Online Social Community Geographic Characterization: Classification and Neighborhood Formation

By

Jiarui Wang

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

S. Felix Wu, Chair

Norman S. Matloff

George A. Barnett

Committee in Charge

2024

This dissertation is dedicated to my family.

Contents

List of Figures

List of Tables

Abstract

Online Social Community Geographic Characterization: Classification and Neighborhood Formation

Over the past decade, online social networks (OSNs) have experienced unprecedented growth, attracting billions of users across the globe. These platforms enable individuals to connect and share content, breaking down the barriers of time and location that limit offline social interactions. Among these, Facebook public pages stand out as a prominent type of OSN community, offering spaces for user discussions, business promotions, and public relations activities. These online social communities interact with each other, forming an online community network.

In the digital realm of online spaces, people's behaviors remain closely linked to location. Geolocation information enables online social communities to make recommendations and promote local businesses and services. This dissertation explores the classification of geolocation information for communities and examines how geolocation contributes to neighborhood formation within online community networks.

The dissertation introduces neighborhood state distribution vectors as novel features for graph neural networks to classify the states of Facebook public pages. It also defines intrastate and interstate Facebook public pages based on high-probability state label outputs from the classification model. Furthermore, it profiles states with varying influences over online communities through an analysis of the classification confusion matrix, interstate page percentages, and the presence of interstate pages across state borders. This approach achieves an improved accuracy (0.88) and F1 score (0.88) compared to previous studies.

Additionally, the dissertation identifies key features that influence link formation and neighborhood structuring within the page graph, employing a methodology that combines node similarity and the topological algorithm GNN for link prediction. The study reveals that the page state location stands out as the most significant single feature for link formation. Furthermore, it is observed that incorporating page node degree and page

city population features alongside the page state location feature improves link prediction accuracy.

Lastly, the dissertation explores city, county, and cluster neighborhood distribution vectors as unique features for page classification. Addressing the challenge of distinguishing among 630 cities with an initial city classification accuracy of 0.6928, a clustering algorithm is developed to leverage the confusion matrix from city classification, constructing a hierarchical city structure. This approach significantly improves city classification accuracy to 0.8014, employing a cluster-city hierarchical classification strategy.

Acknowledgments

First and foremost, I would like to express my gratitude to my academic advisor, Prof. Shyhtsun Felix Wu. This dissertation would not have been possible without his encouragement and guidance. Felix has consistently provided me the flexibility to explore new research areas while guiding me to deepen my understanding and build a solid foundation in my research. From him, I have learned not only research skills but also the attitude towards studying, working, and life in general. His passion for research and kindness towards people will continue to inspire me as I navigate my future, from all perspectives.

I am grateful for the time devoted by Prof. Norman S. Matloff and Prof. George Barnett, who served on my dissertation committees. Their feedback and advice have been critical in shaping this thesis.

I would like to thank my fellow doctoral colleagues and friends in the Davis Sincere Research Lab and our collaborators: Xiaoyun Wang, Chun-Ming Lai, Yeh-Cheng Chen, Jon Chapman, and Yunfeng Hong. Interacting with them has provided me with invaluable learning experiences and moments of joy.

I am extremely grateful to my parents for their unwavering support and love. Because of you, all of this was possible. Finally, I wish to thank my girlfriend for her love and understanding. Her endless encouragement and support were instrumental in the completion of this dissertation.

Chapter 1

Introduction

Over the past decade, online social networks (OSNs) have experienced explosive growth, attracting billions of users across the globe. Platforms like Facebook and Twitter have revolutionized the way individuals form connections and share content, eliminating the limitations of time and place that bound traditional social interactions. These users, forming the core of online social networks, symbolize the commercial potential of these platforms, representing a vast pool of potential customers for diverse products and services. The complex network of social relationships among these users is a fundamental aspect of user networks.

Another significant activity on OSNs involves various entities, including businesses, non-profit organizations, and governmental bodies, utilizing these platforms to further their interests. Along with individual users, these entities create diverse online social communities aimed at catering to specific interests. These communities encompass a wide array of groups, from official pages of corporations and non-profits to user-initiated groups centered around common interests such as neighborhood events, professional connections, and hobbies like animal enthusiasm.

Facebook public pages are among the most popular platforms for online communities, serving as venues for information announcements, user discussions, news dissemination, public relations, and business promotions. These pages interact not only with their followers but also with other Facebook pages, establishing connections through "likes." This creates a network of page likes, which is the focus of this dissertation. In this network,

each node represents a Facebook public page, and outgoing edges from a node signify pages it likes.

In the digital realm, people's behaviors are influenced significantly by location. Individuals often show a preference for local news, are inclined to connect with friends in close proximity, and favor local dining and entertainment options, demonstrating the importance of location-centric activities. On platforms like the Facebook public page, location information enhances the relevance of pages and their services, enabling targeted dissemination of news, personalized product and service recommendations, and timely notifications for emergencies. The location metadata of a Facebook public page, highlighting the primary geographical focus of its activities and user engagement, is a critical attribute.

1.1 Objectives

The aim of this dissertation is multifaceted, targeting three principal objectives to enhance our understanding and utility of Facebook public pages.

