NET 4103/7431 Network Science and Graph Learning – Final Project

Professor: Vincent Gauthier

Submitted by : Arjun Saseendran

Link : https://github.com/saseendran-arjun/network-science-graph-learning-final-project.git

1. Introduction

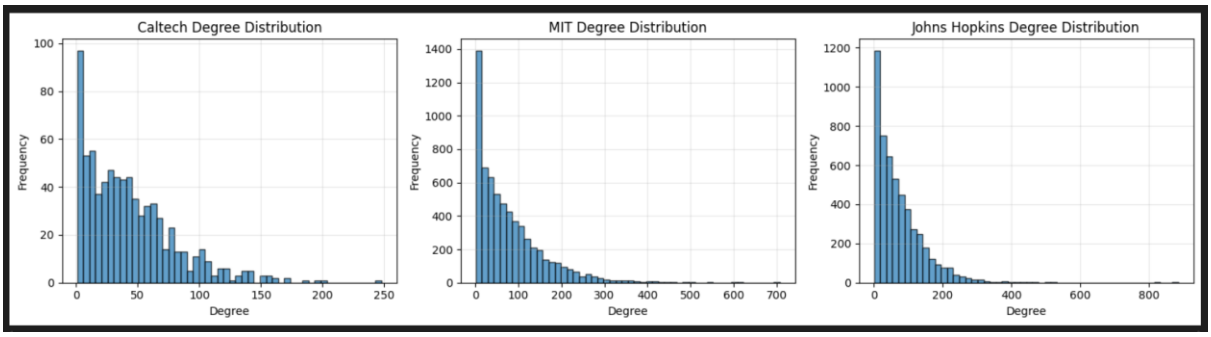
Read the following papers:

a. Assembling the Facebook: Using Heterogeneity to Understand Online Social Network Assembly.
b. Comparing community structure to characteristics in online collegiate social networks.
c. Social Structure of Facebook Networks

The papers examine how Facebook networks evolved and organized across 100 US universities in 2005. The first paper shows networks mature based on adoption rate rather than just time. The other two papers find that friendship groups align most strongly with class year and dorm assignments, though this varies by university.

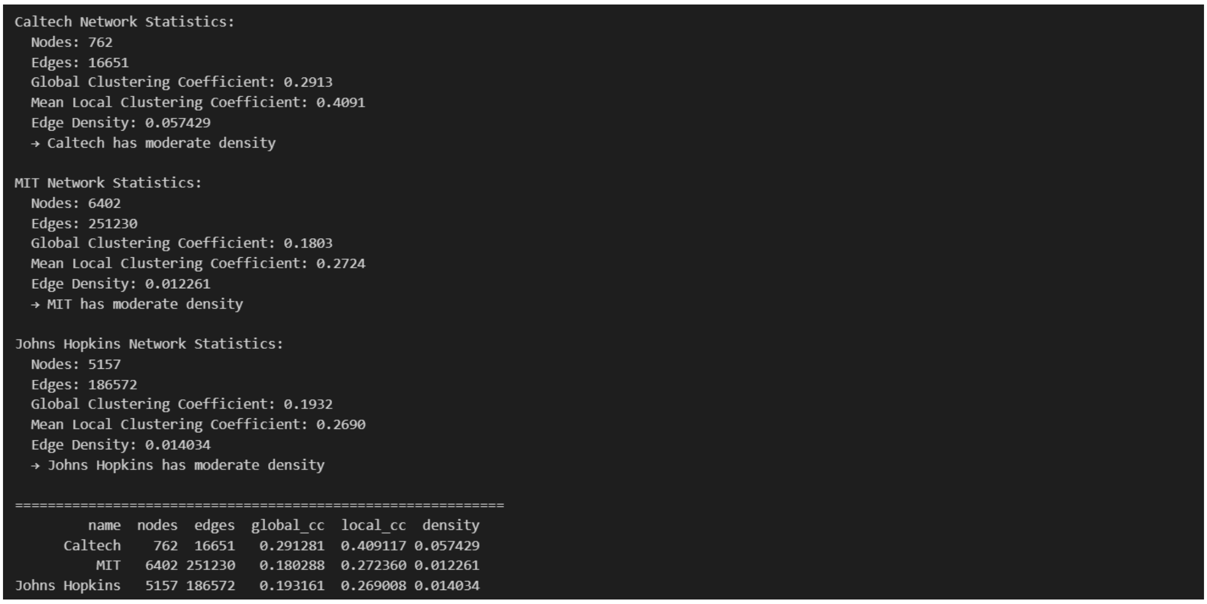
2. Social Network Analysis with Facebook100 Dataset

