Online Social Community Neighborhood Formation

Jiarui Wang[1][0009-0007-2528-7473], George Barnett[1][0000-0002-7511-1886], Norman Matloff[1][0000-0001-9179-6785], and S. Felix Wu[2][0000-0001-6033-5353]

[1] University of California, Davis, Davis CA 95616, USA
{jrwwang,gabarnett,nsmatloff}@ucdavis.edu
[2] National Cheng-Kung University, Tainan City, Taiwan
sfelixwu@gs.ncku.edu.tw

Abstract. Online social networks (OSNs) provide community platforms that engage users. The public page is a popular example. These pages are connected through "like" relationships, creating online community networks and neighborhoods. We investigated the pivotal features influencing link formation and neighborhood structuring within the page graph by exploring a series of potential features, both graph-based and content-based. Our methodology combines node similarity and Graph Neural Networks to perform link prediction. We identified the page state label as the single most accurate predictor in link prediction tasks, which also is the most efficient feature with the smallest number of classes. Moreover, we observe that augmenting the page state label feature with the page node degree and page city population features further enhances link prediction accuracy. Page location label shows a strong effect on pages connecting with their neighbors.

Keywords: Online social networks · Network formation · Location · Link prediction.

1 Introduction

In the past decade, online social networks (OSNs) have witnessed exponential growth, attracting billions of users worldwide. These platforms empower individuals to create profiles, establish connections, and share content, offering unparalleled access without the traditional constraints of time and location associated with offline social groups. Users can effortlessly connect with others globally who share similar interests, fostering the rapid expansion of OSNs.

A prevalent activity on these online social platforms involves individual users setting up personal profiles, connecting with friends or strangers, and sharing content. These individuals form the basis of online social networks, with their numbers indicative of the platforms' business potential, as they represent prospective consumers for a wide array of products and services. This vast user base attracts a variety of entities, including businesses, non-profit organizations, and governmental bodies, all seeking to leverage these platforms for their respective interests. These entities, along with individual users, establish various online social communities to cater to specific interests. These communities range from corporate and non-profit organization pages to user-created groups focusing on shared interests like neighborhood activities, workplace connections, and hobbies such as animal enthusiasts.

Numerous offline groups and communities have established their presence online through information pages or discussion forums. Additionally, the internet has seen the birth of myriad communities and groups that operate exclusively online, without any offline interactions. The rapid growth and sheer volume of these online social communities are remarkable, especially considering their relatively brief history. Unlike their offline counterparts, online communities face no constraints related to time or location, allowing for unlimited connections and interactions with other online entities. This paper delves into the dynamics of connections between various online social communities.

This study specifically focuses on public pages, which serve as a platform for disseminating information, facilitating user discussions, spreading news, and promoting businesses or public relations activities. Like individual users on social media, these pages can like or follow other pages, creating a network of connections among online social communities. This network, in turn, forms a vast graph of online social community interactions. Our research aims to uncover the pivotal factors that influence these connections and the development of neighborhoods within the online social community landscape.

In this study, we aim to contribute to the understanding of online social communities by investigating a range of page features to determine their impact on the formation of page neighborhoods. This is achieved through a methodology that applies link prediction techniques to each individual feature. We identified the page state label as the single most accurate predictor in link prediction tasks, which also is the most efficient feature with the smallest number of classes. Additionally, we find that a combination of features—specifically the page state label, page node degree, and page city population—yields the best performance in link prediction accuracy.

2 Data Description

2.1 Data Acquisition

We use the same public page data as [6]. This dataset encompasses a broad spectrum of page metadata, including identifiers, names, descriptions, categories, as well as geographical data like country and city, alongside relational data such as liked pages. Notably, this collection process ensures the exclusion of any user-specific private information. The methodology for data acquisition relied on snowball sampling[11], initiating from a set of popular public pages and progressively encompassing pages liked by these initial nodes. This approach organically constructs a directed graph representation of the page network.

2.2 Data Cleaning

In this directed page-likes graph, each node represents a page, with outgoing edges indicating pages liked by this page. The graph comprises 61,263,729 pages connected by 789,494,545 edges. However, only 30.8% of these pages, totaling 18,895,994, have location information (country and city) specified by their page managers. We consider location information a key feature for predicting links between pages. Our analysis is centered on the subgraph comprising all U.S. pages, given that the U.S. encompasses the largest number of pages among all countries in our dataset.

The page-likes graph is constructed exclusively from ground truth data, comprising 6,194,277 pages with verified city locations within the United States and connections between them. We exclude 55,069,452 pages and their associated edges either located outside the United States or lacking city location information. Consequently, the resultant subgraph of U.S. pages exhibits disconnected components, primarily due to the exclusion of some connecting pages. The largest connected component encompasses 5,873,395 pages and 84,480,575 edges. Our analysis prioritizes this component due to its significant size relative to others.

All pages within our U.S. graph have their city locations in the United States, as listed by their managers. Among these, 36.6% of the pages are associated with cities that have unique names across all 50 states, making their city and state locations determinable. We refer to these as deterministic pages. Conversely, the remaining 63.4% of pages are linked to cities with names that duplicate across multiple states; we classify these as non-deterministic pages. Our study concentrates on the deterministic pages.

