possible values, ranging from 0 to 1, the maximal density. All courses have about 21- 26% of density, which describes how individuals are connected within networks. That is, students were connected between 21 to 26 times out of 100 within one semester.

Local level

At the global level, we found out these network sociograms (figure 1) can not only represent the overall structure of students' interaction but also give a clue about the important individuals within the network in different roles. As explained above, overall, a sparser, less centralized, and less weighted structure appeared in C4 in comparison to C5 and C6. In the sociograms, node colors represent different groups, and node sizes are degrees.

At a local level, we analyzed the interaction by groups generated by the algorithm. For this paper, only one group was analyzed. The size of the groups is well distributed though the modularity score is low (Table 3). The interactions (edges) within groups varies from 12 to 29. Group 1 and 2 show constant interaction in some pairs, and those interactions are not just within one pair, so we can assume that group 1 and 2 are considered communities (Figure 2). However, group 3 shows a very light weight leaving only one pair communicating frequently, although the nodes in the group are all connected. One interesting node in this graph is 49. This person works as a bridge who might influence the interaction over two clusters. As expected while looking at the graphs, triangles(cliques) were less than 1 in group 3 but over 10 in group 2, so we see some gaps in groups from this algorithm (Table 4).

Conclusion

We first created networks based on their interactions in three courses and compared them to understand the overall network structure in asynchronous discussion within a course. We found slight variances in measures by courses, but they were all within a range, with no extreme outcomes. In visualization, we could still observe the nodes at the fringes though they are connected. After this global level analysis, we used a modularity optimization method to group students and investigated this grouping method to understand whether it provides a legitimate interpretation of interactions. The analysis provides weights and connections among participants, but it does not mean that all group members are equally connected within groups when it comes to outcomes at the local level. Nevertheless, a few groups are still highly connected to be assumed as communities and it can be very influential for the effective group formation. That is, the variations of network measures in groups proved whether the groups were effective.

As the data analysis is only based on student discussions throughout the course without learner attributes, and the algorithm uses randomness and weight, it may be difficult to generalize this algorithm as a reliable way for group formation of learners, but this could be an efficient way for grouping in large numbers (Yassine, 2020). Also, we believe this method is an excellent way to analyze group relations to anticipate group effectiveness and this research will help expand the ways to group students with their interaction patterns in online courses where asynchronous discussion is used as a significant part of learning processes. For future research, we will examine the discourse data aligning with the network data.

References

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology*, *92*(3), 595–615. https://doi.org/10.1037/0021- 9010.92.3.595

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Carolan, B. V. (2013). *Social network analysis and education: Theory, methods & applications*. Sage Publications.

De Dreu, C. K. W., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *Journal of Applied Psychology, 88*(4), 741–749. https://doi.org/10.1037/0021-9010.88.4.741

Froehlich, D. E., Van Waes, S., & Schäfer, H. (2020). Linking quantitative and qualitative network approaches: A review of mixed methods social network analysis in education research. *Review of Research in Education*, *44*(1), 244-268.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821-7826.

Jalan, S., Sarkar, C., Madhusudanan, A., & Dwivedi, S. K. (2014). Uncovering randomness and success in society. *PloS one*, *9*(2), e88249.

Jan, S.K., Vlachopoulos, P. Social Network Analysis: A Framework for Identifying Communities in Higher Education Online Learning. *Tech Know Learn* **24,** 621–639 (2019). https://doi.org/10.1007/s10758-018-9375-y

Kourtellis, N., Alahakoon, T., Simha, R., Iamniti, A., & Tripathi, R. (2013). Identifying high betweenness centrality nodes in large social networks. *Social Network Analysis and Mining*, *3*(4), 899-914.

Kozlowski, S. W., & Bell, B. S. (2003). Work groups and teams in organizations.

Marsden, Peter V. (2004). "*Network Analysis*." Encyclopedia of Social Measurement, edited by Kimberly Kempf-Leonard, 819-825. San Diego: Academic Press.

Magpili, N. C., & Pazos, P. (2018). Self-managing team performance: A systematic review of multilevel input factors. Small Group Research, 49(1), 3–33. https:// doi.org/10.1177/1046496417710500

Menczer, F., Fortunato, S., & Davis, C. (2020). *A First Course in Network Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108653947

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.

Newman, M. (2018). *Networks*. Oxford university press.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Ouyang, F., & Scharber, C. (2017). The influences of an experienced instructor's discussion design and facilitation on an online learning community development: A social network analysis study. *The Internet and Higher Education*, *35*, 34 47. https://doi.org/10.1016/j.iheduc.2017.07.002

Rodríguez, D., Sicilia, M. Á., Sánchez-Alonso, S., Lezcano, L., & García-Barriocanal, E. (2011). Exploring affiliation network models as a collaborative filtering mechanism in e-learning. *Interactive Learning Environments*, *19*(4), 317-331.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre (2008) Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.