Initially, a significant gap is identified in the available location information, with only 30.8% of public Facebook pages specifying their geographical data. Addressing this gap, the dissertation sets out to predict missing locations, focusing on the granularity of sub-locations like U.S. states. This challenge not only seeks to fill the void of missing geographic data but also delves into the impact of such locations on the network of page likes, using pages from the U.S. as the primary dataset for exploration.

The second objective revolves around the network of likes among Facebook pages, which symbolizes the interconnected web of online social communities. With pages featuring varied metadata such as topics, countries, and cities, this dissertation endeavors to identify key elements that drive these inter-page connections. By dissecting the page-likes graph, the research aims to understand the formation of online community neighborhoods and the dynamics of community interactions within this digital ecosystem.

Lastly, the dissertation recognizes the limitations of State or Province-level classification in addressing the needs of services that require more detailed geographic specificity,

such as city-level information for local business suggestions or election campaigns. Acknowledging the complexity of city-level classification as a step beyond state categorization, this work aspires to develop a comprehensive framework for classifying Facebook pages by country, state, and city. This ambition intends to serve a broad spectrum of research and practical applications, enhancing the relevance and precision of location-based analyses in online social communities.

1.2 Research Contributions

1.2.1 Online Social Community Sub-Location Classification

Location attributes of Facebook public pages garner significant attention as they provide insights into the physical footprints of online communities. However, not all pages have their location information specified by their managers, making the prediction of missing locations crucial for further geographic location-related research. The classification into sub-locations, such as states within the United States, presents a substantial challenge and is of significant importance.

In Chapter Chapter 2, we examine the limitations of prior research on sub-location classification, particularly focusing on pages from the United States. We introduce neighborhood state distribution vectors as features and utilize graph neural networks for the state classification of pages. This approach significantly outperforms previous algorithms, achieving improvements in classification accuracy and F1 scores. Additionally, we define intrastate and interstate Facebook public pages based on the high-probability state label output from the classification model and analyze the influence of different states over online communities by examining the classification confusion matrix, interstate page percentages, and the presence of interstate pages across state borders.

This work was published as "Online Social Community Sub-Location Classification" in the proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2023) [1].

1.2.2 Online Social Community Neighborhood Formation

Online community subdivision location classification has illuminated the significance of state location for analyzing Facebook pages within the page-likes graph. Through the analysis of pages' high-probability state labels, we identified whether a page could be categorized as interstate or intrastate. The foundation of this study is the exploration of the neighborhood of nodes within the page-likes graph, prompting inquiries into how "likes" relationships are formed from one page to its neighbors, and identifying the key factors influencing these "likes" relationships.

In Chapter Chapter 3, we explore a variety of page features to ascertain their influence on the formation of page neighborhoods. This exploration employs link prediction techniques applied to each feature individually. Our findings highlight the page state label as the most accurate predictor in link prediction tasks. Furthermore, we discovered that a combination of features—namely, the page state label, page node degree, and page city population—delivers the best results in terms of link prediction accuracy.

A preliminary version of this work was published as "Online Social Community Neighborhood Formation" in the proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2024).

1.2.3 Online Social Community City Classification

Classifying communities at the city level poses a greater challenge than state-level classification due to the extensive number of cities involved. This finer granularity is essential for various services, such as recommendations for local businesses, connections with local friends, and the promotion of local public services or elections. Successfully achieving city classification would facilitate the creation of a comprehensive system capable of categorizing Facebook pages by country, state, and city. Such a development would significantly benefit research efforts and services related to community locations.

In Chapter Chapter 4, we focus on the examination of flat city classification for pages, specifically targeting pages from California. We introduce innovative features for graph neural networks in city classification of pages, such as city, county, and cluster neighborhood distribution vectors. Additionally, we propose a novel city clustering algorithm

and implement a two-stage hierarchical classification method. This approach significantly improves upon the flat city classification methods for pages, offering a more refined and effective classification system.

A preliminary version of this work was published as "Online Social Community City Classification" in the proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2024).

Chapter 2

Online Social Community Sub-Location Classification

2.1 Introduction

The social relationship has long been a subject of academic interest. The social network represents social interactions and relationships. In the early 21st century, the emergence of online social platforms, such as Facebook and Twitter, extended social networks from the physical to the digital world, enabling people to easily make new friends online, akin to offline interactions. This phenomenon is referred to as an online personal social network. There exists a vast body of literature covering various aspects of online personal social networks.

Not only have personal social networks transitioned online, but social communities have as well. People often belong to multiple communities and engage in conversations and activities within these communities offline, which could be neighborhoods, workplaces, or groups with shared interests. Many of these groups or communities maintain online information pages or discussion forums on Facebook and other online platforms. Moreover, numerous communities and groups exist solely online, without any in-person activities. The emergence of a vast amount of online communities in such a relatively short span of history is indeed a remarkable feat.

Facebook is the most popular platform for online communities and has been the focus of numerous research projects. Our research centers on public Facebook pages, which serve

Figure 2.1: The New York Times Facebook Page

as platforms for information announcements, user discussions, news dissemination, public relations, and business promotions. Figure 2.1 presents an example of the New York Times Facebook page. A key attribute of Facebook pages is their location, indicating where the majority of page activities and users are concentrated. Conducting research based on page location, such as targeting highly influential pages in specific areas, holds significant potential.

