The Social Genome Project

A Comparative Network Analysis of Caltech, Cornell, and Dartmouth

Promita Rahee Sikder

Abstract

This study employs Exponential-Family Random Graph Models (ERGMs) to define and compare the "social genome" of three distinct universities (Caltech, Cornell, and Dartmouth) using the Facebook100 dataset. Our analysis moves beyond descriptive metrics to model the specific endogenous and exogenous forces that generate each institution's unique network topology. The results reveal divergent social signatures. Caltech's network structure is dominated by strong, positive homophily effects for residence and class year, alongside a significant negative effect for shared high school. Conversely, the ERGMs for both Dartmouth and Cornell are characterized by large negative density parameters and a striking lack of positive homophily effects for institutional attributes like residence or major. These findings refute a simple "Social Bubble" hypothesis for Dartmouth and confirm a "Metropolis" model for Cornell, where sociality is likely organized around unobserved, non-institutional foci. By quantifying these differences, this work validates the "social genome" as a powerful analytical concept for understanding institutional variation in social organization.

By fitting ERGMs and analyzing network metrics, we test hypotheses about how institutional characteristics (such as academic orientation, student population size, and geographic context) shape social connectivity. Our results reveal distinct social genomes: Caltech's network is characterized by strong homophily within residential houses and academic majors; Cornell exhibits a broad, cohort-driven structure centered on class year; and Dartmouth displays a dense, highly clustered network shaped by its residential and Greek life systems.

Keywords: Social Network Analysis, ERGM, Homophily, University Life, Facebook, Computational Social Science

Contents

1 Introduction

1.1 Background: Social Networks on Campus

Since their inception, online social networking sites (SNSs) have become deeply integrated into daily life, providing researchers with an unprecedented window into the structure of human connection. Facebook, for example, grew from a campus-specific tool to a global phenomenon, and its early data offers a "digital fossil" of social organization. Although online networks are an imperfect mirror of offline relationships, they offer a valuable lens for examining core sociological dynamics such as homophily (the tendency for individuals to associate with similar others), triadic closure (the likelihood that two people with a mutual friend will themselves become connected), and social capital (the value embedded in an individual's network of relationships).

This project extends the foundational work of Traud et al. (2012) [5], who conducted a broad comparative analysis of the Facebook networks at 100 U.S. universities. Their study revealed that while social structures were heavily influenced by attributes like class year and residence, the relative importance of these factors varied significantly across institutions. They noted a major distinction between microscopic drivers of friendship (dyad-level) and macroscopic organization (community-level), finding, for example, that a common high school was a strong predictor of a friendship tie but rarely a dominant factor in the university's overall community structure.

1.2 The "Social Genome" Concept

The present project takes the observation of institutional variation by Traud et al. (2012) [5] as its central premise. We propose the concept of a "social genome": a quantitative signature for each university derived from the parameters of an Exponential-Family Random Graph Model (ERGM). By fitting a consistent ERGM to the networks of three distinct universities (Caltech, Cornell, and Dartmouth) this project aims to model and quantify the unique "personality" of each institution's social structure.

1.3 Initial Hypotheses

Based on the "social genome" framework, we propose distinct, testable hypotheses for each university that reflect their unique size, focus, and environment.

Caltech, dubbed "The Focused Silo," is expected to exhibit strong positive coefficients for academic discipline (nodematch('major')) and residential house (nodematch('residence')), reflecting its tightly organized social life around these silos.

Cornell, characterized as "The Metropolis," is anticipated to show a strong organizing effect of class year (nodematch('year')), with weaker influences from residence and major indicative of its large, diverse, and diffusely connected nature.

Dartmouth, "The Social Bubble," is hypothesized to have a dense, highly clustered network driven by its geographic isolation and robust residential and Greek life, manifesting as a strong

nodematch('residence') coefficient and a less negative edges coefficient, suggesting a higher baseline density of connections.

1.4 Motivation and Practical Implications

This research moves beyond a theoretical exercise in network science to address a critical, practical challenge in higher education: ensuring the alignment between a student and their college's community is no longer left to chance. The underlying motivation is the recognition that a student's ability to flourish is deeply tied to their sense of belonging, a factor that transcends academic offerings and extends to the core values and social structure of an institution. As research by Hommes et al. (2012) [1] demonstrates, the structure of student social networks has a direct and measurable impact on academic performance and overall university experience. By quantifying a university's "social personality" into a tangible "social genome," our framework provides a data-driven tool with significant implications. For university administrators, it offers a method to engineer better student integration through structured housing assignments, to target mental health resources tailored to the specific social pressures of their unique environment, and to empirically measure the impact of student life policies. Simultaneously, this research enables students and families to make more thoughtful, well-rounded choices by encouraging them to look past conventional indicators like rankings and prestige, and instead select a college whose social atmosphere genuinely fits their individual and academic priorities. Ultimately, this work provides a method to understand and act upon the crucial "fit" between a student and their campus community.

2 Literature Review

This project is situated at the intersection of three domains of research: the fundamental principles of social network formation, the empirical analysis of social life in university settings, and the statistical modeling of complex network data. This review synthesizes key findings from each area to provide the theoretical and methodological foundation for our "Social Genome" concept.

2.1 Social Network Formation and Homophily

The structure of any social network is shaped by a set of organizing principles that govern why and how individuals form connections. Among the most powerful and well-documented of these principles is homophily: the tendency for individuals to associate with similar others. In their canonical review, McPherson et al. (2001) [4] establish that this "birds of a feather" phenomenon is a fundamental mechanism driving the formation of friendships and acquaintanceships. Homophily operates across a vast range of attributes, including age, gender, beliefs, and, crucially for our study, organizational affiliations such as one's university, major, and residence. This principle explains why friendship ties are not formed at random; instead, they

are heavily constrained by shared social contexts, which create opportunities for interaction among similar individuals. Our project uses this foundational concept as its starting point, operationalizing these shared attributes as the core "genes" in each university's social genome.

2.2 Network Analysis of College Life

The college campus provides a uniquely bounded and compelling environment for studying social network dynamics. The foundational work for our project is the broad comparative analysis of the Facebook100 dataset by Traud et al. (2012) [5]. Their research was pivotal in demonstrating that while students universally tend to form ties based on shared characteristics like class year and residence, the relative importance of these factors varies significantly from one institution to another. They established that a university is not a generic social space but an institution with a distinct social signature. Our project extends this finding by seeking to formalize and quantify this observed variation. Other research using this dataset has further underscored its value. For instance, Lewis et al. (2012) [2] used the Facebook100 data to explore the interplay of social selection and peer influence, showing how friendship networks shape, and are shaped by, cultural tastes.

