From Pedagogy to Prestige: Predicting Prestige Using Academic Hiring Networks

Tim Poštuvan
EPFL
tim.postuvan@epfl.ch

Veniamin Veselovsky
EPFL
veniamin.veselovsky@epfl.ch

Abstract

The prestige of a university, often associated with its ranking, significantly influences a student's university selection process and future career prospects. In this study, we employ network analysis and machine learning to predict university prestige based on faculty hiring networks across various academic disciplines in the United States. Using an extensive dataset representing these hiring decisions, we begin by exploring the structural properties of the network, comparing its features across eight domains of academia. We further utilize graph-based machine learning models, including a standard message-passing Graph Convolutional Network (GCN) to predict prestige scores for institutions in the network. Our exploration reveals differences between disciplines in terms of prestige and other structural properties, while our predictive models demonstrate the potential of employing faculty hiring networks for estimating university prestige. In particular, we find that a model trained on a specific set of domains generalizes to new domains, and graph structure dramatically improves prestige prediction. This work provides valuable insights into the inter-institutional dynamics of academia and offers a novel approach to evaluating university prestige. Our code is available at https://github.com/vminvsky/nml-project.

1 Introduction

When selecting a university, prestige has long been an important factor in student decision-making (Harahap et al., 2018). A university's ranking plays an important role in a student's future job search and income (Jung and Lee, 2016). Typically, studies have used various institution-level measures to predict the prestige of a university (Cyrenne and Grant, 2009). In this paper, we explore whether the network structure of faculty hiring can be used to recover the prestige scores for the universities in the US faculty hiring dataset (Wapman et al., 2022). We consider the academia as a whole, as well as eight academic disciplines separately. Throughout our analysis, we focus on two main research questions:

RQ1: Does prestige generalize across disciplines? If not, when does it?

RQ2: How well can we predict prestige purely based on the hiring networks?

2 Exploration

We begin our analysis by providing a high-level description of US faculty hiring networks (Wapman et al., 2022). In this section, we first define the graphs and illustrate their quirks and features by exploring their structure. Then, we compare hiring decisions within academia (i.e., academic graph) with various random graphs models. Finally, we finish with an analysis of clusters and variance in prestige.

2.1 Data

US faculty hiring dataset contains a network consisting of all university faculty hiring decisions in the United States between 2011–2020. The nodes represent institutions and directed weighted edges (u, v, w) represent hires (i.e., w graduates from institution u were hired by institution v). We split the networks based on individual domains, while also constructing a graph with information from all domains (i.e., academic graph). In total, there are eight domains included in the dataset: applied sciences, education, engineering, humanities, mathematics and computing, medicine and health, natural sciences, and social sciences. To supplement network structure with node features, we add statistics about institutions that are at the disposal in the dataset, namely non-attrition events, attrition events, and production rank. We then filtered out all nodes that did not have corresponding node fea-

tures from our graph, leaving us with 359 nodes in the academic graph.

2.2 Initial analysis

We begin our analysis by providing graph statistics across the different domains for unweighted graphs (i.e., disregarding weights w) as well as weighted graphs. In particular, we examine the number of nodes, number of edges, density, average degree, clustering coefficient, average shortest path, diameter, and algebraic connectivity on unweighted graphs. Furthermore, we compute the number of edges, average degree, clustering coefficient, and algebraic connectivity also on weighted graphs. The few most important statistics are shown in Figure 1. [1]

The graphs are quite small, having up to 277 nodes and 14451 edges. Nevertheless, all domains have a very high density, ranging from 0.26 to 0.38. This is an important observation with a serious impact on further exploitation analyses. Such high densities imply that over-smoothing may be a persistent problem since representations of nodes will be influenced by representations of nearly all other nodes. Otherwise, it can be observed that there exist large differences between different domains, indicating that graphs of different domains have different structural properties. The average shortest path is 1.75 for the education domain but it drops to 1.625 for social sciences. Additionally, the humanities typically have the highest weighted and non-weighted clustering coefficients, over two times larger than medicine and health. Also, the distribution of prestige differs across domains, with applied sciences averaging a score of 0.39 and engineering peaking at an average score of 0.51. This analysis suggests that there are structural differences across the different domains and generalization among domains might be limited.

