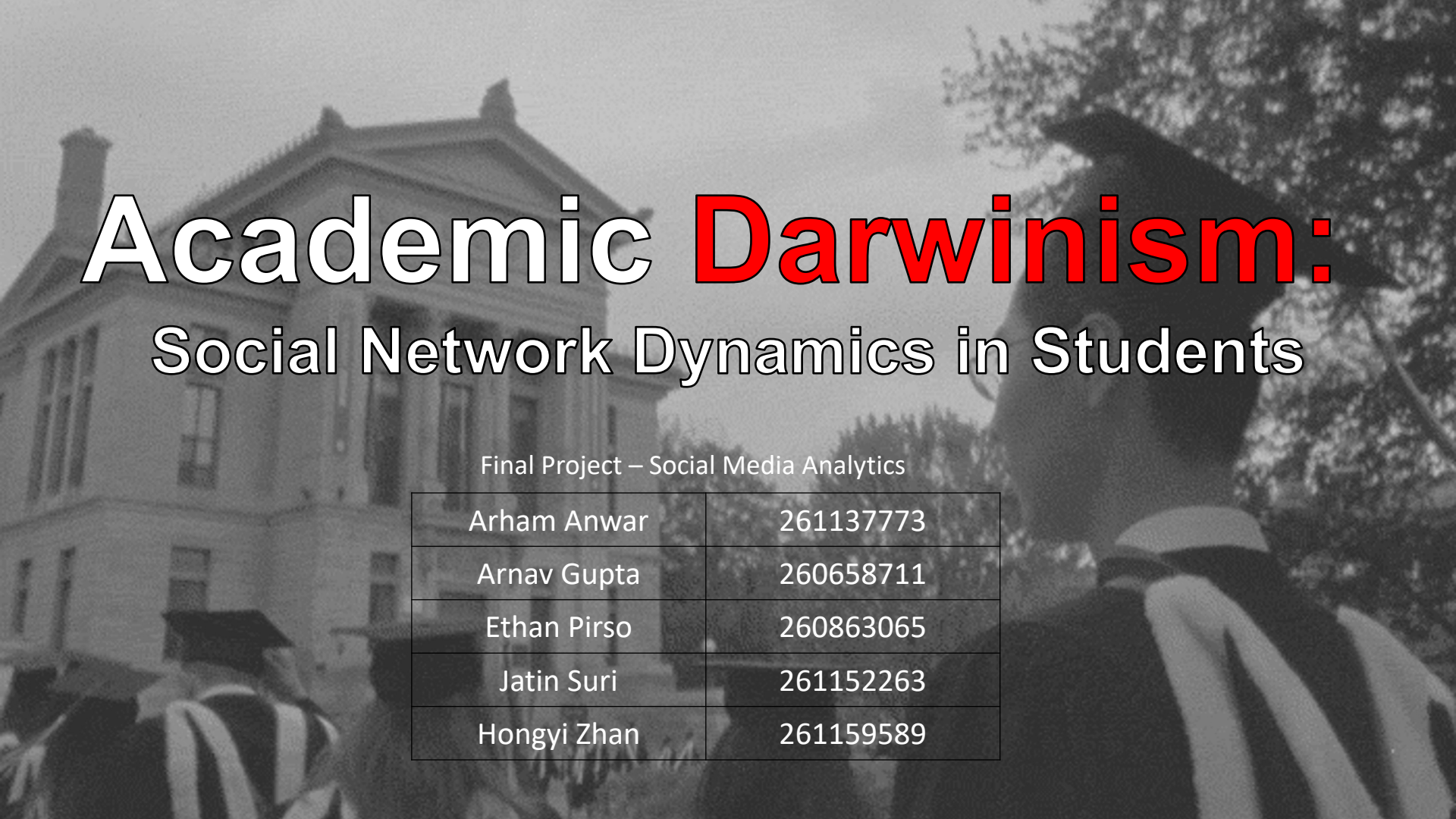


Academic Darwinism:

Social Network Dynamics in Students

Final Project – Social Media Analytics

Arham Anwar	261137773
Arnav Gupta	260658711
Ethan Pirso	260863065
Jatin Suri	261152263
Hongyi Zhan	261159589

Recap: Organic Feedback Mining

what McGill Students say behind the filter

INSY-669-076

Problem Statement

Misalignment Between McGill Actions & Student Interests

There is a stark misalignment between the actions of McGill University and the genuine interests of the students. This discrepancy leads to a lack of enthusiasm and engagement among the student body.

Lack of Organic Feedback Mining

Feedback is usually taken through the official means of course reviews and surveys which prevent unfiltered feedback to be recorded. This leads policy and decision makers misdirected leading to the misalignment

Why it matters:

To foster a more **supporting educational environment**

To enable **visibility** into university's formally uncaptured problems

To **find areas** where the university is doing well and poor

Solution Approach

By implementing an organic feedback extraction process utilizing text analytics, the project aims to capture unfiltered feedback from two streams of opinion. This will provide a comprehensive understanding of the sentiments and viewpoints of the student community.