However, not all pages have their location information specified by their managers. In our dataset of public Facebook pages, only 30.8% (18,895,994 pages) out of a total of 61,263,729 pages have listed their location. Predicting the missing locations is crucial for further geographic location-related research. The classification of sub-locations, such as states within the United States, poses an even greater challenge and significance. In this research, we aim to make the following contributions:

- Investigate the limitations of previous research on sub-location classification, focusing on pages from the United States.
- Introduce neighborhood state distribution vectors as features and utilize graph neural networks for state classification of pages, significantly outperforming previous

algorithms with improvements in classification accuracy and F1 scores to 0.88.

- Define intrastate and interstate Facebook public pages based on the high-probability state label output from the classification model.
- Analyze the influence of different states over online communities by examining the classification confusion matrix, interstate page percentages, and the presence of interstate pages across state borders.

This chapter is organized as follows: Section 2.2 introduces related research on user and page geographic location analysis and the graph neural networks for classification. Section 2.3 describes the data utilized in this study and the ground truth data for verification. Section 2.4 details the investigation into the limitations of previous studies on page sub-location classification. Section 2.6 proposes the features and graph neural networks used for classification. Section 2.7 outlines the experimental setups, data-expansion methods, experiment results, and profiles of pages and states. Finally, Section 2.8 provides a summary of this chapter.

2.2 Related Work

2.2.1 Facebook User Graph Analysis

The Facebook user graph, representing a network of users connected through friendship ties, has been extensively researched by social and computer scientists alike. Studies by Ugander et al. have characterized the global structure of the Facebook user graph, revealing various network properties [2]. Barnett and Benefield explored the determinants of the Facebook user network, identifying proximity and cultural homophily as crucial factors in the formation of friendships on Facebook. Their findings also highlighted that countries with international Facebook friendships often share borders, languages, civilizations, and migration patterns [3].

2.2.2 Facebook Page Location Classification

The Facebook page graph is defined as a network of pages connected when one page likes another. Hong et al. [4] investigated this network, proposing a majority voting

algorithm to classify missing country location information of Facebook pages. This algorithm is effective for country location classification as most pages linked by edges share the same cultural, language, and social context. However, it falls short in more granular classifications, such as state labeling within the United States.

Addressing this shortfall, Lin et al. [5] introduced a Breadth-First Search (BFS)-based machine learning algorithm, utilizing hand-picked anchor pages as seeds for initiating the search [5]. Despite its innovative approach, the algorithm's performance did not meet expectations due to incomplete data coverage, lower performance on the total dataset, and other limitations.

2.2.3 GraphSAGE

Graph neural networks (GNNs) are specialized artificial neural networks designed for graph data processing [6]. GNNs are widely adopted in graph representation learning, where they are trained to generate node embeddings for downstream tasks, such as node classification and link prediction. One of the most popular GNNs is the graph convolutional network (GCN). GCNs work by updating the feature vectors for all nodes in the graph at one iteration. This process, however, requires the entire graph's adjacency matrix to compute the aggregated messages for each GCN layer, leading to high computational costs and significant GPU memory requirements[7][8]. GraphSAGE addresses these scalability issues by learning a function that generates node embeddings through sampling and aggregating features from a node's local neighborhood [9]. The embeddings for a node are produced by aggregating messages first from the node's neighbors and then from the node itself, as described by the following equation [9]:

$$h_v^{(l)} = \sigma(W^{(l)} \cdot CONCAT(h_v^{(l-1)}, AGG(h_u^{(l-1)}, \forall u \in N(v))))$$

Where:

- h : the aggregated message for a node.
- l : the number of layer.
- σ : a nonlinear activation function such as $Relu(\cdot)$ or $Sigmoid(\cdot)$.

- W : a weight matrix.
- AGG : aggregation such as $Mean$ or Sum.
- $N(v)$: the neighborhood of node v.

2.2.4 GraphSAINT

GraphSAINT introduces a novel approach to scale GCNs to large graphs, effectively managing the "neighbor explosion" problem by sampling the training graph to construct mini-batches, rather than sampling nodes or edges across GCN layers [10]. Each iteration builds a complete GCN from the sampled subgraph, ensuring all layers contain a fixed number of well-connected nodes, which enhances both training efficiency and model accuracy. The technique prioritizes nodes with high mutual influence for subgraph sampling, allowing sampled nodes to support each other's learning within the same mini-batch [10]. This method has demonstrated improvements in training speed and accuracy in various experiments.

2.3 Data Description

2.3.1 Data Acquisition

The Facebook public page data utilized in this study was acquired through the Facebook Graph API version 2.8, enabling researchers to collect social data. Similar to the datasets used by Lin et al.[5] and Hong et al[4], our data was also obtained via this API. However, our dataset is slightly larger than that used by Lin et al.[5] and substantially larger and more recent than the dataset utilized by Hong et al.[4]. For our research, we concentrated on specific page meta-information, including page ID, name, location (country and city), and the list of other pages liked by each page. Our data collection employed snowball sampling[11] through the API, beginning with popular seed pages and expanding by incorporating pages liked by these seeds. This method naturally constructs a directed graph where each edge signifies a page liking another page.