Comparison with network models Approximating our graphs with networks can give more insights into graph structure. For this purpose, we try to match the structural properties of academic graph by employing Erdös-Rényi, Watts-Strogatz, and Barabási-Albert models. Results are presented in Table 1. Surprisingly, the Watts-Strogatz graph almost perfectly fits the number of edges, average degree, average shortest path length, and clustering coefficient. Alternatively, the other two graph models poorly approximate the average shortest path length and/or clustering coefficients of our graph. This indicates that our graph is small-world as quintessential for Watts-Strogatz graphs. The claim can be also confirmed by checking small-world property: $\overline{d} \sim \frac{\ln N}{\ln \overline{k}}$, where $\overline{d}$ is the average shortest path length, N is the number of nodes, and $\overline{k}$ is the average degree. For academic graph, the average shortest path length of 1.59 is certainly close to the theoretical value $\frac{\ln 359}{\ln 151.4} = 1.17$.

	# of Edges	Avg. Degree	Avg. Shortest Path Length	Cluster Coeff.
Original	27170	151	1.6	0.76
ER ($p = 0.42$)	27170	151	1.2	0.42
WS ($\beta = 0$)	26925	150	1.7	0.74
BA	27284	152	3.3	0.10

Table 1: Comparison of academic graph with Erdös-Rényi (ER), Watts-Strogatz (WS), and Barabási-Albert (BA) models.

Clustering and spectral analysis For a more concrete understanding of connections between institutions and their spread in the network, we use spectral analysis and spectral clustering. In general, academic graph is well connected with algebraic connectivity of 0.59. On the other hand, spectral clustering gives more fine-grained insights into the data. It leverages the eigenvalues of the adjacency matrix to partition data points into distinctive clusters. In Table 2, we report the five most prestigious universities in five selected clusters on academic graph.[2] The clusters capture similarities among institutions based on geography (cluster 4), primary domain and geography (cluster 13), religion (clusters 4 and 14), and medicine (cluster 3).

Variance in prestige Since our downstream task is to predict the prestige of an institution depending on co-hiring patterns, we explore the difference in prestige between hiring and graduating institutions. In other words, we measure the variance in "prestige score" between the two institutions, which has been referred to as ubiquitous hirings of prestige by previous work (Wapman et al., 2022). Formally, variance in prestige is calculated using the following formula:

$$\mathrm{Var}(u) = \frac{1}{\sum_{v \in \mathcal{N}_u} w(v,u)} \sum_{v \in \mathcal{N}_u} w(v,u)(p_u - p_v)^2, \quad (1)$$

[1] All statistics are available on the GitHub repository.

[2] All clusters are available on our GitHub repository.

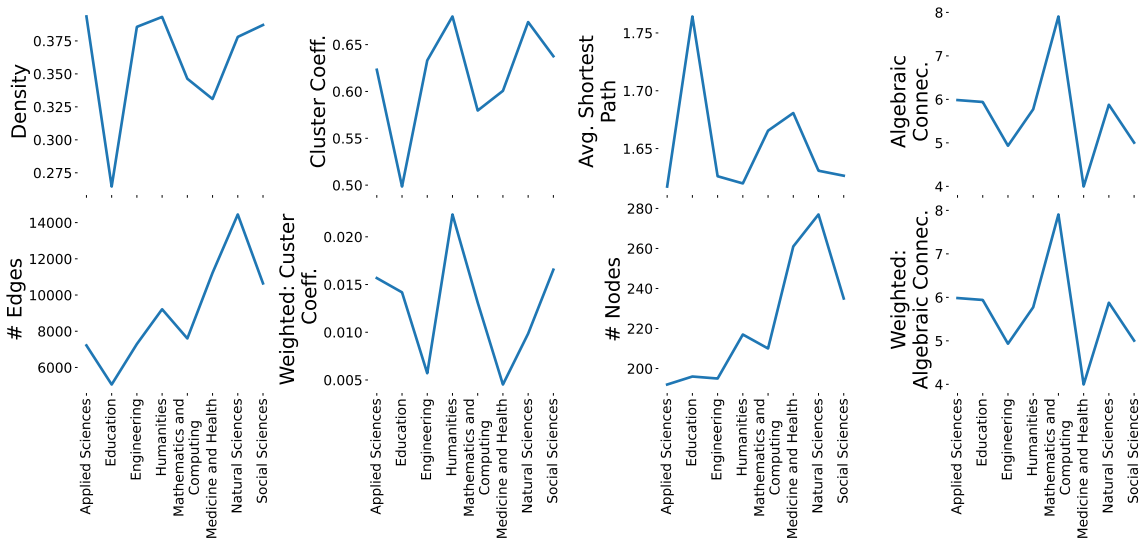

Figure 1: Graph statistics across different academic domains.

