Social Network and Self-representation in Megathread: Group Formation in a Data Science Crowdsourcing Community

Shiyao Wei
Florida State University
sw22b@fsu.edu

Ran Bi
SAS Institute
ran.bi@sas.com

ABSTRACT

Crowdsourcing platforms are evolving into hubs for professional development and informal learning. As crowdsourcing increasingly encourages people to work together, knowledge workers actively participate in these collective endeavors. The initial step in this collaboration journey is the formation of teams. This study explores the group formation process within the crowdsourcing context, using Kaggle—a virtual community for knowledge workers in data science—as a case study. By leveraging competition and conversation data from Kaggle in 2023, we employed social network analysis to observe the emergence of groups within the social space. Additionally, we utilized BERTopic, a text mining program, to identify key themes in users' discourse and examine how individuals self-represent through social engagement. The findings from this research contribute to shaping the future design of online communities, such as Kaggle and MOOCs. Furthermore, we suggest how individual knowledge workers can enhance their participation in online collaboration through crowdsourcing.

Keywords

Group Formation, Kaggle, Crowdsource, Social Network Analysis, Text Mining

1. INTRODUCTION

Knowledge workers have been more engaged in the crowdsourcing economy in recent years [9]. Crowdsourcing initially did not encourage collaboration on single tasks as they are simple. As the task becomes more complex and challenging, collaboration is encouraged within knowledge workers [4, 8] [5]. The new landscape of the crowdsourcing economy makes collaboration with strangers possible in virtual communities, which requires individuals to be able to collaborate online.

Crowdsourcing is one way for professionals to earn money and compile resumes [20]. Meanwhile, online collaboration in crowdsourcing is a possible way for students to gain practical experience and build up pre-career skill sets, such as collaboration and problem-solving abilities [16].

The first stage of online collaboration within the context of crowdsourcing is group formation. Working in a team is necessary in companies and schools, where individuals cannot choose collaborators. In contrast, in virtual communities, individuals have the autonomy to choose their collaborators.

A channel for online asynchronous discussion designed for teammate matching in an online community offers a place where individuals can learn about their potential collaborators and competitors and represent themselves to facilitate the group formation process [11]. However, limited research has investigated how users represent themselves within this space [14, 17]. Identifying this gap in previous studies, we are interested in understanding the impact of this space on facilitating the group formation process and examining how users represent themselves during this process. Kaggle is an online community for knowledge workers focused on data science. As a virtual community, Kaggle itself and other organizations hold data science competitions, allowing strangers to team up to solve data science problems in real life. Due to the complexity of the task, which is also influenced by rewards and other factors, participants choose to engage in the competitions as a team [9].

Each competition on Kaggle features an online community with sections including data, models, discussions, code, and leaderboards. Participants can access data, check models, engage in conversations, upload code, and review leaderboards after running their models. Kaggle staff serve as moderators in these communities, addressing questions from participants. In the discussion thread, most competitions have a designated space called the ”Look for a Team Megathread.”(L4T Megathread). This thread allows individuals to post and respond to group formation information, serving as an example of online asynchronous discussion for team building. Kaggle moderators introduced this mechanism 2022, creating a space for individuals seeking teammates.

This study holds significance for online communities, like Kaggle, aiming to enhance the design to facilitate the group formation process on a larger scale. The study provides insights into how individual participants can effectively utilize the space and represent themselves during group formation.

S. Wei and R. Bi. Social network and self-representation in megathread: Group formation in a data science crowdsourcing community. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 738–743, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729934

2. RELATED STUDY

2.1 Group Formation in Online Community

There are two major ways of forming a team, depending on whether human agency is considered in the process [8]. Automatic group formation is grouping individuals based on their history, performance, and personal characteristics [12]. Current practices in the MOOC community focus on automatic group formation to decrease the workload for instructors [19, 18]. Sanz-Martinez et al. [19, 18] compared an automatic students' activity criteria grouping approach to a baseline grouping function from the platform, indicating that a homogenous group based on prior activity positively impacts student satisfaction and group interactions.