a. Degree Distribution Analysis



The figure shows the degree distributions for Caltech, MIT, and Johns Hopkins networks. All three networks exhibit similar right-skewed distributions, meaning most students have a moderate number of friends (20-80 connections) while a few highly social individuals have hundreds of connections. This is typical of real social networks and indicates a "small-world" structure where most people are connected through a few well-connected hubs. Caltech, being smaller, has a tighter distribution with degrees mostly under 200, while MIT and Johns Hopkins show longer tails extending to 700+ connections. The consistent pattern across all three universities suggests that social behaviour on Facebook follows similar principles regardless of institution size.

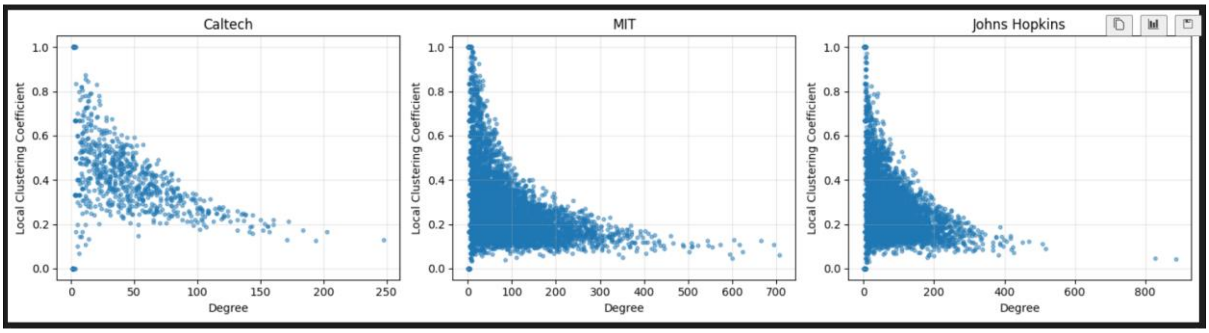
b. Clustering and Density Analysis



```
Caltech Network Statistics:
  Nodes: 762
  Edges: 16651
  Global Clustering Coefficient: 0.2913
  Mean Local Clustering Coefficient: 0.4091
  Edge Density: 0.057429
  → Caltech has moderate density

MIT Network Statistics:
  Nodes: 6402
  Edges: 251230
  Global Clustering Coefficient: 0.1803
  Mean Local Clustering Coefficient: 0.2724
  Edge Density: 0.012261
  → MIT has moderate density

Johns Hopkins Network Statistics:
  Nodes: 5157
  Edges: 186572
  Global Clustering Coefficient: 0.1932
  Mean Local Clustering Coefficient: 0.2690
  Edge Density: 0.014034
  → Johns Hopkins has moderate density

============================================================
         name  nodes  edges  global_cc  local_cc  density
      Caltech    762  16651   0.291281  0.409117 0.057429
          MIT   6402 251230   0.180288  0.272360 0.012261
Johns Hopkins   5157 186572   0.193161  0.269008 0.014034
```

The results present key network metrics for the three universities. MIT and Johns Hopkins are clearly sparse networks with densities around 1.2-1.4%, meaning only about 1% of all possible friendships actually exist. Caltech is notably denser at 5.7%, likely because it's a smaller, more tight-knit community where students know each other better. All three networks show high clustering coefficients (Caltech: 0.29-0.41, MIT: 0.18-0.27, JHU: 0.19-0.27), indicating that "friends of friends are likely friends" - a hallmark of real social networks. The higher clustering in Caltech suggests stronger community structure, possibly due to its residential house system. Despite being sparse overall, the high clustering means these networks have locally dense friendship circles, which is exactly what we'd expect in college social environments where people form tight-knit groups.

c. **Degree vs Clustering**



The figure reveals a clear negative correlation between node degree and local clustering coefficient across all three networks. Students with few friends (low degree) tend to have very high clustering - their friends all know each other, forming tight social circles. In contrast, highly connected students (hubs with 200+ friends) have much lower clustering, often below 0.2, because they bridge different social groups that don't overlap. This pattern suggests that social hubs act as connectors between communities rather than belonging to a single tight group. The relationship is remarkably consistent across all three universities despite their size differences, indicating this is a fundamental property of college social networks where popular students link otherwise disconnected communities.

