Master of Computer Science Thesis

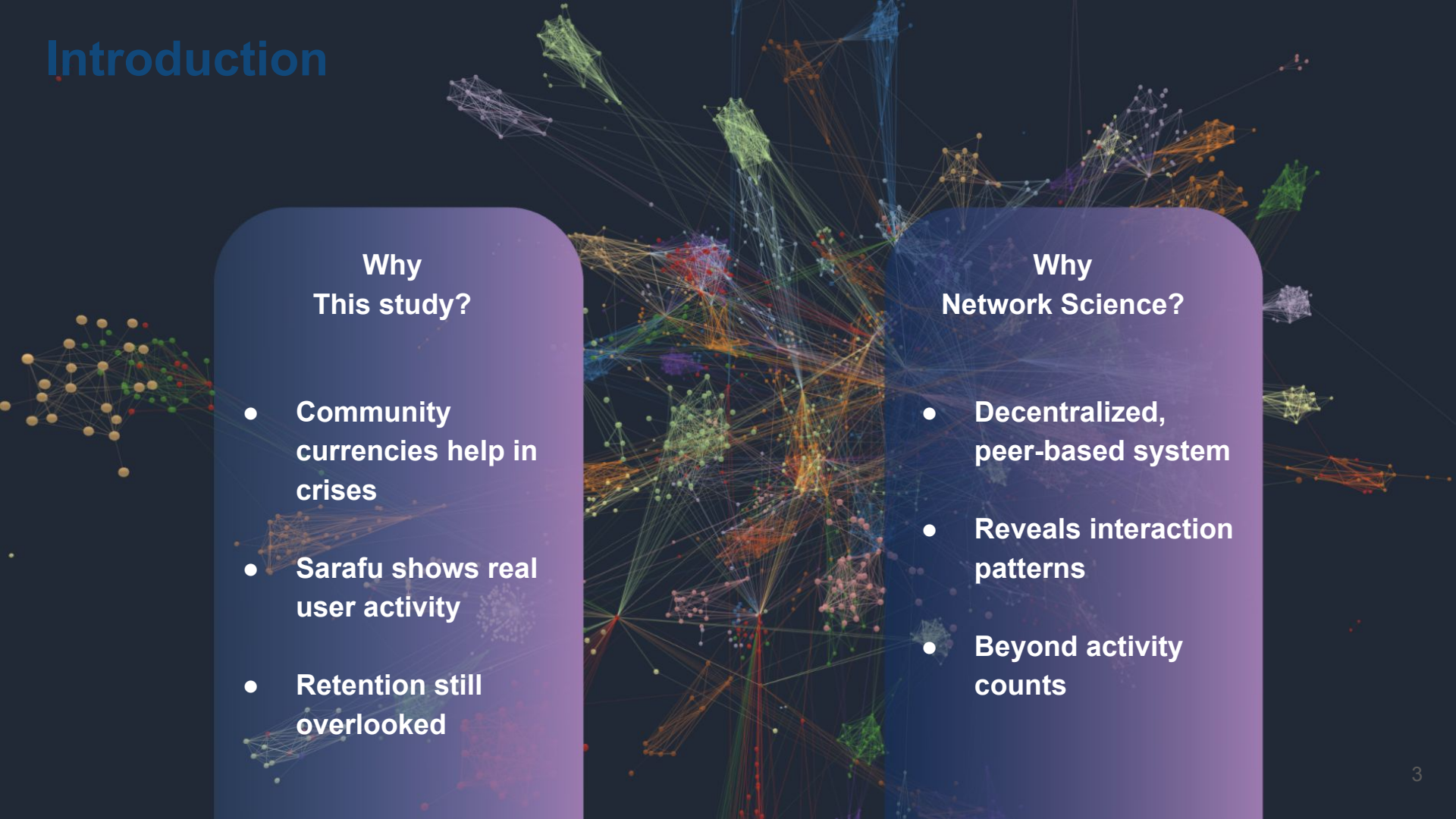
From Global Network Attributes to Temporal Motifs : Predicting User Retention in the Sarafu Network

Supervisors :
Dr. Frank Takes
Dr. Carolina Mattsson

Student : Sadaf Esmaeili Rad

Table of Contents

Introduction



Why This study?

- Community currencies help in crises
- Sarafu shows real user activity
- Retention still overlooked

Why Network Science?

- Decentralized, peer-based system
- Reveals interaction patterns
- Beyond activity counts

Research Questions

Main RQ : **What kinds of features in a community currency network contribute most to user retention?**

1. How do global centrality measures compare to community-level centralities in predicting user retention?
2. Which static structural patterns and demographic attributes are helpful in identifying retained users?
3. What dynamic network aspects such as temporal motifs are useful for predicting user retention?

