Slack ID: CoBird

Track: Data Analysis

Stage: Stage 7 task

NETWORK ANALYSIS ON MOOC USER ACTION DATASET

Students on MOOC platforms exhibit several behaviors and trends. Data from a popular MOOC platform was retrieved and examined in further detail to answer relevant questions that could provide insights for organizers and future students of MOOC courses. The data for this exercise was retrieved from here and was explored using relevant libraries in Python.

Dataset Background

The dataset used for this task contains three files and represents "actions taken by users on a popular MOOC platform", with the nodes representing either users and or the target activities, while edges signify the actions taken by nodes. The data were individually loaded into the Python interpreter and each file was examined for any error or inconsistency.

Exploratory Data Analysis

Each of the files was found to contain no duplicate, no missing or null values, nor other errors that could diminish the integrity of the analysis. Using the merge command, the three files were merged into a single file on their common column, ACTIONID, to proceed with the analysis. Each file and the corresponding final file contained 411,749 distinct entries (rows), with the combined final file having 9 columns.

Research Questions

This project aims to answer six research questions with the dataset given.

1. **Does network position correlate with continued engagement over time or are some users central but disengaged?**

One important feature that was examined from the dataset is to investigate whether the extent of a user's connection within the network correlates with how engaged they are over time. To examine these, three network centrality measures were examined. They are the degree centrality, the

betweenness centrality, and the eigenvector centrality. Using the appropriate lines of code, the network graph of the student's connections and the visualizations of their network centrality score to their level of continuous engagement (that is, how long they get engaged in the course) are presented below.

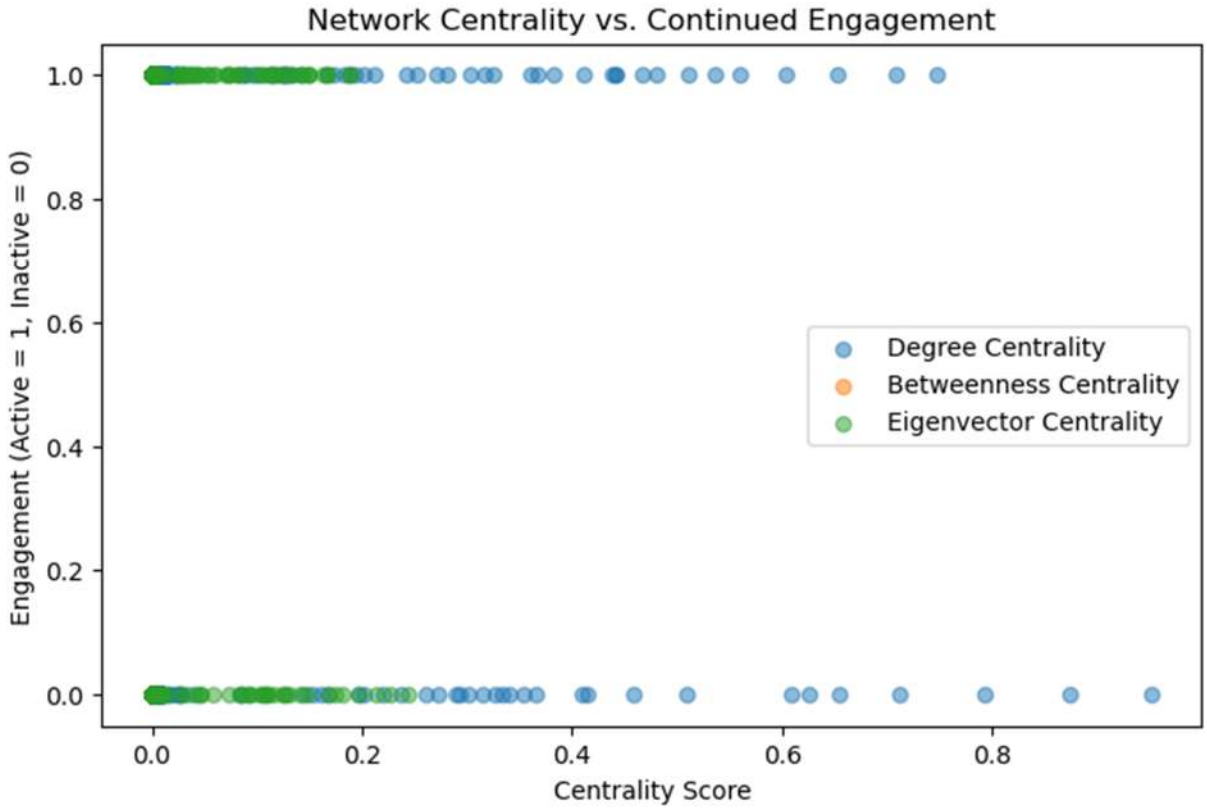


Figure 1: Figure showing Student's Network Centrality in Respect to Continued Interaction