3. Assortativity Analysis with the Facebook100 Dataset

We computed assortativity coefficients for five attributes across all 100 university networks in the Facebook100 dataset: student/faculty status, major, vertex degree, dorm, and gender. Assortativity measures the tendency of nodes to connect with similar nodes—positive values indicate homophily (like connects with like), while negative values suggest heterophily (opposites attract).

Figure 1 - Combined scatter plot showing assortativity vs network size.

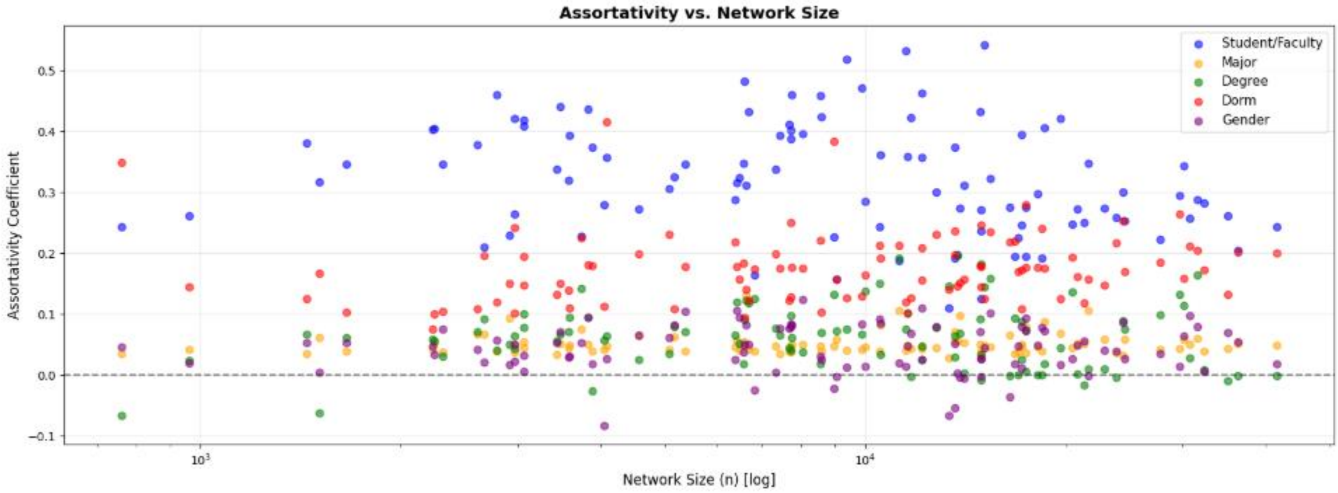


Figure 2 - Density overlay plot showing distribution of assortativity coefficients