Dataset Overview

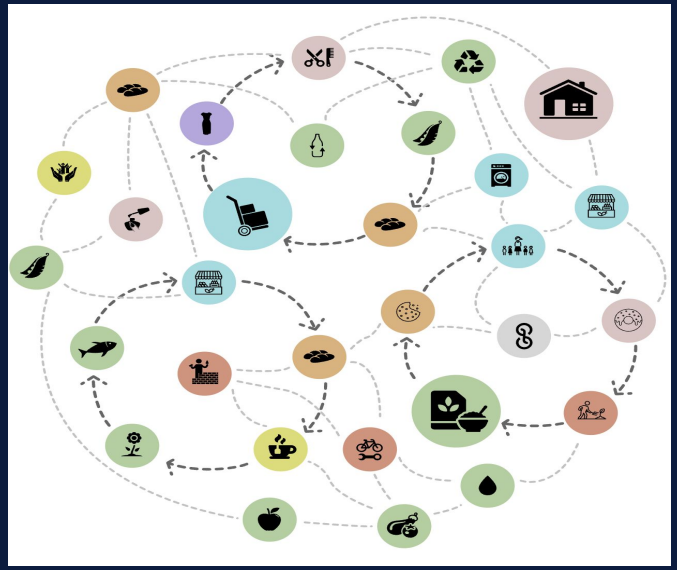
❖ Sarafu

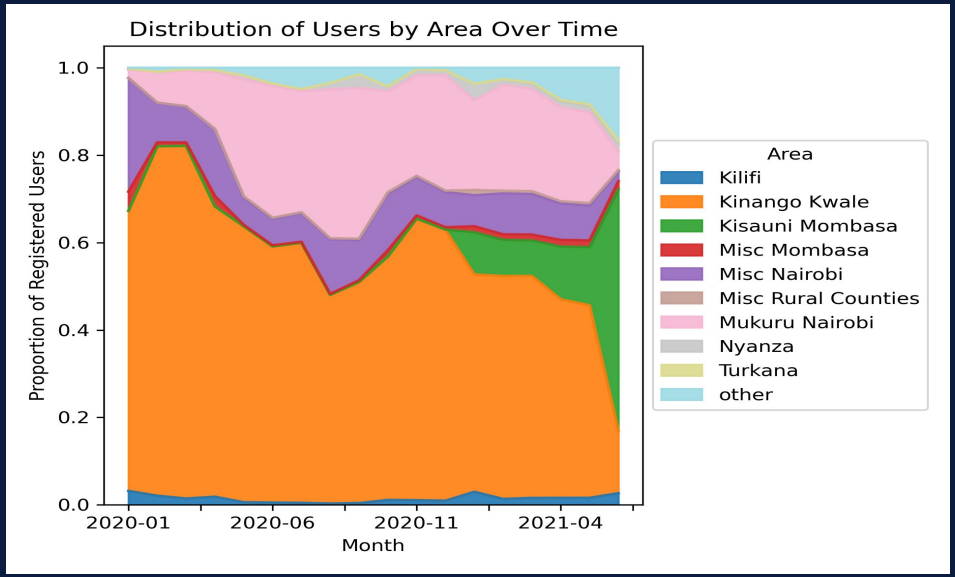
💱 **Transactions:** 422,721 peer-to-peer transfers

👥 **Users:** 40,767 unique users

🕒 **Period:** January 2020 – June 2021

🚻📍 Includes user-level demographics





Methodology

1. Preprocessing
2. Network Construction and Retention Labels
3. Feature Selection
4. Predictive Models

Preprocessing

1. Keep only STANDARD transactions
2. Remove system accounts
3. Exclude 2021 first joiners
4. Require at least one send
5. Drop receive-only users

Final Data Set for Network Analysis : 30,689 users

Network Construction and Descriptives

Directed Graph G = (V, C) from 2020 Sarafu Transactions

Nodes (V): Each node represents a unique user

Edges (C): A directed edge from user *i* to *j* means *i* sent tokens to *j*

Weights: Edges are weighted by total volume transferred in 2020

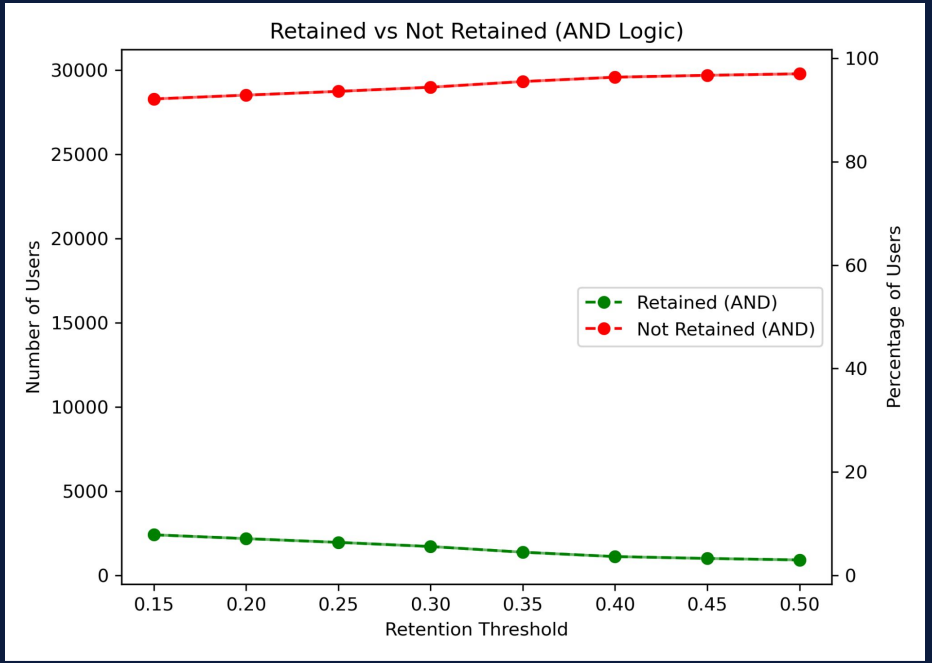
Metric	Value
Number of Nodes	30,689
Number of Edges	118,092
Density	0.000121
Reciprocity	0.4867
Largest SCC size	15,172
Average Shortest Path	2.99

Defining Retention Labels

Retention defined as:

Score =
0.5 × Count + 0.3 × Months + 0.2 × Volume

Retained if:
Score_2021 ≥ threshold × Score_2020



Feature Selection

⚙️ **Macro-level**

- In/Out-degree
- Weighted degrees
- Clustering
- Betweenness
- ⟳ Cyclic Patterns

↓

\+

👥 **Meso-level**

Community Size

↓

△ **Micro-level**

- 3-Node count
- Reciprocal count
- Completion time

↓

👤 **Individual-level**

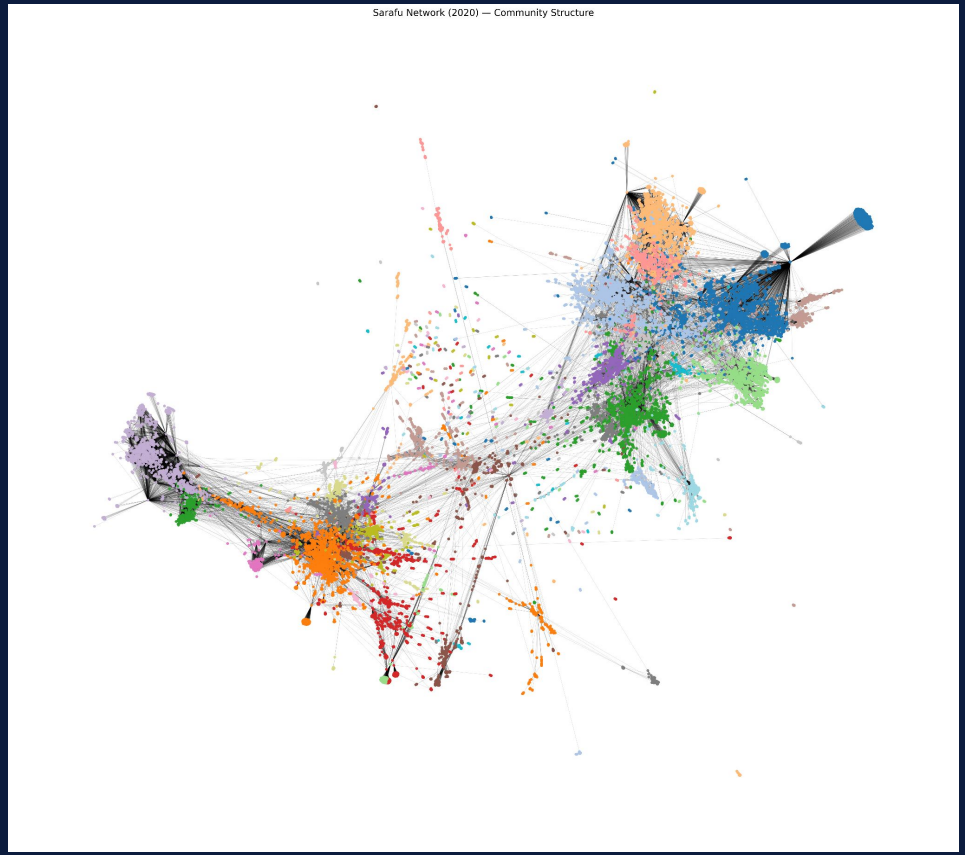
- Gender
- Area
- User role
- Business type