Other than automatic group formation, human agency is another important way of forming a group; for example, group formation decisions are made by individuals. Marlow et al. [14] highlighted that users actively seek information about their potential collaborators to assist their decision of group formation. Lykourentzou et al. [21] shared that participants interact on easy tasks before working on complex tasks as a group, which leads to better online collaborations. These works highlighted the significance of human agency in collaborative environments. Huang et al. [9] highlighted a gap in research, noting that most studies on team performance assume pre-formed transient teams. Our study addresses this gap by exploring how teams naturally form in virtual communities.

2.2 Self-representation and Personal Characteristics

Whether human agency is highlighted or not, the characteristics of group members influence the group formation process. Dissanayake et al. [4] observed a positive correlation between language proficiency and past collaborations with team-up decisions in Kaggle competitions. At the same time, factors such as geographical distance, skill disparity, and tenure disparity exhibited a negative association with team-up decisions. Prior research has focused more on objective parameters, including language [4, 24], geographical distance, skill, and tenure disparity[4], often overlooking subjective aspects of users' self-representation discourse. Self-presentation, defined as how users decide and can present themselves to others through social networks [8], is a significant way of exposing personal and group characteristics. Group formation depends on the information available about the members and their objectives [22]. Users build up an online impression based on historical activities to attract future collaboration from strangers [14].

2.3 Research Questions

From the above literature review, we identify research gaps in the group formation process in the crowdsourcing community and how participants represent themselves in a social space when ready to team up with others. We aim to answer the following research questions:

1. Is Megathread a commonly used mechanism for individuals seeking to form teams?
2. How does Megathread facilitate team formation in the context of competitions?
3. To form a team, how do participants represent themselves and participate in the Megathread discussion?

3. METHODS

3.1 Data Collection

To answer the above research questions, we gathered data from the Kaggle website, a widely visited platform for data science competitions and discussions among professionals in data science and machine learning. Utilizing the Kaggle API [9, 20], we extracted information on competitions and user profiles from January 1st, 2023, to December 31st, 2023. Within the Kaggle dataset, our attention was directed towards content relevant to the ”Look for a Team” (L4T) Megathread. We analyzed all Kaggle competitions held in 2023, identifying 31 such competitions, of which 18 featured L4T Megathread.

3.2 Data Processing

To ensure the integrity of our analysis, competitions designated as research projects for educational purposes (non-competitive) were excluded from consideration. This removes two such competitions from the pool of 18 featuring L4T Megathread. In addition, we examined all L4T Megathread created and posted by Kaggle Officials. There is one L4T Megathread, falsely created by one Kaggle user. It didn't attract much attention, but we could distinguish and exclude it from our dataset.

3.3 Data Analysis

To answer RQ1, we utilized Kaggle API to visualize the percentage of users who utilized L4T Megathread and their group formation status.

Social network analysis is used to answer RQ2. We utilized the Python package NetworkX [7] for network analysis and generated the plot. Each node represents a user, while an edge signifies a comment from one user to another. The width of an edge indicates the number of comments between two nodes. For visualization purposes, we employed a spring layout [1]. We used a directed graph for detailed analysis. We analyzed the interactions among users, including those who comment on whom or who is part of a specific team. We emphasized the teams where members engaged in reciprocal commenting and collaborated successfully to complete the competition.

Text mining is another component of our methods framework. To answer RQ3, we examined the textual content within the L4T posts and comments, employing BERTopic [6], a machine-powered methodology to gain an overview of users' themes. Initially, we engaged in data preprocessing procedures, cleansing the L4T Megathread comments utilizing the Natural Language Toolkit (NLTK) package [13]. Subsequently, we harnessed the transformative capabilities of the BERT model through the BERTOPIC Python package [6], facilitating the identification of emerging topics and thematic patterns embedded within the discourse. To enhance the semantic representations of textual entities, we utilized the 'all-MiniLM-L6-v2' variant of the SentenceTransformer for embedding extraction. Dimensionality reduction was then conducted via the UMAP algorithm. Leveraging the K-Means clustering technique, we delineated coherent