3 Page-Likes Link Prediction

3.1 Link Prediction

Link prediction spans various research fields, including statistics, network science, data mining, and machine learning, focusing on predicting the presence of links between nodes in a network. This task aligns with real-world applications such as predicting social connections in social networks or recommending products in user-product graphs.

From the social network perspective, Liben-Nowell and Kleinberg have developed link prediction techniques based on measures for analyzing the "proximity" of the nodes in a network[12]. The nodes within the "proximity" in the network are similar in some sense, leveraging the concept of homophily. Therefore, these nodes are more likely to interact with each other and be connected by edges. Thus, the most commonly used link prediction algorithms are similarity-based algorithms[14].

Given our data's graph structure, where edges represent "likes" between pages, graph-based algorithms are particularly suitable for link prediction. Graph-based representation learning effectively addresses this by encoding node features and graph topology into vector representations. These vectors are then used to calculate scores indicating the likelihood of edge formation between node pairs. Existing edges (positive edges) are labeled as 1, while non-existing edges (negative edges), introduced through uniform negative sampling, are labeled as 0[13]. Our use of link prediction algorithms aims to identify key factors influencing the formation of page neighborhoods.

3.2 GraphSAINT

Graph Convolutional Networks (GCNs) face scalability challenges due to the necessity of updating all feature vectors within each iteration, making them less efficient for large graphs. To address these limitations, both GraphSAGE and GraphSAINT models adopt node sampling strategies, albeit through differing approaches. GraphSAGE employs uniform sampling to select a fixed number of neighboring nodes for each node in every layer and iteration. Conversely, GraphSAINT samples a sub-graph of the whole graph by nodes' importance as the mini-batch in each iteration, subsequently applying a GCN-like model on this sub-graph. This method effectively reduces the size of the original graph to a more manageable sub-graph, significantly enhancing training efficiency and time compared to GraphSAGE. Our prior research [6] has shown the GraphSAINT model to exhibit superior performance on the page graph, leading us to select GraphSAINT for encoding node representation vectors within the graph.

3.3 Feature Selection

In our study, we delve into the dynamics behind the "likes" relationships among public pages to unveil the mechanisms underlying online social community neighborhoods. This investigation is framed as a link prediction challenge, aiming to identify features that yield precise predictions within a directed page graph.

We introduce an array of candidate features for utilization within graph neural networks to forecast page-likes connections. By evaluating the predictive accuracy of these diverse features, we uncover the pivotal elements influencing the formation of page edges and neighborhoods. These features are categorized into two primary types:

1. **Topology-related features:** These features relate to the page's position and role within the graph's structure, such as its degree, or the network information for its neighborhood, such as state neighborhood distribution.
2. **Community-specific features:** These features relate to the intrinsic attributes of the page community, including the page's category, the population of the page's city, geographic coordinates of the page's city, and labels for both the city and state of the page.

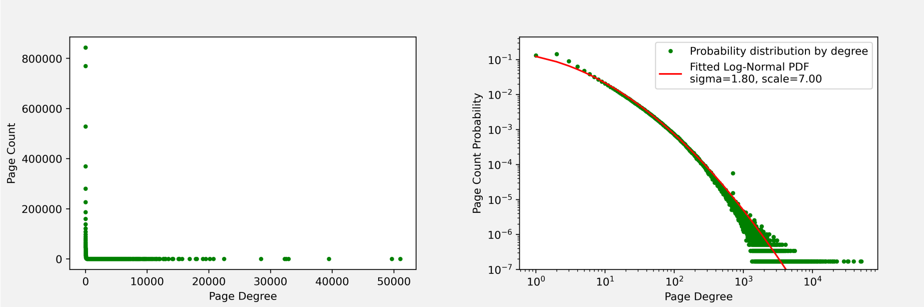
By analyzing the effectiveness of these features in link prediction, we aim to elucidate the foundational factors that drive the establishment of online social community neighborhoods.

Constant Feature Graph neural networks (GNNs) harness both node features and the graph's structural information to facilitate learning. The quality and informativeness of node features are crucial as they encapsulate the attributes of the nodes. Conversely, edge connections unveil the graph's structural intricacies. To enable a baseline comparison, we employ a constant value of 1 as the node feature across all nodes. This approach restricts the model to learning exclusively from the graph's topology and its connections, rendering all nodes indistinguishable based on their features.

The principle of homophily suggests that similar nodes tend to be closer or directly linked within a graph[15]. Our adoption of a uniform feature stems from the hypothesis that pages in proximity within the graph share certain similarities, thereby increasing their likelihood of forming connections. This method provides a foundational comparison, emphasizing the role of graph structure over individual node attributes in predicting linkages.