Institution	Cluster	Prestige
Washington University St. Louis	3	0.68
Albert Einstein College of Medicine	3	0.54
New York Medical College	3	0.49
SUNY Downstate Health Sciences	3	0.59
University of Maryland Baltimore	3	0.53
Union Theological Seminary	4	0.73
Drew	4	0.46
Chicago Theological Seminary	4	0.47
Princeton Theological Seminary	4	0.69
Emory	4	0.56
UPenn	6	0.79
Columbia	6	0.74
Harvard	6	0.91
University of Chicago	6	0.86
Yale	6	0.84
Stanford	13	0.90
Caltech	13	0.90
Princeton	13	0.92
UC Berkeley	13	0.92
MIT	13	0.92
Catholic University of America	14	0.52
Georgetown	14	0.52
Jewish Theological Seminary	14	0.57
Smith College	14	0.65
Alfred	14	0.49

Table 2: Spectral clustering of Academia graph.

where $\mathcal{N}_u$ is the set of directed edges pointing into u, p_u represents the prestige of the institution u, and $w(v, u)$ denotes the number of directed edges from v to u.

To verify that institutions hire from other institutions with similar prestige, we compare the actual variance of academic graph with the variance of the academic graph where prestige scores are randomly permuted. A comparison between the two distributions is illustrated in Figure 2. We find that the random reassignment resulted in a higher variance in prestige (i.e., an average variance of 0.056, while the real graph has an average variance of 0.036). Qualitatively, this means that institutions are more likely to hire faculty from universities of a similar prestige level. This implies that machine learning models that can incorporate graph structure into predictions (e.g., graph neural networks) should more accurately predict prestige than other models that do not consider graph structure (e.g., linear regression).

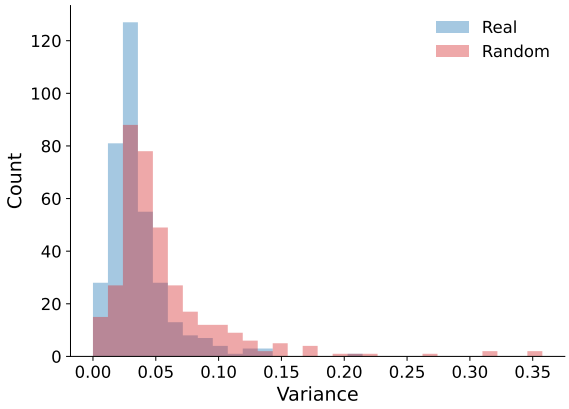

Figure 2: Variance in prestige between hiring institutions and degree-providing institutions on academic graph.

3 Exploitation

Using the observations from the exploration, in this section we attempt to predict the prestige of institutions. First, we examine how well a simple linear regression model on node-level features can capture prestige. Afterwards, we test more powerful

GNN models that include graph structure and apply them to different versions of the graphs.

3.1 Models

Average baseline Since the dataset is relatively new, there have not been any attempts to predict prestige with machine learning models. Therefore, our simplest baseline predicts the prestige of all institutions as the average prestige in the training set. Although the performance of this model should not be particularly good, it serves as an important baseline to verify the correctness of other models.

Linear regression baseline A more sophisticated baseline that already considers node attributes is linear regression. Linear regression provides a good baseline for our GNN models as it does not use graph information. Thus, the benefits of incorporating graph structure information are clearly visible when comparing performances.

GNN models In our study, we use two common GNN architectures: standard message-passing Graph Convolutional Network (GCN) (Kipf and Welling, 2017) and attention-based Graph Attention Network (GAT) (Veličković et al., 2018). As our dataset has edge weights, we also employ the former on weighted graphs. We hypothesize that GCN on weighted graphs would perform better since the aggregation function is going to consider neighboring node representations according to the following weight:

$$w_{\text{GNN}}(i,j) = \frac{w(i,j)}{\sqrt{deg(i) \cdot deg(j)}}, \qquad (2)$$

where and $w(i,j)$ represents the number of directed edges from i to j, and $deg(i)$ denotes the weighted degree of the node i. Equation 2 resembles pointwise mutual information (PMI):

$$pmi(i,j) = \log_2 \frac{p(i,j)}{p(i) \cdot p(j)}, \qquad (3)$$

where $p(i,j)$ is joint probability mass function, and $p(i)$ is probability mass function. Therefore, if an institution i hires more personnel from another institution j (i.e., the edge has a higher weight), the node representation of i is going to influence the node representation of j more. Conversely, if institutions have many hires (i.e., the corresponding nodes have greater degrees), influence is going to be lower.