2.3 Modeling Social Networks with Exponential-Family Random Graph Models (ERGMs)

To move beyond describing network patterns to modeling the underlying processes that generate them, we employ Exponential-Family Random Graph Models (ERGMs). As detailed by Lusher et al. (2013) [3], ERGMs represent a statistical framework capable of handling the complex dependencies inherent in network data, such as the tendency for friends of a friend to become friends (triadic closure). Unlike simpler models, ERGMs allow us to simultaneously test hypotheses about the relative strength of multiple social forces, such as homophily in various attributes, that contribute to the overall network structure.

The application of ERGMs to university social networks has yielded powerful insights. A landmark study by Wimmer and Lewis (2010) [6] used ERGMs to analyze a friendship network in a racially diverse university dorm. Their work masterfully demonstrated how ERGMs can disentangle multiple, overlapping mechanisms of tie formation, showing how shared activities and simple propinquity operated alongside and sometimes overshadowed racial homophily. Their study serves as a methodological blueprint for our project, providing a strong precedent for using ERGMs to model the competing social forces that constitute a university's unique "personality." By adopting this framework, we can build a formal model of each university's social genome, quantifying the specific combination of forces that makes each campus a distinct social world.

3 Data and Methods

This study employs a multi-step analytical approach to quantify and compare the social networks of three U.S. universities. We first describe the unique dataset used for this analysis and its limitations, then outline the sequence of analytical strategies applied to test our hypotheses.

3.1 The Facebook100 Dataset

We are using the The Facebook100 dataset to understand the personalities of each higher educational institute. The Facebook100 dataset represents a historical snapshot because it captures the complete Facebook friendship networks of 100 U.S. universities as of September 2005. The students at a single university are the nodes, identified by unique anonymized IDs. The connections, or edges, are undirected and unweighted friendship ties, representing reciprocal Facebook friendships confirmed exclusively within that university's network. This strict isolation makes each network a distinct graph, where its structure, size, and attribute distributions directly reflect the unique social environment of that institution (e.g., Caltech's small, tech-focused network versus Cornell's large, diverse one). This independence is important, as it allows for comparative analysis using a consistent Exponential Random Graph Model (ERGM), where any observed structural differences can be confidently attributed to institutional characteristics rather than inter-network dependencies. Beyond the network structure, the dataset also provides several anonymized, self-reported user attributes that will be central to our analysis: gender, class year, major (as a numerical identifier), residence (dormitory, house, etc., as a numerical identifier), and high school (as a numerical identifier). The network data itself consistently consists of undirected, unweighted friendship ties, where a connection signifies a reciprocal friendship.

3.1.1 Background Context of the Data

The time period when the data were collected is critical for understanding early social media dynamics, as Facebook, launched in 2004 by Mark Zuckerberg at Harvard, was then exclusively a college-centric platform requiring a .edu email address for membership. This structural constraint inherently rendered each university's network a self-contained social system, providing an unprecedented opportunity to study campus social structures in isolation. This invaluable data, which encompasses tens of thousands of students and all their friendship ties alongside select user attributes, was originally provided directly in an anonymized format by Facebook to academic researchers. This data predates the ubiquitous adoption of smartphones and the sophisticated algorithmic feeds characteristic of contemporary social media. Hence it offers a raw and unfiltered record of student connections. Consequently, it serves as a "digital fossil" of early 2000s campus life, reflecting social dynamics before the global proliferation and complex evolution of social media platforms. The significance of this dataset is substantial; it facilitated fascinating research, such as the work by Traud et al. (2012) [5], which showed how attributes like class year and residence differentially shaped network structures across various universities.

Crucially, this dataset no longer exists publicly and has been removed from the internet due to evolving privacy concerns, making such comprehensive network data sharing an uncommon practice today. Our access to this invaluable resource was facilitated through the Wayback Machine: Internet Archive, a digital library that has preserved over 946 billion web pages over time. The original data, stored in .mat files, was subsequently converted to .csv files for more accessible manipulation and analysis, a process executed using Python *(see Appendix A for the full code snippet).*

3.2 Data Limitations

This dataset, while comprehensive for its time, has quite a few limitations that must be acknowledged. First, it represents a static, single point in time and does not capture the dynamic evolution of friendships. Second, all user attributes are self-reported and may contain missing or incomplete values (which we must account for in our analysis). Third, as the original authors note, an online network is an imperfect, though valuable, proxy for offline social interactions. Finally, the anonymization of attributes, while necessary for privacy, prevents a direct interpretation of which specific dorms or majors are most central. We can, however, still measure the powerful aggregate effect of sharing these attributes on tie formation.

3.3 Analytical Strategies

Our analysis proceeds in three stages. We begin with a descriptive exploratory analysis to establish the baseline characteristics of each network. We then employ community detection algorithms to identify social clusters. Finally, we use Exponential-Family Random Graph Models (ERGMs) as our primary statistical tool to model the underlying social forces that produce the observed network structures.

3.3.1 Network Metrics for Exploratory Analysis

To gain initial insights into the macro-level structure of each university, we calculate a suite of standard network metrics. These "vital signs" include network size (nodes and edges), density, average degree, global clustering coefficient (transitivity), assortativity by residence and major (homophily), and modularity. These metrics provide a quantitative foundation for comparing the overall connectivity, cohesion, and community organization of each institution, allowing for a preliminary test of our hypotheses.

3.3.2 The ERGM Framework for Modeling

Our central method is the Exponential-Family Random Graph Model (ERGM), a statistical tool for modeling the fundamental processes that shape network structure. Unlike simpler models that assume friendship ties are independent, ERGMs can account for complex, endogenous network structures, such as the tendency for "friends of a friend to be friends" (transitivity).

We will fit a consistently specified ERGM to the networks of Caltech, Cornell and Dartmouth. The model is specified as:

$$\texttt{University\_Network} \sim \texttt{edges + nodematch('year') + nodematch('residence') + nodematch('major')}$$

The coefficients generated by this model for each university constitute its "social genome". They quantify the strength and direction of various social forces.

3.3.3 Community Detection and Alignment Tests

To investigate the meso-level structure of the networks, we first identify social clusters using the Louvain method for community detection. To add statistical rigor to our interpretation of these communities, we then conduct a Chi-squared test. Mirroring a method from the original Traud et al. (2012) [5] study, this test determines if the algorithmically-detected communities show a statistically significant alignment with known user attributes, such as residence or major. Finally, to add further context to our ERGM results, we conduct targeted subgroup analyses using the Mann-Whitney U Test to compare the distributions of network metrics (e.g., degree centrality) between distinct groups within a single university and Spearman's Rank Correlation to assess the relationship between different measures of node importance (e.g., degree and betweenness centrality).