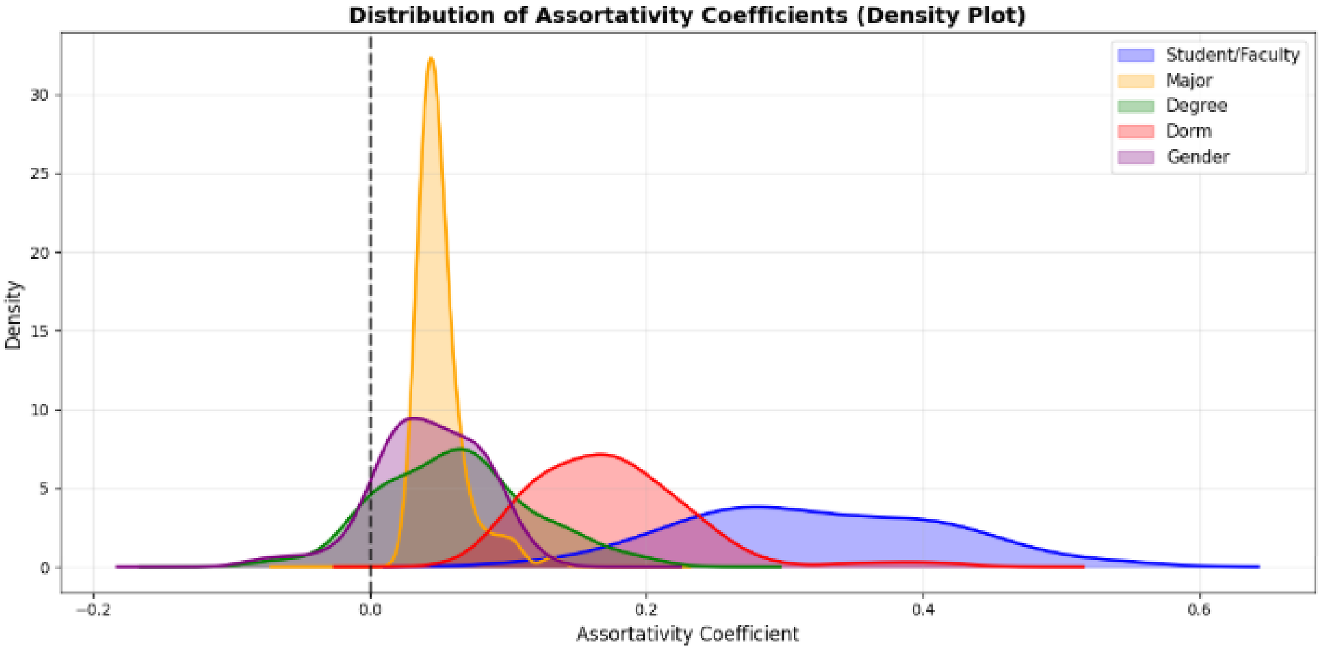


Figure 1 shows assortativity coefficients versus network size (log scale) across all 100 networks, color-coded by attribute. Figure 2 displays the distribution of assortativity values as overlaid density curves. Dashed lines indicate zero assortativity (random mixing).

Student/Faculty Status emerged as the dominant factor with the highest mean assortativity of 0.323 (range: 0.110–0.543), with blue points clustering consistently above zero, reflecting strong institutional segregation between students and faculty.

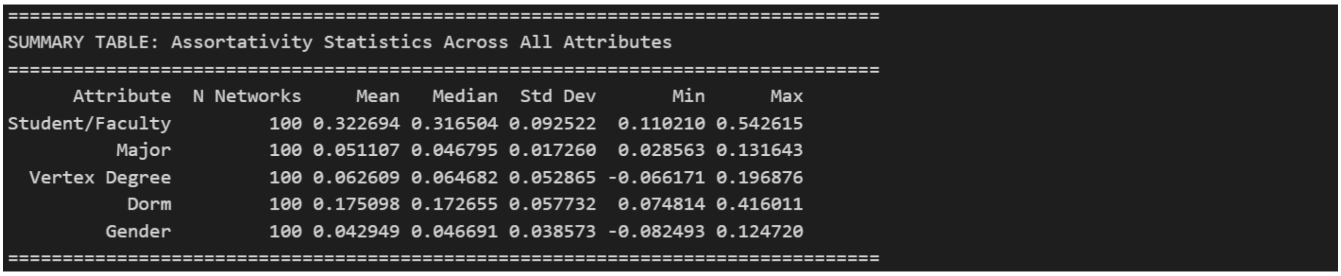
Dorm showed the second-highest assortativity at 0.175 (range: 0.075–0.416), confirming that physical proximity and shared living spaces are powerful drivers of friendship formation—students in the same dormitory encounter each other daily, naturally facilitating stronger bonds.

Vertex Degree displayed modest positive assortativity at 0.063 (range: -0.066–0.197), indicating that popular students tend to befriend other popular students while less-connected individuals cluster together. Surprisingly,

Major exhibited very weak assortativity at only 0.051 (range: 0.029–0.132), suggesting that academic field has minimal influence on friendship patterns—unlike dorm assignments which impose daily physical proximity, students easily maintain friendships across majors through extracurricular activities and residential mixing.

Gender showed the weakest signal at 0.043 (range: -0.082–0.125), with values tightly clustered around zero and some networks even exhibiting slight heterophily (negative values), indicating that gender plays a minimal role in structuring these networks.

The hierarchy is clear: institutional role and physical location drive friendships much more than academic interests or demographics. Essentially, who you see every day matters way more than who you're similar to—proximity beats homophily.



```
================================================================================
SUMMARY TABLE: Assortativity Statistics Across All Attributes
================================================================================
      Attribute  N Networks     Mean   Median  Std Dev       Min      Max
Student/Faculty         100 0.322694 0.316504 0.092522  0.110210 0.542615
          Major         100 0.051107 0.046795 0.017260  0.028563 0.131643
  Vertex Degree         100 0.062609 0.064682 0.052865 -0.066171 0.196876
           Dorm         100 0.175098 0.172655 0.057732  0.074814 0.416011
         Gender         100 0.042949 0.046691 0.038573 -0.082493 0.124720
================================================================================
```