3.2 Data Pre-processing

Feature engineering The dataset contains only a few node attributes: non-attrition events, attrition events, and production rank. While they provide some additional information about nodes, they are not particularly useful for predicting prestige. Concretely, a linear regression model trained only on those three attributes performs worse than the average baseline. Therefore, we enrich node features with additional topological features: betweenness centrality (Brandes, 2001), eigenvector centrality (Bonacich, 1987), clustering coefficient, degree centrality, and closeness centrality (Freeman et al., 2002). We do not use unsupervised node2vec (Grover and Leskovec, 2016) features as they would likely not be able to capture meaningful relationships due to the high density of our graphs.

Graph processing We perform experiments on three versions of the graphs as not all architectures can be applied to weighted graphs. Furthermore, we expect that graph processing is going to play a vital role in performance on the downstream task. We generate the following three graph types: weighted graphs, unweighted graphs, and unweighted sparsified graphs. Weighted graphs are graphs directly constructed from the dataset, where the edge weight represents the number of hires between the corresponding two institutions. Unweighted graphs are exactly the same as the weighted ones, except that all edge weights are set to 1. Since unweighted graphs are very dense, neighboring nodes can not be distinguished well. Therefore, we propose unweighted sparsified graphs, where all edges that have a weight lower than a threshold T are removed, and the remaining ones are treated as in the case of unweighted graphs (i.e., weights of all remaining edges are set to 1).

3.3 Experimental Setting

Implementation All experiments are implemented in the PyTorch Geometric framework (Fey and Lenssen, 2019). For GAT, and GCN on unweighted graphs, the official implementations are used. On the other hand, we implement GCN on weighted graphs ourselves as the official implementation does not consider weighted degrees. All models are trained with the batch size of 1 graph for 500 epochs from random initialization. They are optimized by AdamW (Loshchilov and Hutter, 2019) using the following parameters: β_1 = 0.9, β_2 = 0.999, ϵ = 1e-6, and no weight decay. The learn-

ing rates of linear regression, GCN, GAT, and GCN on weighted graphs are 0.005, 0.002, 0.005, and 0.001, respectively. All GNN architectures have 2 layers and hidden dimension of 64. Since prestige scores are within $[0, 1]$ range, models benefit from a sigmoid activation function on top of regression scores. We employ mean squared error (MSE) as the training loss as well as the evaluation metric. The dataset is split into training/validation/testing sets according to the following proportions: 70% in the training set, 10% in the validation set, and 20% in the testing set. Experiments are conducted in two experimental settings: transductive and inductive. In the transductive setting, nodes of each graph are split into disjoint sets (i.e., each graph independently conforms to aforementioned proportions), while in inductive settings, each set contains nodes in only a certain proportion of graphs (i.e., all nodes within a graph belong only to a single set). We compare models based on testing MSE, where each testing MSE is computed by the best version of the model during training according to validation MSEs. To account for randomness, all experiments are performed five times, and the median is reported. The threshold for unweighted sparsified graphs is set to $T = 30$.

3.4 Results

The results of the experiments in the transductive setting are shown in Table 3 and inductive setting in Table 4. Average baselines show that prestige is not uniformly distributed and is more concentrated. Actual MSE equals ~ 0.04, however, uniform distribution has theoretical MSE of ~ 0.08. In both settings, linear regression already achieves considerably better results by roughly halving the MSE of average baselines. This indicates that already computed topological features are decent predictors of prestige. GNN models can not learn anything on unweighted graphs and perform poorly, no matter their capacity and expressiveness. Probably, heavy over-smoothing issue (Chen et al., 2020) occurs due to very dense graphs (i.e., all nodes are connected to almost half of the others), where all neighbors are considered equally. Sparsifying the graphs immediately results in performance superior to graph-oblivious baselines. In the transductive setting, there is a significant boost for ~ 0.0050, while in the inductive setting, GNN models are only slightly better than linear regression. When sparsifying the graphs, the threshold plays an important role as setting it incorrectly can result in too dense or too sparse graphs. The technique is probably less efficient in the inductive setting as the same threshold is not best for all graphs, and due to the way nodes are split, this is more evident in the testing performance. GCN on weighted graphs outperforms all other models by a large margin. It roughly halves the MSEs of both baselines, while still being significantly superior to other GNN models (i.e., ~ 0.0040 in the transductive setting, and ~ 0.0100 in the inductive setting). This is expected as weighted graphs contain more information than the graphs of the other two types, or not knowing the structure at all. Results support the findings about variance in prestige as graph-aware models perform much better than baselines if applied correctly. Furthermore, the last model highlights that properly pre-processing the data is extremely important.