The result of the correlation between centrality and engagement shows that while the correlation between degree centrality and engagement is 0.579, the correlation between Betweenness centrality and engagement, and between Eigenvector centrality and engagement is 0.024 each. This result implies that there is a moderately positive relationship between having high connections within the network (degree centrality) and engaging for longer periods in the course, r = 0.579. However, having connections with students who have high connections themselves (Eigenvector centrality) or serving as a link between different groups of users/students (betweenness centrality) in the network is weakly related to the duration of a student's engagement.

2. Does the level of user activity influence dropout rates?

One of the objectives of this analysis is to determine whether there is a significant difference in the activity /engagement level of students (users) who dropped out after the action and students who continued in the program. To answer this question, the entire dataset was segregated into two groups, with the first group consisting of students who dropped out and the second comprising students who continued the program. Descriptive statistics on the number of activities engaged in by users from both groups were obtained using the 'describe' function. It was found that the average number of activities engaged by students who didn't drop out was 58.428, SD = 57.96, while for students who dropped out, the average number of activities they engaged in was 53.948, SD = 51.038. The boxplot diagram of the interaction level of both groups of students is also presented below:

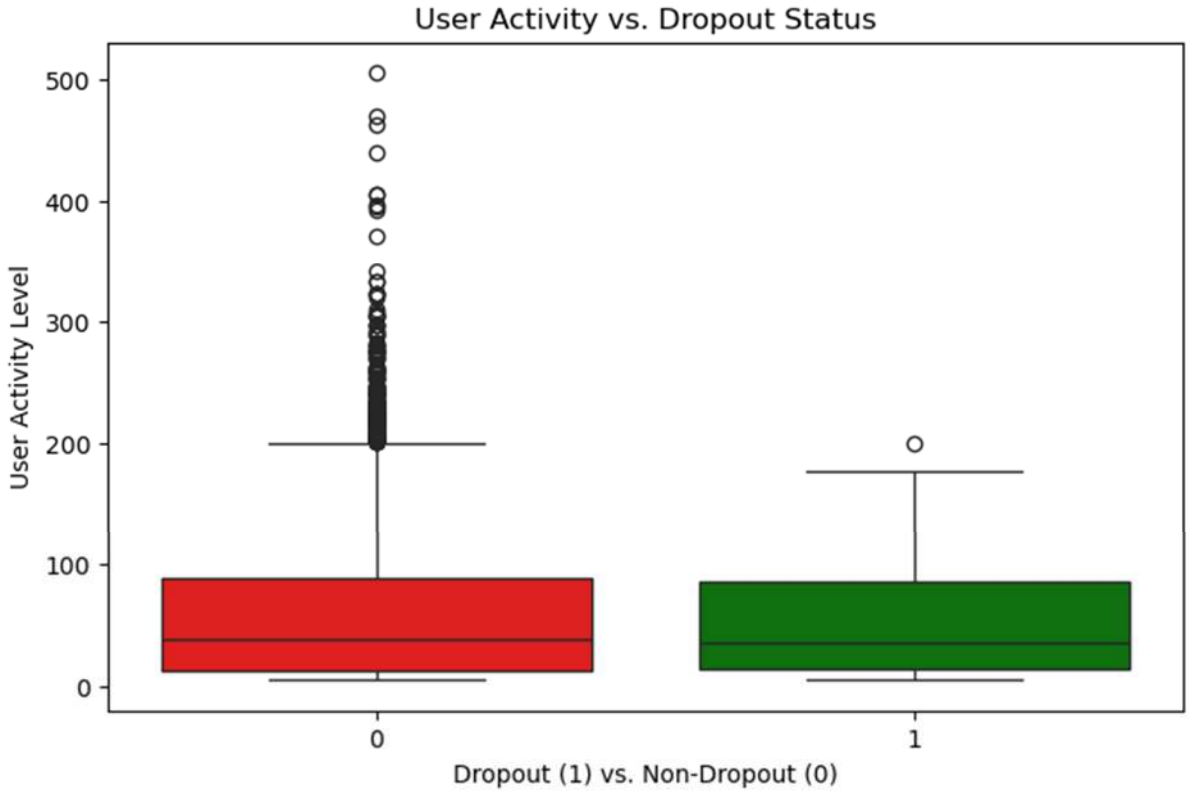


Figure 2: Boxplot of User Activity Level for Dropouts and Continuing Students

The statistical t-test was finally used to examine whether there exists a significant difference in the activity levels of both groups of users. The result of the test showed that at the 95% significance