Weinberger, A., Ertl, B., Fischer, F., & Mandl, H. (2005). Epistemic and social scripts in computer?supported collaborative learning. *Instructional Science*, *33*(1), 1–30. https://doi.org/10.1007/s11251-004-2322-4

Yang, D. Liu, J. Liu (2010). Discovering communities from social networks: methodologies and applications. Handbook of Social Network Technologies and Applications, Springer, pp. 331-346

Yassine, S., Kadry, S., & Sicilia, M. A. (2021). Application of community detection algorithms on learning networks. The case of Khan Academy repository. *Computer Applications in Engineering Education*, *29*(2), 411-424.

Tables and Graphs

	# of students	# of posts	# of topics
C4	53	808	9
C5	50	764	9
C6	51	839	9

Table 1. Course information

	# of nodes	# of edges	Av. weighted degree	Clustering coefficient	Diameter	Av. path length	Graph density
C4	53	314	16.679	0.345	5	2.111	0.228
C5	50	261	14.96	0.282	4	1.996	0.213
C6	51	340	17.294	0.372	4	1.807	0.267

Table 2. Network statistics at a global level

	Group 0	Group 1	Group 2	Group 3	Group 4	Modularity
Sizes (# of nodes)	11	11	10	10	8	.231
Edges	22	29	29	12	14	

Table 3. sizes and edges by groups

Modularity class	Degree	Weighted Degree	Eccentricity	Clustering Coefficient	Triangles	Eigenvector
0	4.00	5.45	2.36	.405	2.73	.646
1	5.27	9.09	2.55	.581	7.91	.721
2	5.80	12.80	2.20	.751	10.80	.817
3	2.40	2.80	3.50	.240	.60	.409
4	3.50	4.50	2.38	.675	2.63	.659