Step 1: **Organic Data Extraction**
Reddit and Rate My Prof

Step 2: **Filter To Issues**
By creating a classifier model

Step 3: **Clustering & Sentiment Analysis**

Text Classification Model

Topic Modeling with LDA:

LDA was used for topic modeling on r/mcgill posts to categorize text and facilitate initial labeling.

Dataset Creation and Classification Model Training:

Labeled datasets were combined to train a classification model to distinguish b/w issue, non-issue text content.

Application of Classification Model:

Classification model was applied to r/mcgill's dataset as well as the RatemyProfs dataset to identify issue-related submissions

Topic Clustering

(r/McGill)

K-Means Clustering on classified data:

Used to identify distinct clusters of issues, facilitating the extraction of actionable insights on r/mcgill.

MDS Plotting:

Multidimensional scaling (MDS) visualized the relationships between different issue topics

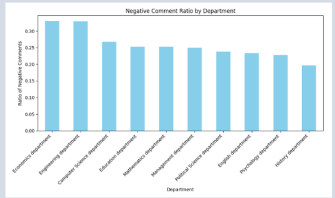
Cluster Name	Count
it-to-anyon-and-the-is-doe-know-my-but	932
you-to-your-do-the-and-of-are-if-what	833
the-to-of-in-is-and-on-for-it-are	693
remov-admiss-to-for-student-the-in-question-transfer-of	595
grade-the-my-to-transcript-for-final-and-is-on	557
for-is-thi-student-the-in-no-to-of-and	500
we-the-to-http-and-com-you-of-in-www	387
user-by-delet-remov-zoom-fit-fine-finish-first-fix	291
campu-on-near-to-place-where-the-remov-is-there	154
cap-friday-real-lock-mandatori-talk-rant-your-onli-to	58

Table A.2: Cluster names and their counts for general issues

Sentiment Analysis:

Sentiment Analysis on RateMyProfessors Data:

Employed to gauge overall sentiment & intensity of issues within each department, especially focusing on mental health issues



Top 10 Dept with highest proportion of Negative Comments on RateMy Prof

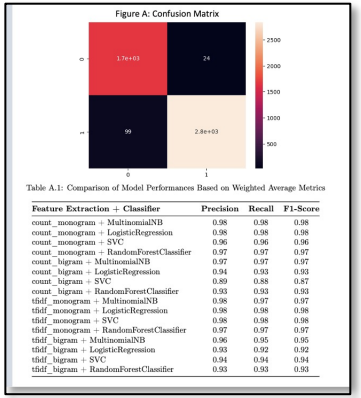


Table A.1: Comparison of Model Performances Based on Weighted Average Metrics

Feature Extraction + Classifier	Precision	Recall	F1-Score
count_monogram + MultinomialNB	0.98	0.98	0.98
count_monogram + LogisticRegression	0.98	0.98	0.98
count_monogram + SVC	0.96	0.96	0.96
count_monogram + RandomForestClassifier	0.97	0.97	0.97
count_bigram + MultinomialNB	0.97	0.97	0.97
count_bigram + LogisticRegression	0.94	0.93	0.93
count_bigram + SVC	0.89	0.88	0.87
count_bigram + RandomForestClassifier	0.93	0.93	0.93
tfidf_monogram + MultinomialNB	0.98	0.97	0.97
tfidf_monogram + LogisticRegression	0.98	0.98	0.98
tfidf_monogram + SVC	0.98	0.98	0.98
tfidf_monogram + RandomForestClassifier	0.97	0.97	0.97
tfidf_bigram + MultinomialNB	0.96	0.95	0.95
tfidf_bigram + LogisticRegression	0.93	0.92	0.92
tfidf_bigram + SVC	0.94	0.94	0.94
tfidf_bigram + RandomForestClassifier	0.93	0.93	0.93

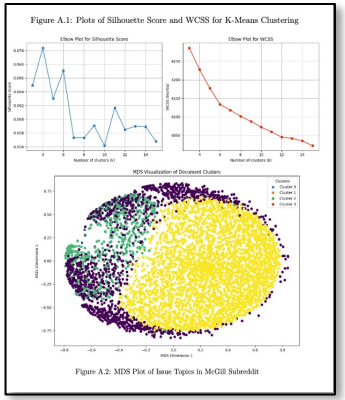
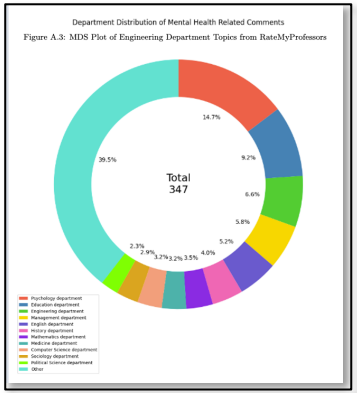