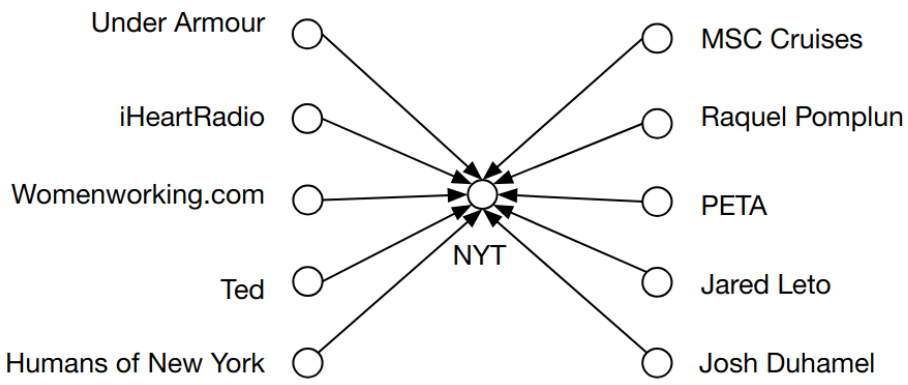


Figure 2.2: The New York Times Facebook Page Graph

2.3.2 Data Structure

The graph is a pivotal data structure in Social Network Analysis (SNA), exemplified by the well-documented Zachary's karate club network[12]. In this network, each node symbolizes an individual in the club, and edges represent friendships between members. Similarly, we constructed a page-likes graph as a directed graph, where edges illustrate the "likes" relationship with inherent directionality. In this graph, every node is a Facebook page, with outgoing edges to nodes representing pages it likes. Figure 2.2 depicts a simplified example of the New York Times page graph, showcasing how it is liked by other pages. Our dataset comprises a total of 61,263,729 pages and 789,494,545 "page likes" relationships, with 6,194,277 pages explicitly labeled with cities within the United States.

2.3.3 Data Cleaning

We constructed a page-likes graph utilizing only the ground truth data, comprising 6,194,277 pages that are located within the United States and include edges connecting these pages. A total of 55,069,452 pages, either located outside the United States or lacking city location information, were excluded from our analysis. This exclusion is justified by our focus on sub-location classification within the United States, coupled with the impracticality of processing an excessive number of pages and edges not relevant to our study's scope.

The generated subgraph of ground truth U.S. pages contains disconnected components due to the exclusion of some unknown-location pages that connect the U.S. pages. The largest connected component comprises 5,873,395 pages. Our analysis focuses on this largest connected component, as the other components are relatively small and thus less significant for our study. The challenge of state location classification for Facebook public pages is effectively reframed into a more manageable problem: analyzing a directed graph where each node represents a Facebook public page labeled with a specific state. In this graph, each edge originates from one page and points to other pages that are liked by the initiating page. Our objective is to enhance state classification accuracy within this page-likes graph, specifically within the largest connected component of the ground truth U.S. pages.

2.3.4 State-known and State-unknown Pages

Table 2.1: State-known and State-unknown Pages

Page ID	**Page Name**	**City Name**	**States Has City Name**
5281959998	The New York Times	New York	NY
48842713792	Barack Obama	Washington	DC, UT, IL, MO, PA, IN, NC, IA, NJ, GA, WV, KS, LA, OK, CA, AR, NE, VA

The U.S. ground truth pages are listed with their country and city locations but not their state locations. This distinction is significant because many cities share names across different states. Within the largest connected subgraph, which comprises 5,873,395 pages, there are two categories of U.S. pages:

- **State-known Pages**: These pages amount to 2,147,399 and are associated with cities that have unique names within the United States, accounting for 36.6% of the pages in the largest connected subgraph. We refer to this group of pages as **Dataset A**.

- **state-unknown Pages**: These pages amount to 3,725,996 and are associated with

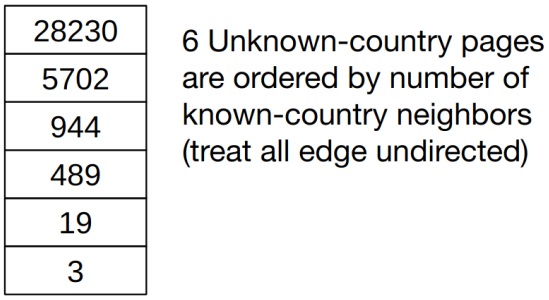


Figure 2.3: Country-known neighbor max heap of pages

cities that share their names with cities in other states within the United States, accounting for 63.4% of the pages in the largest connected subgraph. We refer to this group as **Dataset B**.

Table 2.1 presents examples of state-known and state-unknown pages, categorized based on whether their city names are unique to one state or shared across multiple states. State-known pages can be directly used as ground truth data because their state locations are unambiguous. In contrast, state-unknown pages, whose state associations are unclear, cannot be directly used for this purpose. Nonetheless, excluding state-unknown pages, which constitute a significant portion (63.4%) of the U.S. ground truth pages, would severely diminish the graph's connectivity and result in the loss of valuable information. Therefore, our approach incorporates both Dataset A and B in the page graph to maintain connectivity.