Table 4. Network statistics by the modularity class

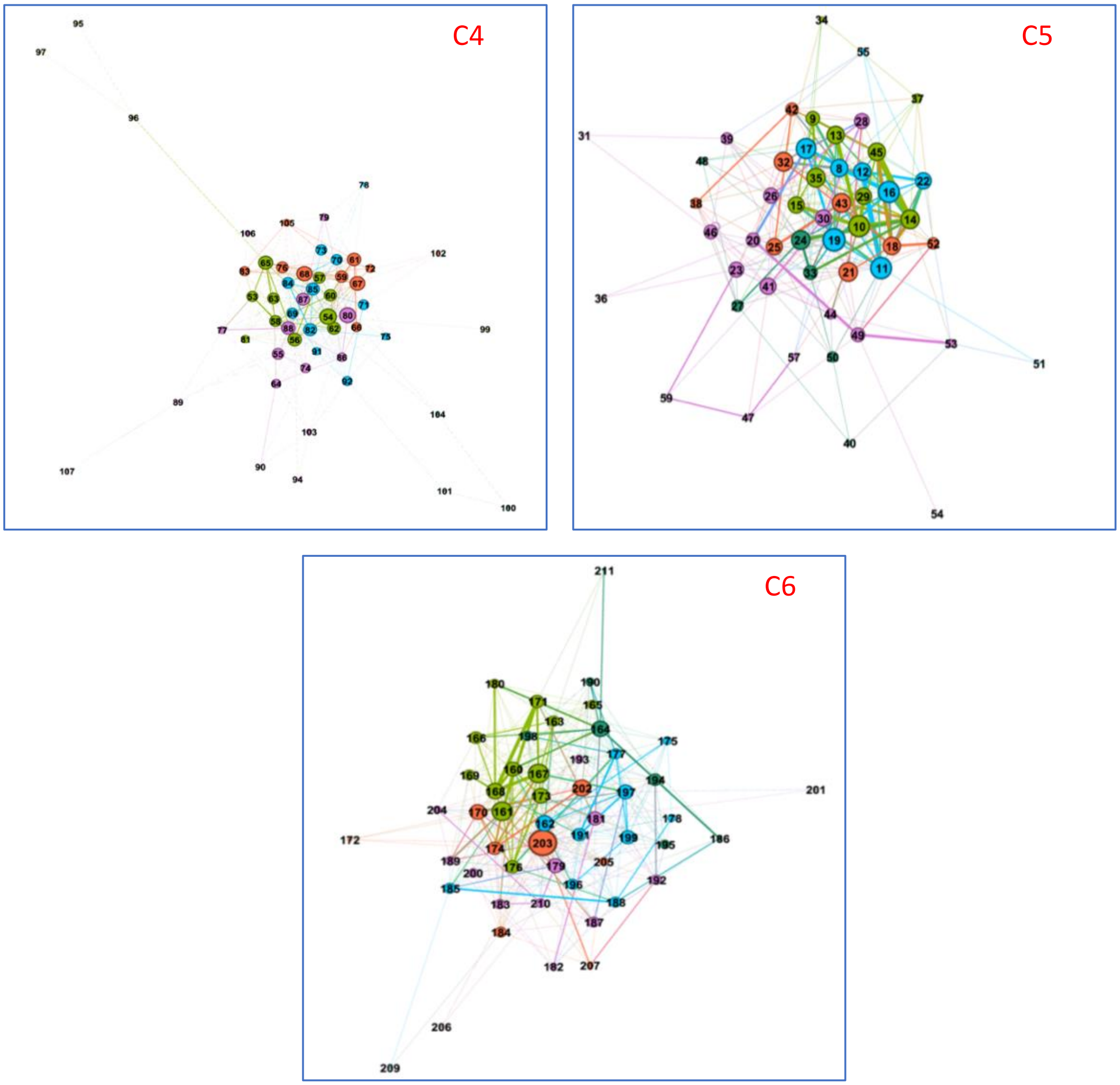

Figure 1. Course network sociograms

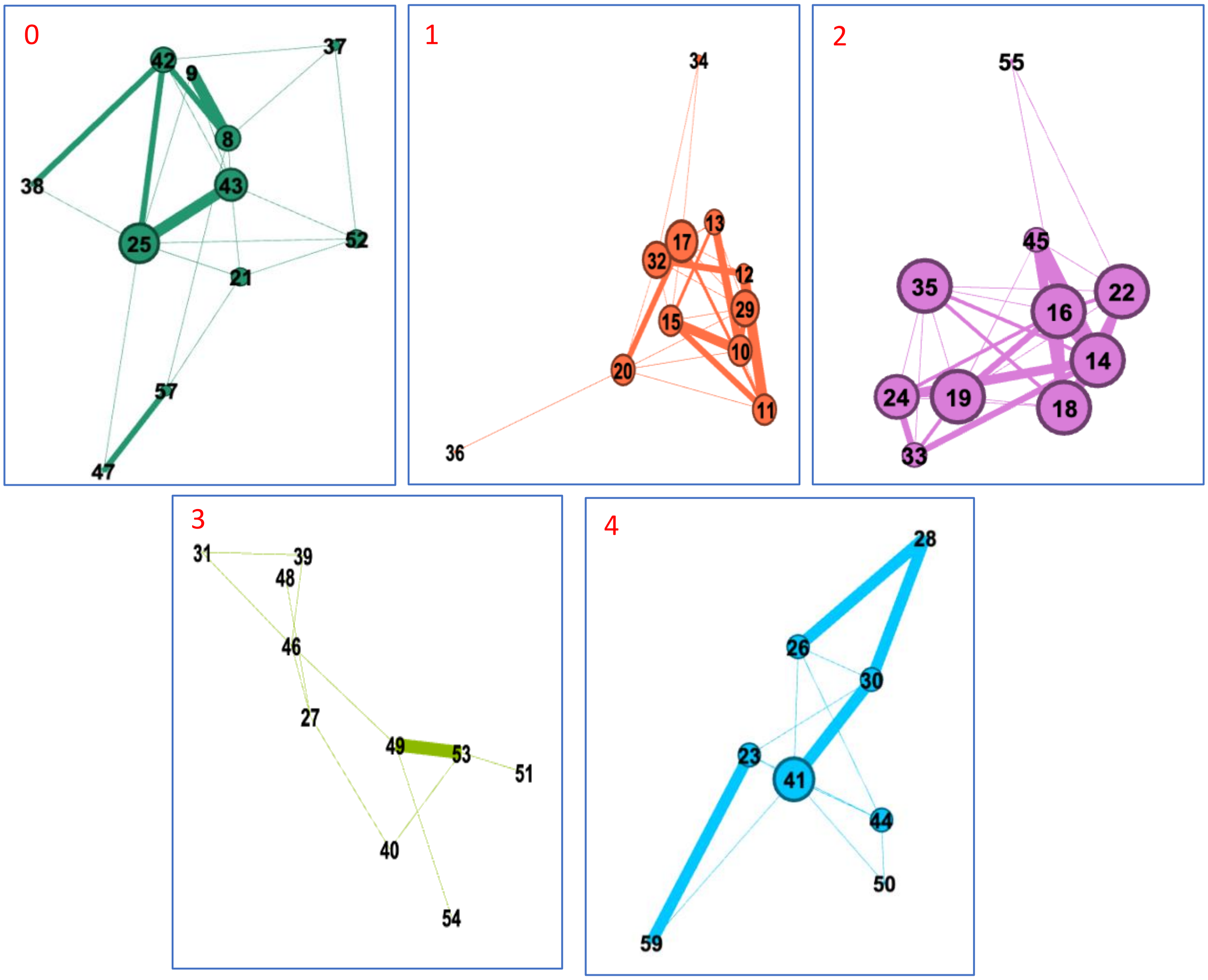

Figure 2. Group networks