level, **there is no statistical difference in the average number of activities engaged by those who dropped out and those who did not drop out of the course**, t = 0.665, p = 0.508

A network diagram of both groups of students and the target activities they engage in is provided below.

3. **How do User Interactions form a Learning /community cluster, and does the community cluster influence retention rates in the program?**

It is of great interest to know whether the interactions of students in the program show one learning cluster or whether the students follow multiple clusters. Such knowledge is important for instructors to know whether information flow in the course follows one singular pattern or whether the course allows for a diversity of ideas and participants. Knowledge of how learning clusters influence dropout rates also assists prospective students in knowing what type of activity to avoid or be involved in for greater chances of course completion. To achieve this, two nodes were created with the UserID (representing each student) and the TargetID variable (repressing course activities), while the interaction between these nodes formed the edges. The **Louvain modularity** algorithm was afterward used to cluster the interaction patterns into similar clusters to determine the learning /interaction clusters among the students using the lines of code below:

```
# Detect communities using Louvain method
partition = community_louvain.best_partition(G.to_undirected())  # Convert to undirected for community detection

# Add communities to a dataframe
df['Community'] = df['USERID'].map(partition)
```

The below network graph shows that three community clusters can be obtained from the pattern of interaction of the students with one another and with the course activities.

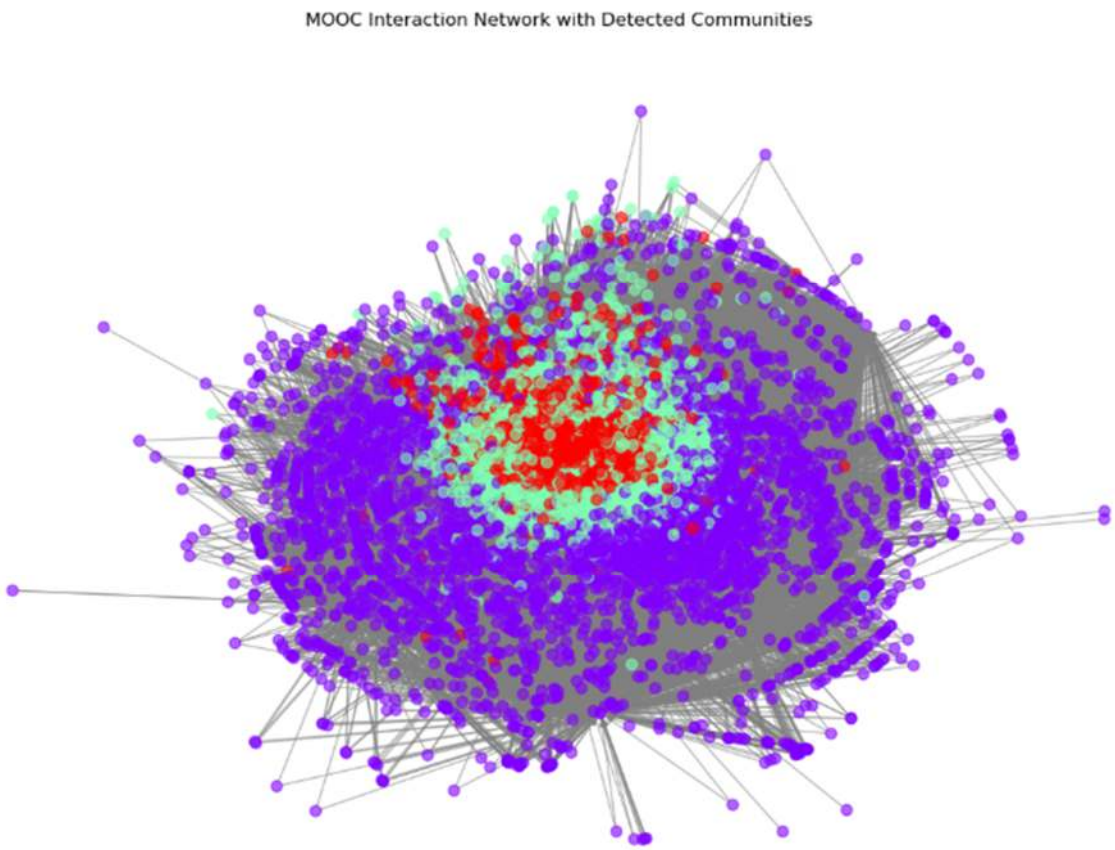


Figure 3: Network Graph of Interactions Among the Three Learning/Community Clusters

Further analysis of the dropout rate according to the community clusters reveals that students with greater interaction levels (community ID = 2) have lower dropout rates (less than 0.005) compared to students in community clusters with the least interactions (community ID = 0) with dropout rates of about 0.04. See the graph of community ID and dropout rates below

Table 1: Dropout Rates by Community Cluster

```
Community-wise Dropout Rates:
 Community
2    0.001404
1    0.006703
0    0.039406
Name: LABEL, dtype: float64
```

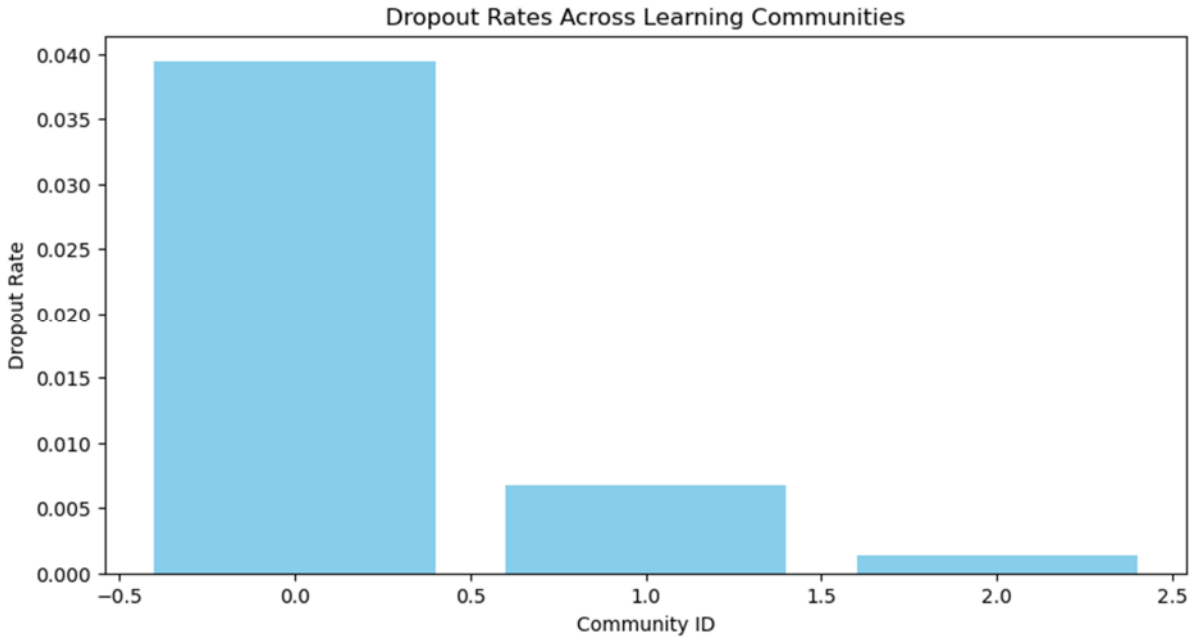


Figure 4: Bar Chart of Dropout Rates Among the Three Clusters of User Interaction

4. **What course activities (Target ID) are most frequently interacted with by students in the MOOC program?**

It is of great interest for course lecturers and platform admins to know the course activities that the students enrolled in their courses interact with the most. Such knowledge will enable them to know which course activities need better instruction clarity and need to be modified to encourage greater participation of students. For this task, only the records of students who did not drop out of the course were used. As such, the number of times each course activity (TARGET ID) was performed by these students was collated and sorted. See figure below.