Page Degree The degree of a page, defined as its number of neighbors, signifies its connectivity within the page-likes graph. The degree values range from a minimum of 1 to a maximum of 51,045. To visualize this distribution, we present the degree distribution across pages in Figure 1a and 1b. Figure 1a uses linear scale and Figure 1b uses logarithmic scale. The linear scale plots show an axis-aligned pattern, while the logarithmic scale plots show a heavy-tailed pattern. This pattern aligns with the degree distribution observed in other real-world networks, such as the MSN messaging network[16], indicating adherence to a log-normal distribution.

Node degree is an often used feature in network analysis. Therefore, we propose the degree of the page node as one candidate feature. The page graph is a directed graph. Hence, we use both normalized inward degree and normalized outward degree as features.



(a) Page Degree Distribution

(b) Page Degree Distribution on Log Scale

Page's Category Public pages categorize their topics as assigned by their managers, encompassing over a thousand distinct categories within the page-likes graph. For instance, the top 20 categories are enumerated in Table 1, extracted directly from the page metadata without modification. Despite the presence of duplicate categories, their impact on prediction accuracy is minimal. According to the theory of homophily[15], pages sharing the same category are more likely to form connections. Given the impracticality of employing one-hot vectors due to the extensive number of categories, binary encoding is utilized. This method efficiently compresses category data into eleven binary digits, significantly reducing memory usage while maintaining accuracy comparable to one-hot encoding[17].

Table 1: Top 20 Categories of Public Page

Page Category	number
Local Business	1,086,041
Non-Profit Organization	230,240
Professional Service	178,543
Restaurant	171,568
Real Estate	127,801
Company	124,187
Community	111,223
Education	109,761
Religious Organization	98,712
Shopping & Retail	95,284
Medical & Health	86,503
Shopping/Retail	84,897
Organization	79,744
Artist	79,668
Musician/Band	69,365
Arts & Entertainment	69,270
Public Figure	69,026
School	63,274
Community Organization	63,272
Nonprofit Organization	57,952

Page's State Label In our previous study[6], neighborhood location information has emerged as a pivotal feature for classifying pages within the page-likes graph, aligning with the principles of homophily theory[15]. This theory suggests that pages within the same geographical state are more likely to establish connections than those across diverse states. Consequently, we advocate for the incorporation of a page's state label as a crucial feature for enhancing link prediction accuracy. Our analysis is confined to deterministic pages, whose state identities are verifiable, thereby ensuring the reliability of our predictions. To represent the geographical state of each page, we employ a 51-dimensional one-hot encoding scheme, accommodating the 50 states and Washington, D.C.

Page's State Neighborhood Distribution In our previous study[6], we employed state neighborhood distribution (SND) vectors as node features for classifying the states of pages, yielding significant accuracy improvements. These vectors represent the distribution of a page's neighbors across different states, offering a nuanced perspective beyond mere state labels. While direct state labels provide definitive location information, neighborhood distribution vectors offer predictive insights based on the proximity and connections of pages within the graph. Although not as unequivocally accurate as state labels, these vectors serve as an informative feature, suggesting the potential state affiliation of a page based on the geographic distribution of its connections.

Page's City Label Inspired by the insights gained from analyzing page state neighborhood distributions and aligned with the principles of homophily theory[15], our investigation extends into more granular location data of pages—their city locations. City-level data offer a finer granularity than state-level information, suggesting that pages within the same city may exhibit even tighter connections than those merely within the same state. However, the extensive variety of cities in our dataset, numbering in the tens of thousands, presents a much more complex challenge for classification compared to the 51 state-level categories. This analysis is confined to deterministic pages, as their city affiliations are unequivocally determined, in contrast to non-deterministic pages. Given the vast number of city categories, binary encoding serves as an efficient method to encode city information, mitigating the increase in feature dimensions associated with one-hot encoding methods.

Page's City Population Observations from the dataset reveal a pattern where popular pages from major urban centers, such as New York and Los Angeles, exhibit higher connectivity, including links to pages from smaller municipalities. We posit that the population size of a city could serve as a pivotal feature in link prediction models. The underlying hypothesis is that larger cities, with their denser populations, host a broader array of activities and enterprises, casting a wider sphere of influence that captivates the attention of individuals from less populous areas. This dynamic is proposed to facilitate the formation of connections between pages representing large urban areas and those from smaller cities.

Page's City GPS Coordinates In our previous study [6], our analysis revealed a notable trend among interstate pages, particularly those associated with cities situated along state borders. These pages demonstrated substantial connections to pages from proximate neighboring cities across state lines, suggesting a potential influence of geographical proximity on the establishment of page neighborhoods. Consequently, we propose incorporating the latitude and longitude of cities—specifically for deterministic pages—as features to explore the extent to which geographic location factors into the formation of these online community networks.

4 Evaluating Page Features

4.1 Experimental Setup

Link prediction inherently presents a binary classification challenge, necessitating a focus on accurately distinguishing between positive and negative edges. Consequently, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) serves as a critical metric for evaluating classifier performance, offering insights beyond mere accuracy by assessing the model's ability to differentiate each class effectively.

