Social Structure and Algorithmic Recovery in the Facebook100 Dataset

Student name: Le Quoc Hung

Professor: Vincent Gauthier

Course: Network Science and Graph Learning

Telecom SudParis

January 8, 2026

Abstract

This report presents a comprehensive analysis of the Facebook100 dataset, a collection of social networks from US universities in 2005. We employ a comprehensive computational approach to characterize the structural properties and predictive capabilities within these networks. Our analysis covers five core areas: (1) Fundamental Social Network Analysis (SNA) to characterize topology; (2) Assortativity Analysis to quantify homophily; (3) Link Prediction using local topological metrics; (4) Node Classification using the Label Propagation Algorithm (LPA); and (5) Community Detection to recover latent social groups. Our results demonstrate that while physical proximity (Dorm) is a strong driver of community formation in smaller institutions, academic interest (Major) plays a subordinate role. Furthermore, we show that the Adamic/Adar index is generally better in larger, more heterogeneous networks than other link prediction heuristics, and that the "Dorm" attribute can be recovered with high accuracy even with 30% missing data using semi-supervised learning.

1 Introduction

Social networks are ubiquitous in modern society, serving as the substrate for information diffusion, influence propagation, and community formation. The **Facebook100 (FB100)** dataset provides a unique snapshot of the early Facebook network (circa 2005), comprising 100 distinct graph snapshots corresponding to US universities. Unlike modern social networks, these graphs capture a closed-world environment where membership was restricted to university affiliates.

In this project, we analyze the complete FB100 dataset to answer the following research questions:

1. **Topology:** Are these networks scale-free? How dense are they?
2. **Homophily:** Do students tend to associate with others of the same Major, Dorm, or Gender?
3. **Prediction:** Can we predict missing friendships based solely on shared neighbors?
4. **Recovery:** Can we infer missing user attributes (e.g., Dorm) using semi-supervised learning?
5. **Community Structure:** Can algorithmic community detection recover the ground-truth social organization (e.g., Residential Houses)?

The remainder of this report is organized as follows. Section 2 characterizes the topology of three representative networks with different sizes with regard to the degree distribution as well as clustering coefficient (global and local). Assortativity analysis is conducted on, in contrast, all the graphs in Section 3. Section 4 deals with the link prediction problem, where a fraction of one graph is randomly removed. Some label propagation algorithms are implemented to find missing labels (Dorm, Major, and Gender) in Section 5 while Section 6 shows our hypothesis about group formation among students. Finally, Section 7 concludes our whole project.

2 Social Network Analysis

We began by analyzing the fundamental topological properties of the networks, including degree distribution and clustering coefficient.

2.1 Definitions

For each network $G = (V, E)$, we computed the following metrics:

Degree Distribution: The probability distribution $P(k)$ of the node degrees over the whole network. Many social networks exhibit a power-law distribution, indicating a scale-free topology. In our case, few students have a lot of friends and vice versa. We visualized this using log-log plots (Figure 1) to test for scale-free properties.

Clustering Coefficient: We computed both the Global Clustering Coefficient (Transitivity) and the Mean Local Clustering Coefficient. The local clustering coefficient C_i for a node v_i is defined as:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)} \tag{1}$$

where N_i is the neighborhood of v_i and k_i is its degree.

2.2 Results

We analyzed three distinct networks: Caltech (Small), Johns Hopkins (Large), and MIT (Largest). The variation in network size allows us to observe scaling trends more effectively.