Figure A.2: MDS Plot of Issue Topics in McGill Subreddit



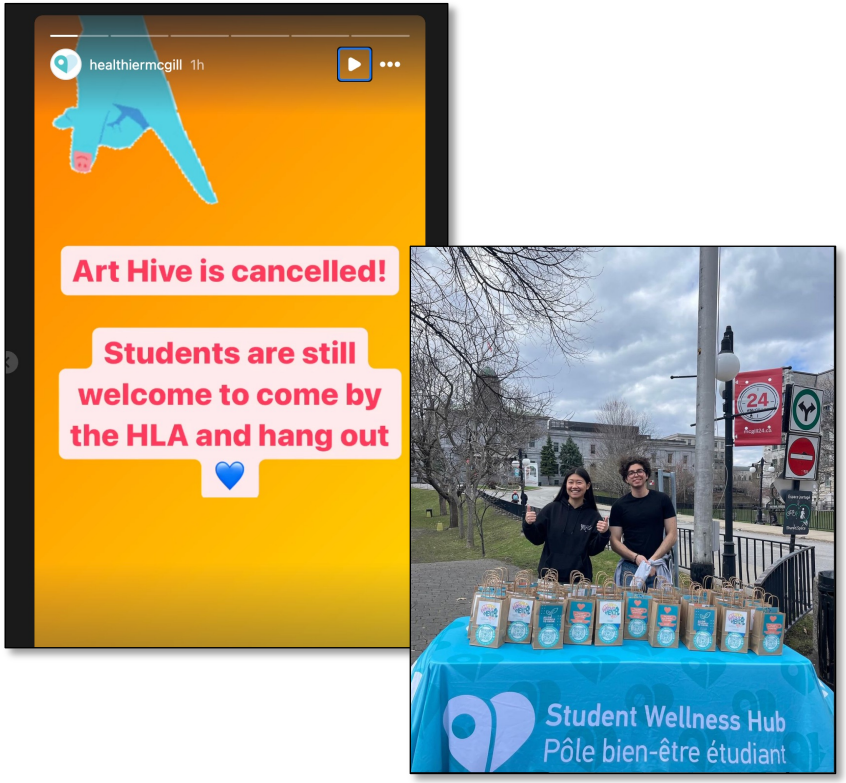
From our previous project in Text Analytics, we have observed that unmoderated online platforms like Reddit are able to provide a more candid view of student experiences and we were able to identify core issues.

Wellness Activities

Student Wellness Hub Weekly Events

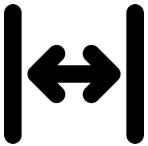
April 22- April 26

Monday, April 22nd	Tuesday, April 23rd	Wednesday, April 24th	Thursday, April 25th	Friday, April 26th
GRAD BREAKFAST CLUB 10:30AM-11:00AM ANIMAL THERAPY 1:00PM-2:30PM	ART HIVE 10:00AM-1:00PM	ANIMAL THERAPY @ MAC CAMPUS - CENTENNIAL 12:00PM-1:30PM WARM-UP WEDNESDAYS FOR ARTS STUDENTS 2:00PM-3:00PM	MAC GRAD SUPPORT GROUP 12:00PM-1:00PM ANIMAL THERAPY 1:00PM-2:30PM EXAM CARE PACKAGE GIVEAWAY 2:30 PM	



Key Takeaway: *Low turnout of wellness activities leads to events being cancelled at times*

Problem Statement



Impeded dissemination of resources hindering performance & retention

Inadequate communication strategies within student social networks impede the dissemination of support resources, hindering academic performance and retention

Social Mobility Restrictions limiting University potential

Information diversity bottlenecks due to alike staying in cliques impeding university performance potential in terms of school projects, projects and extra curricular-activites

Why it matters:

To foster a more **supporting educational environment,** optimizing dissemination of resources & team building strategies

tailored communication strategies: **bridging groups** for homophily-driven networks and engaging influencers as **service ambassadors**

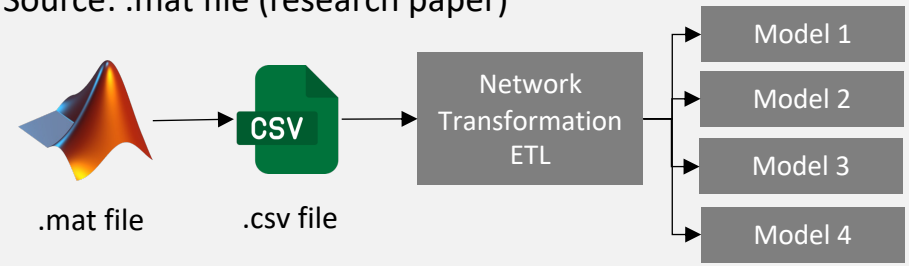
Data Strategy

Model building on Proxy data School Network Interaction Data with capability to expand to real McGill Students Network

Proxy Data Sourcing:

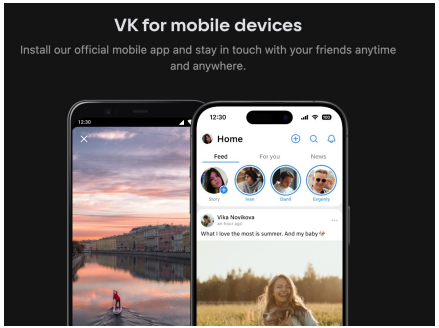
Considerations: Realism, Scale, Unfiltered/Organic