Table 2: Highest AUC-ROC on test set on different Positive/Negative edge ratio with page state label as feature

Ratio	1:1	1:5	1:10
AUC-ROC	0.8898	0.9175	0.9125

In the experiments, we employ a two-layer GraphSAINT model with random walk sampling to encode node features and topology. Each layer, implemented via the PyTorch Geometric (PyG) framework[19], contributes to a GNN layer[18]. The outputs from both layers are concatenated as the input to a linear layer, which outputs the node embedding vector. A dot product function, renowned for its efficacy in computing embedding similarities, acts as the decoder. Given the sparse nature of the page graph, actual edges are significantly outnumbered by potential non-existent edges. Because the total number of negative edges is enormous, we use negative sampling to sample a certain number of negative edges in the training[13]. We optimize the ratio of positive to negative edges at 1:5 for training and testing purposes. This specific ratio demonstrates superior AUC-ROC performance compared to alternative ratios, as evidenced in Table 2. The training process has 2000 epochs, necessitating approximately 20 hours to complete.

4.2 Single Feature

In this section, each experiment isolates a proposed feature as the sole node attribute. Comparative analysis reveals the page state label as the superior node feature, distinguished by the highest scores in Table 3, the most stable training loss curve, and the most consistent testing AUC-ROC score curve in Figure 3a.

Performance The graph topology can affect the link formation between two nodes based on whether they are within their proximity. The Graph Neural Network algorithm automatically uses the graph topological information to learn the embeddings for the nodes in the graph. Inputting node features into Graph Neural Network adds node information to the graph topological information for

Table 3: Summary of feature analysis across the entire dataset, ordered by average AUC-ROC

Single Page Feature	Avg. AUC-ROC	Avg. TPR	Avg. TNR
State Label	0.9308	0.8485	0.8832
SND	0.9306	0.8481	0.8729
City Label	0.9107	0.8113	0.8581
Constant Number 1(baseline)	0.9075	0.8109	0.8508
Page Category	0.9041	0.8040	0.8493
City GPS Coordinates	0.8964	0.7844	0.8609
Node Degree	0.7632	0.5832	0.9672
City Population	0.6849	0.5140	0.9043

generating node embeddings. It could be better or worse. Therefore, we need a baseline of the classifier performance, which is performed only on the graph topological information. We assign constant number 1 to all nodes as their features. Since all nodes have the same feature 1, the Graph Neural Networks only use the topological information in the training and testing.

Table 3 presents the prediction results for each feature. The results are averaged values of 3 runs. Column AUC-ROC represents how well the algorithm classifies the positive and negative edges. Columns TPR and TNR represent the true positive rate and true negative rate of the optimal threshold in the ROC curve for edge predictions. The table shows that the feature page state label has the best performance. Page state label, city label, and state neighborhood distribution features have better performance than the constant number 1 feature. These node features add useful node information to the graph topological information for the edge prediction. The rest of the features perform worse than the baseline feature constant number 1. Their node information interferes with the topological information, which causes the algorithm to perform worse on the prediction.

The marginal advantage of the page state label feature over the page state neighborhood distribution feature may stem from its direct and definitive representation of state labels. While the state neighborhood distribution offers insights into a page's state association, it does not achieve the exact correspondence of the actual state labels. Notably, the page state neighborhood distribution feature encompasses 306 dimensions, in contrast to the page state label feature's more concise 51 dimensions.

Three categorical features, state label, city label, and category, demonstrate superior performance. We select features based on the homophily phenomenon in networks, which suggests that nodes are more likely to connect within the same class. Table 4 shows that the model performs better on intra-class edges for all features. Therefore, the higher the intra-class edge ratio, the better the model's performance. This explains why the state label exhibits the best performance. Page location label shows a strong effect on pages connecting with their neighbors.

Table 4: Categorical feature comparison for positive edges

Feature	class#	intra-class edge		inter-class edge	
		ratio	accuracy	ratio	accuracy
State Label	51	0.7354	0.8512	0.2645	0.8321
City Label	12196	0.5080	0.8332	0.4919	0.7971
Constant baseline	1	-	0.8109	-	0.8109
Page Category	1412	0.1098	0.8155	0.8901	0.8034

