Multimodal Network and Textual Analysis of Reddit Engagement During Exam Seasons

STUDENT MENTAL HEALTH CRISIS DEEPENS DURING EXAMS

October 30, 2024

Report: Mental Health Disrupts Studying for Most Students

Students were also more likely to feel negative emotions than positive ones while studying and completing their assignments, according to data collected by Kahoot!

By Johanna Alonso

Student's wrong results suicide prompts support call

19 November 2024

Share Save

Chris Dearden
BBC News

FEATURE

Student mental health is in crisis. Campuses are rethinking their approach

Amid massive increases in demand for care, psychologists are helping colleges and universities embrace a broader culture of well-being and better equipping faculty to support students in need

By Zara Abrams Last updated: October 12, 2022 Date created: October 1, 2022 14 min read
Vol. 53 No. 7
Print version: page 60

OUR GOAL

Understand factors driving student interaction online during critical academic periods.

Insights into mental health and stress management topics relevant to student communities

Potential applications for platform moderation, targeted content strategies.

TEXT ANALYSIS APPROACH

STEP 1

Data Preparation

STEP 2

Network Construction

STEP 3

Centrality & Visualization

STEP 4

Community Detection

STEP 5

Engagement Analysis

STEP 6

Sentiment Analysis

STEP 7

Graph Neural Network (GNN)

STEP 8

results comparaison

Step 1: Data Preparation

Data Description

- 5375 comments and 1967 submissions from Mcgill Reddits Post
- Variables:
  - Submission Category:
    - Orginal Post
      - title
      - selftext
      - author
      - timestamp
      - subreddit
      - score
  - Comments Category:
    - User reply
      - body
      - author
      - parent_id
      - timestamp
      - subreddit
      - score

- 5699 comments and 2355 submissions from Concordia Reddits post
- Variables:
  - Submission Category:
    - Orginal Post
      - title
      - selftext
      - author
      - timestamp
      - subreddit
      - score
  - Comments Category:
    - User reply
      - body
      - author
      - parent_id
      - timestamp
      - subreddit
      - score

Data Preparation

- Filtering by subreddit and exam months (April, December)
- Selecting relevant columns (e.g., author, post ID, time, text, score)
- Cleaning and normalizing text fields
- Building reply-based user interaction pairs for network construction

Step 2: Network Construction

Building User Interactions Networks

Built an interaction graph for the r/Concordia subreddit:

- Nodes = Users (2,634 total)
- Directed Edges = Replies between users (8,099 total)

Analyzed user engagement via degree distributions:

- Out-degree: Measures user activity (how often they reply). Identifies active participants.
- In-degree: Measures user influence/popularity (how often they receive replies). Highlights key users.

Key Finding:

- No isolated nodes were found. Indicates every user in the dataset participated in at least one interaction (replying or being replied to).
- Represents a fully participatory interaction network.

Built an interaction graph for the r/McGill subreddit:

- Nodes = Users (5,323 total)
- Directed Edges = Replies between users (21,095 total)

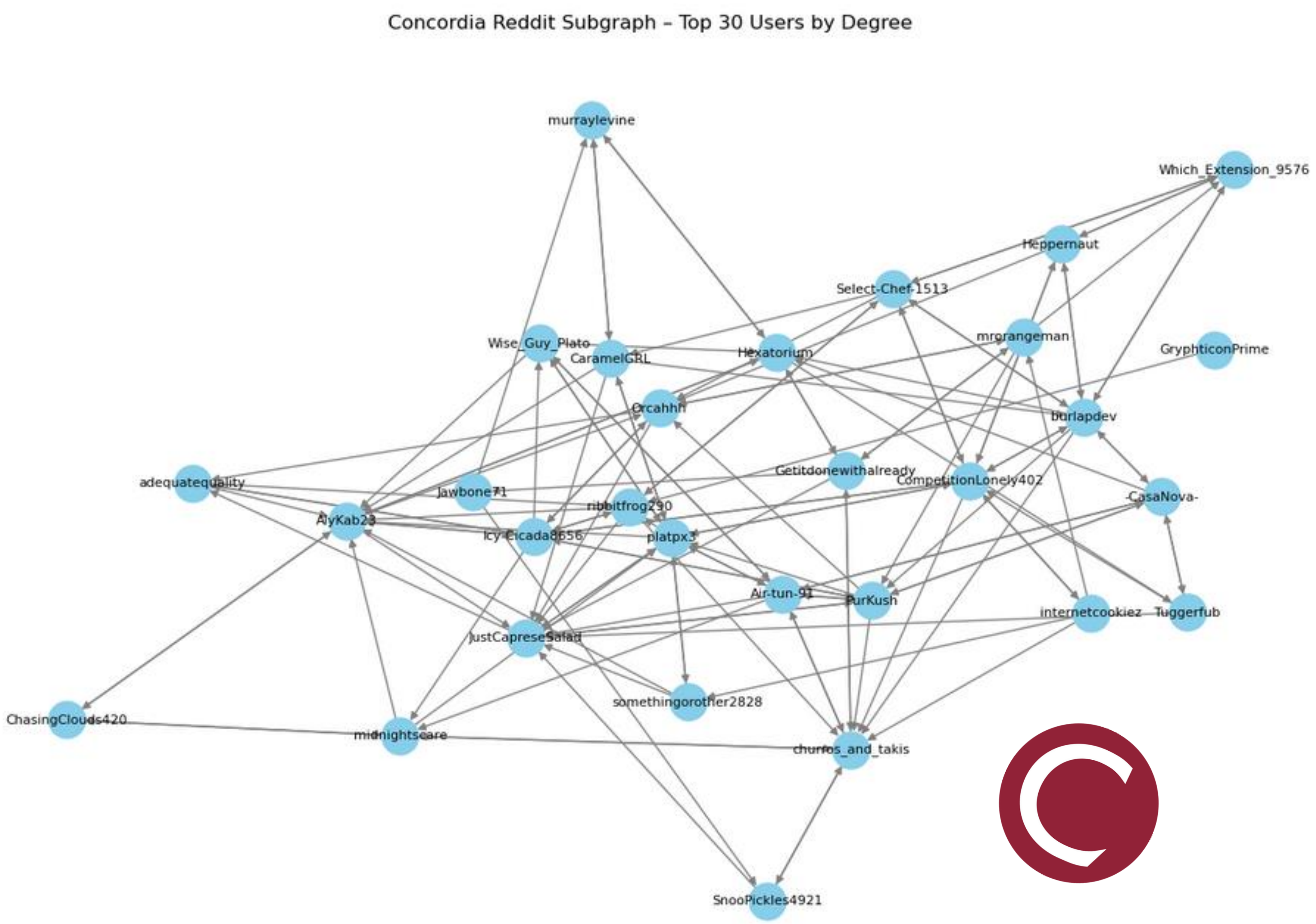
Analyzed user engagement via degree distributions:

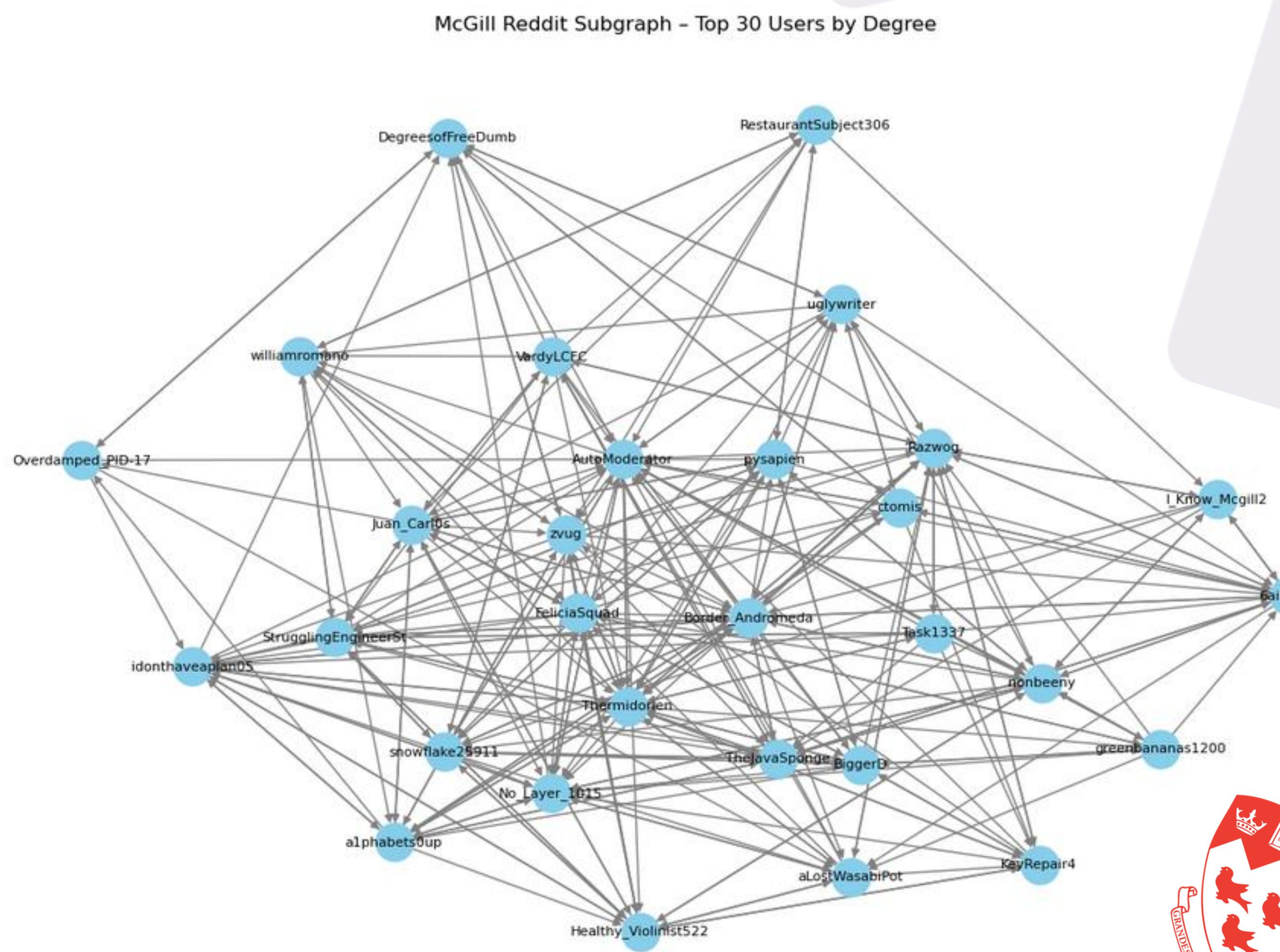
- Out-degree: Measures user activity (how often they reply). Identifies active participants.
- In-degree: Measures user influence/popularity (how often they receive replies). Highlights key users.

Key Finding:

- No isolated nodes were found.Indicates every user in the dataset participated in at least one interaction (replying or being replied to).
- Represents a fully participatory interaction network.

Step 2: Network Construction





The subgraphs visualization reveals a dense core and key users (like 'AutoModerator') acting as community hubs, which helps understand discussion flow.

Our computations

- Degree: # of direct connections (active users)
- Closeness: How close a user is to all others (reach)
- Betweenness: How often a user lies on shortest paths (bridge)
- Eigenvector: Influence based on who you’re connected to
- Katz: Influence including distant neighbors (up-and-coming users)

Step 3: Centrality & Visualization

User	Degree	Closeness	Betwenness	Eigenvector	Katz	Page Rank
PurKush	0.097227	0.252425	0.091338	0.292360	0.027978	0.007893
CompetitionLonely402	0.065325	0.231656	0.049678	0.122849	0.024583	0.004482
JustCapreseSalad	0.057349	0.264271	0.058117	0.243568	0.029663	0.009610
Air-tun-91	0.049753	0.238794	0.048296	0.172353	0.024063	0.004788
burlapdev	0.039499	0.222563	0.028982	0.078937	0.022274	0.002007
Orcahhh	0.035321	0.243511	0.032683	0.143446	0.025407	0.005595
adequatequality	0.034182	0.226971	0.022256	0.098747	0.022876	0.002881
AlyKab23	0.030384	0.243111	0.023817	0.170118	0.025347	0.004561
Hexatorium	0.029624	0.227256	0.023830	0.089879	0.023585	0.003437
Heppernaut	0.026965	0.219029	0.014918	0.066306	0.022057	0.003597