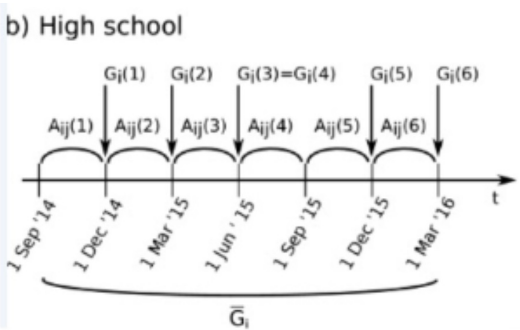
Source: .mat file (research paper)

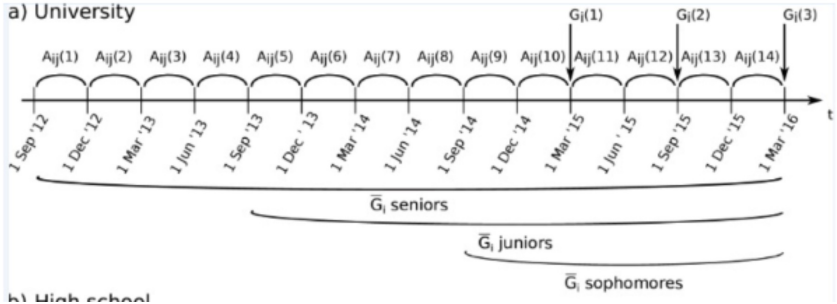


Data Description:

- Dataset containing info about the detailed evolution of a **friendship network of 6,000 students across 42 months** based on social interactions between Russian students from a public high school and a university
- The dataset contains **information on the students' academic performance at several time points** during their studies together with the detailed information about the evolution of their friendship networks







Student networks (6000 students) from VK, Russia's largest social site, formed by mutual likes within a timeframe used as proxy

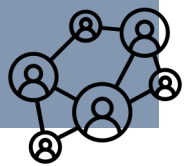
Solution Approach

We analyze student friendship networks over time, integrating social data with academic performance (GPA). Method includes - tracking network changes over time, assessing homophily, predicting GPA based on network features, & identifying student communities to find intervention points

Step 1:

Network Construction

& Time Evolution of Network



Step 2:

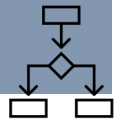
Homophily Tests

a. Reference Paper
b. Homophily Index
c. Hypothesis Testing

Step 3:

Regression & SHAP Analysis

Random forest to predict academic performance based on network attributes



Step 4:

Community-Influencer Identification

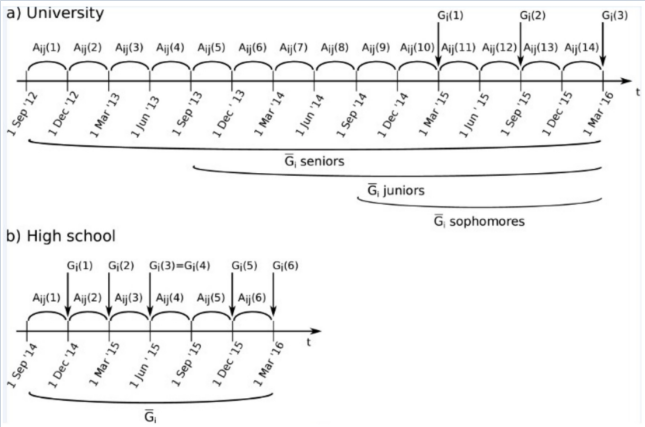
a. Greedy Algorithm to find communities
b. Hits Algorithm to find key nodes

1. Network Construction

Network Construction:

- Construct friendship networks for high school and university students
- Nodes represent students,
- Edges represent exchanged "likes”

“Friendship links approximated from "likes" on each other's pages. i.e., Link formed if at least one "like" exchanged within a timeframe.”



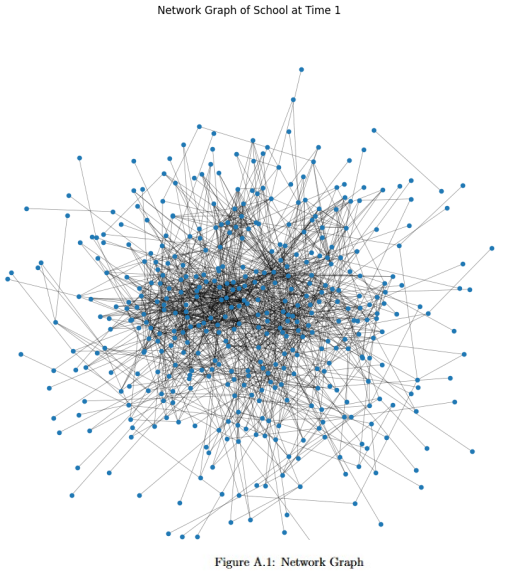
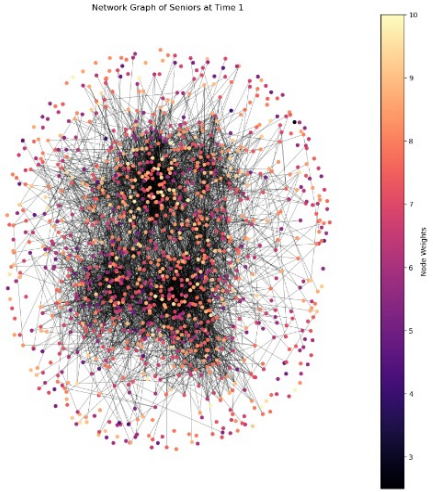