Community Detection

Leiden Algorithm

50 Iterations → Consensus Labeling

Metric	Value
Number of communities	317
Average community size	1,859
Minimum community size	2
Maximum community size	3,629



Cyclic Pattern Detection

Cyclic

SCC ≥ 2 → Binary 1

Acyclic

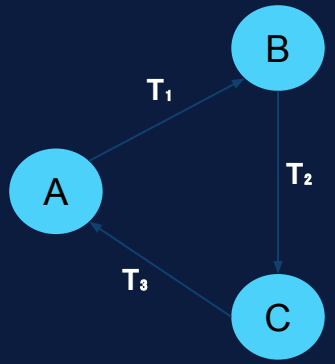
Not in SCC → Binary 0

Motifs Detection

Exhaustive enumeration of all directed subgraphs in 2020 network,

focused on two key motifs:

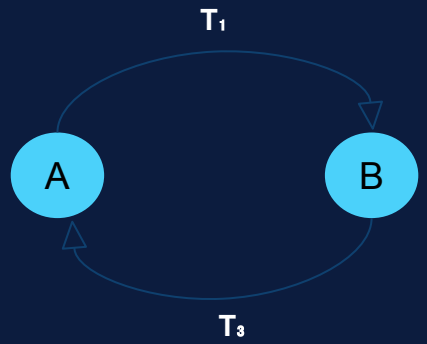
3-Node cycles



$T_1 < T_2 < T_3$

Avg. closure time: $T_3 - T_1$

Reciprocal



$T_1 < T_2$

Avg. response time: $T_2 - T_1$

Predictive Models

I. Global centralities on full network

II. Community-based centralities and size

III. Global and community features combined

IV. Adds motifs and cyclic status

V. Adds demographic attributes

Classifier ➡ XGBoost

Dataset split:

70% training, 15% validation, 15% test

Split	Users	Edges	Nodes
Train	21,483	304,544	29,684
Validation	4,603	89,872	13,334
Test	4,604	96,562	14,324

Predictive Models

Handling Class Imbalance

Step 1: Resample training set → SMOTE-ENN (val/test unchanged)

Step 2: Scale loss → retained errors count more

Step 3: Optimize threshold → maximize F1 from precision–recall curve

Predictive Models

Hyperparameter Optimization

Optuna → Bayesian search (30 trials)

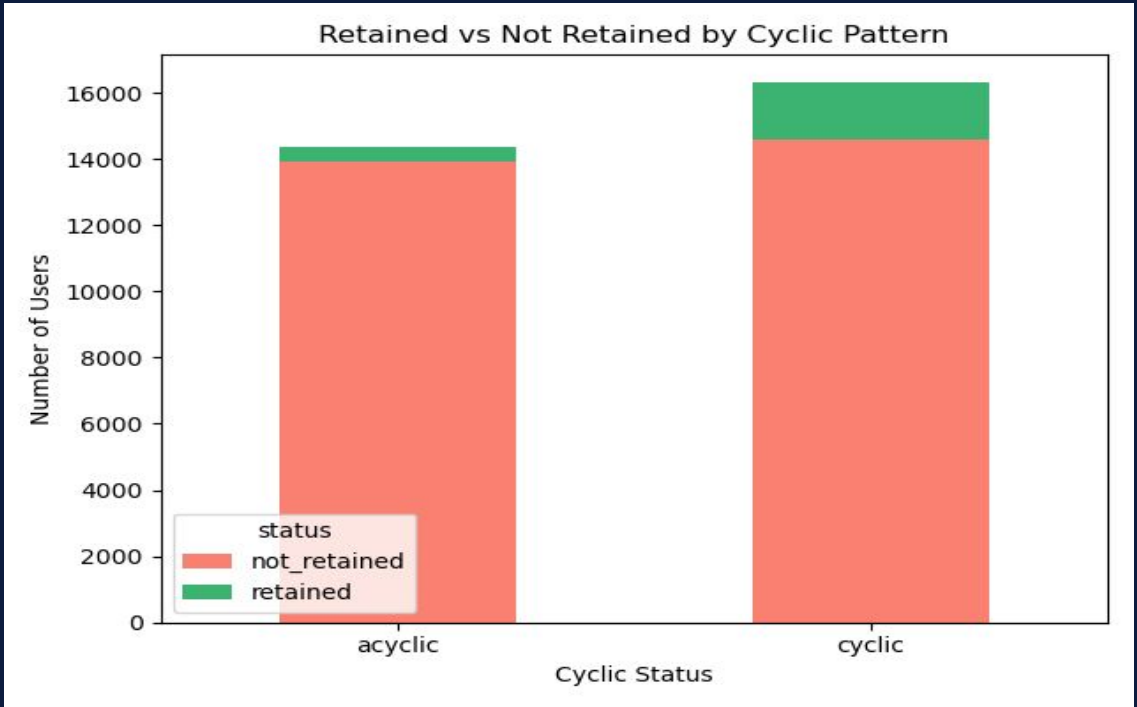
- Boosting rounds → number of trees
- Learning rate → step size
- Max depth → tree complexity
- Subsample → sample ratio
- Colsample → feature ratio