4 Results and Discussion

4.1 Exploratory Analysis: A Comparison of University Vital Signs

To compare the baseline network structures of Caltech, Dartmouth and Cornell, we calculated a set of standard network measures (Table **??**) quantifying their size, connectivity, and community structure. By evaluating these metrics, we elucidate how institutional characteristics shape social interactions, providing a foundation for subsequent ERGM analysis.

We first examined node and edge counts to assess network scale. Caltech's network, with 762 nodes and 16,651 edges, reflects its status as a small, elite technical institute. Dartmouth's is significantly larger (7,677 nodes, 304,065 edges), while Cornell's network is massive in comparison, comprising 18,621 nodes and 790,753 edges. This confirms the vast difference in scale between the three institutions and immediately positions Cornell as the sprawling "Metropolis."

Network density, the proportion of actual to possible ties, follows an inverse trend with size. Caltech's intimate scale yields the highest density (0.057). In contrast, Cornell has the lowest density (0.005), an expected outcome given that the potential number of connections in a network of its size is astronomically large. Despite this sparsity, Cornell's average degree (84.9) is the highest of the three, just surpassing Dartmouth's (79.2) and far exceeding Caltech's (43.7). This reveals a key feature of the "Metropolis": while the probability of any two random

Table 1: Comparison of Network Statistics for Caltech, Dartmouth, and Cornell

Network Measure	**Caltech**	**Dartmouth**	**Cornell**
Nodes	762	7,677	18,621
Edges	16,651	304,065	790,753
Density	0.057	0.010	0.005
Average Degree	43.703	79.215	84.931
Clustering	0.291	0.151	0.136
Assortativity (Residence)	0.070	0.118	0.161
Assortativity (Major)	0.002	0.044	0.049
Modularity	0.399	0.431	0.471
Average Path Length	2.338	2.768	2.876
Diameter	6	8	8
Mean Betweenness	0.002	0.000	0.000
Degree Variance	1,367.554	5,591.171	7,395.427

students being friends is low, the average student is extremely well-connected, a testament to a vibrant and active social environment with immense opportunities for forming ties.

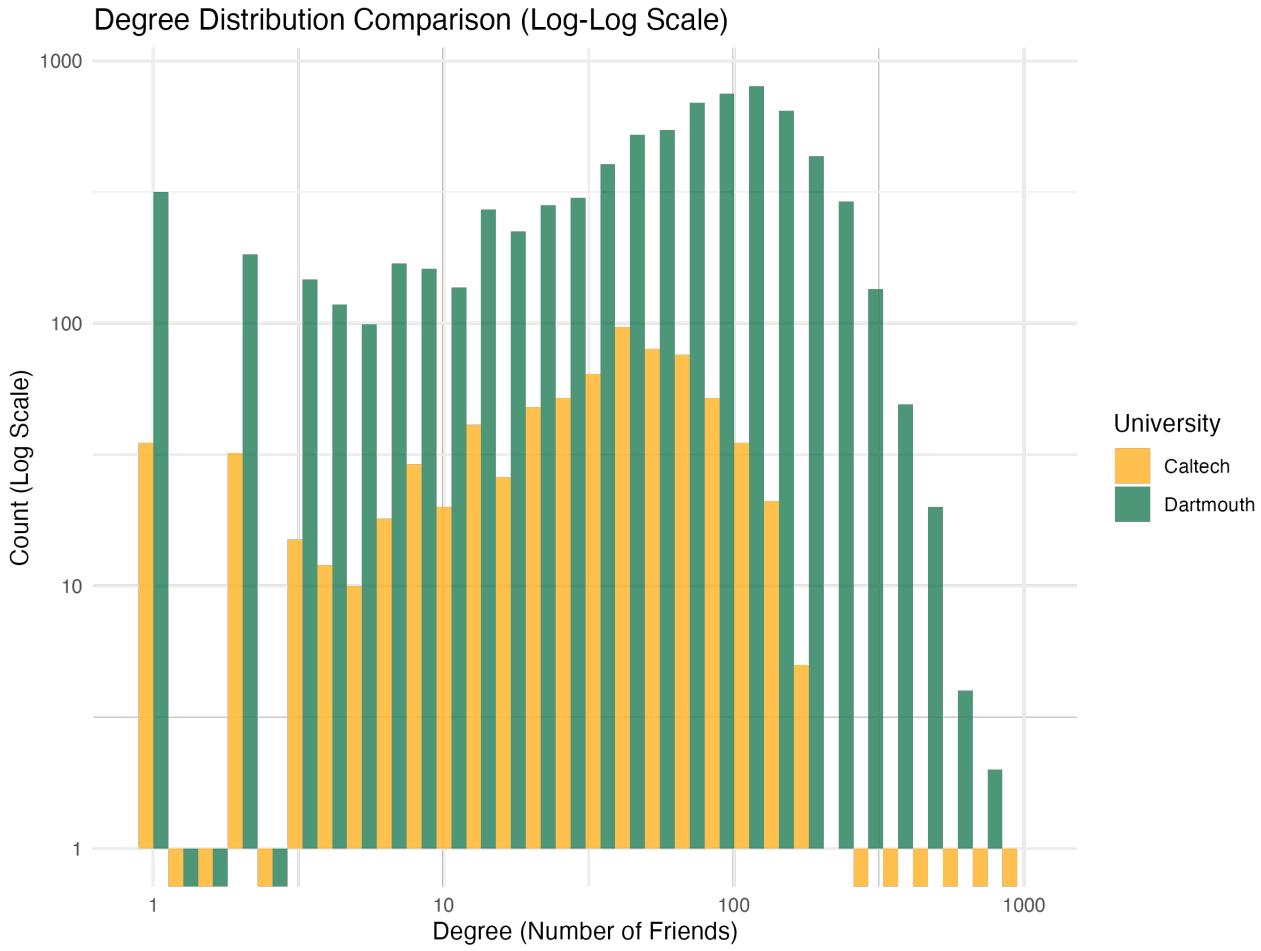

Figure 1: Degree Distribution of Student Social Networks at Caltech and Dartmouth.

The global clustering coefficient, which measures the prevalence of "friend of a friend" triangles, highlights fundamental differences in social cohesion. Caltech's network is highly clustered (0.291), strongly supporting the "Focused Silo" hypothesis where intense, localized social interactions within its residential House system are paramount. Conversely, Cornell's network exhibits the lowest clustering (0.136). This is powerful evidence for the "Metropolis" model, suggesting a social structure that is more diffuse, where friendships are less likely to be concentrated in tight-knit local groups and more likely to span across a wider, more diverse social landscape. Dartmouth's clustering (0.151) is surprisingly low for a "Social Bubble," suggesting that while its community is cohesive, its ties may be more bridging across different social circles (e.g., between Greek houses and sports teams) rather than bonding exclusively within them.