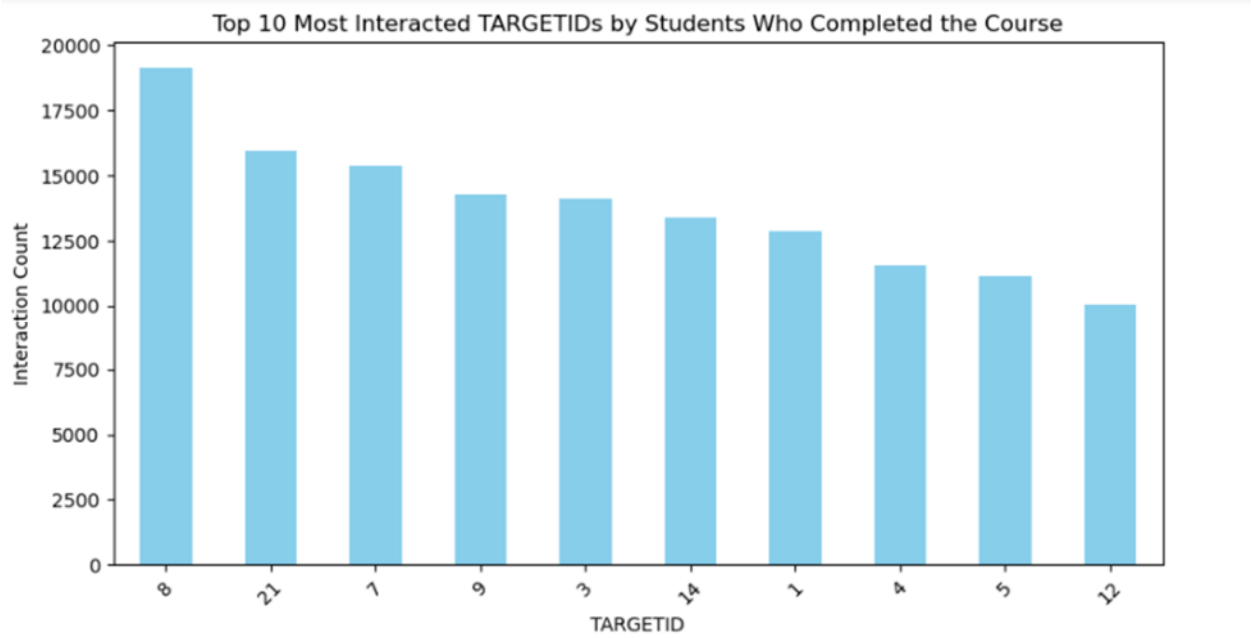


Figure 5: Bar Chart of 10 TargetID/ Course Activities with the Most Student Interaction

From the above figure, it was realized that activities with Target IDs 8, 21, 7, 9, and 3 are the most engaged in while course activities with Target IDs 93, 92, 94, 90, and 95 are the least performed.