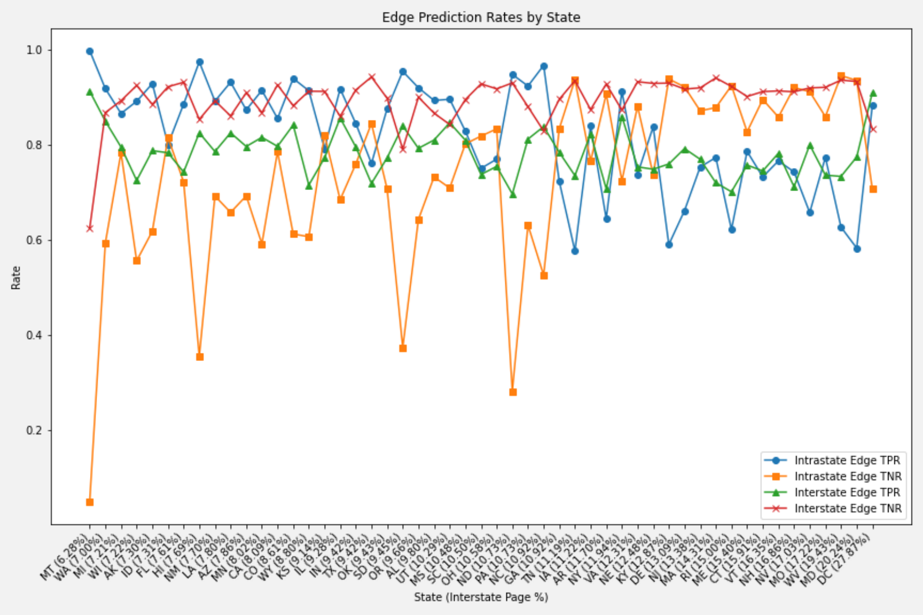


Fig. 2: Edge prediction rates by state

Learning Curve The learning curves in the training process offer insights into each feature's performance on the page graph data. Training loss curves and testing AUC-ROC score curves for all features are presented in Figure 3, with each subplot applying a consistent log scale for training loss and a linear scale for testing AUC-ROC scores. Among these, the page state label feature, as illustrated in Figure 3a, displays the most stable and conventional loss curve and AUC-ROC score, indicating its superior fit for the page graph data and effectiveness in link prediction. In contrast, features like the page state neighborhood distribution (Figure 3b), page city label (Figure 3c), constant number 1 (Figure 3d), page category (Figure 3e), and page city geographic coordinates (Figure 3f) exhibit unstable training loss curves, particularly in their plateau phases. Both the page node degree and page city population features demonstrate atypical loss and AUC-ROC curves, further distinguishing the page state label feature's distinct advantage.

Table 5: Combine one feature with page state label feature analysis across the entire dataset

Feature	Feature	AUC-ROC	TPR	TNR
State Label	Node Degree	0.9317	0.8463	0.8850
State Label	City Population	0.9289	0.8436	0.8793
State Label	- (baseline)	0.9308	0.8485	0.8832
State Label	Category	0.9243	0.8358	0.8701
State Label	City GPS Coordinates	0.9241	0.8365	0.8623
State Label	Constant Number 1	0.9228	0.8298	0.8688
State Label	City Label	0.9131	0.8178	0.8595

4.3 Combined Feature

Table 6: Combine more features with page state label feature analysis across the entire dataset

Feature	Feature	Feature	Feature	Feature	AUC-ROC	TPR	TNR
State Label	Degree	Population	-	-	0.9394	0.8590	0.8857
State Label	Degree	-	-	-	0.9317	0.8463	0.8850
State Label	-	-	-	- (baseline)	0.9308	0.8485	0.8832
State Label	-	Population	-	-	0.9289	0.8436	0.8798
State Label	Degree	Population	Category	Constant 1	0.9159	0.8168	0.8608
State Label	Degree	-	Category	-	0.9126	0.8141	0.8541

Performance Initially, we explored the combination of two features, focusing on the page state label feature due to its superior performance. We paired it with other features to assess potential enhancements in accuracy. Table 5 reveals that combining the page state label feature with the page node degree feature improves performance beyond the baseline established by the sole use of the page state label feature. When the page state label feature is combined with the page city population feature, the performance is similar to the baseline. However, integrating other features with the page state label feature leads to a decrease in performance compared to the baseline.

Further experimentation led us to combine three features: page node degree, page city population, and page state label, which collectively exhibited the highest performance, as depicted in Table 6. The table also illustrates that merging the page category, page city GPS coordinates, constant number 1, and page city label features with the page state label feature resulted in suboptimal performance. While exhaustive combinations of these less effective features were not explored, a few examples are provided in Table 6 for illustrative purposes.

There are two types of edges in the page graph: interstate edges and intrastate edges[6]. Intrastate edges connect pages within the same state, while interstate edges link pages from different states. We evaluated the accuracy of predicting interstate and intrastate edges for each state, as well as for all states combined, using an algorithm that incorporates a feature combination of page state label, page node degree, and page city population. The results are presented in Table 7.

We detail the true positive rates (TPR) and true negative rates (TNR) for both intrastate and interstate edges across various states, as presented in Table 7. Given the uniform and random sampling of negative edges within the graph, intrastate and interstate negative edges constitute 4.85% and 95.15%, respectively, of all negative edges, as shown in Table 8. Conversely, intrastate and interstate positive edges represent 73.54% and 26.46%, respectively, of all positive edges. The distribution of intrastate edges, split into 75.22% positive and 24.78% negative, contrasts with interstate edges, which are divided into 5.27% positive and 94.73% negative, according to Table 9. This disparity in data distribution likely influences the observed discrepancies in TPR and TNR values for intrastate and interstate edges, underscoring the complexity of accurately predicting link formations within the page graph.