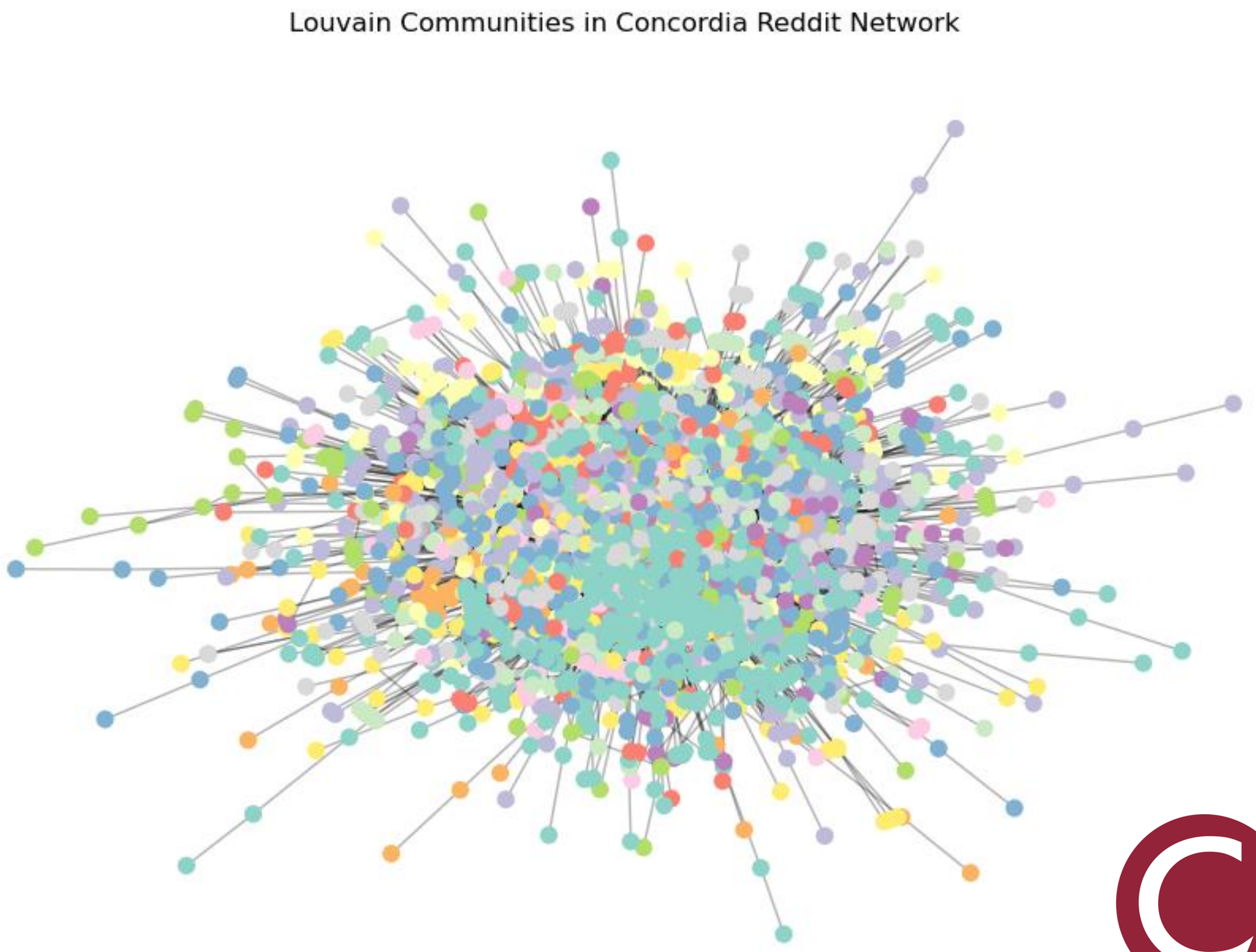
User	Degree	Closeness	Betweness	Eigenvector	Katz	Page Rank
Thermidorien	0.127208	0.310865	0.126049	0.126049	0.126049	0.126049
AutoModerator	0.111424	0.327943	0.083623	0.293152	0.043068	0.016299
snowflake25911	0.102217	0.291015	0.082139	0.204149	0.030639	0.012889
TheJavaSponge	0.060691	0.282348	0.038027	0.158340	0.022797	0.005012
Razwog	0.043593	0.279169	0.025008	0.148765	0.024079	0.004655
uglywriter	0.040962	0.275567	0.024412	0.119405	0.023358	0.004643
No_Layer_1015	0.039083	0.289991	0.019939	0.150079	0.023454	0.004044
BiggerD	0.037392	0.265339	0.018641	0.087028	0.018539	0.002362
zvug	0.036452	0.273505	0.020415	0.094027	0.018931	0.002396
nonbeeny	0.036452	0.276621	0.023678	0.112816	0.019911	0.002966

Community Detection with Louvain Method

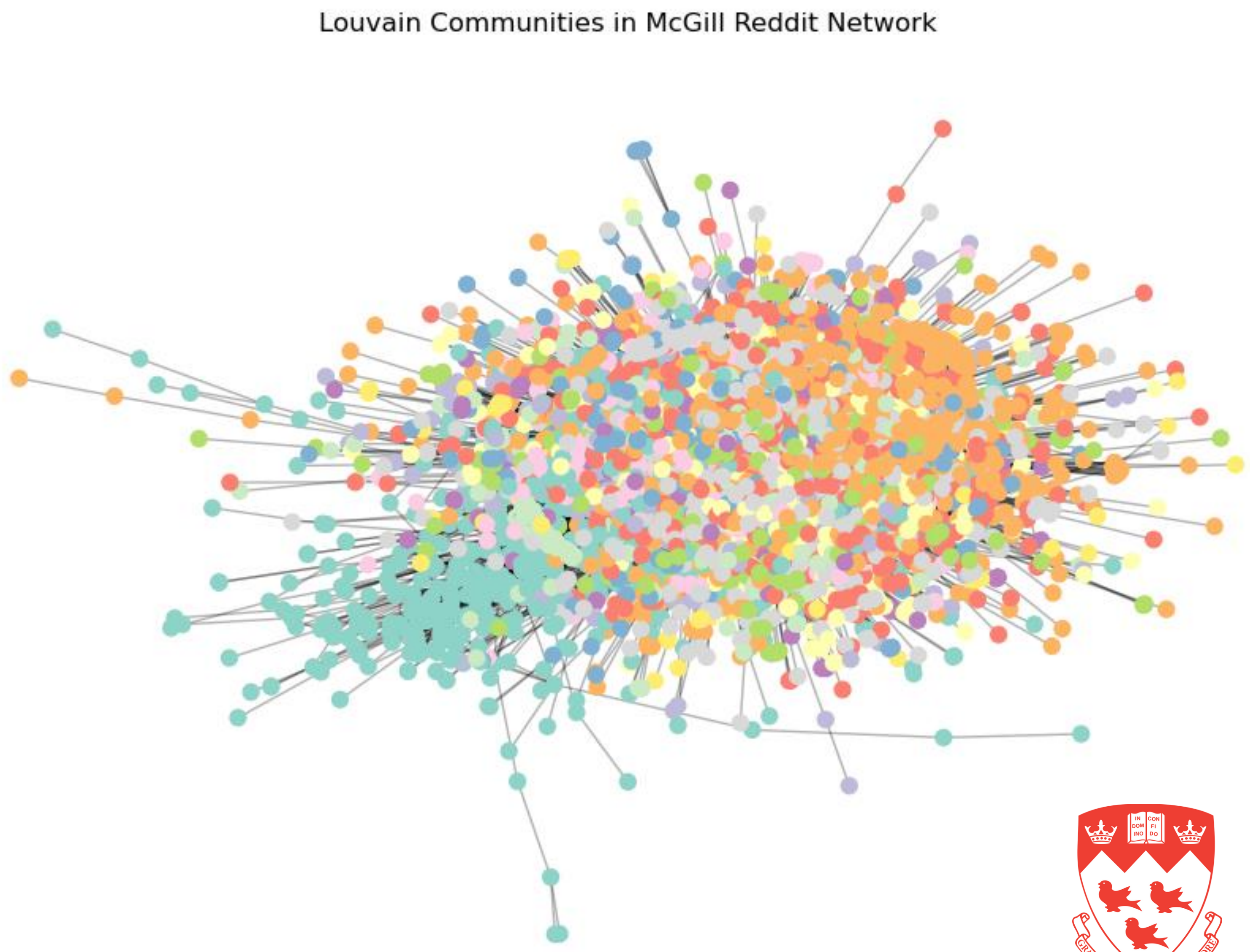
- Data Preparation:Extracted the largest weakly connected component from the interaction graph.
- Converted this component to an undirected graph (required for Louvain).
- Result: Identified 20 distinct communities.
- Interpretation: Suggests the presence of multiple distinct discussion groups or subtopic clusters within the subreddit.