Figure A.1: Network Graph

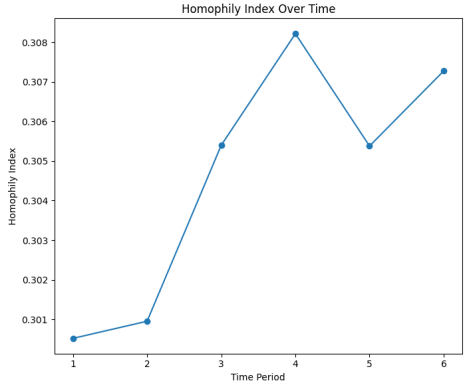


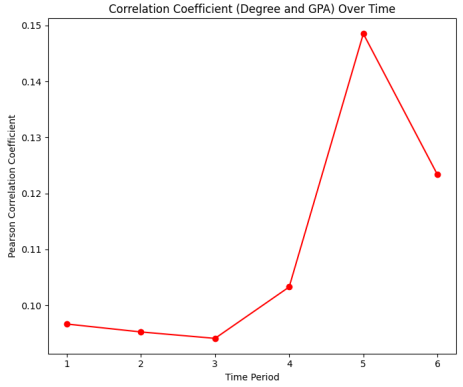
2. Homophily Testing

A. Homophily Index & Pearson Coefficient:

Quantifies the extent to which students with similar GPAs are more likely to be friends than those with dissimilar GPAs.

The Pearson correlation coefficient calculated b/w students' GPA & their degree centrality within the network. This correlation measures whether students with higher GPAs tend to have more connections (higher centrality)

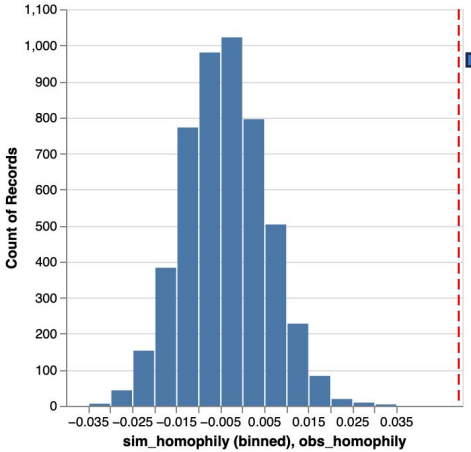




B. Homophily Hypothesis Testing:

Permute node attributes in a graph to assess structure sensitivity, focusing on mixed edges with differing attributes.

(If the fraction of cross-gender edges is significantly less than 2pq, then there is evidence of homophily)



Red line measure as the different between the expected percentage of mixed edges and the observed percentage of mixed edges.

P value <<0

2. Homophily Testing

C. Relational autocorrelation:

We used the auto correlation theorems mentioned in paper cited below and tested the homophily gain. The concept is

If a homophily effect is present in the data, the autocorrelation will increase when we consider the link changes from time t to time t + 1: (Homophily Gain)

$$C(X_t, G_{t+1}) > C(X_t, G_t)$$

Concept Reference:

Timothy La Fond and Jennifer Neville. "Randomization Tests for Distinguishing Social Influence and Homophily Effects." In: Apr. 2010, pp. 601–610. doi: 10.1145/1772690.1772752.

Autocorrelation Increase:

The increase in autocorrelation from 0.298 at time t to 0.309 at time t+1 suggests that the network became more homophilous over time. This means that nodes increasingly connected with other nodes that are similar to themselves.

Also, 2014 links being added and only 961 dissolved, suggest that the network's structure is evolving in a way that possibly increases homophily, leading to a more homogeneously connected network over time.