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-Score
- ROC AUC

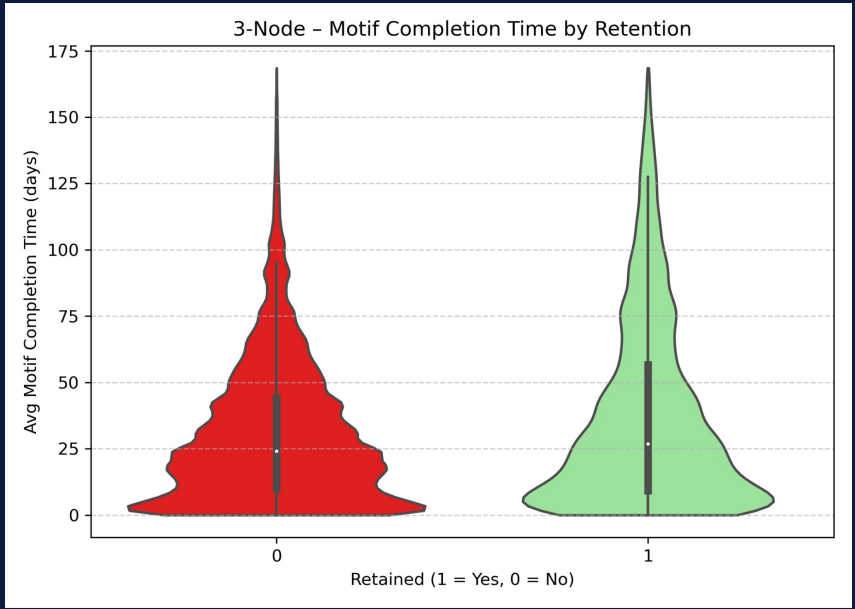
Cyclic Patterns vs Retention

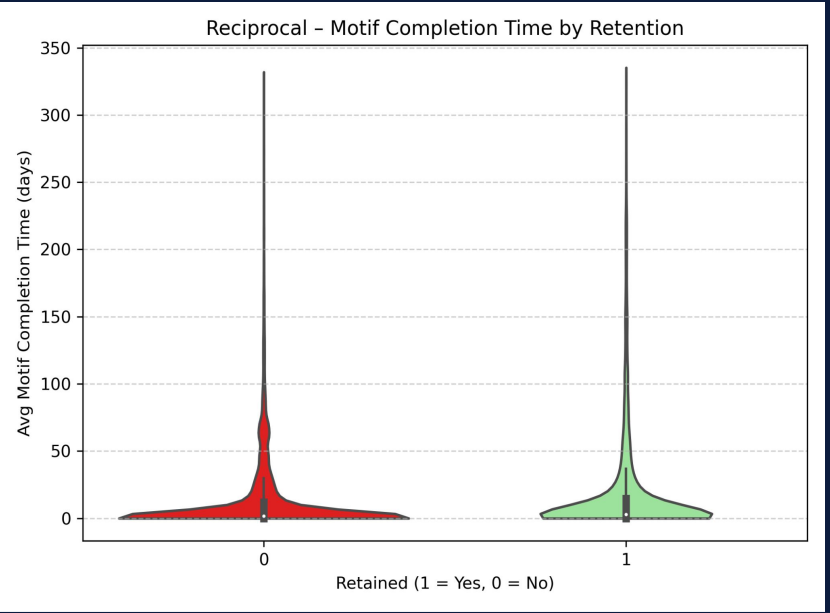
Cyclic users were more likely to be retained



Motifs vs Retention

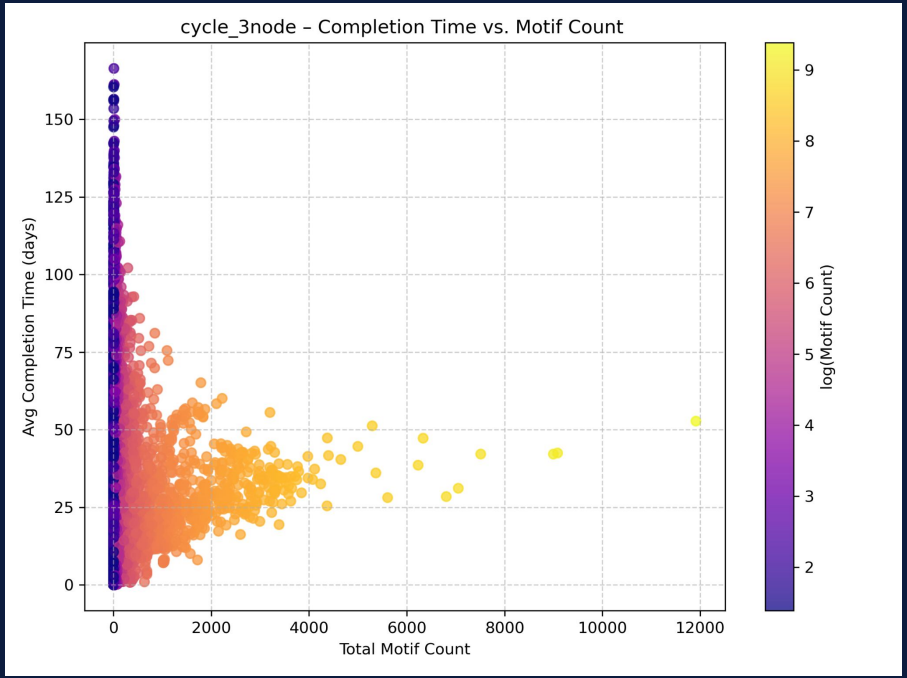
Faster Motif Completion Among Retained Users