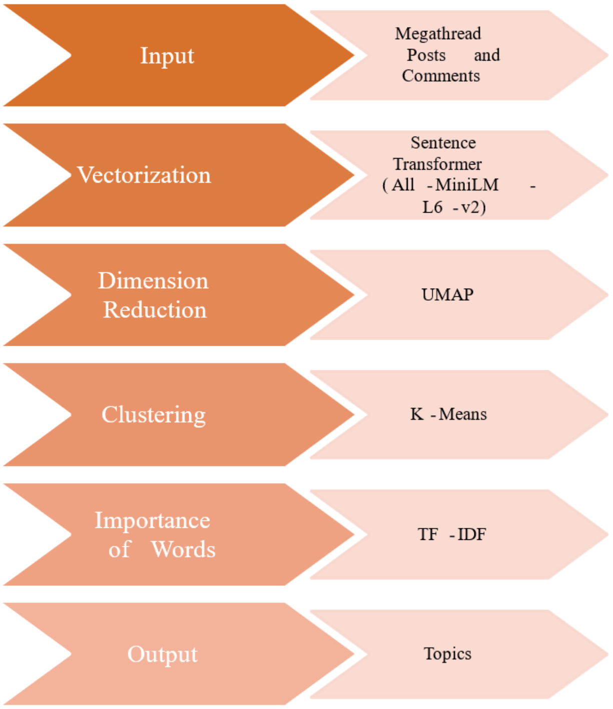

Figure 1: NLP pipeline of utilizing BERT model

topic clusters. Moreover, we capitalized on the CountVectorizer for tokenization and subsequently computed TFIDF scores to construct robust topic representations.

4. RESULTS

4.1 RQ1: Is Megathread a popular mechanism for individuals who want to team up?

Since the launch of Megathread, 127 Megathreads have been created, with 10,248 replies and 1,669 upvotes. In 2023, in reward-based competitions with Megathread, 669 individuals engaged in Megathread discussion and participated in 335 teams. Of these teams, 103 are multiplayer teams, and 28 have two or more players from Megathread, i.e., only around 6% of participants who posted successfully teamed up with others posted in the Megathread. Compared to a total of 38,652 teams in 2023 competitions, Megathread supported limited online group formation and collaboration.

4.2 RQ2: How does a team emerge in Megathread conversation?

To examine the meso-level of group formation, we used two competitions in 2023 as cases to visualize their group formation. In competition #1 (Fig. 3), we illustrate how one team emerges in the network of teammate matching conversation. In this network, N20 refers to a Kaggle staff member, the moderator of the space. All individuals looking to team up replied under this thread. Following the ”looking for teammate” posts, other users responded, expressing their interest in teaming up. The directions between the two nodes indicate that users interacted with each other. Nodes with one arrow pointing in different directions indicate that a participant replied to others’ posts. Nodes with one arrow directed to suggest that they received replies from others. Nodes without edges indicate that their comments were deleted. Despite multiple rounds of conversations in the community, only two green nodes (N1 and N2) successfully formed a team. In competition #2, as shown in Fig. 4, four users (N28, N30, N39, N44) teamed up after a round of conversation in the Megathread. Users did not team up under one thread in this group formation case. Instead, under the first call for teammate posts by N28, two individuals (N30 and N44) replied to the post and successfully teamed up with the original poster. They found the fourth teammate (N39) after the fourth teammate replied to another post by the original poster (N28).

We visualized the social network based on all Megathreads from competitions in 2023, as shown in Fig. 5. Different colors of lines represent different competitions they participated. A cross-competition communication network grew out of the Megathread discussion.