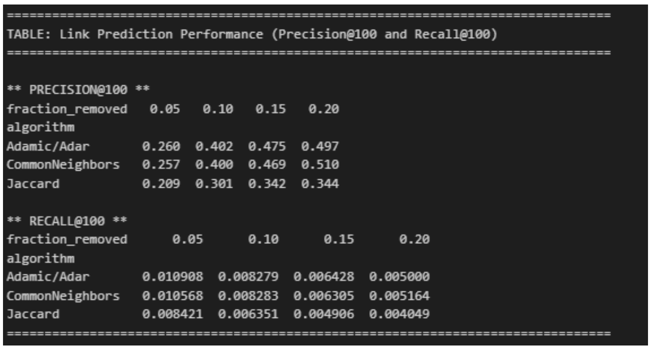
4. Link prediction

a. This paper establishes the link prediction framework we implemented, comparing topology-based proximity measures on co-authorship networks. The authors show that measures like Katz and Adamic/Adar significantly outperform random prediction, confirming that network structure encodes information about future connections.

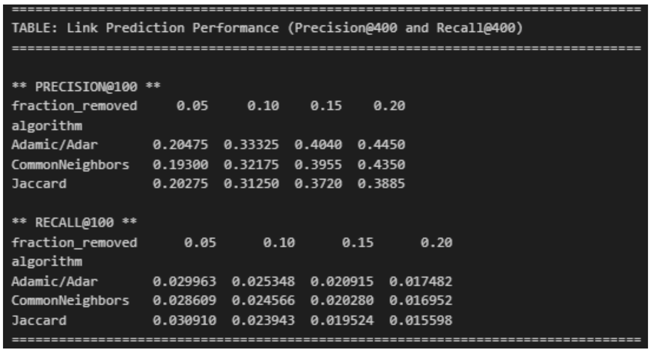
b. **Implementation :** We evaluated the ability to predict missing edges in social networks, a fundamental task in recommendation systems, using three topology-based heuristics across 10 Facebook university networks We implemented three topological heuristics:

1. Common Neighbors: $|\Gamma(u) \cap \Gamma(v)|$
2. Jaccard Coefficient: $|\Gamma(u)\cap\Gamma(v)| \ / \ |\Gamma(u)\cup\Gamma(v)|$
3. Adamic/Adar: Penalizes connections via high-degree hubs.
   $AA(u,v) = \Sigma\ 1/\log|\Gamma(z)|$ for $z \in \Gamma(u) \cap \Gamma(v)$

We randomly removed fractions $f \in [0.05, 0.1, 0.15, 0.2]$ of edges from each network to create a training graph. We then computed prediction scores for all non-existent node pairs and evaluated Precision@k and Recall@k for $k \in \{50, 100, 200, 300, 400\}$. Unlike balanced evaluation approaches, we ranked all $\sim|V|^2/2$ candidate pairs, representing a realistic cold-start recommendation scenario.



```
================================================================================
TABLE: Link Prediction Performance (Precision@100 and Recall@100)
================================================================================

** PRECISION@100 **
fraction_removed   0.05   0.10   0.15   0.20
algorithm
Adamic/Adar       0.260  0.402  0.475  0.497
CommonNeighbors   0.257  0.400  0.469  0.510
Jaccard           0.209  0.301  0.342  0.344

** RECALL@100 **
fraction_removed      0.05      0.10      0.15      0.20
algorithm
Adamic/Adar       0.010908  0.008279  0.006428  0.005000
CommonNeighbors   0.010568  0.008283  0.006305  0.005164
Jaccard           0.008421  0.006351  0.004906  0.004049
================================================================================
```



```
================================================================================
TABLE: Link Prediction Performance (Precision@400 and Recall@400)
================================================================================

** PRECISION@100 **
fraction_removed     0.05     0.10    0.15    0.20
algorithm
Adamic/Adar       0.20475  0.33325  0.4040  0.4450
CommonNeighbors   0.19300  0.32175  0.3955  0.4350
Jaccard           0.20275  0.31250  0.3720  0.3885

** RECALL@100 **
fraction_removed      0.05      0.10      0.15      0.20
algorithm
Adamic/Adar       0.029963  0.025348  0.020915  0.017482
CommonNeighbors   0.028609  0.024566  0.020280  0.016952
Jaccard           0.030910  0.023943  0.019524  0.015598
================================================================================
```