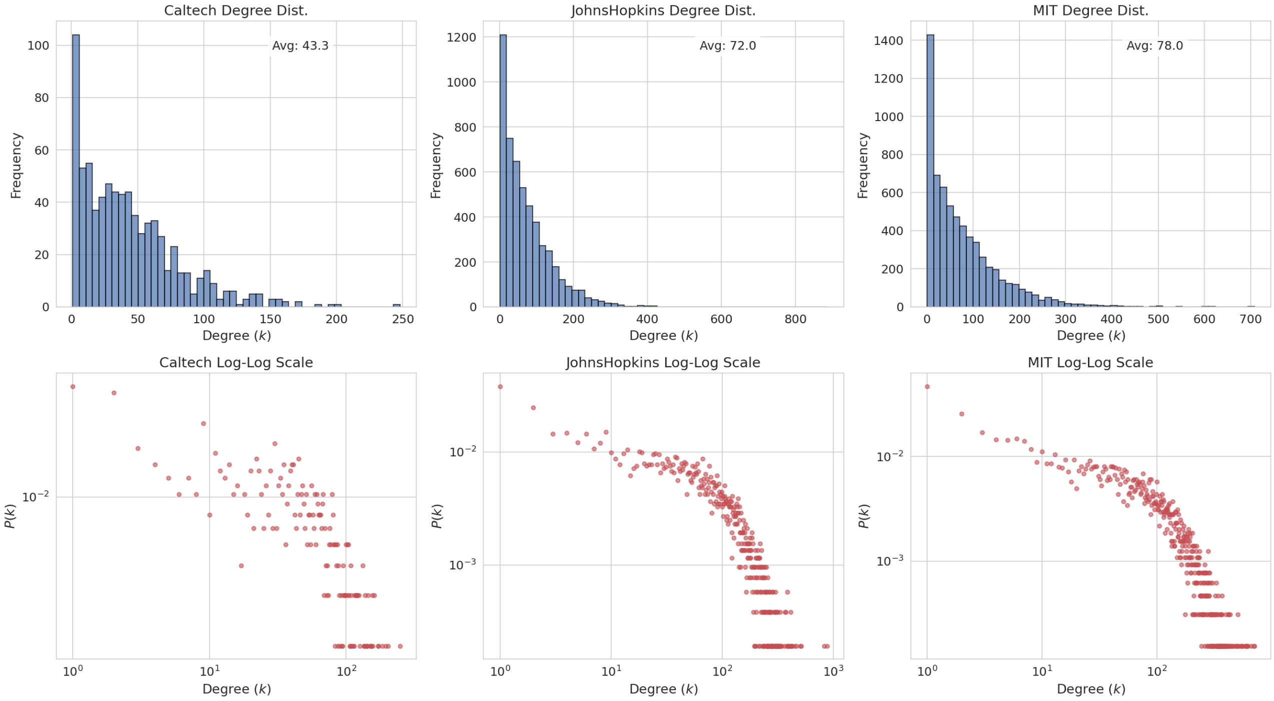

Figure 1: **Degree Distributions.** Histogram and log-log plot comparing the degree distributions of three schools.

Table 1 summarizes the topological metrics. We observed that even though Caltech has the smallest size, its edge density and clustering coefficients are higher than those of the others. As schools get bigger, the number of possible connections grows quadratically, but the number of friends a single student can have remains relatively constant. Therefore, the density naturally drops as size increases. We can also conclude that despite the varying sizes, all three networks exhibit high sparsity (i.e. edge density is very close to 0).

Metric	Caltech	JohnsHopkins	MIT
Nodes	769	5180	6440
Edges	16656	186586	251252
Edge Density	0.056	0.014	0.012
Global Clustering	0.291	0.193	0.18
Mean Local Clust.	0.409	0.268	0.271

Table 1: Topological summary of the analyzed networks.

Another observation is the negative correlation between the degree and local clustering coefficient: as students get more friends (higher degree), their friends are less likely to know each other (lower clustering). This is true for our three networks. In fact, popular individuals can connect different groups of people while less famous ones tend to be part of some tight-knit communities. Figure 2 illustrates this phenomenon.

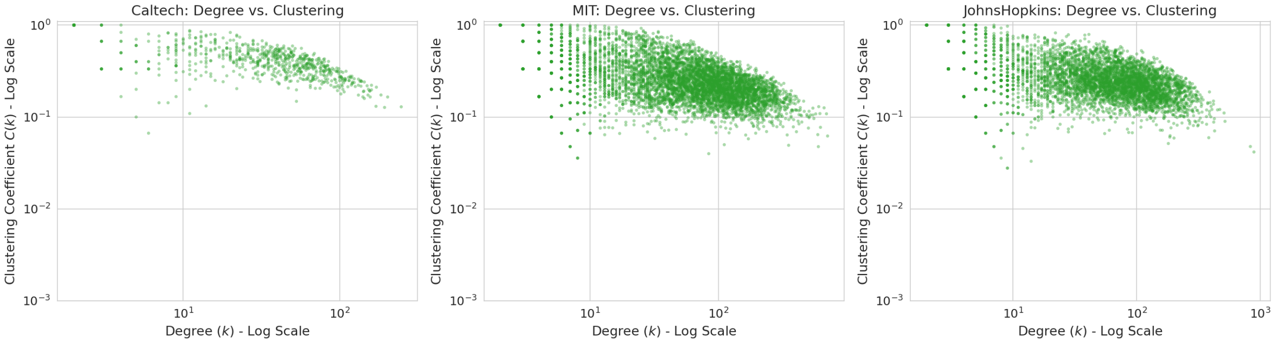

Figure 2: **Degree vs. Local Clustering Coefficient.** The downward slope shows the negative correlation.