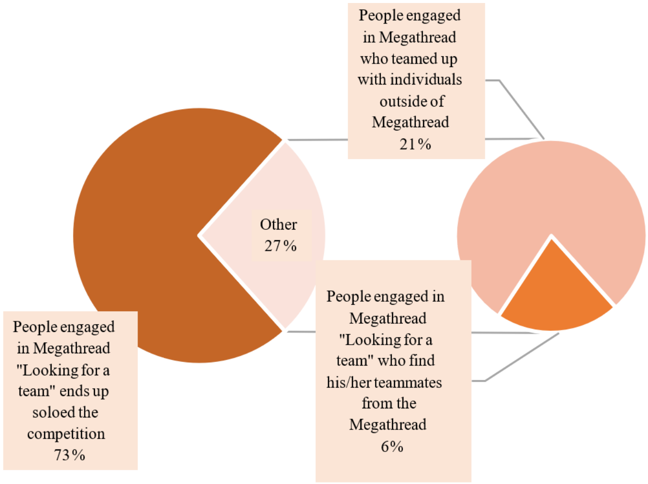

Figure 2: Overview of team-up percentage in 2023 competitions

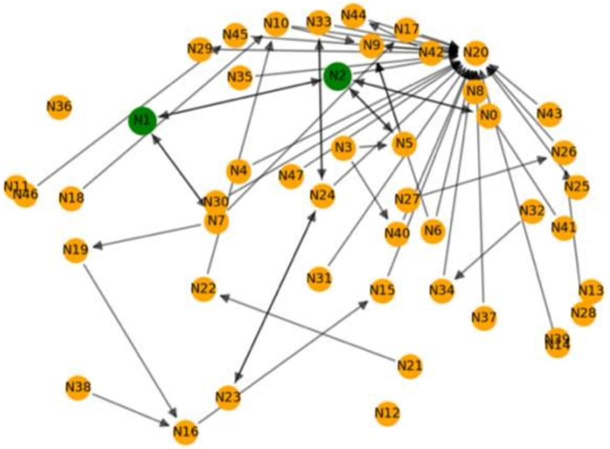

Figure 3: Social network analysis of Megathread conversation of one competition*

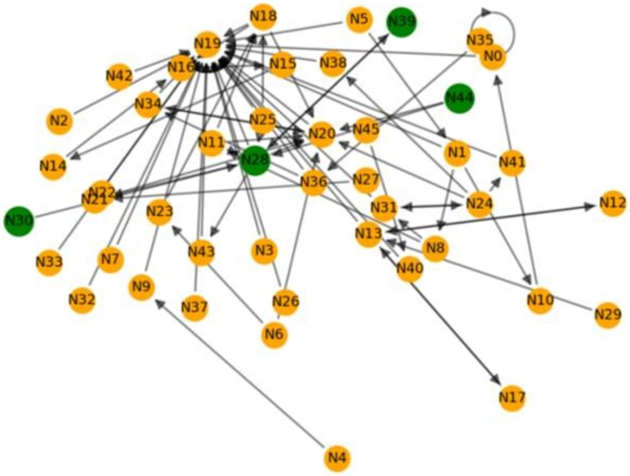

Figure 4: Social network analysis of Megathread conversation of one competition**

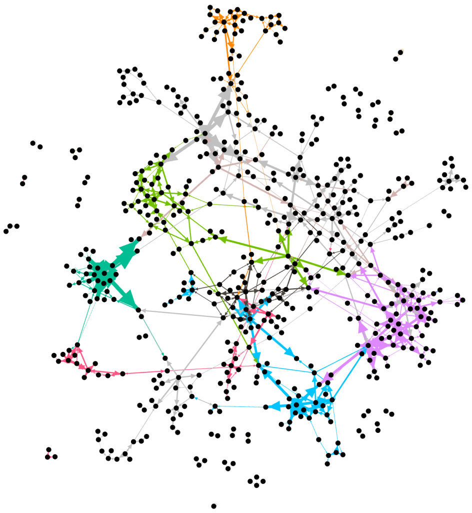

Figure 5: Social network analysis of Megathread conversation of all competition

4.3 RQ3: How do participants represent themselves in Megathread?

From the initial analysis of BERT's results, we identified several predominant themes from users' interactions. Participants repeatedly mentioned ”learning” within the context of the competition, indicating their passion for acquiring new knowledge. Additionally, under the first topic, we observed that participants frequently used the word ”new,” suggesting that they are newcomers to Kaggle, and most of the participants who posted there were novices in the community. In the second topic, once participants proceed to the next step, they utilize additional communication tools such as ”email” and ”Discord” to facilitate team communication. Participants occasionally express their appreciation to potential teammates in the third topic, using phrases like ”thks” (thanks). In the fourth topic, the terms 'joining,' 'effort,' and 'contribution' appear to form a collective theme. In the fifth topic, ”issue” and ”problem” emerge as negative signals in group formations. In the last topic, Kaggle staff posts an announcement every time a Megathread is created.