Tables above summarize performance across 10 networks. Adamic/Adar consistently achieves the highest precision (avg P@100=0.402), followed by CommonNeighbors (0.400) and Jaccard (0.301). The gap between Adamic/Adar and Jaccard widens in larger networks, demonstrating the importance of hub-weighting in heterogeneous social structures. Recall values remain low (< 0.01 at k=100) due to the extreme class imbalance: removed edges represent < 0.01% of all candidate pairs in the full search space.

c. **Evaluating a link predictor:**

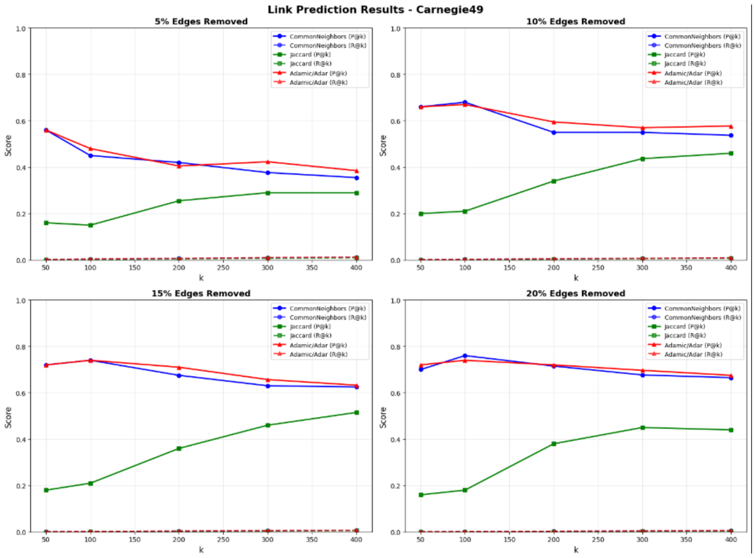


Figure above shows precision/recall curves for Carnegie49, a mid-sized network with ~6,000 nodes. Adamic/Adar and CommonNeighbors exhibit nearly identical performance, with overlapping precision curves across all removal fractions. At 10% removal, both achieve P@100 ≈ 0.68, while Jaccard lags at P@100 ≈ 0.22. The convergence of AA and CN suggests that Carnegie49 lacks extreme degree heterogeneity—when most students have similar degrees, the hub-weighting advantage of Adamic/Adar diminishes. Jaccard's poor performance (consistently 40-50 percentage points lower) confirms its unsuitability for social networks where union-based normalization over-penalizes high-degree nodes.

Based on the evaluation across 10 university networks, we draw the following conclusions:

1. Adamic/Adar (AA): Best overall performance, especially in networks with high degree variance. Down-weights connections through hubs, making it robust for heterogeneous social structures.
2. Common Neighbors (CN): Matches AA in mid-sized networks with moderate degree heterogeneity (e.g., Carnegie49). Fastest to compute (simple set intersection), making it the practical baseline.
3. Jaccard (JC): Consistently worst-performing. Over-normalization by union size makes scores negligibly small in scale-free networks, unsuitable for friend recommendation systems.

5. Label Propagation

a. The assigned paper by Zhu and Ghahramani surveys node classification methods in social networks, where the goal is predicting unknown node attributes using network topology and partial labels. The paper presents two algorithm families: iterative classifiers using node features and neighbourhood information, and random-walk-based methods that propagate labels through network paths based on homophily (connected nodes are similar). We implemented Zhou et al.'s label propagation algorithm, which uses the iterative update Y^(t+1) = α × T × Y^(t) + (1-α) × Y^(0) to balance neighbour influence with original labels, converging quickly while remaining scalable to large networks.

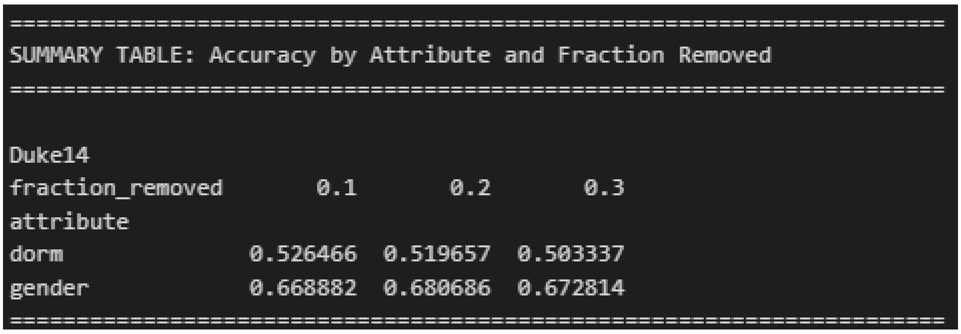
b. **Implementation** : We implemented the semi-supervised label propagation algorithm following Zhou et al.'s formulation. The algorithm represents the network as an adjacency matrix A, computes a row-normalized transition matrix T where T[i,j] represents the probability of moving from node i to node j, and initializes a label matrix Y where Y[i,c] gives the probability that node i has label c.