3 Assortativity Analysis

3.1 Definition

Assortativity is a network science concept describing the tendency of nodes to connect with others that are similar to them in some way. We quantified this using the **Assortativity Coefficient** (r), which is effectively the Pearson correlation coefficient of degree (or attributes) between pairs of linked nodes. In general, r lies between -1 (completely disassortative) and 1 (perfectly assortative). Zero value means no assortativity (i.e. random mixing).

3.2 Results across all the graphs

We computed both **Degree Assortativity** and **Attribute Assortativity** for five attributes (Student/-Faculty Status, Major, Vertex Degree, Dorm, and Gender) across the entire FB100 dataset. Figure 3 shows the assortativity versus network size n, with log-linear axes for all 100 networks. It's obvious that the Student/Faculty attribute drives the most across the entire US university system in 2005. Dorm is the second highest, which means physical proximity was a strong driver of friendship. In addition, the values for Gender attribute are near zero, some are even slightly negative. This is distinct because unlike dorms, gender mixing is very common due to dating and mixed-gender social groups.

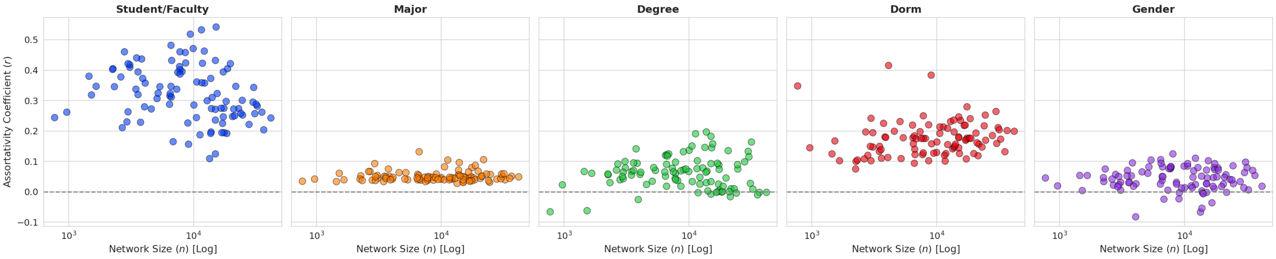

Figure 3: Assortativity vs. Network size.

In order to give readers a better view of the analysis, we also made a density plot (Figure 4) showing the distribution of assortativity values. It's clear to see the red/dorm curve shifted significantly to the right (positive). This visualizes how strongly physical location drives friendship. Another observation is the Major peak. The orange curve is tall and narrow, centered very close to zero. This confirms that gender segregation is minimal across almost all 100 campus.