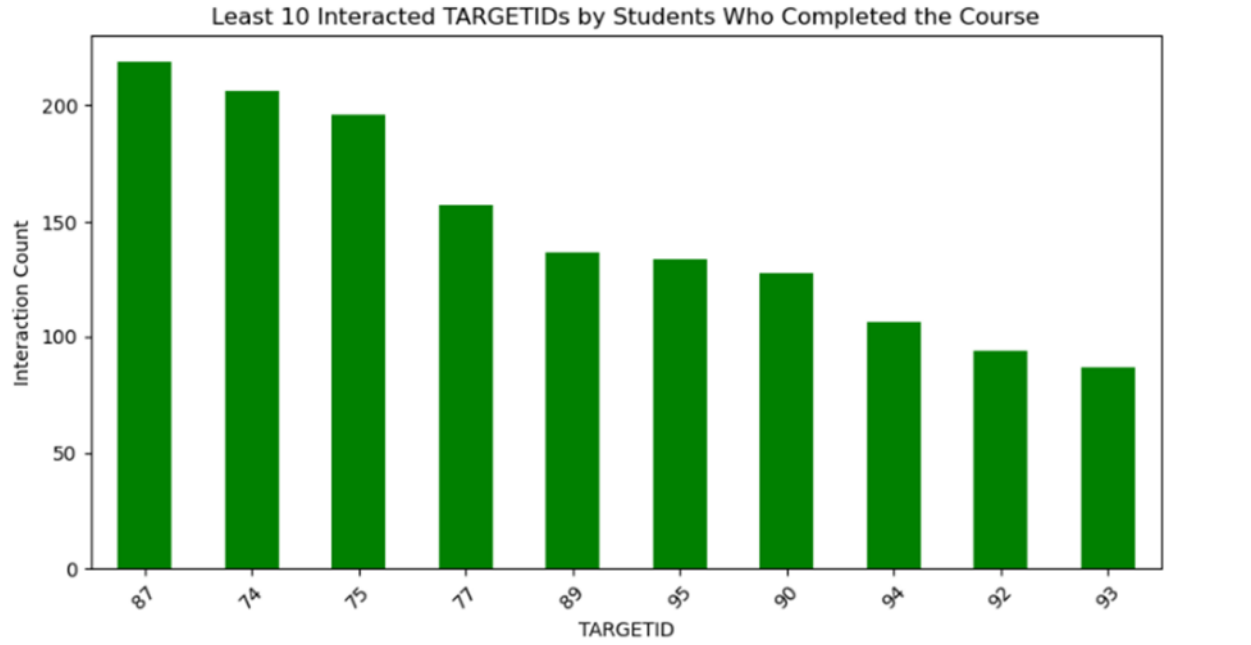


Figure 6: Bar Chart of 10 TargetID/ Course Activities with the Least Student Interaction

5. **Do students' interactions with the course wane or remain the same over time?**

This research question aims to explore whether student's behavior/interactions with the course activity remain fairly the same over the length of the course. Knowledge of a significant deviation in students' devotion to the course activity can be used to further interrogate whether the duration of time devoted to the course is a predictor of completion rate. To answer this question, the time duration spent on course activities by each student weekly was computed, and the standard deviation was calculated for each user. Users with very low weekly standard deviation in interaction duration are classified as consistent while users with high standard deviation were classified as intermittent. The below code was used to perform the engagement classification using Python's "qcut" function.

```
if engagement_stats['std'].nunique() > 2:
    num_bins = min(4, engagement_stats['std'].nunique())  # Ensure valid bins
    bin_labels = ['Consistent', 'Moderate', 'Intermittent', 'Late'][:num_bins - 1]  # Labels must be one fewer than bins
    engagement_stats['Engagement_Type'] = pd.qcut(engagement_stats['std'], q=num_bins, labels=bin_labels, duplicates='drop')
else:
    engagement_stats['Engagement_Type'] = 'Consistent'  # Default if all std values are the same
```

The result of the classification revealed that 50% of the users had engaged consistently with the course, 25.37% engaged moderately with the course activities, and 24.6% engaged with the course activities intermittently as seen from the screenshot below.

```
Percentage counts per engagement type:
Engagement_Type
Consistent      50.007095
Moderate        25.372499
Intermittent    24.620406
```