Step 3: Community Detection

Detected 19 communities in concordia network



Detected 20 communities in McGill network



PRE-PROCESSING TEXT

TOKENIZE

Splits text into individual words for more granular processing, enabling the model to analyze each token's contribution to toxicity

LEMMATIZE

Reduces words to their root form, minimizing redundant variants and improving model consistency.

STOP WORDS

Stop words are commonly removed from text to filter out frequent yet insignificant words (like "the", "is", "in"), improving processing efficiency and focusing analysis on meaningful terms.

Step 4: Relating Communities to Topics & Engagement

Topic Analysis (via Post Content)

- **Overwhelming Academic Focus:** Keywords like "class," "course," "final," and "exam" dominate discussions across almost all communities, highlighting core academic activities and concerns.
- **Identifiable Subject Clusters:** Specific academic areas emerge, with communities clearly focused on Communications ("comm"), Engineering ("engr"), Math, and Economics ("econ").
- **Informal & Supportive Interaction:** Frequent use of informal terms ("im," "wa") and help-seeking keywords ("anyone," "tutor," "discord") indicates a casual, peer-support-driven environment.

- **Strong Academic Focus:** Discussions are dominated by academic keywords like "course," "exam," "final," "admission," and "program," reflecting core student concerns.
- **Specialized Community Interests:** Distinct topics identify specialized communities focusing on areas such as Admissions/Programs (C2), Computer Science (C5), Math (C8), and Housing/Rent (C1).
- **Help-Seeking & Informal Tone:** Frequent use of terms like "anyone," "help," and "im" points to a supportive, conversational environment where students actively seek advice.
- **Contextual Themes Evident:** Keywords also reveal specific contextual discussions, such as COVID-related academic concerns ("covid" - C13) and general support needs.

Step 4: Relating Communities to Topics & Engagement

Engagement Analysis (via Scores)

Average post score by community:

Community 0: 6.25
Community 1: 5.13
Community 2: 4.64
Community 3: 8.78
Community 4: 5.98
Community 5: 6.46
Community 6: 7.97
Community 7: 3.30
Community 8: 11.74
Community 9: 3.92
Community 10: 3.56
Community 11: 12.37
Community 12: 3.87
Community 13: 9.61
Community 14: 10.82
Community 15: 12.48
Community 16: 2.77
Community 17: 6.39
Community 18: 11.18

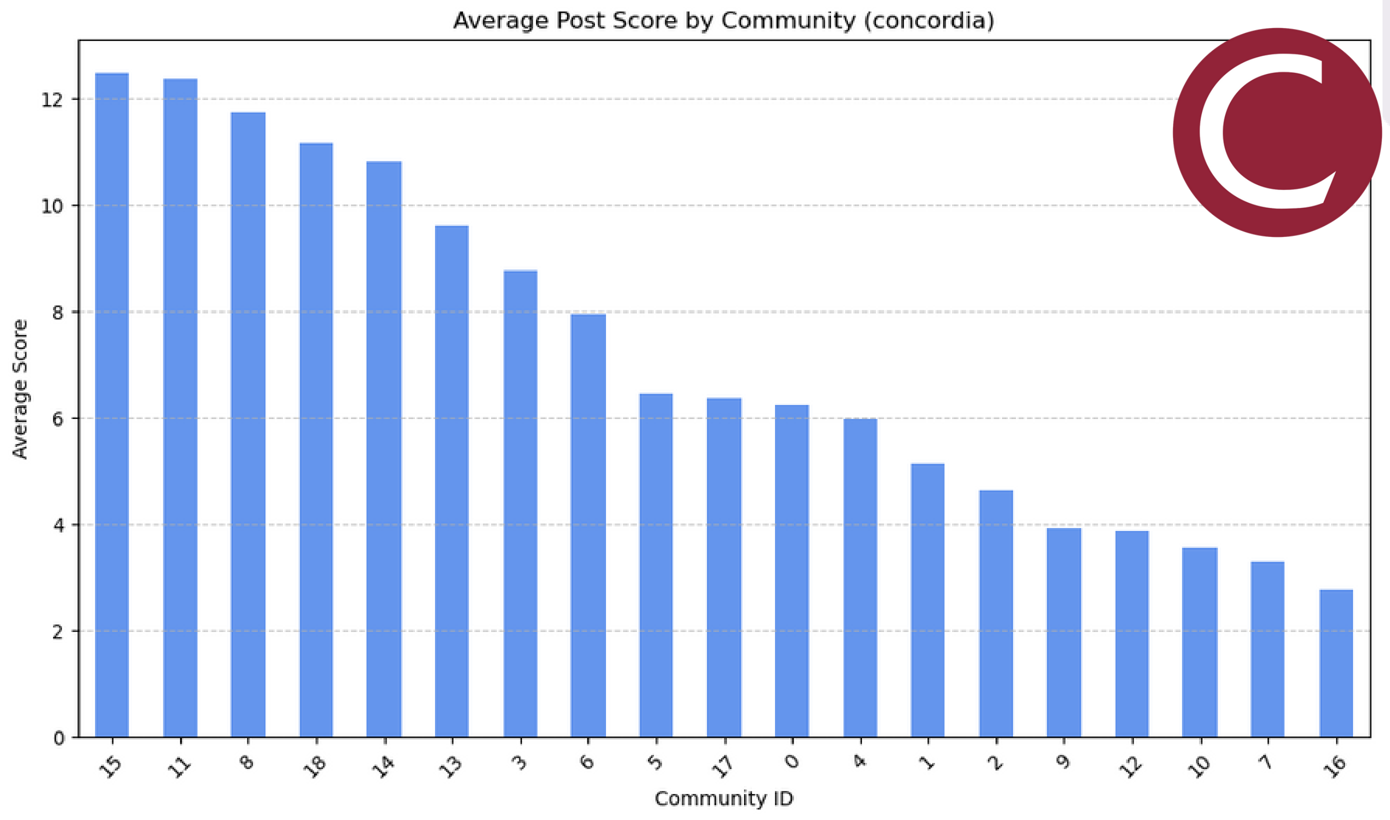
Average post score by community:

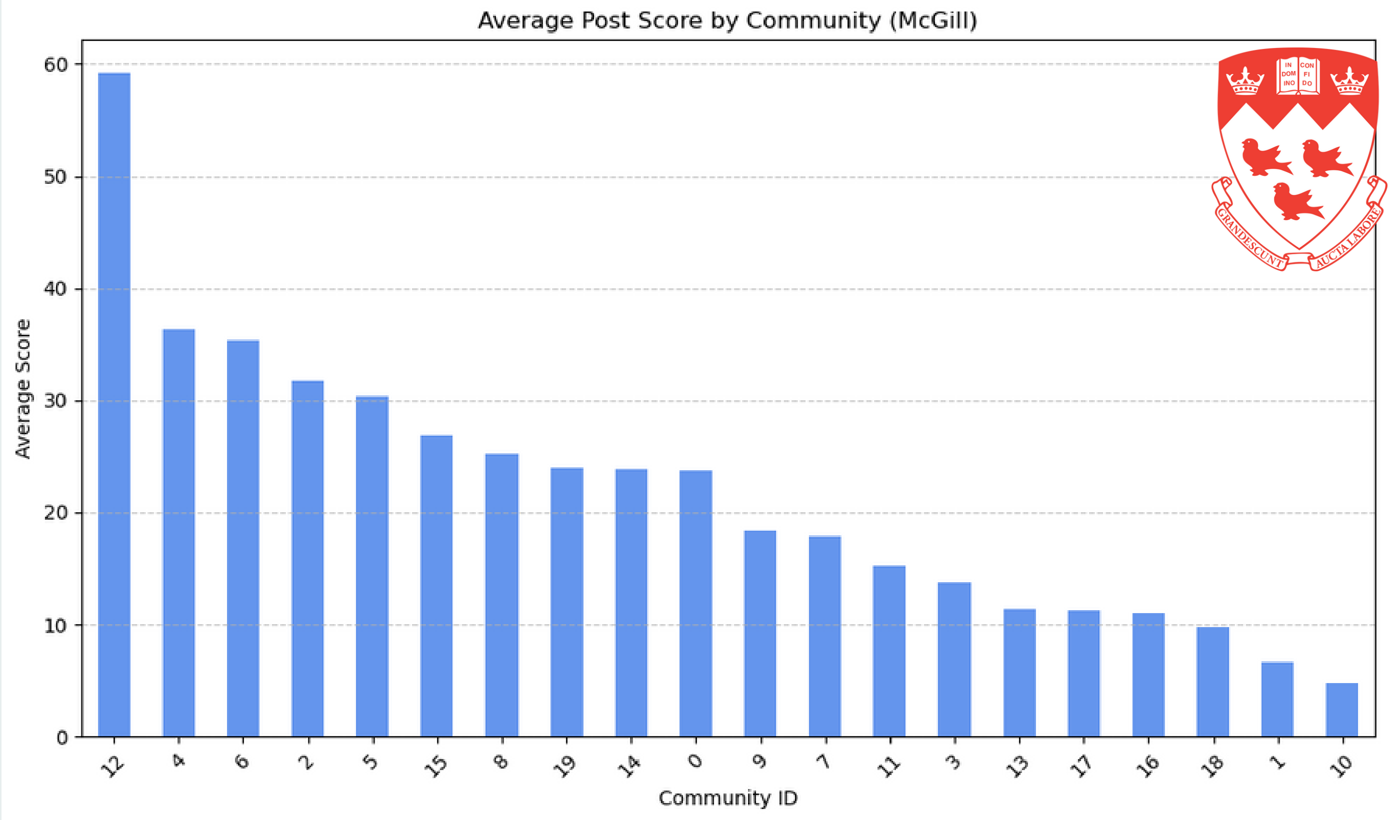
Community 0: 23.71
Community 1: 6.62
Community 2: 31.67
Community 3: 13.73
Community 4: 36.35
Community 5: 30.32
Community 6: 35.36
Community 7: 17.85
Community 8: 25.27
Community 9: 18.40
Community 10: 4.80
Community 11: 15.28
Community 12: 59.17
Community 13: 11.43
Community 14: 23.93
Community 15: 26.92
Community 16: 11.06
Community 17: 11.24
Community 18: 9.81
Community 19: 24.00

Step 4: Relating Communities to Topics & Engagement

Average Engagement per Community

For the Concordia dataset, Community 10 shows the highest average post score, indicating strong engagement, contrasting with lower scores in communities like 15 and 9. This variation likely reflects differences in topic relevance, writing style, or the overall sentiment within each community's discussions.