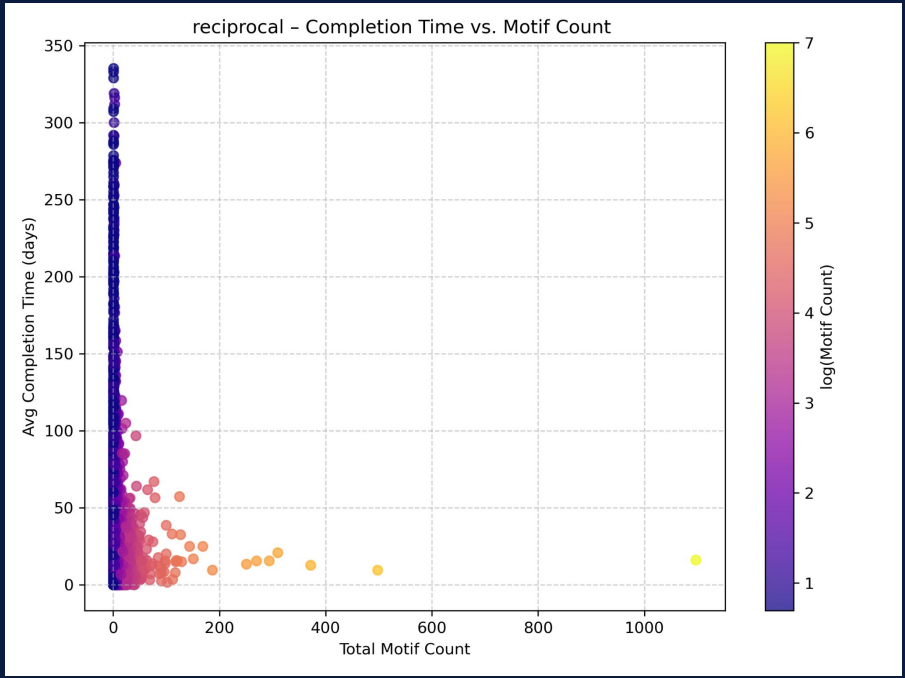




Motifs vs Time

There is a link between frequency and speed



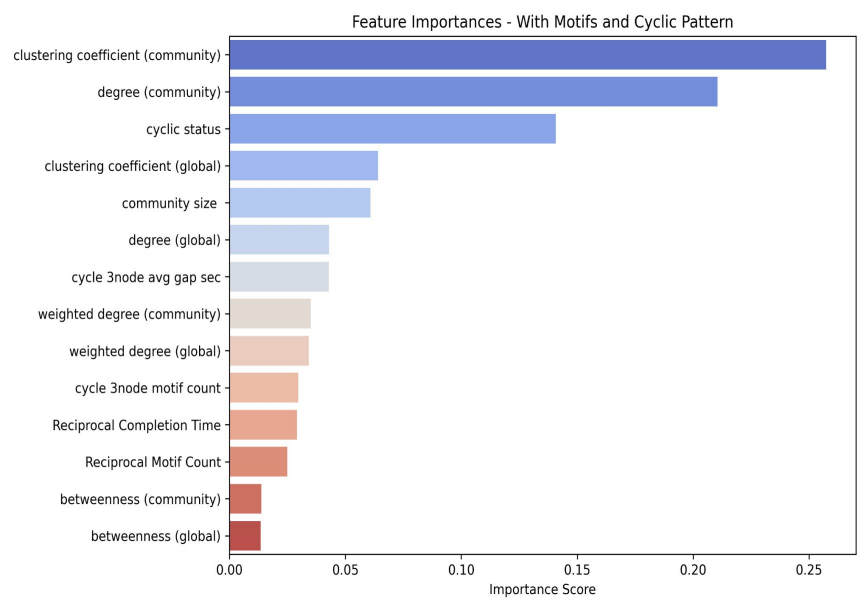


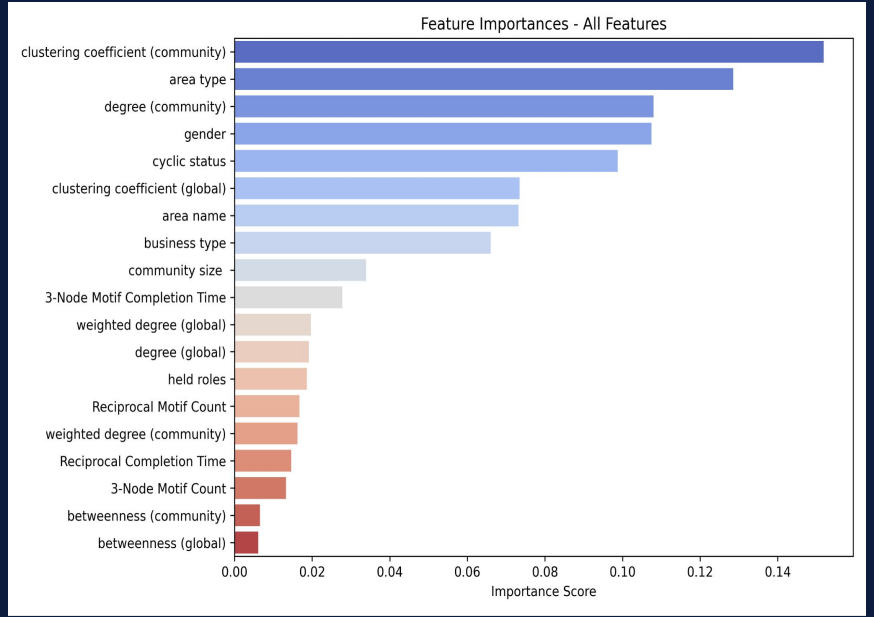
Model Evaluation

Peak Performance with Motifs & Cyclicity

Model	Accuracy	Precision	Recall	F1-score	ROC AUC
I	0.81	0.16	0.39	0.22	0.70
II	0.89	0.30	0.43	0.35	0.77
III	0.90	0.30	0.35	0.33	0.77
IV	0.90	0.34	0.48	0.40	0.82
V	0.90	0.33	0.45	0.38	0.81

Feature Importance





Limitations and Future Work

- Short observation window
- Class imbalance persists
- Model choice limited

- Richer feature representations
- Alternative modeling approaches
- Compare across currency systems

Conclusion

Sub-RQ1

- Community centrality stronger than global

Sub-RQ2

- Cyclic users more likely retained

Sub-RQ3

- Motif timing linked to engagement

Main RQ

- Retention shaped by interaction patterns

Thank you !

Any Questions?