Model	MSE
Average	0.0401
Linear regression	0.0207
GCN + unweighted graphs	0.2158
GAT + unweighted graphs	0.2369
GCN + unweighted sparsified graphs	0.0169
GAT + unweighted sparsified graphs	0.0153
GCN + weighted graphs	**0.0116**

Table 3: Results of prestige prediction in the transductive experimental setting.

Model	MSE
Average	0.0420
Linear regression	0.0243
GCN + unweighted graphs	0.2360
GAT + unweighted graphs	0.2490
GCN + unweighted sparsified graphs	0.0218
GAT + unweighted sparsified graphs	0.0223
GCN + weighted graphs	**0.0122**

Table 4: Results of prestige prediction in the inductive experimental setting.

4 Discussion and Conclusion

In this work, we study the recently released US faculty hiring dataset. We present an exploration of the hiring graphs as well as an exploitation of their structural ability in predicting prestige. We find that the graphs are clustered into coherent groupings based on geography, the primary domain of institutions, and prestige; that different domains

exhibit different macro-structures; and that institutions indeed hire from institutions with a similar prestige by inspecting the variance in prestige.

We now answer the questions we posed at the outset of the paper. First, in **RQ1** we asked if prestige is universal across different domains. To answer the question, we conducted experiments in the inductive experimental setting where domains are split into disjoint sets: models are trained on a subset of domains and evaluated on the remaining ones. This setting permitted us to explore how well structural concepts of "prestige" generalized to other domains. We found that prestige does generalize across domains, achieving better scores than linear regression. In **RQ2** we wanted to see the gain of including structural features for predicting academic prestige. In this analysis, we compared a GNN approach with a more simple linear regression baseline. The results show the GNN model, which exploits local graph structure, was capable to perform more than two times better than linear regression, which uses solely graph-level features.

Weaknesses Our work also has a few weaknesses, which need to be considered when readings the findings. First, our approach does not explore which causal mechanisms are at play in forming prestige. Instead, this ML-focused approach explores the correlations and can not tell us if the network dynamics cause prestige to change. More explicitly, we can not rule out the possibility that it is not the network dynamics causing prestige but instead prestige causing the existing network dynamics.

Another concern we have is that the underlying measure of prestige is not accurate. This compound metric might not be defined by the levels of "prestige" we truly care about. For example, one study found that if you took two universities one prestigious and one not, then there were often little differences in academic rigour (Cyrenne and Grant, 2009). Consequently, prestige is really a complex metric and we require a better and more interpretable measure for its estimation.

The final problem that arose when conducting the analysis in this paper was that our models were not particularly robust. Sometimes the models did not manage to converge, which is also the reason that we report the median of five runs.

References

Phillip Bonacich. 1987. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182.

Ulrik Brandes. 2001. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3438–3445.

Philippe Cyrenne and Hugh Grant. 2009. University decision making and prestige: An empirical study. *Economics of Education Review*, 28(2):237–248.

Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*.

Linton C Freeman et al. 2002. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1:238–263.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

D Harahap, Ratih Hurriyati, Vanessa Gaffar, and Dita Amanah. 2018. The impact of word of mouth and university reputation on student decision to study at university. *Management science letters*, 8(6):649–658.

Jisun Jung and Soo Jeung Lee. 2016. Influence of university prestige on graduate wage and job satisfaction: the case of south korea. *Journal of Higher Education Policy and Management*, 38(3):297–315.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

K Hunter Wapman, Sam Zhang, Aaron Clauset, and Daniel B Larremore. 2022. Quantifying hierarchy and dynamics in us faculty hiring and retention. *Nature*, 610(7930):120–127.