Table 1: Topic Representation

Topic	Count	Representation
1	702	Team, data, experience, learning
2	695	Join, email, invite, sent
3	68	Team, love, thks, possible
4	56	Join, efforts, contribute, want
5	37	Issue, tell, problem, try
6	19	Competition, info, considered, adherence

Table 2: Team Performance

Competition	Levels of Teammates	#Comments	Rank
Predict Student Performance from Game Play	Contributor, Novice	57	456/2051
Vesuvius Challenge - Ink Detection	Contributor(3), Expert	67	529/1249

5. DISCUSSIONS

5.1 Social Dynamics of Megathread

The initial analysis of this pilot study investigates how the group formation process is facilitated and succeeds within a crowdsourcing community. Drawing parallels with collaborative studies in Massive Open Online Courses (MOOCs), it underscores online environments' challenges in forming effective teams [23]. Despite active involvement in the initial teaming-up phase with strangers in the Megathread, a consistent trend emerges, revealing that they ultimately form teams with acquaintances [9].

However, our pilot study introduces a novel perspective by shedding light on the role of Megathread in providing an expansive platform. This proves beneficial for individuals who face challenges in finding collaborators in real-life scenarios and are keen on engaging with others for collaborative efforts.

Megathread, as revealed by our study, serves as a platform for initiating group formation in the crowdsourcing community. The analysis, utilizing BERT's results, indicates that participants often migrate to other social media platforms like Slack or Discord or exchange email addresses for sustained communication. In this context, Megathread functions as an entry-level social space, catalyzing users to establish connections and present themselves.

Furthermore, we observed that participants who initially teamed up in Megathread continued their collaboration by working on another competition. This finding proves that as an initial social space, Megathread could scaffold future collaboration with others. This discovery resonates with the team date study [21], emphasizing the significance of fostering connections in online collaboration.

5.2 Self-representation of Participants

As participants gear up for collaborative efforts with strangers on the Megathread, they consciously portray themselves as eager learners, friendly individuals, and committed team players. This behavior aligns with established online impression management strategies [14, 20], emphasizing the importance of showcasing one's expertise in online collaboration [17]. As the crowdsourcing economy evolves towards a more collaboration-oriented paradigm, we suggest that knowledge workers consistently cultivate and augment their online professional profiles [14, 17] to attract more future collaboration opportunities.

A distinctive personal characteristic identified through BERTopic is that participants who ask for a team-up are relatively "new" to the space. This may stem from their limited data science knowledge or unfamiliarity with platforms like Kaggle within the online domain. Those falling into the former category tend to articulate their commitment to learning, while the latter cohort often shares their expertise and backgrounds, seeking connections with potential collaborators. Online communities are encouraged to consider implementing support structures for newcomers, offering guidance on online collaboration etiquette.

Drawing inspiration from Lave and Wenger's legitimate peripheral participation [10], the community could curate an archive of successful group collaboration to encourage newbies' observational learning in the space.

Professional development is a significant aspect of users' self-representation within the crowdsourcing community. When users mention "learning," their motivations extend beyond the pursuit of awards and monetary gains and express their desire to refine their professional skills. In this way, Kaggle competitions could be regarded as a project-based learning experience for students in related majors [3, 16, 15].

5.3 Limitations

This pilot study has certain limitations that warrant consideration. Firstly, in the social network analysis, the data limitation led to the cumulation of all competitions to illustrate collective group formation in 2023 on Kaggle. However, a more refined approach treating each competition as a distinct entity, attributed to different Kaggle staff, will allow us a more precise visualization of outcomes and facilitate the examination of cross-competition group formation behaviors. To address this limitation, future studies should explore the collection of additional data sources.

Furthermore, using BERTopic to analyze discourse in Megathread may limit our understanding. BERTopic, as an unexplainable black-box program, may not capture nuanced aspects of the discourse effectively[2]. To enhance our knowledge, especially in addressing RQ3, it is recommended that we consider incorporating hand coding in future research. This approach combines machine understanding with human interpretation, providing a more comprehensive analysis of the intricacies in Megathread discourse.

6. ACKNOWLEDGMENTS

We thank Dr. K. Bret Staudt Willet for reviewing and for the support during the ideation of this project.