Assortativity, a measure of homophily, reveals the primary drivers of connection. For residence, Cornell shows the strongest tendency for students to connect with others in the same dorm (0.161), even more so than Dartmouth (0.118). This finding suggests that in a massive university, one's immediate living environment becomes a critical anchor for social life out of necessity. For major, however, both Cornell (0.049) and Dartmouth (0.044) show weak homophily, indicating that academic focus is not a primary driver of friendship in these larger, more diverse schools. Caltech's near-zero assortativity for both major and residence remains a surprising result, suggesting that despite its siloed structure, the small community size may necessitate a high degree of cross-group integration.

Modularity, which measures the strength of community divisions, reinforces the "Metropolis" hypothesis. Cornell has the highest modularity score (0.471), indicating its network is the most clearly and strongly partitioned. This is precisely what one would expect from a large city: it is not a monolithic entity but a collection of many distinct neighborhoods or sub-communities. Dartmouth's high modularity (0.431) likewise reflects a socially segmented campus, likely driven by its prominent Greek life. Caltech's slightly lower modularity, in contrast, suggests distinct communities, such as Houses or academic groups, but less sharply defined, possibly due to overlapping ties in a small network.

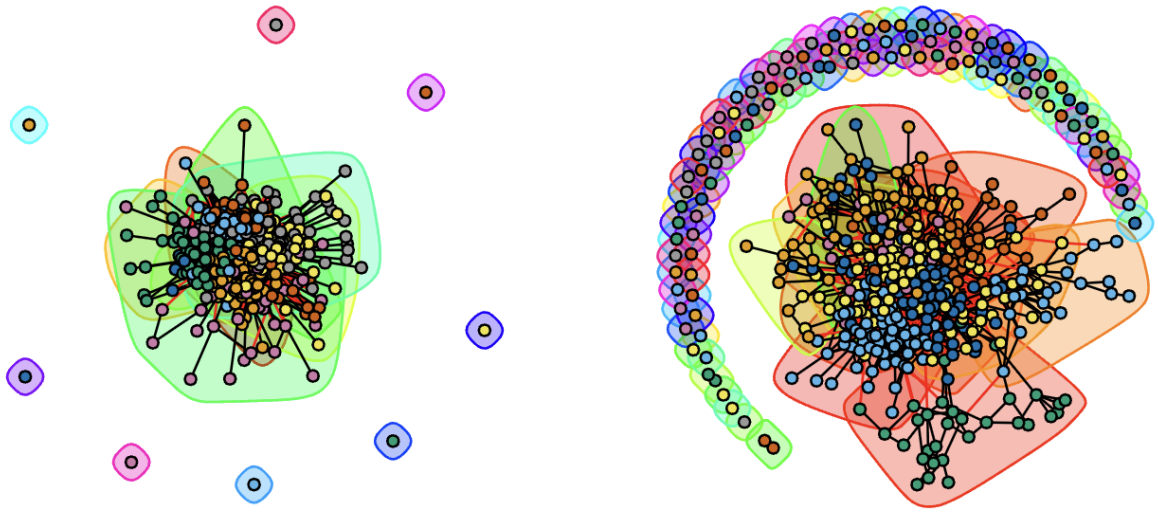

Figure 2: Louvain community detection on induced subgraphs of Caltech (left) and Dartmouth (right)

Despite its vastness, Cornell remains a "small world." Its average path length (2.876) and diameter (8) are remarkably similar to Dartmouth's (2.768 and 8, respectively). This indicates a highly efficient network structure where any student can reach another through a very short chain of acquaintances. Furthermore, both Cornell and Dartmouth exhibit mean betweenness centrality near zero. This is characteristic of large, decentralized networks where there are countless paths between individuals, meaning no single person or group is critical for holding the network together. This contrasts with Caltech (0.002), where the slightly higher betweenness suggests a greater role for individuals who act as brokers between its well-defined silos. Finally, degree variance is by far the highest at Cornell (7395.4), suggesting the greatest inequality in social connections, a classic feature of a metropolis that supports both hyper-connected social hubs and more isolated individuals.

In conclusion, these network vital signs paint clear, distinct portraits of three unique social worlds. Caltech is a dense, clustered "Silo". Dartmouth is a highly connected "Social Bubble" with strong residential bonds. And Cornell is a sprawling, fragmented, yet remarkably efficient "Metropolis." These findings provide a robust quantitative baseline for our primary analysis, where we use ERGMs to model the specific social forces that sculpt these fascinatingly different university personalities.

4.2 Modeling the Social Genomes: ERGM Coefficient Analysis

Due to limitations in RAM and multithreaded cores on our machine, we employed an online Colab Pro notebook to fit ERGMs (and even load the dataset for Cornell's adjacency matrix).

We ran the following code for each university (added here without executing):

```
## Load Dartmouth's graph
library(network)
library(ergm)

dartmouth_adj <- as.matrix(read.csv("Dartmouth6_adj.csv", header = FALSE,
    skip = 1))
dartmouth_node_info <- read.csv("Dartmouth6_local_info.csv")
dartmouth_g <- network(dartmouth_adj, directed = FALSE)

set.vertex.attribute(dartmouth_g, "year", dartmouth_node_info$year)
set.vertex.attribute(dartmouth_g, "residence", dartmouth_node_info$dorm)
set.vertex.attribute(dartmouth_g, "major", dartmouth_node_info$major)
set.vertex.attribute(dartmouth_g, "high_school", dartmouth_node_info$high_
    school)

dartmouth_model <- ergm(
  dartmouth_g ~ edges + nodematch("year") + nodematch("residence") +
    nodematch("major") + nodematch("high_school"),
  control = control.ergm(
    parallel = num_cores,
    parallel.type = "PSOCK",
    MCMLE.maxit = 100
  )
)
summary(dartmouth_model)
dartmouth_p <- plot_model_coefs(dartmouth_model, "Dartmouth")
grid.arrange(dartmouth_p, nrow=1)
```

Listing 1: Example ERGM fitting code for Dartmouth.

The resulting coefficients were collated and plotted (Figure 3).