Appendix

Global vs Community

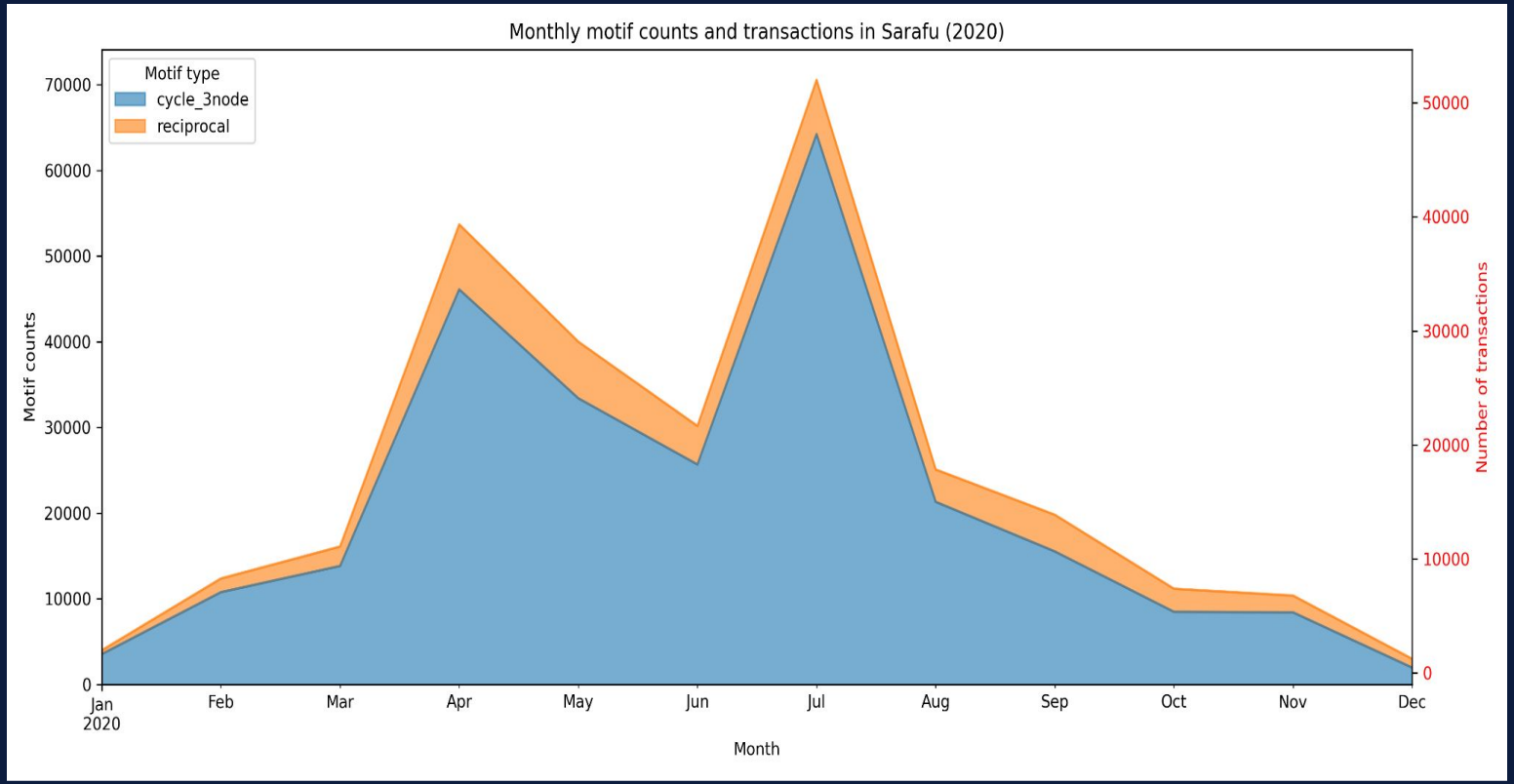
Source	Not Retained (mean ± std)	Retained (mean ± std)
Degree (Global)	1.39 ± 0.87	2.02 ± 1.10
Degree (Community)	1.38 ± 0.85	2.01 ± 1.08
Weighted Degree (Global)	6.09 ± 1.72	6.77 ± 2.25
Weighted Degree (Community)	6.02 ± 1.74	6.72 ± 2.24
Clustering (Global)	0.25 ± 0.29	0.34 ± 0.28
Clustering (Community)	0.25 ± 0.30	0.34 ± 0.28
Betweenness (Global)	0.000068 ± 0.00140	0.000195 ± 0.00108
Betweenness (Community)	0.0020 ± 0.0207	0.0047 ± 0.0243

Demographics

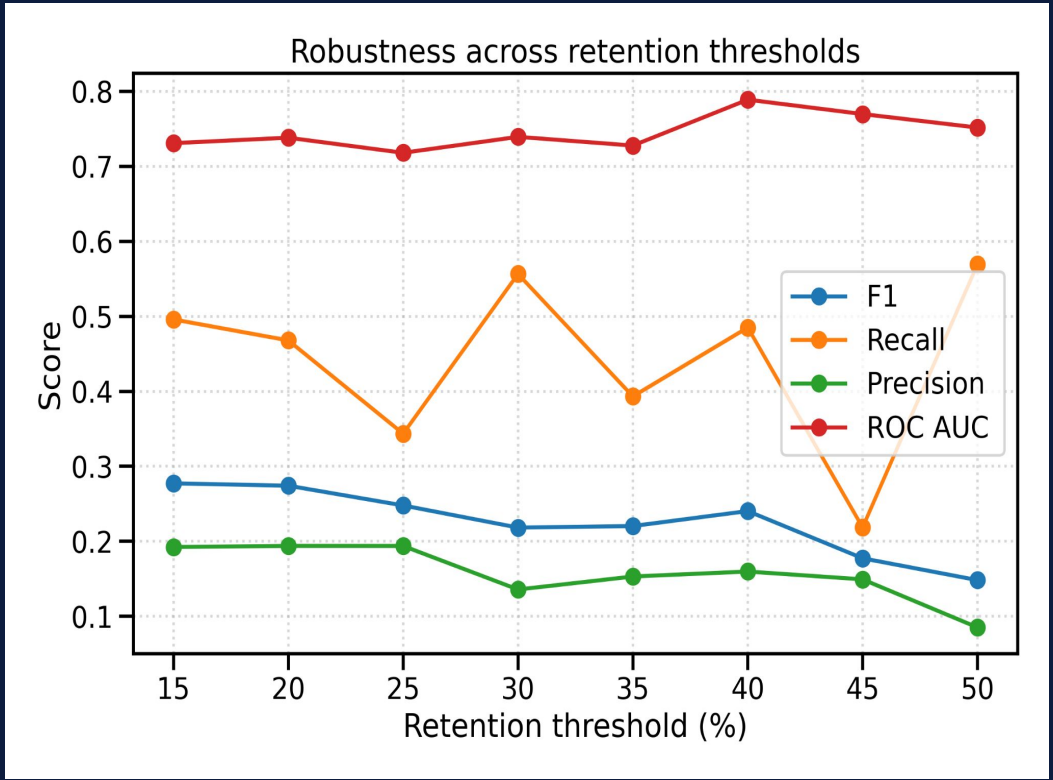
Table 7: Retention and non-retention rates by demographic attributes

Category	Group	Retention Rate	Non-Retention Rate
Gender	Female	8.6%	91.4%
	Male	5.8%	94.2%
	Unknown	7.4%	92.6%
Area	Kinango Kwale	8.2%	91.8%
	Mukuru Nairobi	3.4%	96.6%
	Misc Nairobi	6.1%	93.9%
	Nyanza	30.6%	69.4%
	Kilifi	17.9%	82.1%
Role	Beneficiary	7.0%	93.0%
	Group Account	41.0%	59.0%
Business Type	Farming	7.4%	92.6%
	Food	7.9%	92.1%
	Government	22.7%	77.3%
	Savings	32.1%	67.9%
	Shop	5.6%	94.4%

Motifs


Monthly motif counts and transactions in Sarafu (2020)
Motif type
cycle_3node
reciprocal
Motif counts
Number of transactions
Month
0
10000
20000
30000
40000
50000
60000
70000
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