2.4 Algorithm Comparison

2.4.1 Majority Voting

2.4.1.1 Algorithm description

Hong et al.[4] addressed the challenge of identifying the geographic country location of Facebook pages using a majority voting algorithm. This algorithm initiates by placing all pages with unknown country information into a max heap, ordered by the number of known-country neighbors each page has. Figure 2.3 illustrates a simplified example of this heap, with the top page being the one with the highest number of known-country

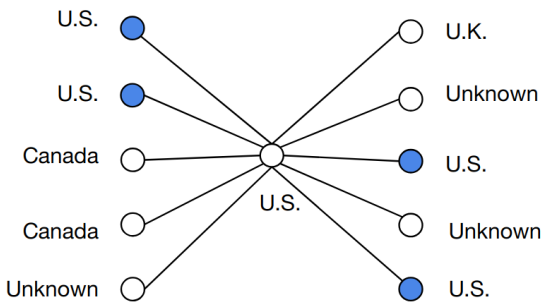


Figure 2.4: Majority voting for the New York Times page

neighbors. The algorithm then determines the most frequent country among the known-country neighbors of the top page and assigns this country to the page. Figure 2.4 depicts the application of majority voting for the New York Times page, demonstrating how the most common country label among its neighbors is determined. Finally, the algorithm removes the top page from the heap and repeats the process for the next page. This method requires only a dataset of pages and their neighbor relationships, bypassing the need for a traditional graph structure.

The majority voting algorithm was evaluated on a subset of 8 million pages with known nationality information, including edges between two known-country pages. The test set was created by designating 50.24% of all pages as unknown-country pages through random selection. The algorithm achieved an accuracy of 90% in nationality labeling for pages.

2.4.1.2 Drawbacks of Majority Voting

Nevertheless, the majority voting algorithm underperforms in state location labeling for U.S. pages. For this evaluation, 50% of known-state U.S. pages were randomly chosen as unknown for the test set. The algorithm's accuracy for state location labeling was only 59.4%[13], highlighting the greater difficulty of subregion labeling compared to country labeling. This challenge is primarily due to the language and culture distance. Pages from different countries, typically using different languages, are unlikely to interact with each other due to language barriers. It's challenging for users to engage with content in unfamiliar languages. Conversely, within a single country, despite the presence of various

subregions, a common language facilitates interactions across these areas. Consequently, pages are more inclined to like or be liked by other pages from different subregions when they share the same language and culture.[13]:

The same language and culture result in a denser and more interconnected graph for subregion-labeled pages than for country-labeled pages. This complexity poses challenges for the majority voting algorithm in accurately classifying pages by subregions.

2.4.2 BFS-based Machine Learning

Lin et al.[5] proposed a BFS-based machine learning algorithm to address the subregion labeling problem, noting the inadequacy of the majority voting algorithm for this task[5]. This algorithm requires a graph data structure of pages. Lin et al.[5] utilized state-known U.S. pages as ground truth from each state and grouped all state-unknown U.S. pages and all non-U.S. pages, totaling 12,685,090 pages, labeled as "other." The distinctions between state-known and state-unknown pages are detailed in Section 2.3.4.

2.4.2.1 Anchor Page

The BFS-based algorithm necessitates associating pages with features to facilitate machine learning classifier training and prediction. Lin et al.[5] introduced the state distance vector (SDV) to denote the hop distances from each page to a hand-selected anchor page in every state. The success of this algorithm hinges on the choice of anchor pages, ideally situated at each state's cluster centroids. However, due to the overlap and entanglement of pages across states, identifying clear cluster boundaries is challenging.

The study argued that the anchored pages have to be as local as possible. This means that anchored pages should have most neighbors from its state, not too many neighbors from other states. In this study, the pages of the local government departments, parks, and state universities are good examples of anchored pages. The pages of famous sports teams are bad choices since they are likely to be liked by pages from other states. These two kinds of pages above are close to the centroid and the boundary of the clusters, respectively. The authors chose "OnlyInYourState.com" pages as the anchor page for every state, without any proof that these hand-picked anchor pages are close to the centroids of the clusters. For example, the page "Only_In_Alabama" is the anchor page for the state of Alabama.

The authors advocate that anchor pages must be as localized as possible, meaning these pages should predominantly have neighbors from their own state, rather than an excessive number from other states. Accordingly, pages associated with local government departments, parks, and state universities serve as ideal examples of anchor pages due to their localized nature. Conversely, pages representing well-known sports teams are considered unsuitable choices, as they tend to attract likes from across state boundaries. These examples illustrate pages that are, respectively, proximal to the centroids and on the periphery of the clusters. The authors selected 'OnlyInYourState.com' pages as the anchor for each state, despite the absence of empirical evidence confirming that these selected pages are near the cluster centroids. For instance, 'Only_In_Alabama' has been designated as the anchor page for Alabama."