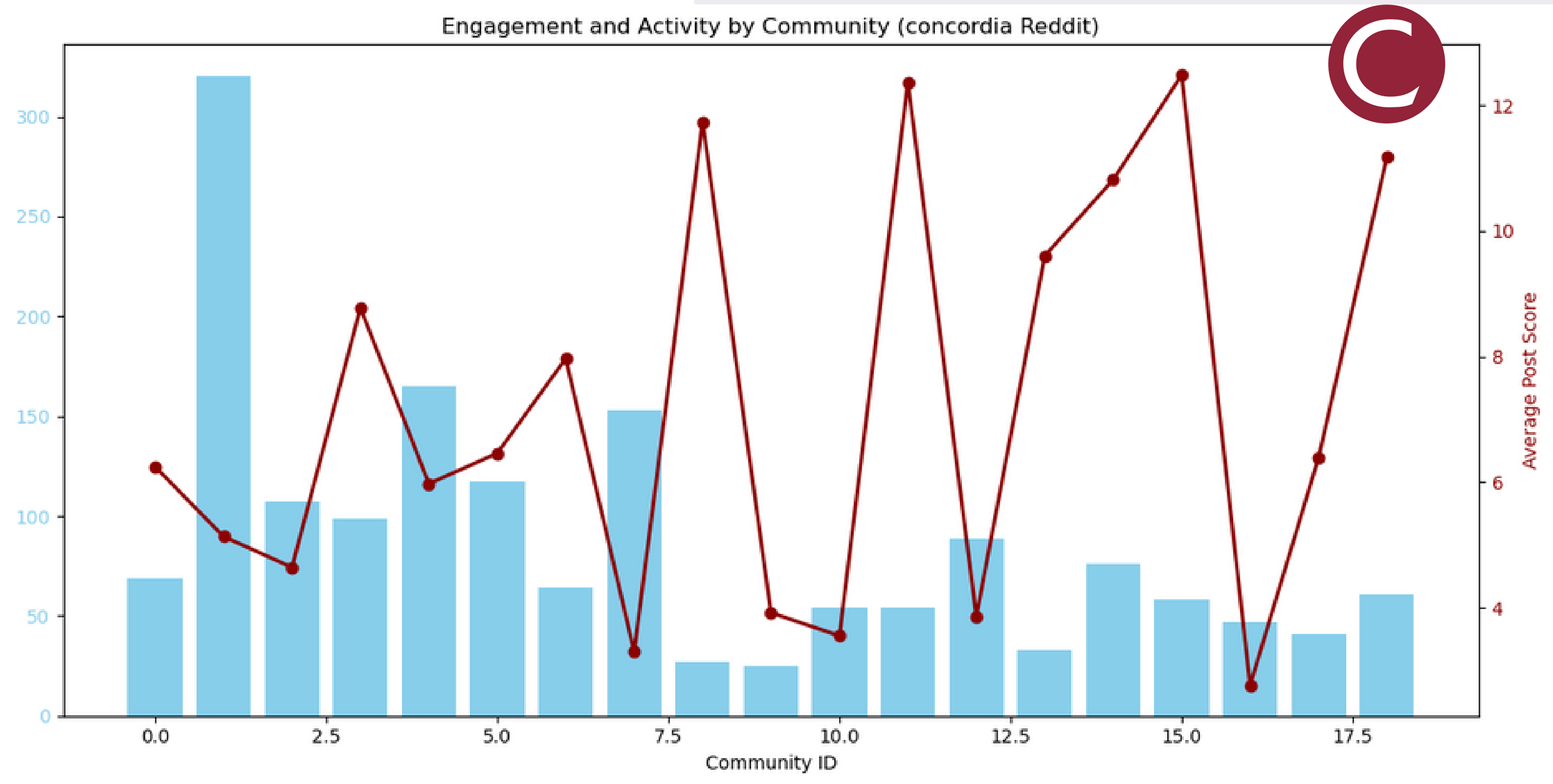


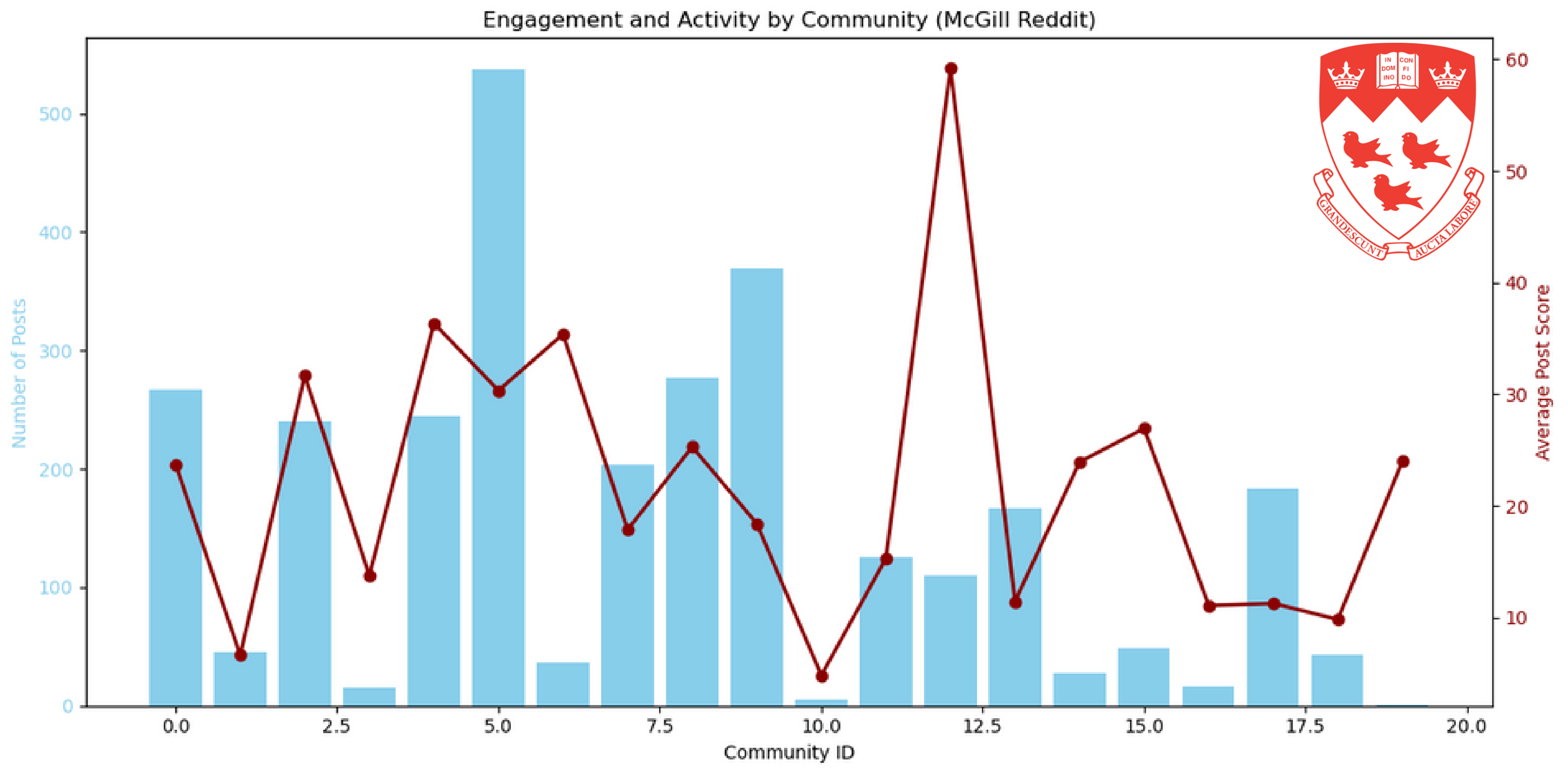
The bar chart reveals significant differences in average post engagement, highlighting communities 16 & 18 as highly resonant, unlike lower-scoring groups (e.g., 2, 12, 9). These findings indicate varying content appeal and provide direction for investigating the specific themes and dynamics within each community.

Step 4: Relating Communities to Topics & Engagement

Overlay Avg Score & Number of Posts

For Concordia communities, high activity (e.g., Community 17, 12) doesn't align with the highest engagement, achieved by lower-volume Community 10. This suggests engagement isn't tied to post volume; smaller communities can create content with higher average impact or resonance.