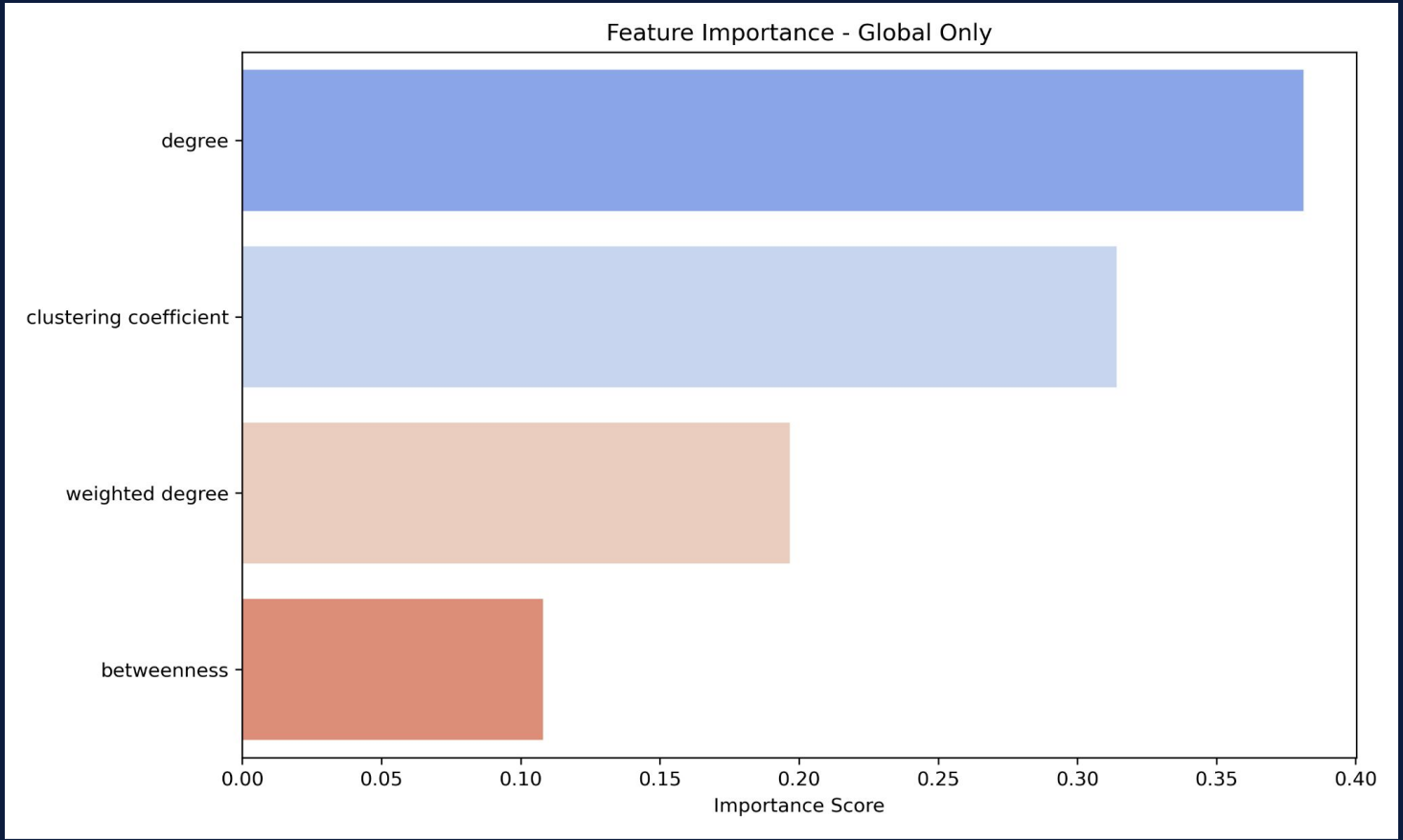
Robustness Check



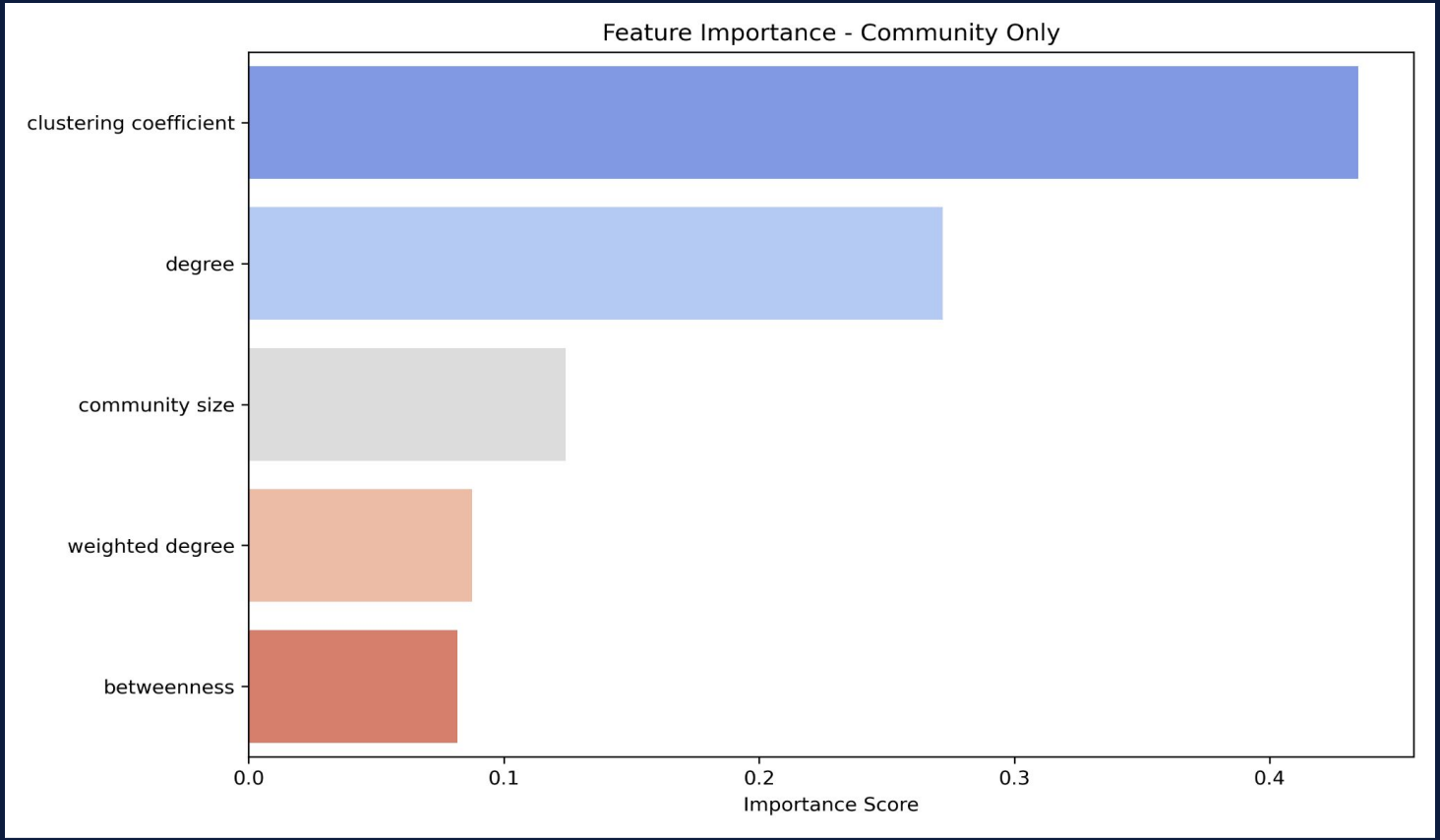
Feature Importance

Feature	Analytical Level
clustering coefficient (community)	Meso-level
cyclic status	Macro-level
degree (community)	Meso-level
area type	Individual-level
business type	Individual-level
clustering coefficient (global)	Macro-level
area name	Individual-level
gender	Individual-level
community size (community)	Meso-level
3-Node Motif Completion Time	Micro-level
Reciprocal Motif Count	Micro-level
weighted degree (community)	Meso-level
degree (global)	Macro-level
3-Node Motif Count	Micro-level
held roles	Individual-level
Reciprocal Motif Completion Time	Micro-level
weighted degree (global)	Macro-level
betweenness (community)	Meso-level
betweenness (global)	Macro-level

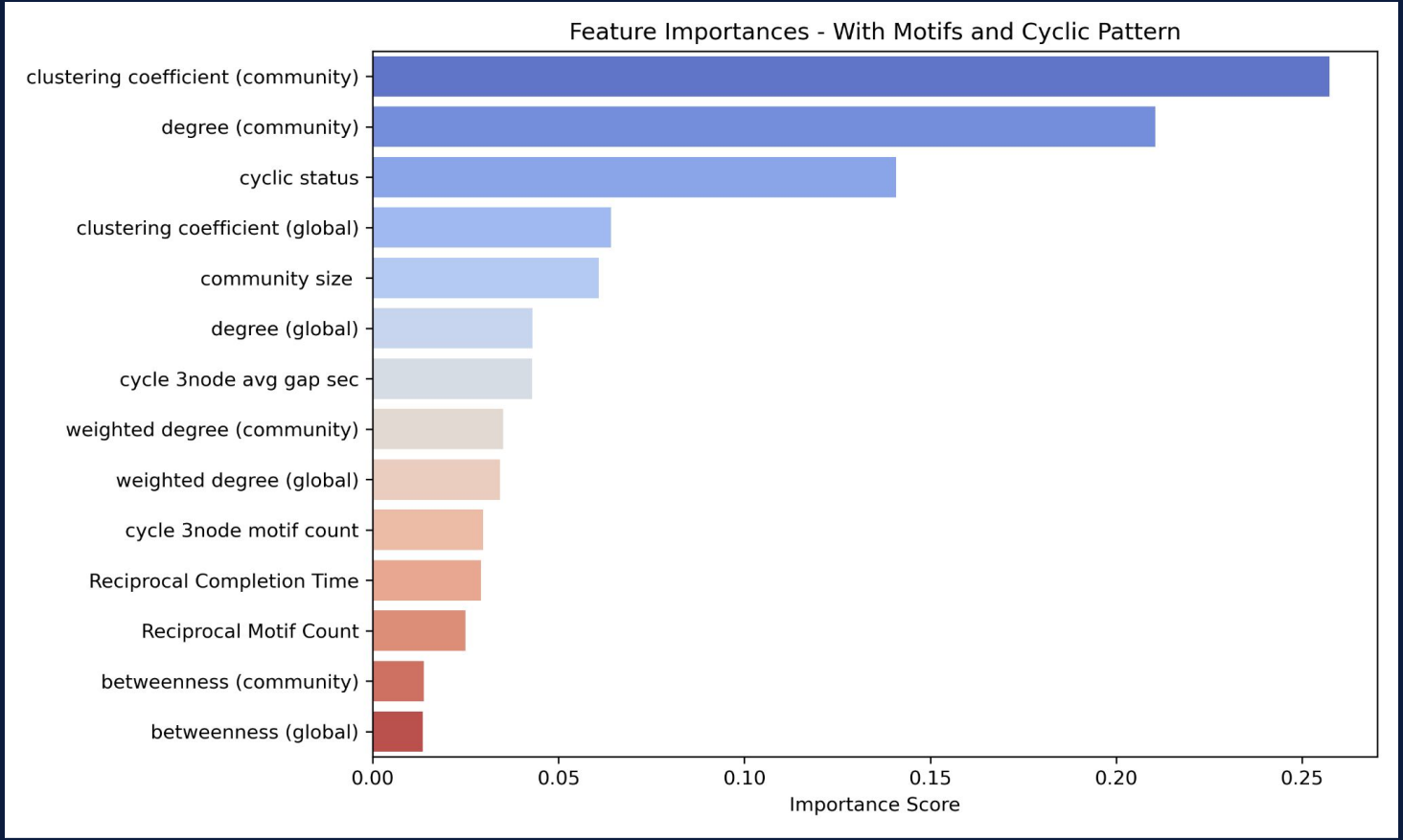
Feature Importance



Feature Importance



Feature Importance



Macro Level Features

Feature Name	Feature Type	Temporal Scope	Variable Type
In-degree	Structural Connectivity	Static	Numeric
Out-degree	Structural Connectivity	Static	Numeric
Weighted In-degree	Transaction Volume	Static	Numeric
Weighted Out-degree	Transaction Volume	Static	Numeric