2.4.2.2 Algorithm description

Lin defined the state distance vector (SDV). Every page's SDV vector has 102 dimensions, which are 51 pairs of hop distance from 51 anchors (50 states, but California has two anchors). Each pair has two distances: the distance to the anchor page in the inward direction and the distance to the anchor page in the outward direction. The following equation shows SDV for each page[5]:

$$SDV(Page) = [[IHOP(Page, Anchor_i), OHOP(Page, Anchor_i)] : i \in 1, ..., N_{\text{number of anchors}}]$$

Definitions:

- $IHOP(Page, Anchor_i)$ denotes the inward hop distance from $Page$ to $Anchor_i$.
- $OHOP(Page, Anchor_i)$ denotes the outward hop distance from $Page$ to $Anchor_i$.

The minimum hop distance, $MHOP$, for each page is defined as the smallest non-zero distance from all reachable anchor pages [5]:

$$MHOP = \min(\min(IHOP), \min(OHOP))$$

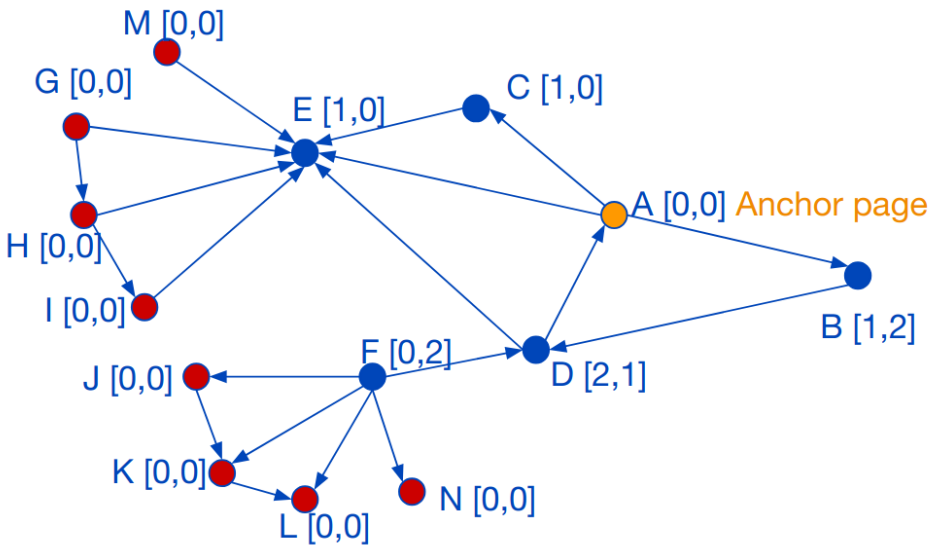


Figure 2.5: Example graph with unreachable pages from an anchor page

This essentially means $MHOP$ is the shortest distance from a page to any reachable anchor, disregarding unreachable (zero-distance) paths.

The BFS-based algorithm proceeds in five steps:

1. Since the anchor pages have been selected, the initial step involves generating feature vectors, namely the state distance vector (SDV), for every page using a breadth-first search (BFS). For instance, the BFS begins at the anchor page "Only_In_Alabama," labeling each encountered page with its hop distance from the anchor while following only the outward edges. Subsequently, the BFS is initiated again from the same anchor page, this time labeling pages based on hop distance while following only inward edges. This process is repeated for all designated anchor pages.

2. The next step involves removing all pages whose SDVs are zero vectors. Effective machine learning features must possess the quality to distinguish between data points distinctly. The SDV performs inadequately in representing every page since ideally, each page should be accessible from at least one anchor page, necessitating at least one non-zero dimension in its SDV for differentiation. However, only 41.5% (1,009,135) of U.S. pages and 46.1% (5,842,776) of other pages are accessible from at least one anchor page in practice. A significant majority of pages are unreachable

from any anchor page, rendering their SDVs as zero vectors. This reachability issue is predominantly caused by the "breaking nodes" in a directed graph.

This reachability issue is predominantly caused by 'breaking nodes' in a directed graph. Figure 2.5 illustrates this scenario, where Page A acts as the anchor. Page D, reachable from Anchor A, has an inward path [A, B, D] with a distance of 2 and an outward path [D, A] with a distance of 1. Pages B, C, D, E, and F are accessible from Anchor A, whereas Pages G through N are not. Specifically, Page E only has inward edges, and Page F only has outward edges, making them 'breaking nodes' within the graph. These nodes serve as the termini of paths originating from Page A, thereby obstructing any potential paths to Pages G through N."

3. The author initially calculates the minimum hop distance ($MHOP$). Subsequently, pages with an $MHOP$ value less than or equal to the local tendency threshold ($N_{\text{local threshold}} = 3$) are classified as pages from the United States. Conversely, pages with an $MHOP$ value greater than or equal to the global tendency threshold ($N_{\text{global threshold}} = 5$) are classified as belonging to the 'other' category. The choice of these thresholds is made arbitrarily. After applying these criteria, the dataset is reduced to 541,407 U.S. pages from an initial total of 1,009,135 U.S. pages that were accessible in the earlier step. Furthermore, the 'other' category is narrowed down to 3,107,168 pages from the preceding total of 5,842,776 accessible pages, with only the first 100,000 'other' pages being retained for analysis.