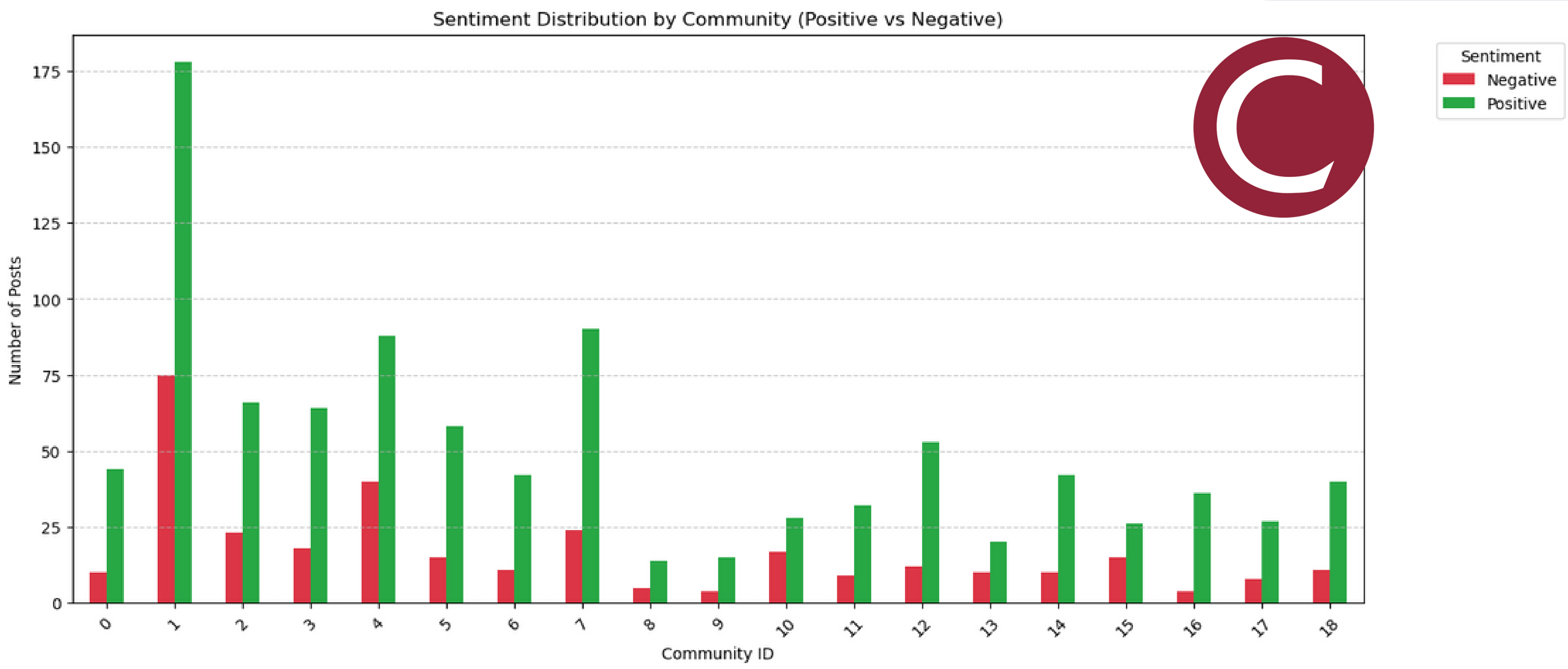


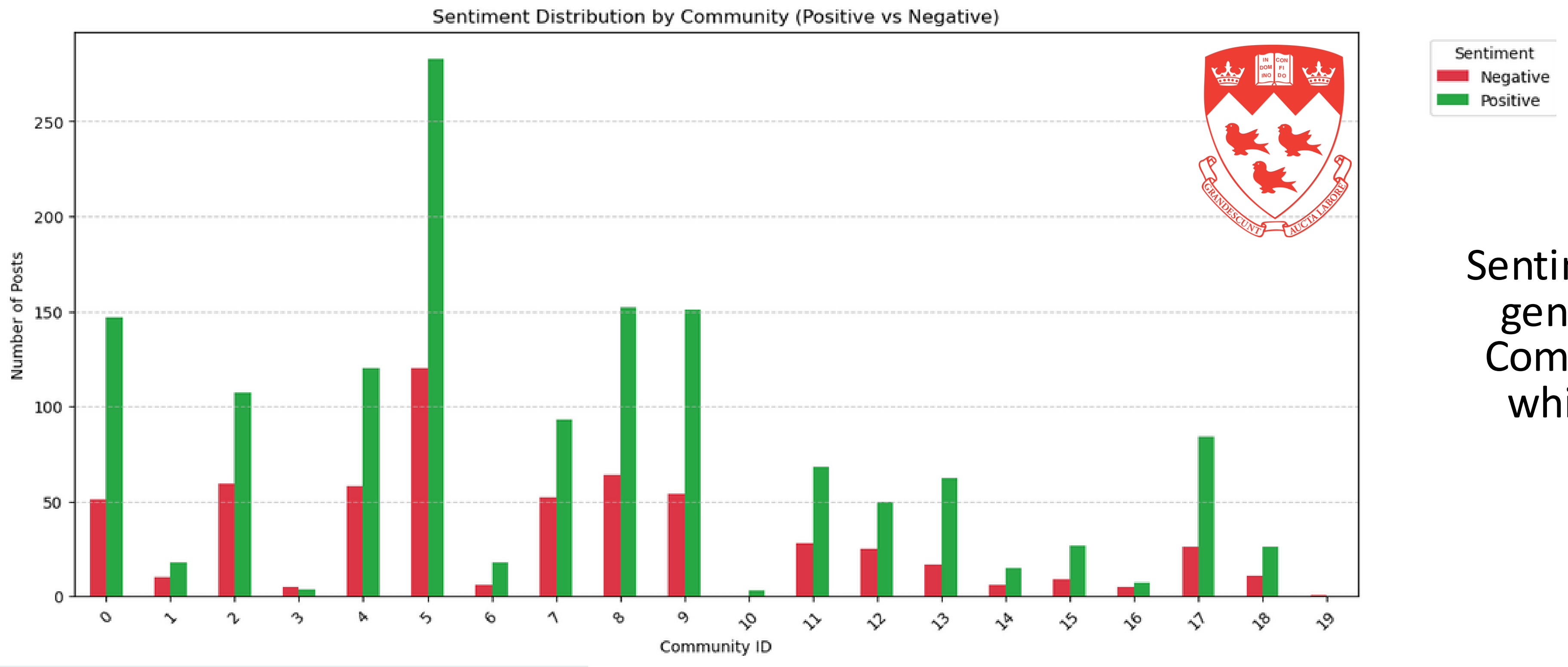
This chart shows some highly engaging McGill communities (e.g., 16, 18, 4) have low post volumes, while high-volume communities (e.g., 0, 3) show moderate engagement. This highlights that post activity doesn't always equate to impact; smaller communities can foster high-quality discussions.

Step 4: Textual Analysis

Sentiment Analysis

Positive sentiment strongly predominates across Concordia communities, with high positive post volumes notably in C17, C2, C12, and C15.





Sentiment analysis shows positive posts generally outnumber negative ones; Community 3 has the highest volume, while 0, 4, & 10 are also active with positive skew.

Step 4: Textual Analysis

Topic Modeling

Topic 1	math	course	transfer	final	comp	class	econ	elective	jmsb	im
Topic 2	prof	moodle	rplace	mid	pm	free	covid	teacher	arent	final
Topic 3	gpa	grade	course	final	class	got	cgpa	semester	internal	retake
Topic 4	tutor	fee	insurance	tip	mark	survey	form	pay	private	asap
Topic 5	coen	le	grey	nun	guy	concordia	online	aware	final	note
Topic 6	poll	housing	view	association	application	campus	link	response	discord	dorm
Topic 7	course	summer	elective	engineering	program	student	im	concordia	coop	class
Topic 8	class	im	exam	know	student	course	like	concordia	semester	time
Topic 9	final	comm	exam	engr	grade	pas	phys	thought	midterm	acco
Topic 10	removed	permit	study	caq	student	disc	international	application	letter	university

- **Topic modeling on Concordia posts identified 10 key themes:** reflecting student focus areas and administrative interactions.
- **Academic life is central:** themes cover course/program specifics (Math, JMSB, Engineering), performance monitoring (GPA, grades, exams), and interactions with instructors/platforms (profs, Moodle).
- **Practical aspects are also prominent:** including student services/finances (tutoring, fees), campus life (housing, associations), co-op, and specific administrative processes (study permits, CAQ applications for international students).

Topic 1	summer	course	im	mcgill	student	research	job	experience	minor	psyc
Topic 2	comp	removed	monday	transfer	news	course	allowed	student	good	class
Topic 3	removed	gym	biol	mcgill	final	hour	library	engineering	fieldhouse	room
Topic 4	removed	final	mcgill	mgcr	admission	econ	application	psyc	anat	exam
Topic 5	math	removed	course	science	major	art	final	chem	computer	faculty
Topic 6	final	removed	grade	thought	course	class	exam	semester	taken	thanks
Topic 7	concordia	removed	mcgill	protest	history	logo	rplace	positive	english	dm
Topic 8	removed	phgy	mcgill	account	aid	bursary	textbook	financial	money	lease
Topic 9	removed	cloudberry	mycourses	calculator	housing	membership	cat	aria	mcgill	dose
Topic 10	exam	im	final	like	dont	class	time	course	know	feel

- **Topic modeling (LDA) on McGill posts identified 10 key discussion themes:** Topics range from specific courses (psych, comp sci, econ) and academic administration (transfers, admissions) to campus facilities (gym, library).
- **Core student experiences are heavily represented:** Major themes include navigating courses/exams , managing finances and finding resources .
- **Broader campus dynamics and sentiment are visible:** Discussions also cover student life, emotional responses , and potentially sensitive/moderated content ("removed" appears frequently).