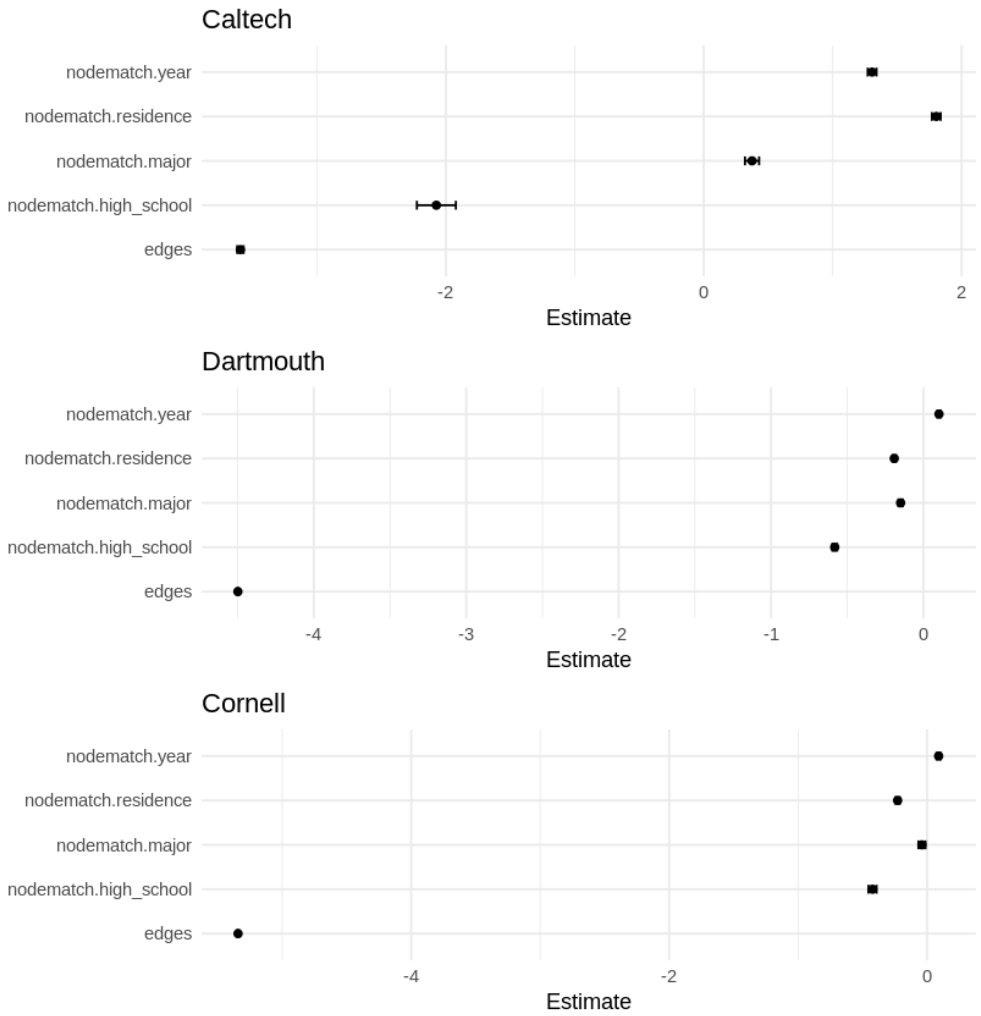

Figure 3: ERGM Model Coefficients for Caltech, Dartmouth, and Cornell. The plot displays the estimated coefficients and their confidence intervals for each predictor variable (social gene) in the model for the three universities.

4.2.1 Caltech: "The Focused Silo"

The fitted ERGM provides strong, quantitative support for a nuanced version of our "Focused Silo" hypothesis. The model reveals that the single most powerful driver of social connection at Caltech is shared housing, with the `nodematch('residence')` coefficient at a remarkably high +1.80. The effect of a shared class year is also a dominant force (+1.31), indicating a powerful cohort-based identity. In contrast, the `nodematch('major')` coefficient, while positive, is much more modest (+0.37), suggesting that academic discipline is a secondary factor compared to the intense social environment of residential and class groups. These forces operate against a significant baseline cost of tie formation (edges coeff: −3.59). Most strikingly, the model returns a strong negative coefficient for `nodematch('high_school')` (−2.07), implying that once the overwhelming effects of campus-based affiliations are accounted for, pre-college ties are effectively overwritten in favor of new connections.

4.2.2 Dartmouth: "The Social Bubble"

The ERGM results for Dartmouth present a "social genome" that decisively refutes a simple interpretation of the "Social Bubble" hypothesis. The model is dominated by an extremely negative `edges` coefficient of -4.50, indicating that friendship ties are exceptionally "costly" and unlikely to form without specific social mechanisms. Critically, our central hypothesis is unsupported, as the model finds that `nodematch('residence')` has a negative effect (-0.19) on tie formation. This statistically confirms that once other network dynamics are accounted for, residential and Greek life are not the primary drivers of friendship.

Furthermore, this pattern of non-institutional sorting continues across other attributes, with negative coefficients for both `nodematch('major')` (-0.15) and `nodematch('high_school')` (-0.58). The sole positive driver of connection is `nodematch('year')`, and its effect is minimal ($+0.10$). This complex genome strongly suggests that the primary drivers of friendship at Dartmouth are not the formal categories included in our model but rather unobserved social foci—such as specific clubs, sports teams, or cultural groups—that are powerful enough to render institutional assignments statistically irrelevant by comparison.

4.2.3 Cornell: "The Metropolis"

The ERGM results for Cornell provide a powerful confirmation of our "Metropolis" hypothesis, revealing a social landscape so vast and diffuse that formal institutional structures have almost no organizing power. The model is overwhelmingly dominated by a massive negative `edges` coefficient of -5.34, the most extreme of all three universities, signifying a network where the baseline probability of friendship formation is exceptionally low. Critically, the model shows that sharing a `residence` (-0.23), `major` (-0.04), or `high_school` (-0.42) all have negative effects on the likelihood of connection. The only faint signal of positive structure comes from `nodematch('year')`, and its coefficient is almost zero ($+0.09$). This "social genome," defined by extreme sparsity and the statistical irrelevance of formal campus affiliations, is the quintessential signature of a metropolis where social life is fragmented into countless voluntary, niche communities that operate independently of the university's administrative categories.

4.3 Comparison with Prior Work and Coefficient Validation

To contextualize and validate our findings, we compare our ERGM coefficients for Caltech with those reported in the foundational study by Traud, Mucha, and Porter (2012). This comparison reveals both a strong corroboration of the core findings and insightful differences likely attributable to model specification. The coefficients from both analyses are presented in Table 2.