We visualize the data from Table 7 using a line chart in Figure 2 for an intuitive understanding of the predictive statistics. The x-axis represents states ordered by their increasing interstate page percentages. Displayed are the true positive rates (TPR) and true negative rates (TNR) for both intrastate and interstate edges. Notably, the high TPR for intrastate edges (blue line) corresponds to states with lower interstate page percentages, whereas states with higher interstate page percentages exhibit lower intrastate edge TPRs. This pattern suggests that pages with numerous out-of-state connections are more often involved in interstate edges, while intrastate pages, primarily linked within their own state, tend to form intrastate edges. Consequently, a lower interstate page percentage implies a higher number of intrastate pages and edges, resulting in increased intrastate edge TPRs. However, some states show anomalously low intrastate edge TNRs (orange line), attributed to a notably smaller number of intrastate negative edges than average, a byproduct of random sampling. This discrepancy likely contributes to the observed outliers.

5 Related Work

5.1 User Social Network Analysis

The analysis of user social networks has received more focus than that of community networks in the fields of network science and social network analysis. Ugander et al. explored the global structure of the Facebook user network, identifying a range of network properties[1]. Barnett and Benefield[3] discovered that proximity and cultural homophily significantly influence Facebook friendship ties, noting that countries with international Facebook friendships often share borders, languages, and cultural traits[3].

Table 7: Link Prediction Performance by State

State	Intrastate Edge		Interstate Edge		State	Intrastate Edge		Interstate Edge	
	TPR	TNR	TPR	TNR		TPR	TNR	TPR	TNR
AL	0.8939	0.7331	0.8107	0.8672	AK	0.9293	0.6189	0.7883	0.8856
AZ	0.8742	0.6932	0.7972	0.9101	AR	0.6455	0.9074	0.7083	0.9284
CA	0.8560	0.7862	0.7974	0.9266	CO	0.9402	0.6131	0.8423	0.8828
CT	0.7332	0.8960	0.7457	0.9125	DE	0.6622	0.9225	0.7927	0.9178
FL	0.8862	0.7226	0.7434	0.9322	GA	0.7234	0.8347	0.7847	0.8964
HI	0.9750	0.3560	0.8258	0.8542	ID	0.8014	0.8158	0.7839	0.9227
IL	0.9182	0.6853	0.8571	0.8606	IN	0.8454	0.7603	0.7961	0.9160
IA	0.8410	0.7661	0.8238	0.8747	KS	0.7926	0.8218	0.7730	0.9130
KY	0.5909	0.9390	0.7604	0.9304	LA	0.9322	0.6585	0.8258	0.8613
ME	0.7867	0.8281	0.7580	0.9023	MD	0.5835	0.9360	0.7749	0.9338
MA	0.7740	0.8788	0.7210	0.9413	MI	0.8666	0.7839	0.7960	0.8934
MN	0.9151	0.5924	0.8163	0.8683	MS	0.8294	0.8026	0.8104	0.8958
MO	0.7741	0.8600	0.7369	0.9215	MT	0.9986	0.0511	0.9133	0.6254
NE	0.8387	0.7380	0.7493	0.9298	NV	0.6582	0.9123	0.8005	0.9203
NH	0.7436	0.9211	0.7125	0.9121	NJ	0.7538	0.8722	0.7703	0.9204
NM	0.8925	0.6930	0.7865	0.8933	NY	0.9129	0.7251	0.8600	0.8739
NC	0.9677	0.5267	0.8389	0.8298	ND	0.9486	0.2824	0.6974	0.9308
OH	0.7705	0.8342	0.7552	0.9183	OK	0.8764	0.7080	0.7732	0.8985
OR	0.9206	0.6436	0.7933	0.9008	PA	0.9242	0.6318	0.8127	0.8808
RI	0.6227	0.9249	0.7019	0.9234	SC	0.7511	0.8189	0.7383	0.9288
SD	0.9551	0.3744	0.8402	0.7917	TN	0.5773	0.9375	0.7359	0.9354
TX	0.7614	0.8455	0.7194	0.9433	UT	0.8965	0.7098	0.8469	0.8433
VT	0.7668	0.8589	0.7828	0.9140	VA	0.7386	0.8817	0.7542	0.9333
WA	0.9186	0.5946	0.8503	0.8684	WV	0.6275	0.9462	0.7342	0.9368
WI	0.8921	0.5569	0.7257	0.9254	WY	0.9143	0.6074	0.7161	0.9134
DC	0.8838	0.7079	0.9104	0.8340	**Total**	0.8737	0.7392	0.8087	0.8976

Table 8: Positive/Negative edges percentage distribution

	Positive Edges	Negative Edges
Intrastate	73.54%	4.85%
Interstate	26.46%	95.15%
Total	100.00%	100.00%

Table 9: Intrastate/Interstate edges percentage distribution

	Positive Edges	Negative Edges	Total
Intrastate	75.22%	24.78%	100%
Interstate	5.27%	94.73%	100%

5.2 Link Prediction

Link prediction has been a popular research area for the past decades. In social network link prediction, researchers typically employ three methodologies: similarity, probabilistic, and algorithmic approaches [4]. The similarity approach leverages graph-measures and content-measures (attributes of nodes or edges). Among algorithmic methods, deep learning has emerged as a particularly popular technique. In our study, we employ both similarity and algorithmic approaches to predict links.