7. REFERENCES

[1] N. Akhtar. Social network analysis tools. In *2014 Fourth International Conference on Communication Systems and Network Technologies*, pages 388–392, 2014.

[2] R. Bi and S. Wei. Exploring the Implementation of NLP Topic Modeling for Understanding the Dynamics of Informal Learning in an AI Painting Community. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 434–437. International Educational Data Mining Society, July 2023.

[3] W. Chow. A pedagogy that uses a kaggle competition for teaching machine learning: an experience sharing. In *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, pages 1–5, 2019.

[4] I. Dissanayake, N. Mehta, P. Palvia, V. Taras, and K. Amoako-Gyampah. Competition matters! self-efficacy, effort, and performance in crowdsourcing teams. *Information & management*, 56(8):103158, 2019.

[5] M. L. Gray, S. Suri, S. S. Ali, and D. Kulkarni. The crowd is a collaborative network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 134–147, New York, NY, USA, 2016. Association for Computing Machinery.

[6] M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[7] A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[8] A. M. Harris, D. Gómez-Zará, L. A. DeChurch, and N. S. Contractor. Joining together online: The trajectory of cscw scholarship on group formation. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[9] K. Huang, J. Zhou, and S. Chen. Being a solo endeavor or team worker in crowdsourcing contests? it is a long-term decision you need to make. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022.

[10] J. Lave and E. Wenger. *Situated Learning: Legitimate Peripheral Participation*. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press, 1991.

[11] X. Li, Y. Bai, and Y. Kang. Exploring the social influence of the kaggle virtual community on the m5 competition. *International Journal of Forecasting*, 38(4):1507–1518, 2022.

[12] Y.-T. Lin, Y.-M. Huang, and S.-C. Cheng. An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization. *Computers & Education*, 55(4):1483–1493, 2010.

[13] E. Loper and S. Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[14] J. Marlow, L. Dabbish, and J. Herbsleb. Impression formation in online peer production: activity traces and personal profiles in github. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, page 117–128, New York, NY, USA, 2013. Association for Computing Machinery.

[15] B. Memarian and T. Doleck. Data science pedagogical tools and practices: A systematic literature review. *Education and Information Technologies*, pages 1–23, 2023.

[16] J. Polak and D. Cook. A study on student performance, engagement, and experience with kaggle inclass data challenges. *Journal of statistics and data science education*, 29(1):63–70, 2021.

[17] R. Ren, B. Yan, and L. Jian. Show me your expertise before teaming up: sharing online profiles predicts success in open innovation. *Internet Research*, 30(3):845–868, 2020.

[18] L. Sanz-Martínez, E. Er, A. Martínez-Monés, Y. Dimitriadis, and M. L. Bote-Lorenzo. Creating collaborative groups in a mooc: a homogeneous engagement grouping approach. *Behaviour & Information Technology*, 38(11):1107–1121, 2019.

[19] L. Sanz-Martínez, A. Martínez-Monés, M. L. Bote-Lorenzo, J. A. Munoz-Cristóbal, and Y. Dimitriadis. Automatic group formation in a mooc based on students' activity criteria. In *Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12–15, 2017, Proceedings 12*, pages 179–193. Springer, 2017.

[20] M. Twyman, G. Murić, and W. Zheng. Positioning in a collaboration network and performance in competitions: a case study of kaggle. *Journal of Computer-Mediated Communication*, 28(4):zmad024, 2023.

[21] F. Vinella, R. Mosch, I. Lykourentzou, and J. Masthoff. The impact of digital nudging techniques on the formation of self-assembled crowd project teams. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 265–275, 2022.

[22] X. Wang, Z. Zhao, and W. Ng. Ustf: A unified system of team formation. *IEEE Transactions on Big Data*, 2(1):70–84, 2016.

[23] M. Wen, K. Maki, S. Dow, J. D. Herbsleb, and C. Rose. Supporting virtual team formation through community-wide deliberation. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19, 2017.

[24] Q. Zhang, K. L. Peck, A. Hristova, K. W. Jablokow, V. Hoffman, E. Park, and R. Y. Bayeck. Exploring the communication preferences of mooc learners and the value of preference-based groups: Is grouping enough? *Educational Technology Research and Development*, 64:809–837, 2016.