Model Selection

Predictive Modeling Objective

The models aim to predict whether a user would achieve high engagement, defined as having an above-median average post score, using features derived from network structure, text sentiment, engagement history, and topic profiles.

- **Input Features: A multimodal feature set including:**
  - **Network Centrality Metrics (degree, closeness, betweenness, eigenvector, Katz).**
  - **User Average Sentiment Score (VADER-derived).**
  - **User Average Historical Post Score.**
  - **User Average Topic Distribution (LDA-derived).**

Option 1: GCN

- Efficiently utilizes graph topology (user connections, neighborhoods) to directly learn complex interaction patterns.
- Automatically creates powerful user representations (embeddings) by combining node and neighbor information for predictive tasks.

Option 2: GAT

- Employs attention mechanisms to selectively aggregate features by learning neighbor importance.
- Generates potentially more powerful node embeddings by dynamically focusing on relevant interactions.

Option 3: Graph Sage

- Designed for inductive learning: Efficiently generates embeddings for new users or data without full retraining.
- Scales to large graphs by sampling neighbors and learning universal aggregation functions.

Step 7: Graph Neural Networks

GCN

EPOCHS

Epoch 0, Loss: 0.7979
Epoch 10, Loss: 0.6443
Epoch 20, Loss: 0.6320
Epoch 30, Loss: 0.6230
Epoch 40, Loss: 0.6146
Epoch 50, Loss: 0.6046
Epoch 60, Loss: 0.5938
Epoch 70, Loss: 0.5822
Epoch 80, Loss: 0.5716
Epoch 90, Loss: 0.5626

Accuracy :0.5994

EPOCHS

Epoch 0, Loss: 0.8290
Epoch 10, Loss: 0.6457
Epoch 20, Loss: 0.6063
Epoch 30, Loss: 0.5733
Epoch 40, Loss: 0.5500
Epoch 50, Loss: 0.5316
Epoch 60, Loss: 0.5177
Epoch 70, Loss: 0.5069
Epoch 80, Loss: 0.4988
Epoch 90, Loss: 0.4919

Accuracy :0.6885

Step 7: Graph Neural Networks

GAT

EPOCHS	EPOCHS
Epoch 0, Loss: 0.7119	Epoch 0, Loss: 1.1276
Epoch 10, Loss: 0.6176	Epoch 10, Loss: 0.6231
Epoch 20, Loss: 0.5782	Epoch 20, Loss: 0.5603
Epoch 30, Loss: 0.5287	Epoch 30, Loss: 0.5307
Epoch 40, Loss: 0.4675	Epoch 40, Loss: 0.5138
Epoch 50, Loss: 0.4024	Epoch 50, Loss: 0.4978
Epoch 60, Loss: 0.3430	Epoch 60, Loss: 0.4816
Epoch 70, Loss: 0.2861	Epoch 70, Loss: 0.4623
Epoch 80, Loss: 0.2429	Epoch 80, Loss: 0.4389
Epoch 90, Loss: 0.2048	Epoch 90, Loss: 0.4139
Accuracy :0.5577	**Accuracy :0.6758**

Step 7: Graph Neural Networks

Graph Sage

EPOCHS	EPOCHS
Epoch 0, Loss: 0.6997	Epoch 0, Loss: 0.9494
Epoch 10, Loss: 0.6099	Epoch 10, Loss: 0.5759
Epoch 20, Loss: 0.5757	Epoch 20, Loss: 0.5044
Epoch 30, Loss: 0.5371	Epoch 30, Loss: 0.4782
Epoch 40, Loss: 0.4935	Epoch 40, Loss: 0.4608
Epoch 50, Loss: 0.4504	Epoch 50, Loss: 0.4440
Epoch 60, Loss: 0.4103	Epoch 60, Loss: 0.4287
Epoch 70, Loss: 0.3745	Epoch 70, Loss: 0.4148
Epoch 80, Loss: 0.3397	Epoch 80, Loss: 0.4022
Epoch 90, Loss: 0.3114	Epoch 90, Loss: 0.3909
Accuracy :0.5321	**Accuracy :0.7140**

Comparison of Model Results

McGill (r/mcgill)

- Network Context: 5,323 users, 21,095 interactions.
- Optimal Model: GraphSAGE.
- Highest Test Accuracy: ~71.4%.
- Interpretation: The combined features demonstrated significant predictive capability for user engagement. GraphSAGE's effectiveness suggests that localized neighborhood features are strong indicators within this larger network.

VS

Concordia (r/Concordia)

- Network Context: 2,634 users, 8,099 interactions.
- Optimal Model: GCN.
- Highest Test Accuracy: ~59.9%.
- Interpretation: The features provided moderate predictive power. GCN's relative success may indicate that global graph structure offers slightly more signal in this smaller network, although overall predictability was lower.

- Performance Discrepancy: A notable performance gap exists, with the optimal McGill model achieving >11% higher accuracy than the optimal Concordia model (~71.4% vs. ~59.9%).
- Differing Optimal Models: GraphSAGE excelled for McGill, while GCN was best for Concordia. This highlights that the ideal GNN architecture can be context-dependent.

Potential Reasons for the Performance Difference

- Network Scale & Structure: McGill's larger, denser network may offer more stable engagement patterns, differing structurally from Concordia's smaller network.

- Model Architecture Suitability: GraphSAGE's local sampling likely suits McGill's larger scale, while GCN's global view may be marginally better for Concordia's smaller graph.

- Feature Informativeness: The selected features (centrality, sentiment, history, topics) appear to correlate more strongly with engagement (post scores) at McGill than at Concordia.

- Data Characteristics: Subtle differences in data distributions (e.g., scores, topics) or noise levels could impact model performance differently between the datasets.

CONCLUSION

- Multimodal Analysis Success: This project successfully integrated network analysis, NLP (sentiment, topic modeling), and Graph Neural Networks (GNNs) to provide a comprehensive view of student engagement on r/McGill and r/Concordia during exam periods.
- Key Insights Uncovered:
  - Distinct user communities and dominant discussion themes (largely academic, plus support, finance, specific courses) were identified in both subreddits.
  - Sentiment remained generally positive despite the high-stress exam context.
  - User engagement (average post score) varied significantly across communities and did not directly correlate with post volume.
- Predictive Modeling Potential & Variability:
  - GNNs demonstrated the feasibility of predicting high-engagement users using combined network, text, and historical features.
  - Predictive accuracy varied significantly between platforms (McGill ~71.4%, Concordia ~59.9%), highlighting the influence of network characteristics (scale, structure) and feature relevance in specific community contexts.

The study offers valuable insights into student online interactions during critical academic times, with potential applications for understanding student well-being, community moderation, and the context-dependent nature of social network analysis.

TH@#K YOU