5.3 Online Social Community Location Classification

Public pages represent a prominent platform for online communities, with each page embodying a distinct social community. While page managers have the option to label their pages with country and state/province locations, many pages lack this geographical information. Hong et al.[5] explored the page graph—a network where pages can "like" each other—and introduced a majority voting algorithm for inferring the missing country locations of pages. This method proved effective for country-level classification, leveraging shared cultural, linguistic, and social contexts among pages from the same country.

Nonetheless, the majority voting approach showed limitations in more granular subdivision location classifications, such as state labeling within the United States. We introduced the concept of neighborhood state distribution vectors and applied Graph Neural Networks for the classification of pages' subdivision locations, achieving notable accuracy[6]. This methodology offers insights into a page's influence across different states.

5.4 Graph Neural Networks

Graph neural networks (GNNs) are a subset of artificial neural networks designed for processing graph-structured data[7]. Graph convolutional networks (GCNs) are one type of GNN that are often used in graph representation learning. These representations aim to encapsulate the graph's topological structure in low-dimensional vectors, facilitating tasks such as node classification and link prediction. Nonetheless, GCNs' reliance on full graph adjacency matrices makes them computationally intensive, particularly for sizable graphs, leading to significant GPU memory demands and prolonged training durations[8][9].

To mitigate these challenges, node sampling techniques have been developed to adapt GCNs for larger graphs. GraphSAINT, specifically, introduces an inductive learning strategy through graph sampling, enhancing both the efficiency and accuracy of training. It generates mini-batches by sampling sub-graphs from the entire graph for each iteration. This approach ensures that nodes influencing each other significantly are likely to be included in the same mini-batch, allowing for mutual support within the mini-batch and circumventing the need for broader graph traversal[10]. Such innovations significantly curtail the computational burden associated with GCNs, concurrently bolstering accuracy[10].