The iterative update rule Y^(t+1) = α × T × Y^(t) + (1-α) × Y^(0) propagates label information through network edges, where α = 0.85 controls the trade-off between neighbour influence and original label retention. At each iteration, labelled nodes are clamped to their true labels to prevent drift, and the algorithm terminates when changes fall below a tolerance threshold (10^-4) or after 100 iterations. The implementation uses NetworkX for graph operations and NumPy for efficient matrix computations, with predictions made by taking the argmax of each node's label distribution.

c. **Experimental Setup :** We selected the Duke14 network from the Facebook100 dataset, containing 9,885 nodes and 506,437 edges in its largest connected component. The network includes three node attributes: dorm (6,995 nodes, 135 unique values), gender (9,038 nodes, 2 values), and major (insufficient valid data for analysis). For each attribute, we randomly removed 10%, 20%, and 30% of known labels, used the remaining labels to train the label propagation algorithm with α = 0.85, and evaluated accuracy on the removed labels. Each experiment was repeated with a fixed random seed for reproducibility.

d. **Results** : Table 2 presents the accuracy of the label propagation algorithm across different attributes and removal rates. Gender classification achieved strong performance (66.9-68.1% accuracy) with remarkable stability across all removal rates. Dorm prediction was more challenging due to 135 possible values but still achieved 50.3-52.7% accuracy, representing a significant improvement over random guessing (0.74%). The algorithm converged quickly in all cases: 12-18 iterations for gender and 32-35 iterations for dorm. Major attribute analysis was not possible due to insufficient valid data in the Duke14 network.

Table 2: Accuracy of Label Propagation Algorithm



```
======================================================================
SUMMARY TABLE: Accuracy by Attribute and Fraction Removed
======================================================================

Duke14
fraction_removed       0.1       0.2       0.3
attribute
dorm               0.526466  0.519657  0.503337
gender             0.668882  0.680686  0.672814
======================================================================
```

Attribute	10% Removed	20% Removed	30% Removed
Dorm	0.527	0.520	0.503
Gender	0.669	0.681	0.673

e. **Analysis and Conclusions**

Gender achieves higher accuracy (67%) than dorm (50-53%) primarily because it's a simple binary classification problem, whereas dorm requires distinguishing among 135 different options. Even though dorm has stronger network clustering (assortativity 0.21 vs. 0.05 for gender from Question 3), predicting the exact dorm among so many choices is inherently harder than predicting male/female. Both attributes show only modest accuracy decline as we remove more labels (1-5%), demonstrating that the algorithm successfully leverages network structure, with dorm still achieving 67× better performance than random guessing and gender remaining remarkably stable around 67-68% accuracy across all scenarios.

6. Communities detection with the FB100 datasets

a. **Research Question :** We investigate whether weak gender homophily at the individual edge level accumulates into significant gender segregation at the community level. While Question 3 showed that gender assortativity is weak (~0.05), meaning individual friendship choices show only minimal same-gender preference, we hypothesize that these small biases compound through network effects to produce measurably imbalanced communities. This tests whether group-level social structure can emerge from individually weak preferences.

b. **Implementation :** We applied the Louvain community detection algorithm to identify natural groupings in four university networks (Caltech36, Reed98, Haverford76, Duke14). For each detected community, we computed gender composition and tested whether the observed male/female ratio significantly differed from the network's overall 50-50 baseline using binomial tests (p<0.05). To measure overall community-gender alignment beyond random expectation, we computed z-Rand scores following the methodology in Traud et al. [2], which standardizes the number of same-community, same-gender node pairs against a null model with fixed marginals. Higher z-Rand scores indicate stronger alignment between community structure and gender attributes.