Parameter	Value
Autocorrelation at time t	0.2979
Autocorrelation at time $t+1$	0.3092
Number of link additions (k)	2014
Number of link dissolutions (m)	961
k equals m	False
Increase in autocorrelation	True

Table A.1: Summary of Network Dynamics Over Time

Key Takeaway: *All three tests confirm homophily dominance in network*

3. Regression

Random Forest Model:

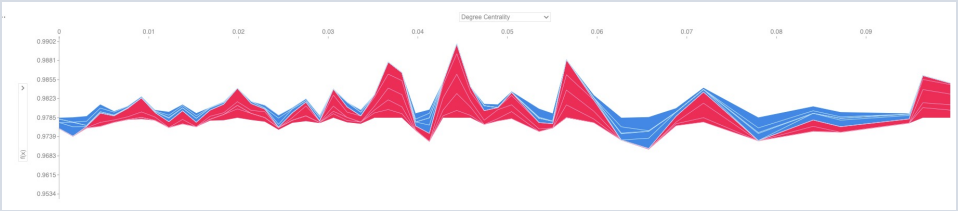
A Random Forest Regression Model was developed to predict GPA using centrality features as predictors.

Please refer figure on right for global feature importance

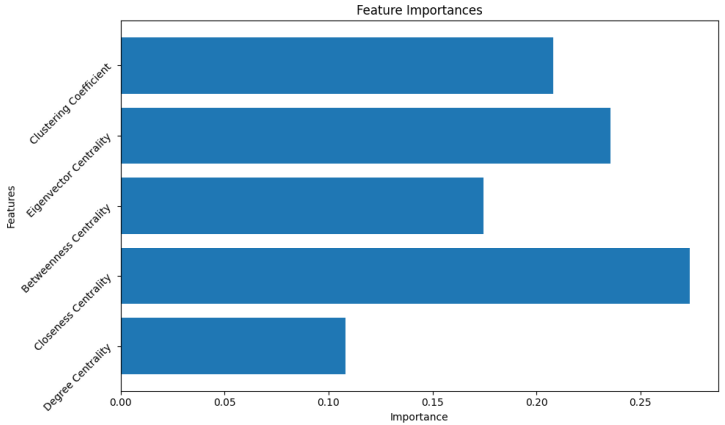
Explainability AI With SHAP:

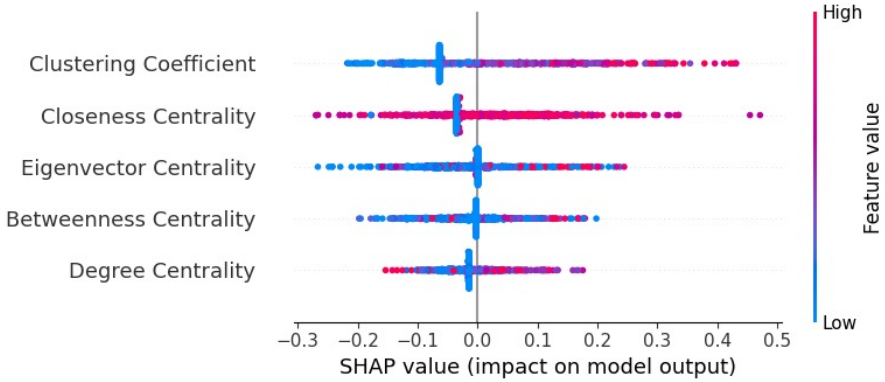
Subsequently, the Explainity AI model was scrutinized to extract valuable business insights from its SHAP values.

Please refer SHAP value plot on right



```
Mean Squared Error: 0.3262658181981827
```





Key Takeaway: *Resources can be strategically & preemptively pushed to students with low centrality scores*

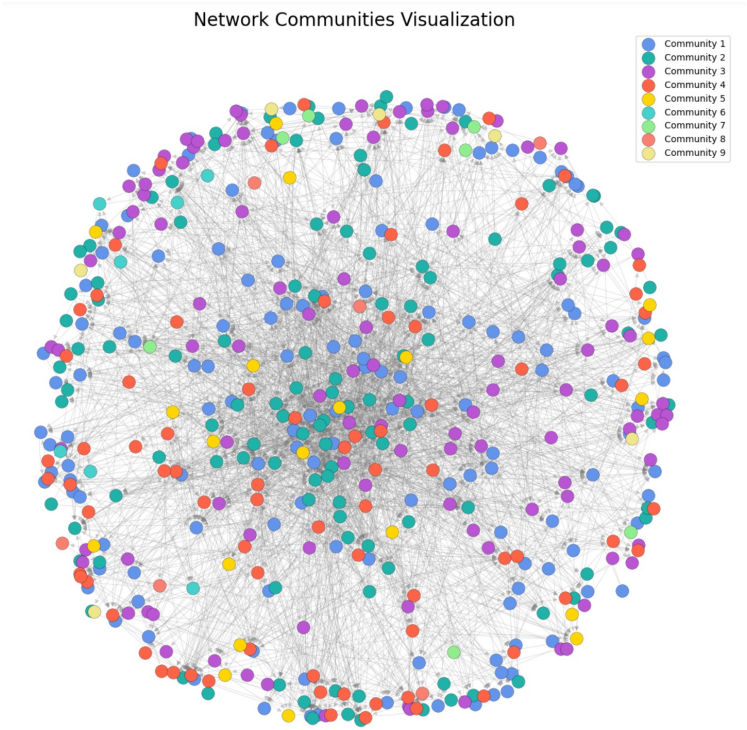
4. Community Detection

Greedy Modularity:

Greedy Modularity algorithm, designed to optimize the modularity of a network using the GPA as weights, allows us to effectively identify & analyze the community structure within the student networks over time. Results: clear delineations of student clusters

HITS Algorithm - Identifying core nodes:

Once we identified the communities, we applied the HITS algorithm to identify the core students in the homophily groups, categorized into "hubs" and "authorities." Here is the table for top for Authority notes in each community ->



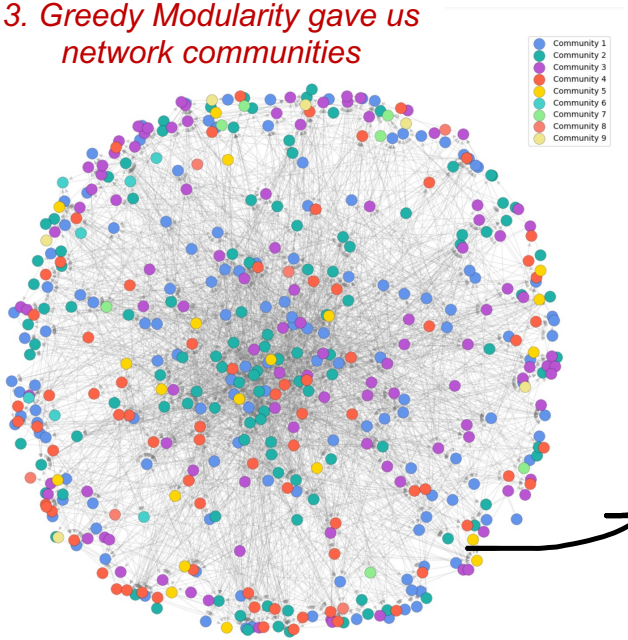
Community Number	Top Node Number	Authority Score
1	169	0.0399
2	98	0.0392
3	28	0.0470
4	33	0.0505
5	408	0.1672
6	224	0.3111
7	601	0.2822
8	412	4.8262
9	202	0.5774

1. *Started with Network Graph of School (Over Time)*

$$C(X_t, G_{t+1}) > C(X_t, G_t)$$

2. *Homophily detection*

3. *Greedy Modularity gave us network communities*

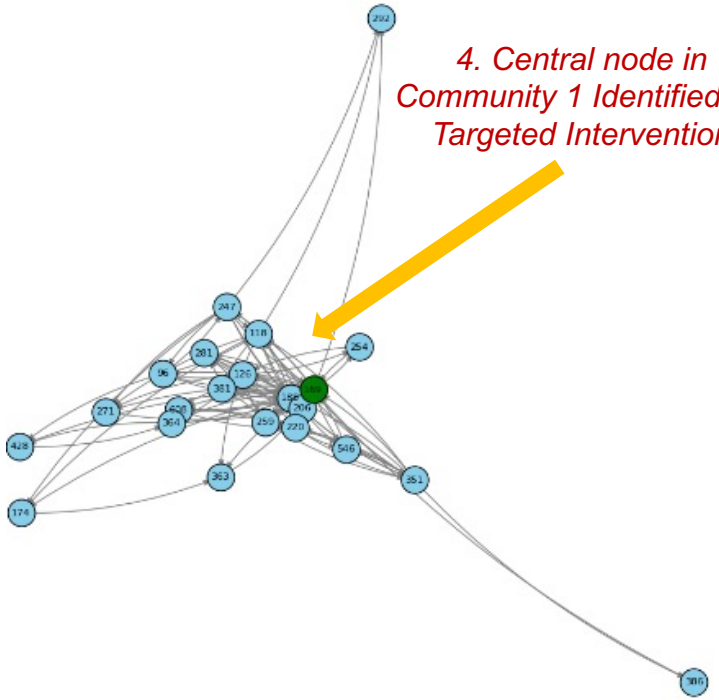


Directed Network for Node 169 with highest authority score in Community 1

Figure A.8: Authority Network for Node 169

Directed network for Node 169 with highest authority score in Community 1

4. *Central node in Community 1 Identified for Targeted Intervention*



Key Takeaway: *Information dissemination can be done through the identified core nodes*

Conclusion

Study revealed significant insights into the structural dynamics of student social networks and their influence on academic performance

Findings

Homophily Dominance:

Students prefer to gradually reorganize their social networks according to their performance levels, rather than adapting their performance to the level of their local group.

Prediction of GPA with Network Centrality (Low RSME):

Students with high network centrality attributes are scoring higher GPA; Influence & performance walking hand-in-hand

Identification of Communities & Influencers:

Communities and influencers within communities are identified which gives transparency into POCs for dissemination and targets for resource utilization drives

Actionable Insights

Tailored Communication Strategies:

Develop targeted communication strategies that consider the network's homophily characteristics to ensure information reaches all groups within the student body.

Group Formation Strategy for Increased social mobility:

Network centrality can be used to find students who need help the most; Random Grouping can be used to elevate class performance.