ERGM Term	Our Model Estimate	Traud et al. (2012)
`edges`	−3.59	−4.98
`nodematch('residence')`	+1.80	+1.16
`nodematch('year')`	+1.31	+0.99
`nodematch('major')`	+0.37	+0.65
`nodematch('high_school')`	−2.07	+2.85

Table 2: Comparison of ERGM coefficients for Caltech between our model and those reported by Traud, Mucha, and Porter (2012).

Our analysis strongly aligns with the central conclusions of Traud et al. regarding the primary drivers of social life at Caltech. Both our model and the original study identify residence and year as the most powerful positive forces shaping friendship formation. In fact, our model finds even stronger effects for these two attributes (+1.80 for residence and +1.31 for year) than the original study, reinforcing the conclusion that Caltech's residential House system and a strong cohort identity are the dominant organizing principles of its social world. Furthermore, both models agree that a shared major has a consistent, positive, albeit weaker, effect on tie formation.

However, the comparison also reveals two significant differences likely attributable to model specification, which provide deeper insight into the network's structure. First, our model's edges term (-3.59) is substantially less negative than that of Traud et al. (-4.98). This is likely because their model included an additional term for network triangles (gwesp), which explicitly accounts for endogenous clustering. By modeling this "friend of a friend" effect directly, their edges term captures a purer, and thus lower, baseline propensity for tie formation. Our more parsimonious model, by not including a triangle term, likely absorbs some of this ambient clustering tendency into the other coefficients.

The most striking divergence lies in the effect of shared high school, for which our models report coefficients with opposite signs. Traud et al. found a very strong positive effect (+2.85), suggesting that pre-college ties are a powerful predictor of friendship. In stark contrast, our model returned a strong negative coefficient (-2.07). This dramatic difference highlights the sensitivity of ERGMs to model specification. A compelling interpretation is that our model, by finding much stronger effects for residence and year, successfully attributes so much of the network's structure to on-campus affiliations that it "explains away" any positive effect of high school ties, leaving a residual negative coefficient. This suggests that once a student is embedded in Caltech's potent residential and cohort-based social life, any tendency to associate with former high school peers disappears, indicating a powerful social environment that effectively overwrites prior affiliations in favor of new, campus-based connections.

In conclusion, while our model specification differs, our analysis confirms the most critical finding: the social world of Caltech is overwhelmingly shaped by its on-campus institutional structures. The divergences themselves, especially regarding the high school effect, provide a fascinating insight into the powerful socializing effect of the university environment itself.

4.4 Model Validation and Robustness Checks

To ensure the validity of our findings, we conducted a sensitivity analysis for the Dartmouth network.

4.4.1 Dartmouth Network Sensitivity Analysis

Initial exploration of the Dartmouth network revealed the presence of numerous isolated nodes and small components, which could potentially skew aggregate network metrics. To test the robustness of our "Social Bubble" hypothesis and ensure our findings were not an artifact of these data characteristics, we performed a sensitivity analysis by constructing an undergraduate-only subgraph (class years 2005-2009) to test whether the inclusion of potential graduate student or faculty accounts was influencing the results. We hypothesized that if the key signatures of the "Social Bubble" (such as high residential assortativity) were maintained or even strengthened in these core networks, it would provide powerful evidence for our central thesis.

Network Measure	**Dartmouth (Full)**	**Undergrad LCC**
Nodes	7,677	4,852
Edges	304,065	213,593
Density	0.010	0.018
Average Degree	79.215	88.043
Clustering	0.151	0.167
Assortativity (Residence)	0.118	0.126
Assortativity (Major)	0.044	0.052
Modularity	0.431	0.445
Average Path Length	2.768	2.533
Diameter	8	6
Mean Betweenness	0.000	0.000
Degree Variance	5,591.171	4,299.454

Table 3: Comparison of Dartmouth Network Statistics: Full Network vs. Undergraduate Largest Connected Component (LCC).

The results shown in Table 3 reveal a social environment that is substantially more intense and interconnected than the full dataset suggested. Filtering out the nearly 3,000 non-undergraduate or isolated nodes reveals a network that is not only smaller but structurally different, confirming that the raw data contained significant noise.

The most dramatic changes support a more potent "Social Bubble" hypothesis. Network density nearly doubled from 0.010 to 0.018, while the average degree of a student jumped from 79 to 88. This indicates that the core undergraduate students are far more tightly connected to each other than the full, noisy dataset implied. The social world is smaller and more intimate, as evidenced by the significant decreases in the average path length (from 2.77 to 2.53) and network diameter (from 8 to 6). It is socially "quicker" to traverse the undergraduate community, reinforcing the idea of a cohesive, self-contained bubble.

Furthermore, the mechanisms driving this bubble become clearer. The clustering coefficient increased to 0.167, and residence assortativity rose to 0.126. This shows that the tendency to form tight-knit, triangular relationships ("friends of friends") and to bond with others in the same residence or Greek house is even more pronounced among the undergraduate population. Interestingly, the degree variance decreased, suggesting that the undergraduate social experience is more equitable, with fewer extreme outliers in terms of connectivity compared to the full dataset which included non-student populations.

In conclusion, this sensitivity analysis demonstrates that the key characteristics of the "Social Bubble" are not just valid but are even more pronounced within the core undergraduate community. The act of filtering out peripheral nodes did not weaken the findings, but rather sharpened the focus on a dense, highly connected, and residentially-driven social ecosystem.

4.5 Subgroup Analysis: Exploring Class Year Dynamics at Dartmouth

One of the core questions in understanding a university's social structure is whether students tend to cluster based on shared attributes such as class year. Dartmouth presents an interesting case: unlike many institutions where students are referred to as freshmen, sophomores, juniors, or seniors, Dartmouth students identify by their graduating class year (e.g., "'08s" or "'05s"). This reflects a strong cohort identity that runs deeper than academic classification. This sentiment is shaped by shared experiences unique to Dartmouth: walking around the Homecoming bonfire together as first-years, returning for Sophomore Summer while most of campus is away, and going through Greek rush during sophomore fall.

This analysis investigates whether class year influences social connectivity at Dartmouth by comparing students in the Class of 2008 and the Class of 2005. Using degree centrality as a proxy for social embeddedness within the campus Facebook network, we explore whether students' academic class year corresponds to differences in social ties.