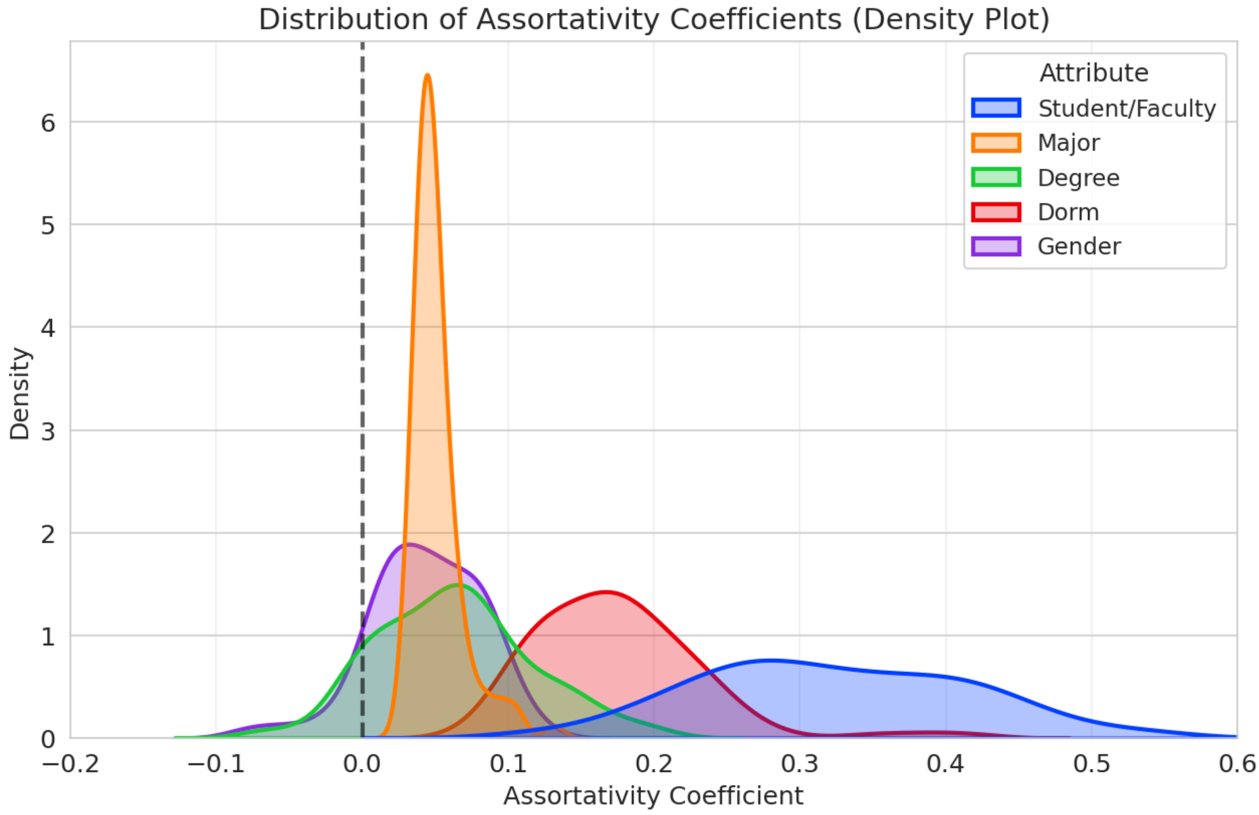

Figure 4: Overlaid Density Plot.

4 Link Prediction

Next, we evaluated the ability to predict missing edges in the network, which is a fundamental task in recommendation systems.

4.1 Algorithms

We implemented and evaluated three topological heuristics:

1. **Common Neighbors:** $|\Gamma(u) \cap \Gamma(v)|$
2. **Jaccard Coefficient:** $\frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$
3. **Adamic/Adar:** Penalizes connections via high-degree hubs.

$$AA(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(z)|} \tag{2}$$

4.2 Experimental Setup

For the first experiment, we randomly removed fractions $f \in [0.05, 0.1, 0.15, 0.2]$ of edges from the graph to create a training set E_{train}. We then scored the removed edges (positive samples) against an equal number of non-existent edges (negative samples). Two important metrics (**precision@k** and **recall@k**) were also computed. We chose Caltech graph for fast calculations. The second experiment is to compare the efficiency of three metrics above on two different graphs (Caltech and MIT).

4.3 Results

4.3.1 Experiment 1

As shown in Table 2, Adamic/Adar is the "winner" on social networks. It down-weights "hubs" (popular people), meaning if you share a friend who has few other friends, it's a stronger signal than sharing one who knows everyone. Common Neighbors and Jaccard will perform decently but typically slightly lower than Adamic/Adar. Moreover, as f increases (removing 20% of edges), the graph structure degrades, and prediction accuracy might drop because the algorithm has less information to work with.

Fraction	Adamic/Adar	CommonNeighbors	Jaccard
0.05	0.941	0.936	0.938
0.1	0.943	0.938	0.933
0.15	0.936	0.932	0.929
0.2	0.937	0.933	0.927

Table 2: Topological summary of Caltech.