4. To rectify the uneven distribution of page data across different states, the author implemented a re-sampling strategy to ensure balance among each state. Given that Nevada had the fewest pages, totaling 824, following the application of step 3, the dataset was standardized to include the first 800 pages from each of the 51 categories—encompassing 50 states plus one "other" category. Consequently, this approach yielded 40,800 pages, which were subsequently utilized to train and evaluate the classifiers in the ensuing step.

5. The dataset of 40,800 pages was divided into 80% for training (32,640 pages) and

20% for testing (8,160 pages). Subsequently, Naive Bayes, AdaBoost, and Random Forest classifiers from the scikit-learn package were employed for the training and testing phases.

2.4.2.3 Drawbacks of BFS-based Machine Learning

- **Limited Reachability from Seed Anchors:** Only 41.5% of U.S. pages (1,009,135 out of 2,430,356) are accessible from at least one anchor within the graph. This algorithm can not be applied to those unreachable pages.

- **Random Accessibility of Pages:** The aforementioned reachability issue indicates that the choice of anchor pages randomly influences the set of pages that can be classified. This randomness introduces a significant lack of predictability and control within the classification process.

- **Questionable Centrality of Anchor Pages:** The methodology for selecting anchor pages does not ensure they are situated at the centroids of state clusters. Given the arbitrary nature of this selection and the substantial overlap among state clusters, anchor pages might not adequately represent their respective clusters.

- **Arbitrary Threshold Selection:** The criteria for establishing local and global thresholds lack explicit justification, rendering the basis for these critical parameters as arbitrary. Consequently, the reliability of results derived from these thresholds is questionable, as altering them could lead to markedly different outcomes.

- **Ambiguity in the "Other" Page Classification:** The "other" category amalgamates both international pages and domestic U.S. pages lacking specific local details. This aggregation complicates the algorithm's ability to discern and accurately classify pages within this broad and heterogeneous category.

2.5 Missing Data Imputation

2.5.1 Missingness Mechanisms

Analysis with missing data has been an active research field for recent decades in statistics and is gaining increasing attention in the machine-learning community. Rubin introduced

three types of missing data mechanisms [14], which are: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). MCAR means that the probability of a data point having a missing value for an attribute does not depend on either the observed data or the missing data in the dataset. MAR means that the probability of a data point having a missing value for an attribute depends on the observed data in the dataset, but not on the missing data. NMAR means that the probability of a data point having a missing value for an attribute depends on the missing data itself, such as the value of the missing data, and not on the observed data in the dataset.

2.5.2 Methods of Missing Data Imputation

There are various missing data imputation methods proposed. Traditional methods include listwise deletion, where data with missing values are discarded, and single imputation methods such as mean/mode imputation, regression imputation, and stochastic regression imputation [14, 15]. Some modern missing data techniques, like maximum likelihood estimation [16] and multiple imputations [17], are also employed. In the machine learning community, machine learning and deep learning algorithms are applied to predict or impute the missing values [18, 19, 20, 21], including methods like K-nearest neighbors [22], multilayer perceptron [23], generative adversarial network [24], and autoencoder [25, 26].

It is worth mentioning that Graph Neural Networks (GNNs) are recently applied to general data missing problems [27, 21, 28], not necessarily limited to network data. For example, all data points and all features are formulated into data nodes and feature nodes in a bipartite graph [27]. If a feature value of a data point is present, an edge with weight is created from the data point to the feature. Thus, missing data imputation is formulated to predict the edge weights between data nodes and feature nodes. The sizes of the datasets in this study range from 314 data points with 6 features to 45,000 data points with 9 features. However, this method is not suitable for our case. First, our dataset contains about 6 million data points with 306 features, which raises scalability issues. Second, transforming the networks into a data/feature bipartite graph results in

the loss of all network topology information.

2.5.3 Missing Data in Social Networks

In social network studies, missing data is also a frequent issue. One of the main causes of missing data in social networks is survey non-response [29]. This issue can lead to missing nodes/edges and attribute values. Most influential network data imputation studies focus on the actor/ties (node/edge) non-response data imputation, thus reconstructing the missing parts of the networks [29, 30]. Imputation of missing attribute data in networks is less popular, as many of the missing data imputation methods mentioned above can be applied.

2.5.4 Case Study: Facebook Public Page

In our Facebook public page study, we also encounter missing data issues. Many page managers did not provide the country/state/city location on the Facebook platform, which occurs case by case and depends on neither the observed data nor the missing data in our dataset. We can classify this as MCAR (Missing Completely at Random). Our focus is on state-level location data within the U.S., which offers finer granularity than country-level location data. Therefore, we exclude all pages with unknown country data and non-U.S. pages. Excluding non-interested nodes from the network and studying an interested subgraph is a convention in social network research [31], especially given our data scale of dozens of millions of nodes.

Within the U.S. page subgraph, all pages are labeled with a city name or community name by their page managers, but not explicitly with the state name. Many cities share the same names across different states, causing ambiguity and resulting in these pages being categorized as city-unknown and state-unknown. The probability of missing state labels depends on the observed city label in the dataset. We classify this scenario as Missing at Random (MAR).