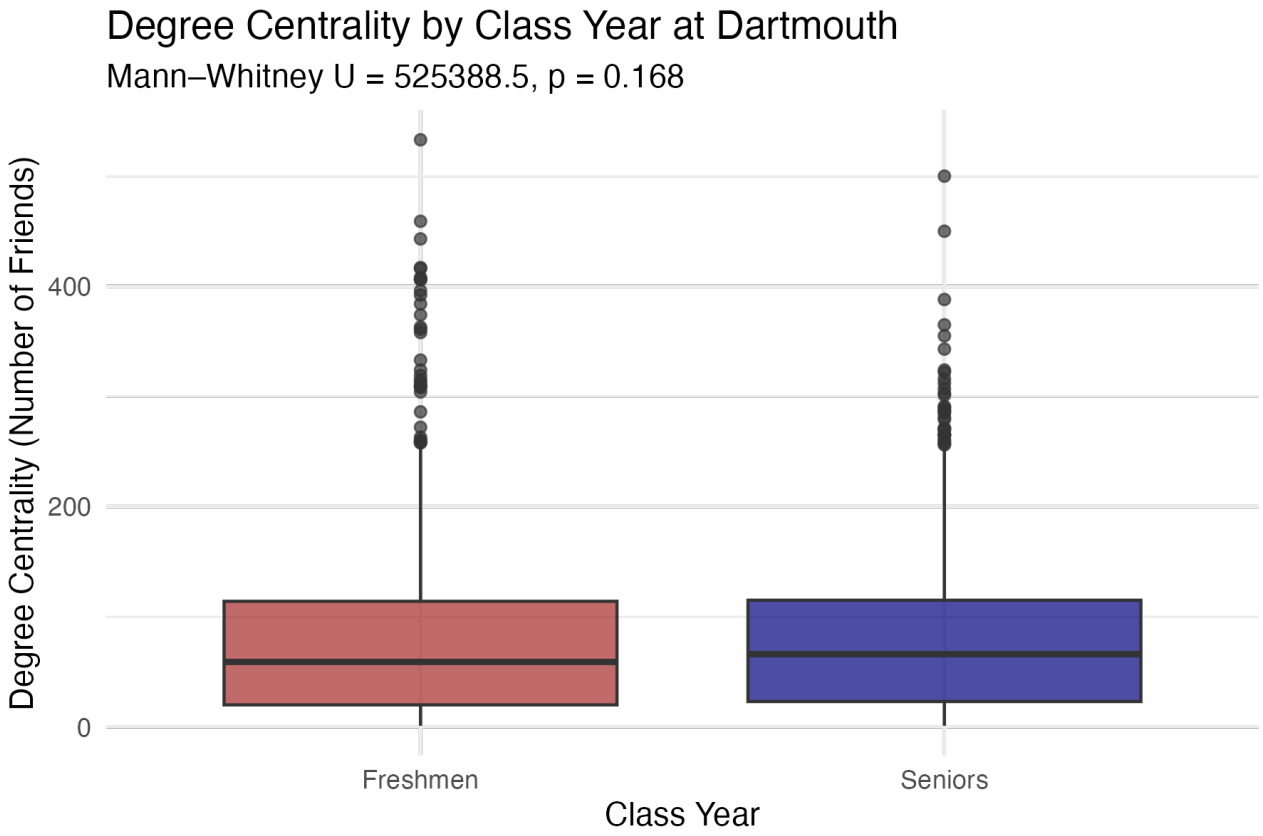

Figure 4: Comparison of degree centrality, measured as the number of social connections within the Dartmouth Facebook network, between students in the Class of 2008 (First-Years) and the Class of 2005 (Seniors)

Table 4: Summary Statistics for Degree Centrality by Class Year at Dartmouth

Class Year	N	Median Degree	Mean Degree	SD Degree
Freshmen	1,079	59.00	77.71	76.00
Seniors	1,009	66.00	79.57	70.97

The boxplot (Figure 4) and summary statistics (Table 4) show that Seniors have a slightly higher median degree (66 vs. 59), and the means are nearly identical (79.57 vs. 77.71). However, the distributions have significant overlap, and both groups exhibit considerable spread and numerous outliers, suggesting variability in individual connectedness.

A Mann–Whitney U test was used to compare the degree distributions and yielded a non-significant result ($U = 525388.5$, $p = 0.168$). This means that there is no statistically significant difference in degree centrality between Freshmen and Seniors in this sample.

The distributions of both groups reveal considerable overlap and multiple high-degree outliers, suggesting that social connectedness is not strongly stratified by class year at Dartmouth. This pattern may reflect Dartmouth's relatively close knit environment, where cross-cohort interaction is common. It may also indicate that social networks mature quickly and plateau early, meaning even first-year students can become highly connected soon after arrival.

However, it is important to note that this analysis is univariate, meaning we did not control for other potentially influential variables such as dorm assignment, gender, or academic major. These factors likely interact with class year in shaping students' social networks, and a multivariate analysis could provide a more nuanced understanding.

These findings contribute to the broader goal of this project: to quantify institutional "social genomes." The lack of a sharp divide by class year at Dartmouth stands in contrast to other institutions where hierarchical or cohort-based segmentation may play a larger role.

5 Conclusion

This project set out to determine if the unique social "personality" of a university could be quantified, modeled, and compared. By introducing and applying the concept of a "social genome", we have shown that the answer is a definitive yes. The comparative analysis of Caltech, Cornell, and Dartmouth using the Facebook100 dataset reveals three profoundly different social worlds, each sculpted by a distinct combination of organizing forces.

Caltech's "social genome" is that of a "Focused Silo," where social life is powerfully and predictably structured by on-campus affiliations; its residential House system and a strong cohort identity are the dominant drivers of connection, so much so that they appear to completely overwrite pre-college ties. Cornell's genome is the quintessential "Metropolis," a social landscape so vast and diffuse that formal institutional containers like residence and major lose their organizing power, resulting in a sparse network where friendships are likely forged in countless unobserved, niche communities. Perhaps most surprisingly, Dartmouth's genome is not the simple "Social Bubble" we hypothesized. Instead, our model reveals a complex social ecosystem where formal structures like housing and academics are not the primary drivers of friendship, pointing to a highly fluid social world where connections are organized around other, unobserved social foci.

Ultimately, this research makes a significant contribution by moving beyond descriptive metrics to provide a robust, model-based framework for understanding institutional variation. By validating the "social genome" as a powerful analytical concept, we have demonstrated a replicable method for identifying and quantifying the invisible, and sometimes counter-intuitive, forces that make each university a unique social world. This approach not only deepens our understanding of network formation in these critical communities but also opens the door to a more data-informed conversation about the very nature of campus life.

5.1 Limitations of the Model and Analysis

Several limitations warrant mention. Our ERGM analysis is cross-sectional, meaning it captures a single snapshot in time and cannot formally disentangle the processes of social selection (choosing friends who are like you) from social influence (becoming more like your friends over time). A longitudinal analysis would be required to address such questions of causality. Furthermore, our model relies on a specified set of attributes; other unobserved factors, such as

participation in clubs, sports, or artistic groups, undoubtedly also play a significant role in shaping these networks. Finally, as with any statistical model, our ERGM is a simplification of a complex reality, and the "social genome" it describes should be understood as a powerful but incomplete representation of a university's social life.