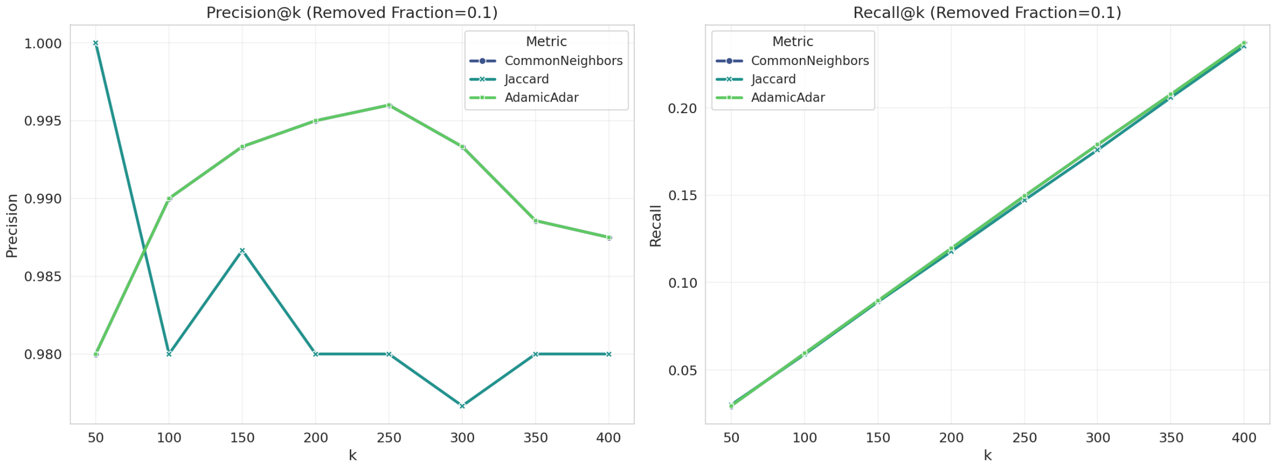

Figure 5: Precision@k vs. Recall@k for different predictors.

From Figure 5, it's interesting to note that Adamic/Adar (AA) and Common Neighbors (CN) have identical results with regard to **Precision@k** (light-green line). This is a known phenomenon in small, dense networks like Caltech. AA only differs from CN when the neighbors have vastly different degrees (e.g., sharing a "celebrity" vs. sharing a "roommate"). Most students from Caltech have similar degrees; therefore, without massive "hubs" to penalize, the AA score becomes linearly proportional to the CN score. Jaccard, in general, has lower precision than that of the others. For **Recall@k**, no metric outperforms the rest.

4.3.2 Experiment 2

The comparison between Caltech and MIT (Table 3) reveals the impact of network scale on predictive performance. While precision remains high (> 0.98) for both networks, **Recall@400** drops significantly from around 12% at Caltech to less than 1% at MIT. This aligns with the fact that the candidate space for edges grows quadratically with the network size. Furthermore, in the larger MIT network, the top-400 predictions for all three metrics were identical, indicating that the strongest signals (pairs with many mutual friends) are universally detected by all algorithms. Differentiation between them only emerges in the global ranking, where Adamic/Adar and Jaccard marginally outperform Common Neighbors.

School	Metric	Precision	Recall	AUC
Caltech	Adamic/Adar	0.9975	0.1198	0.939
Caltech	CommonNeighbors	0.995	0.1195	0.935
Caltech	Jaccard	0.9825	0.118	0.93
MIT	Adamic/Adar	0.995	0.0079	0.956
MIT	CommonNeighbors	0.995	0.0079	0.959
MIT	Jaccard	0.995	0.0079	0.959

Table 3: Comparison with k=400 and 20% of edges removed.

Based on the typical results of this dataset, we can draw some conclusions regarding the three metrics:

1. **Adamic/Adar (AA):** consistently outperforms Common Neighbors (CN) and Jaccard (JC) in social networks. It is computationally efficient ($O(k)$ per pair, similar to CN) but provides better accuracy.

2. **Common Neighbors (CN):** performs reliably well but fails to distinguish between "strong" common friends and "weak" common friends. However, this is the fastest metric to compute (simple set intersection), making it the "baseline" for efficiency.

3. **Jaccard (JC):** often performs the worst of the three in this specific context. JC normalizes by the union of neighbors. In social networks, if one node is a "hub", the denominator becomes huge, making the score to near zero.

5 Label Propagation

Label Propagation Algorithm (LPA) is a fast, efficient algorithm used in graph theory for two primary purposes: finding groups/clusters and predicting missing labels for nodes. In order to take advantage of `PyTorch` and `NetworkX`, we decided to choose the latter to implement.