Clustering Coefficient	Local Cohesion	Static	Numeric
Betweenness Centrality	Bridging Role	Static	Numeric
Cyclic Status	Structural Pattern	Static	Boolean

Meso Level Features

Feature Name	Feature Type	Temporal Scope	Variable Type
Community Size	Group Structure	Static	Numeric
In-degree	Structural Connectivity	Static	Numeric
Out-degree	Structural Connectivity	Static	Numeric
Weighted In-degree	Transaction Volume	Static	Numeric
Weighted Out-degree	Transaction Volume	Static	Numeric
Clustering Coefficient	Local Cohesion	Static	Numeric
Betweenness Centrality	Bridging Role	Static	Numeric

Micro Level Features

Feature Name	Feature Type	Temporal Scope	Variable Type
3-Node Motif Count	Motif Participation	Static	Numeric
3-Node Motif Completion Time	Motif Timing	Temporal	Numeric
Reciprocal Motif Count	Motif Participation	Static	Numeric
Reciprocal Completion Time	Motif Timing	Temporal	Numeric

Individual Level Features

Feature Name	Feature Type	Temporal Scope	Variable Type
Gender	Demographic Attribute	Static	Categorical
Area	Demographic Attribute	Static	Categorical
User Role	Demographic Attribute	Static	Categorical
Business Type	Demographic Attribute	Static	Categorical