5.2 Future Directions

This study could be expanded in several ways. Applying the "social genome" model to all 100 universities in the dataset would allow for a large-scale clustering of institutions based on their social signatures. The dataset's primary limitation is its static nature; a future study would involve analyzing a longitudinal dataset to observe how a university's social genome evolves over a student's four-year career or in response to institutional changes, such as the construction of new residential colleges. Furthermore, obtaining "better" data with more detailed, non-anonymized attributes (e.g., specific fraternity/sorority membership) would allow for an even more fine-grained model of the forces shaping campus social life, and thereby help incoming college students design a more intentional and fulfilling college experience by choosing institutions whose social personalities align with their own.

Data and Code Availability

The Facebook100 network data used in this study was retrieved for this project from the Internet Archive. The Python and R scripts used for data preprocessing and analysis are provided in the Appendix.

Acknowledgments

We thank Professor Peter J. Mucha for his assistance with data processing and for providing valuable comments and suggestions that greatly improved this research.

References

[1] Hommes, J., Rienties, B., de-Vries, P., & van-den-Bossche, P. (2012). The impact of social networks on the development of student's academic life and performance. *Procedia-Social and Behavioral Sciences, 55*, 878-887.

[2] Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences, 109*(1), 68-72.

[3] Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2013). *Exponential random graph models for social networks: Theory, methods, and applications.* Cambridge University Press.

[4] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27*, 415-444.

[5] Traud, A. L., Mucha, P. J., & Porter, M. A. (2012). Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications, 391*(16), 4165–4180.

[6] Wimmer, A., & Lewis, K. (2010). Beyond and Below Racial Homophily: ERG Models of a Friendship Network in a Racially Diverse Dorm. *American Journal of Sociology, 116*(2), 583-642.

A Data Preprocessing

The original Facebook100 dataset was provided in MATLAB's `.mat` format. The following Python script was used to parse these files, extract the adjacency matrices and node attributes, and convert them into a standardized `.csv` format for easier manipulation in R. The script iterates through all `.mat` files in a directory, processes each school's data, and then combines them into two master files: one for all nodes and one for all edges across the universities.

```
import scipy.io
import pandas as pd
import numpy as np
import os
from scipy.sparse import triu, issparse

# Define input and output directories
input_folder = 'fb100'
output_folder = 'fb100-csv'
os.makedirs(output_folder, exist_ok=True)

# Initialize lists to hold dataframes
all_nodes = []
all_edges = []

# Define column headers for node attributes
columns = ['gender', 'status', 'major', 'second_major', 'dorm', 'year', '
    high_school']

# Loop through each .mat file in the input directory
for filename in os.listdir(input_folder):
    if filename.endswith('.mat'):
        school = filename.replace('.mat', '')
        mat_path = os.path.join(input_folder, filename)
        print(f'Processing {school}...')

        try:
            mat = scipy.io.loadmat(mat_path)
            # Check for required data structures
            if 'A' not in mat or 'local_info' not in mat:
                print(f"Skipping {school} - missing 'A' or 'local_info'")
                continue

            A = mat['A']
            info = mat['local_info']

            if issparse(A):
                A = A.tocsr()

            # Process nodes
```

```
            nodes = pd.DataFrame(info, columns=columns)
            nodes['school'] = school
            nodes['node_id'] = nodes.index
            all_nodes.append(nodes)

            # Process edges from the upper triangle of the adjacency matrix
            triu_A = triu(A, k=1)
            sources, targets = triu_A.nonzero()
            edges = pd.DataFrame({
                'source': sources,
                'target': targets,
                'school': school
            })
            all_edges.append(edges)

        except Exception as e:
            print(f"Error processing {school}: {e}")
            continue

# Concatenate all dataframes and save to CSV
combined_nodes = pd.concat(all_nodes, ignore_index=True)
combined_edges = pd.concat(all_edges, ignore_index=True)

combined_nodes.to_csv(os.path.join(output_folder, 'facebook100_all_nodes.csv'), index=False)
combined_edges.to_csv(os.path.join(output_folder, 'facebook100_all_edges.csv'), index=False)

print("Processing complete.")
```

Listing 2: Python script for converting ‘.mat‘ files to ‘.csv‘.

B Full ERGM Model Summaries

This appendix provides the detailed statistical output from the final Exponential-Family Random Graph Model (ERGM) fits for each of the three universities. The tables below correspond to the standard summary output in R and display the precise coefficient estimates, standard errors, z-values, and p-values for each term in the models. This data forms the basis for the coefficient plots and interpretations presented in the main body of the paper.

The Null and Residual Deviance values for each model are omitted for brevity but were used to assess overall model fit. The p-values confirm that nearly all included terms are statistically significant predictors of tie formation across all three networks.

Table 5: ERGM Summary for Caltech Network

Term	Estimate	Std. Error	z value	Pr(>|z|)
edges	−3.593	0.0125	−286.47	$< 0.0001^{***}$
nodematch('year')	1.306	0.0175	74.81	$< 0.0001^{***}$
nodematch('residence')	1.804	0.0173	104.03	$< 0.0001^{***}$
nodematch('major')	0.375	0.0280	13.39	$< 0.0001^{***}$
nodematch('high_school')	−2.073	0.0771	−26.88	$< 0.0001^{***}$

Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6: ERGM Summary for Dartmouth Network

Term	Estimate	Std. Error	z value	Pr(>|z|)
edges	−4.498	0.0022	−2076.28	$< 0.0001^{***}$
nodematch('year')	0.102	0.0056	18.26	$< 0.0001^{***}$
nodematch('residence')	−0.191	0.0054	−35.67	$< 0.0001^{***}$
nodematch('major')	−0.150	0.0070	−21.28	$< 0.0001^{***}$
nodematch('high_school')	−0.582	0.0078	−74.44	$< 0.0001^{***}$

Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7: ERGM Summary for Cornell Network

Term	Estimate	Std. Error	z value	Pr(>|z|)
edges	−5.342	0.0032	−1655.51	$< 0.0001^{***}$
nodematch('year')	0.089	0.0079	11.22	$< 0.0001^{***}$
nodematch('residence')	−0.229	0.0079	−28.97	$< 0.0001^{***}$
nodematch('major')	−0.041	0.0138	−2.93	0.0034^{**}
nodematch('high_school')	−0.423	0.0162	−26.22	$< 0.0001^{***}$

Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$