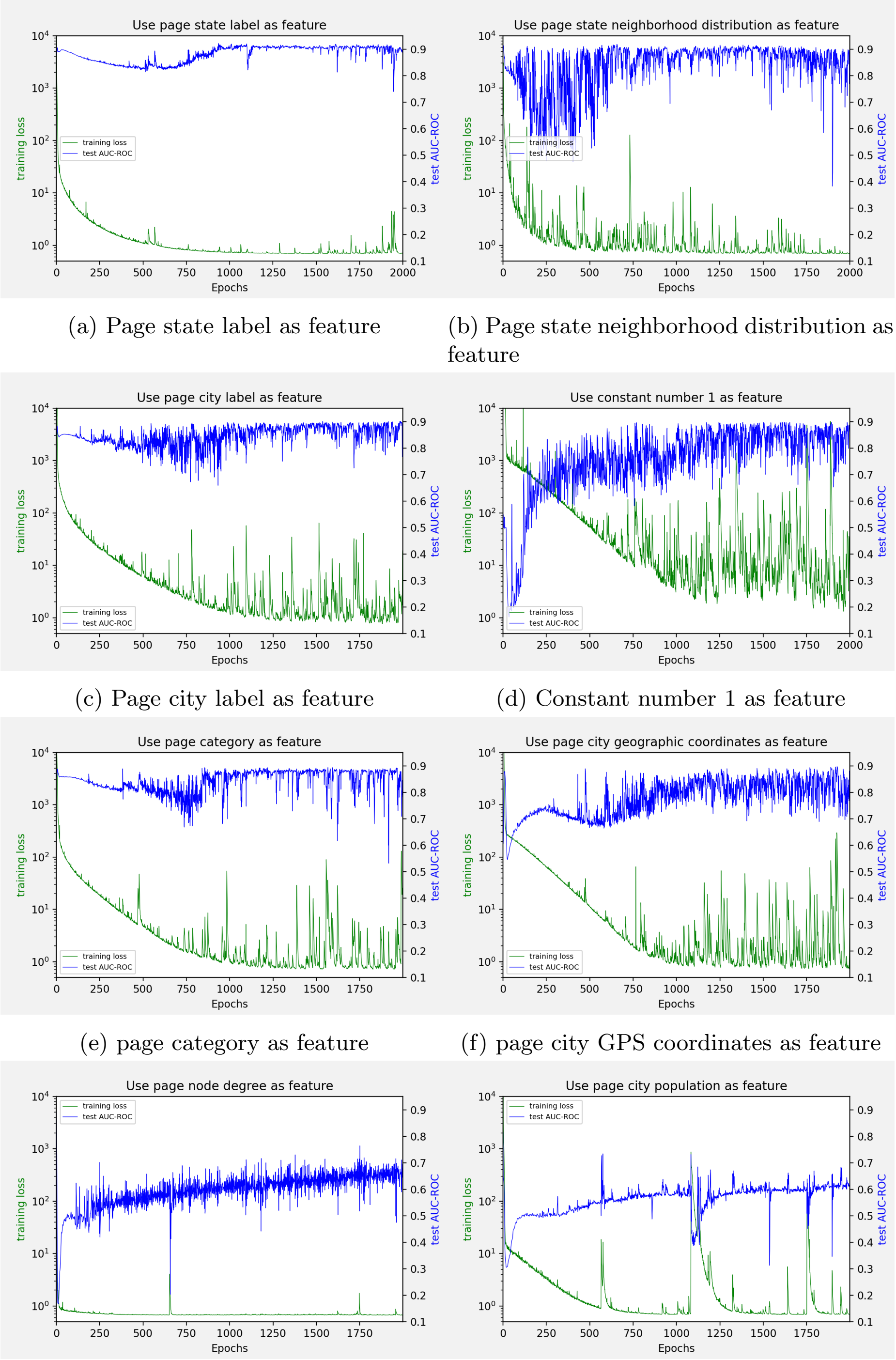
6 Conclusion

In this paper, we investigated the pivotal features that influence link formation and neighborhood structuring within the page graph. Initially, we explore a series of potential features, both graph-based and content-based, that may impact link connectivity. Subsequently, we present our methodology, combining the node similarity and GNN to perform the link prediction. Through meticulous experimentation with both individual and combined features, we ascertain that the page state label emerges as the most influential single feature for link formation. Moreover, we observe that augmenting the page state label feature with page node degree and page city population features further enhances link prediction accuracy. Ultimately, our analysis reveals a correlation between the true positive rate of intrastate positive edges and the interstate page percentage, underscoring the nuanced dynamics of link formation within the page graph. Page location label shows a strong effect on pages connecting with their neighbors.

References

1. Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow, “The anatomy of the facebook social graph”, arXiv preprint arXiv:1111.4503, 2011.
2. Wikipedia contributors, “List of social platforms with at least 100 million active users — Wikipedia, the free encyclopedia,” 2023, [Online; accessed 30-May-2023].
3. George A Barnett and Grace A Benefield, “Predicting international Facebook ties through cultural homophily and other factors”, New Media & Society, vol. 19(2), pp. 217-239, 2017.
4. N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, and N. B. Anuar, “Applications of link prediction in social networks: A review,” Journal of Network and Computer Applications, vol. 166, p. 102716, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804520301909
5. Yunfeng Hong, Yu-Cheng Lin, Chun-Ming Lai, S. Felix Wu, and George A. Barnett, “Profiling facebook public page graph”, 2018 International Conference on Computing, Networking and Communications (ICNC), pp. 161–165, 2018.
6. Jiarui Wang, Xiaoyun Wang, Chun-Ming Lai, and S. Felix Wu. 2024. Online Social Community Sub-Location Classification. In Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM ’23). Association for Computing Machinery, New York, NY, USA, 276–280. https://doi.org/10.1145/3625007.3627504
7. Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini, “The graph neural network model”, IEEE Transactions on Neural Networks, vol. 20(1), pp. 61-80, 2009.
8. Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun, “Spectral networks and locally connected networks on graphs”, Proc. Int. Conf. Learn. Representations, pp. 1-14, 2014.
9. Mingyu Yan, Zhaodong Chen, Lei Deng, Xiaochun Ye, Zhimin Zhang, Dongrui Fan, and Yuan Xie, “Characterizing and understanding GCNs on GPU”, IEEE Computer Architecture Letters, vol. 19(1), pp. 22–25, 2020.
10. Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna, “Graphsaint: Graph sampling based inductive learning method”, arXiv preprint arXiv:1907.04931, 2019.

11. Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong, "Statistical properties of sampled networks", Phys. Rev. E, vol. 73, pp. 016102, Jan 2006.
12. D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," Journal of the American Society for Information Science and Technology, vol. 58, no. 7, pp. 1019–1031, 2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20591
13. Z. Yang, M. Ding, C. Zhou, H. Yang, J. Zhou, and J. Tang, "Understanding negative sampling in graph representation learning," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1666–1676. [Online]. Available: https://doi.org/10.1145/3394486.3403218
14. E. A. Yilmaz, S. Balcisoy, and B. Bozkaya, "A link prediction-based recommendation system using transactional data," Scientific Reports, vol. 13, no. 1, p. 6905, 2023.
15. M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annual Review of Sociology, vol. 27, pp. 415–444, 2001. [Online]. Available: http://www.jstor.org/stable/2678628
16. J. Leskovec and E. Horvitz, "Planetary-scale views on an instant-messaging network," 2008.
17. K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," International journal of computer applications, vol. 175, no. 4, pp. 7–9, 2017.
18. C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and leman go neural: Higher-order graph neural networks," 2021.
19. Matthias Fey and Jan Eric Lenssen, "Fast graph representation learning with pytorch geometri", arXiv preprint arXiv:1903.02428, 2019.



(a) Page state label as feature

(b) Page state neighborhood distribution as feature

(c) Page city label as feature

(d) Constant number 1 as feature

(e) page category as feature

(f) page city GPS coordinates as feature

(g) page node degree as feature

(h) page city population as feature

Fig. 3: Comparison of page features based on state and city