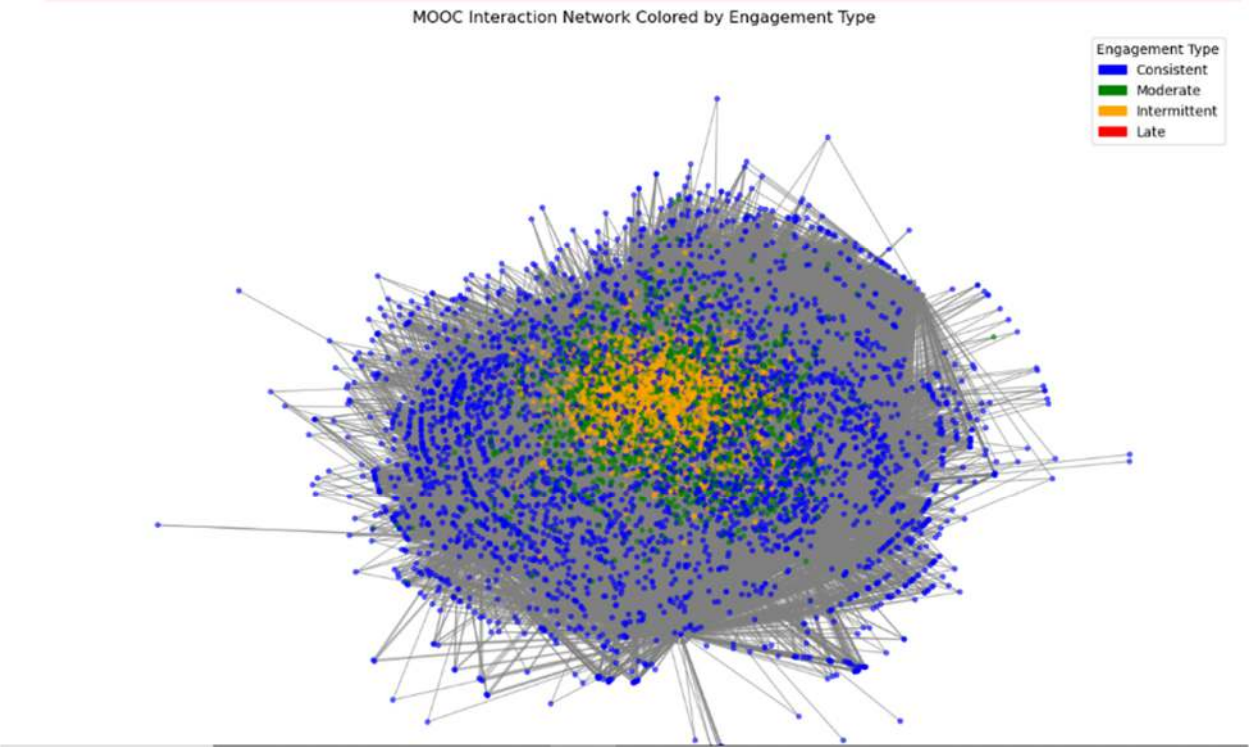
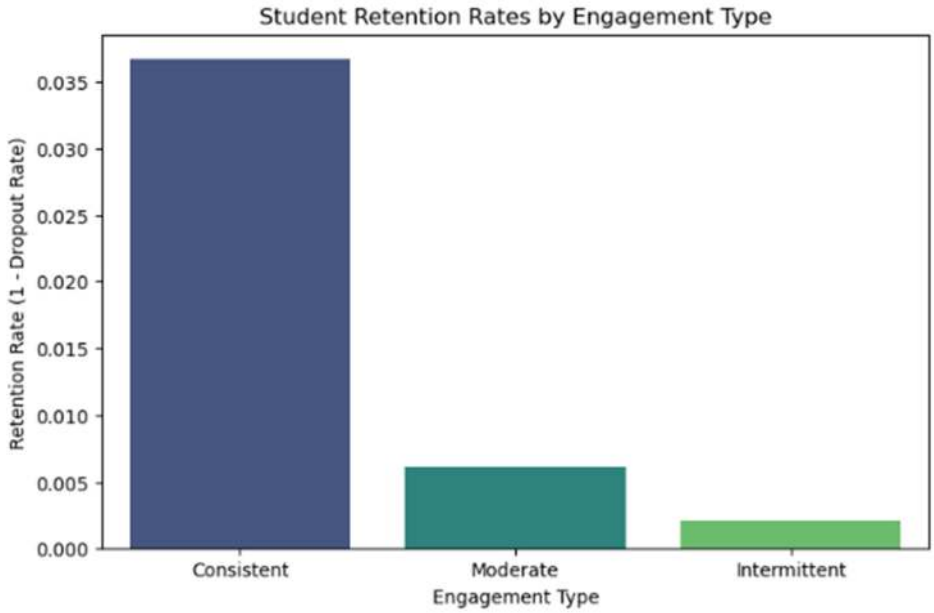


Figure 7: Network Graph of Users by Engagement Type

6. **Does student engagement classification affect retention rates?**

Finally, an attempt was made to investigate whether retention rates in the course were significantly different between students who had consistent, moderate, and intermittent levels of engagement as defined in the previous research question. A chi-square test of significance was used to explore this question at the 95% level of statistical significance. As shown in the bar chart below, the retention rates for students with Consistent, Moderate, and Intermittent engagement types are 0.0367, 0.00617, and 0.00222 respectively.




```
Retention rates for each engagement type:
 Engagement_Type
Consistent      0.036708
Moderate        0.006169
Intermittent    0.002221
Name: LABEL, dtype: float64
Chi-square test result: chi2=7127.574667904879, p-value=0.0
There is a statistically significant difference in dropout rates among Consistent, Moderate, and Intermittent engagement groups.
```

Figure 8: Chart of Retention rates by Engagement Type and Chi-Square Test of Significance Result

The result of the Chi-square test further shows that a statistically significant difference exists in the retention rate of students according to the engagement type they fall into, X2 = 7127.57, p = 0.00.