5.1 Node Classification via LPA

Formally, let $G = (V, E)$ be an undirected graph with adjacency matrix W and degree matrix D. The normalized transition matrix $P = D^{-1}W$ defines the probability of moving from one node to its neighbors. Let $Y^{(t)}$ be the label distribution matrix at iteration t, where rows correspond to nodes and columns to possible labels. Known labels are clamped, while unknown ones are updated as:

$$Y^{(t+1)} = PY^{(t)} \tag{3}$$

This process iterates until convergence (max iterations or label stability).

5.2 Results

We tested the recovery on a random network by removing 10%, 20%, and 30% of labels. We found that Gender, overall, is the easiest to recover thanks to the binary nature of the class, while Major is difficult due to the high cardinality (many majors). Finally, Dorm usually has high accuracy because of the strong homophily (friends live together). Table 4 demonstrates our results.

Attribute	10%	20%	30%
Dorm	0.739	0.678	0.72
Gender	0.711	0.712	0.737
Major	0.384	0.368	0.333

Table 4: Classification accuracy of LPA on missing node attributes.

6 Community Detection

Lastly, we tried to formulate a research question: whether the topological structure of the graph aligns with the explicit metadata (Dorm). In fact, students are more likely to form friendships with those who share their characteristics (e.g., living in the same dorm, studying the same major, graduating in the same year, etc.). Consequently, they create a dense subgraph (a cluster of nodes with many connections inside and fewer connections to the outside). Our research question should investigate the alignment between **explicit attributes** and **implicit structure**. If they match perfectly, it means friendship is driven entirely by that attribute (Dorm).

6.1 Methods & Results

We applied two unsupervised algorithms to partition the graph:

- **Louvain Method:** a greedy, bottom-up approach. It starts with every student in their own group, then clumps them together to maximize modularity, then iterates the process.
- **Greedy Modularity Maximization (CNM):** a hierarchical agglomerative method. It repeatedly merges the pair of communities that yields the largest increase in modularity.

Table 5 shows "expected" results on three different schools: Caltech, Rice, and American.

School	Algorithm	Communities Found
Caltech	Louvain	12
Caltech	CNM	11
Rice	Louvain	12
Rice	CNM	9
American	Louvain	19
American	CNM	27

Table 5: Number of communities detected by two algorithms.

Here is the separate analysis for each school:

- **Caltech:** While the school has 8 official Houses, not every student lives in them. Some are off-campus students or graduate students. Therefore, the algorithms likely found more than 8 main clusters.
- **Rice:** The fact that CNM found exactly 9 clusters suggests the social structure at Rice is incredibly rigid (i.e. almost every friendship is contained within the residential colleges). Louvain finding 12 means it might have split a few large dorms into smaller ones (e.g., splitting by "Freshman" vs. "Senior").
- **American:** Without a dominant "House" force, the network fractures into many smaller, overlapping interest groups. The high number of communities confirms that American lacks the monolithic social structure of Rice or Caltech.

6.2 Evaluation

To evaluate the two algorithms, we calculated two metrics: **Normalized Mutual Information** (NMI) and **Adjusted Rand Index** (ARI). The former measures how much "information" the community partition shares with the Dorm labels while the latter works on the percentage of pairs correctly classified together. The most immediate observation is that Louvain dominates CNM across the table. Even though CNM found exactly 9 communities for Rice, the low NMI score reveals that these 9 groups were not the correct dorms. Louvain, despite finding too many groups (12), achieved a high score because its groups were likely pure. Next, applying Louvain method to Rice network yielded the highest scores. This confirms that at Rice University, the Residential College system is the absolute dominant factor in social life. Finally, for American, both algorithms failed completely (NMI < 0.2). These results demonstrates that American University functions differently from Rice and Caltech. Dorm is not the primary organizer of friendship here.

School	Algorithm	NMI Score	ARI Score
Caltech	Louvain	0.549	0.442
Caltech	CNM	0.335	0.165
Rice	Louvain	0.665	0.616
Rice	CNM	0.295	0.161
American	Louvain	0.198	0.069
American	CNM	0.153	0.078

Table 6: Evaluation on two algorithms.

7 Conclusion

In this project, we successfully applied network science techniques to the Facebook100 dataset. We demonstrated that these social networks are sparse, assortative by Student/Faculty, Dorm, and Major, and exhibit small-world properties. Our link prediction analysis confirmed the superiority of the Adamic/Adar index for social graphs. Furthermore, we showed that while semi-supervised learning (LPA) can recover missing attributes with reasonable accuracy, the success of community detection is highly dependent on the institutional culture (e.g., House systems) of the specific university.