c. **Results and Evaluation :**

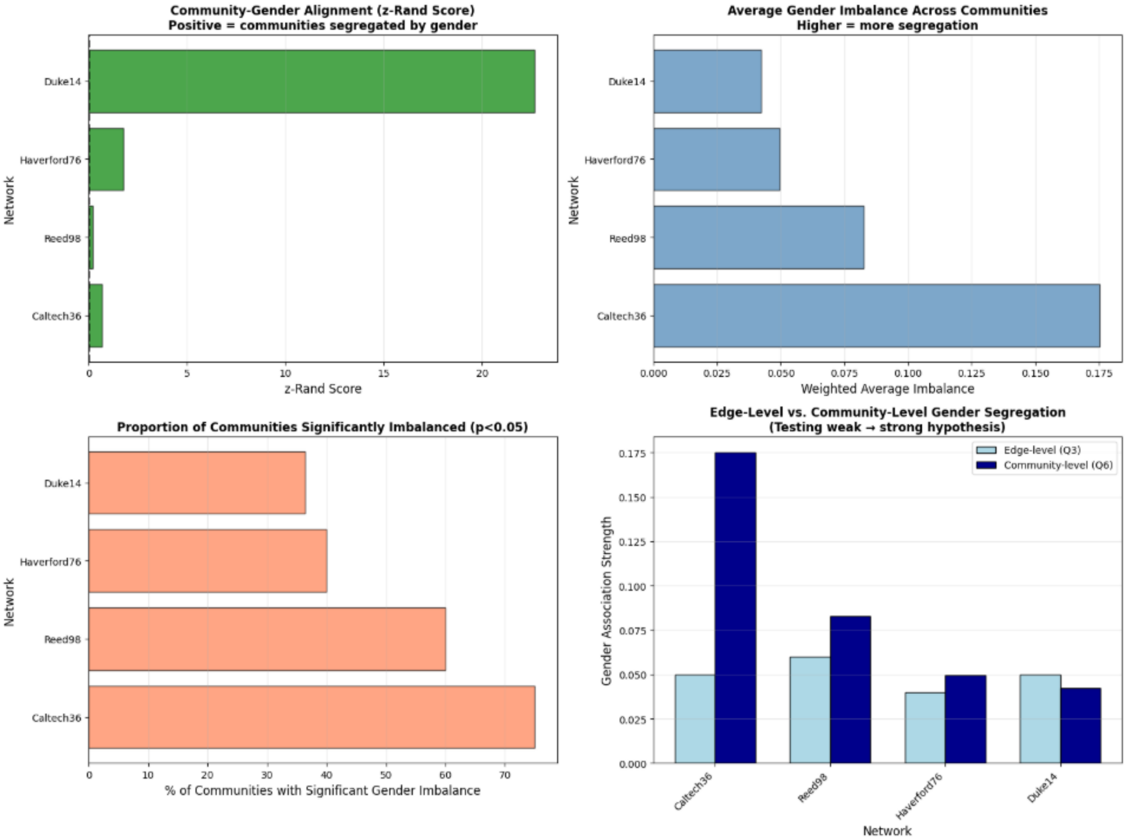


Table: Gender Segregation at Community Level

Network	Size	Communities	z-Rand	Avg Imbalance	% Significant Communities	% Nodes in Significant
Caltech36	762	8	0.69	0.175	75.0%	85.3%
Reed98	962	6	0.20	0.083	60.0%	74.0%
Haverford76	1446	5	1.76	0.050	40.0%	50.7%
Duke14	9885	12	22.68	0.042	36.4%	52.4%

All networks show positive z-Rand scores, confirming that communities are more gender-segregated than random. Average imbalance measures the weighted mean deviation from 50-50 gender ratio across communities, with values ranging from 4.2% (Duke) to 17.5% (Caltech). The proportion of statistically significant communities (p<0.05) ranges from 36-75%, indicating that a substantial fraction of detected communities exhibit non-random gender composition.

7. Conclusion

Our analysis confirms the hypothesis that weak individual-level gender preferences accumulate into measurable community-level segregation. Despite edge-level gender assortativity of only ~0.05 from Question 3, we observed significant gender imbalances in detected communities across all four universities, with z-Rand scores ranging from 0.69 to 22.68. The results reveal an interesting size-dependent pattern: smaller networks like Caltech show larger imbalances (17.5% average deviation) affecting most communities (75% significant), while larger networks like Duke exhibit smaller imbalances (4.2%) but with overwhelming statistical confidence (z-Rand=22.68). This demonstrates that even minimal same-gender friendship preferences can produce substantial group-level structure through network clustering effects. The finding has practical implications for understanding social segregation: significant demographic separation can emerge organically from individually weak biases rather than requiring strong explicit preferences, with the effect being most reliably detected and measured in large-scale social networks.