PoC Leads for Outreach Amplification:

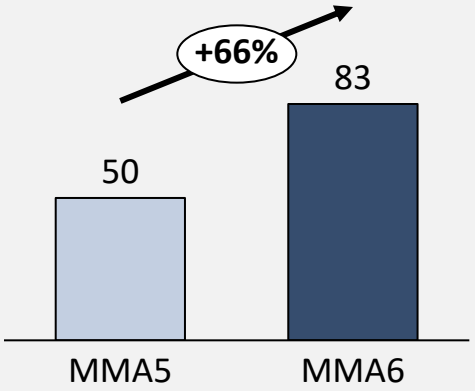
Engage identified core nodes within the homophily group networks leveraging their positions to enhance outreach and engagement.

Business Value and Expected Impact

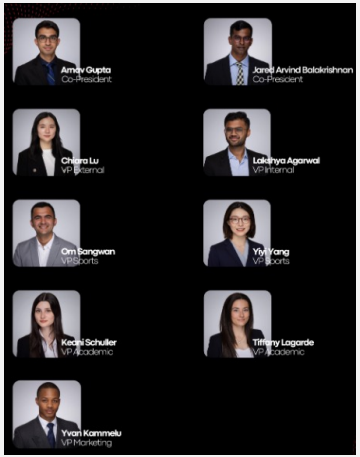
- Tailored strategies ensure inclusive communication, fostering engagement and cohesion among diverse student groups.
- Targeted support initiatives enhance academic outcomes, reducing dropouts by addressing specific student needs effectively.
- Engaging influential nodes amplifies outreach efforts, maximizing dissemination impact & promoting community involvement.

Recommendation Examples

1



66% Increase in Batch Strength
Increase Batch Size demands changes in network optimization for dissemination of resources
This year MMA6 had a council of 9 instead of 2 class reps led to a significant event participation increase!



3x

Event Participation

Thank you

Questions & Answers

Appendix

Greedy Modularity Algorithm

Modularity:

This is a metric that quantifies the quality of a division of a network into communities. High modularity indicates that there are dense connections between the nodes within communities and sparse connections between nodes in different communities.

Greedy Algorithm:

This method applies a greedy optimization approach to maximize modularity. It starts with each node in its own community and progressively merges communities that result in the greatest increase in modularity. This stepwise optimization continues until no further improvement in modularity can be achieved.

Steps:

- Start with Each Node as a Community
- Calculate Possible Gains
- Identify the pair of communities that results in the highest increase in modularity.
- Combine the two communities into a single community. Update the network structure accordingly, treating the merged entity as a single community.
- After merging the two communities, it's necessary to update the modularity change calculations for any community pair that involves the newly formed community.
- If the merge resulted in an increase in the overall modularity of the network, go back to Step 2 and consider the next set of possible merges.

Greedy Modularity Algorithm

Calculation of Modularity Change ΔQ

When considering merging two communities within a network, you calculate the change in a metric called modularity, denoted as ΔQ. This value helps to decide whether merging the two communities will lead to a more optimal community structure according to the network's connectivity.

The formula to calculate the modularity change ΔQ when considering the merger of communities i and j is:

$$\Delta Q = \left(\frac{e_{ij}}{m} - \frac{d_i d_j}{2m^2} \right)$$

Where:

- e_{ij} is the number of edges between communities i and j.
- m is the total number of edges in the entire network.
- d_i and d_j are the sum of the degrees of all the nodes in communities i and j, respectively.

Explanation of Terms

- **Edges between communities (e_{ij})**: This term counts how many connections exist between the nodes in community i and community j.
- **Total edges in the network (m)**: This is the sum of all edges in the network, which acts as a normalizing factor in the formula.
- **Sum of degrees (d_i and d_j)**: These are the sums of the connections each node in community i and community j has to any other node in the entire network. This term adjusts the expected number of edges between communities i and j if the network were randomly connected.

Greedy Modularity Algorithm

A *greedy algorithm*, which iteratively joins nodes if the move increases the new partition's modularity.

Step 1. Assign each node to a community of its own. Hence we start with N communities.

Step 2. Inspect each pair of communities connected by at least one link and compute the modularity variation obtained if we merge these two communities.

Step 3. Identify the community pairs for which ΔM is the largest and merge them. Note that modularity of a particular partition is always calculated from the full topology of the network.

Step 4. Repeat step 2 until all nodes are merged into a single community.

Step 5. Record for each step and select the partition for which the modularity is maximal.

Ref- https://www.youtube.com/watch?v=YNCpqP7q_TE&ab_channel=RabeehAbbasi

HITS Algorithm

Hyperlink Induced Topic Search (HITS):
Link Analysis algorithm.

Hubs:
Hubs are nodes with a high number of outgoing links to authorities, indicating they serve as great pointers to valuable resources.

Authorities:
Authorities are nodes with a high number of incoming links, suggesting they are valuable sources of information.

Once the communities were established, the HITS algorithm was applied to these communities to identify which nodes within them are core nodes—either as hubs or authorities.