Comparing our dataset with those used in other research on data imputation and network data imputation, we identified significant differences. First, our dataset is of a very large scale, containing 2 million complete data points and 3 million with missing labels,

whereas the datasets in other studies typically comprise only a few tens of thousands of data points or nodes at most [32, 29, 28, 27, 21, 20, 18]. Medium to small networks are sensitive to missing data [32] since less complete data will introduce more bias. We argue that our 2 million complete data points provide more robustness against missing data comparing small datasets. Second, all the datasets used in the aforementioned studies on missing data are simulated from complete datasets without any missing data, allowing these studies to verify the results of data imputation experiments. In contrast, our dataset consists of large-scale real data, for which there is no ground truth to verify the accuracy after data imputation.

Since there are no missing nodes or edges within the U.S. page subgraph, we focused solely on imputing the missing labels. We leveraged the network structure and the concept of homophily in networks. Node distance is naturally defined by the connecting edges in the network. Our most-neighbors labeling method is a type of K-NN imputation. However, we could not verify the imputation performance due to the absence of ground truth. The complete data points, missing label data points, and the combination of these two could exhibit three different data distributions.

2.6 Facebook Page State Classification

Given the limitations of both the majority voting algorithm and the BFS-based machine learning approach, there is a pressing need for innovative methods that address these challenges. Considering the interconnected nature of pages within a graph, where edges signify 'likes' between pages, leveraging graph-based algorithms emerges as a logical solution. Graph neural networks (GNNs), in particular, stand out as an ideal choice for sub-location classification within country borders. GNNs have the capability to understand the topological relationships between pages, utilizing this knowledge effectively to classify pages based on their sub-locations.

2.6.1 Neighborhood State Distribution Vector

Machine learning algorithms need features associated with the pages to perform training and classification. We know the distances from each page to each state anchor page, also

called state distance vector (SDV) in Lin's BFS-based machine learning algorithms, are not a good feature because about 60% of the pages in the data have zero vectors as their SDVs.

Given the limitations in the majority voting and BFS-based machine learning algorithms, we propose the utilization of neighborhood state distribution vector (NSD) as an alternative feature set. Unlike state distance vectors (SDVs), every page within the connected graph possesses a nonzero number of neighbors, ensuring that NSDs are always nonzero and, therefore, potentially more informative. NSDs quantify the proportion of a page's neighbors from each state relative to its total number of neighbors, providing a more robust feature for machine learning classification.

To capture a comprehensive picture of the local neighborhood, we incorporate NSDs calculated within both one-hop and two-hop distances, accounting for neighbors in inward, outward, and undirected edge directions. This multi-dimensional approach allows for a richer representation of page connections, improving the potential for accurate classification. The NSD vector for a page is defined as follows:

$$
\begin{aligned}
&NSD(Page) = [\\
&[INSD_1(Page, State_i), ONSD_1(Page, State_i), UNSD_1(Page, State_i), \\
&\ INSD_2(Page, State_i), ONSD_2(Page, State_i), UNSD_2(Page, State_i)] : \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i \in 1, ..., N_{number\ of\ states}]
\end{aligned}
$$

Where:

- $INSD_k(Page, State_i)$ denotes the inward neighborhood state distribution for $State_i$, calculated within a k-hop distance from the $Page$.
- $ONSD_k(Page, State_i)$ denotes the outward neighborhood state distribution for $State_i$, calculated within a k-hop distance from the $Page$.
- $UNSD_k(Page, State_i)$ denotes the undirected neighborhood state distribution for $State_i$, calculated within a k-hop distance from the $Page$.

Furthermore, each element of the NSD for a page, whether $INSD$, $ONSD$, or $UNSD$, is defined as the ratio of neighbors from state i within a j-hop distance, normalized by the total number of neighbors across all states within the same hop distance:

$$XNSD_j(Page, State_i) = \frac{XNeighbor_{ij}}{\sum_{i=1}^{N_{number\ of\ states}} XNeighbor_{ij}},$$
$$i \in \{1, ..., N_{number\ of\ states}\},$$
$$j \in \{1, 2\},$$
$$X \in \{I, O, U\},$$

Where:

- i denotes the ith state.
- j denotes the one-hop or two-hop distance.
- X denotes one of three edge directions, inward I, outward O, or undirected U.
- $XNeighbor_{ij}$: the total number of neighbors from State i within j hop distance from the $Page$ for inward I, outward O, or undirected U edge direction.

For instance, Figure 2.6 exemplifies the inward one-hop NSD for The New York Times, revealing a distribution vector of [0.4, 0.1, 0.1, 0.4] for states CA, FL, MD, and NY, respectively, with other states omitted for brevity.

For pages with a limited number of one-hop neighbors, the state distribution may exhibit bias. This bias arises because the state distribution is derived by dividing the count of neighbors from each state by the page's total number of neighbors. Consequently, a page with few neighbors can have its state distribution disproportionately affected by a few neighbors from a single state. To mitigate this bias, we also include two-hop neighbors in our analysis, thereby increasing the neighbor count for each page and diminishing the potential for bias in the state distribution. However, we refrain from extending our

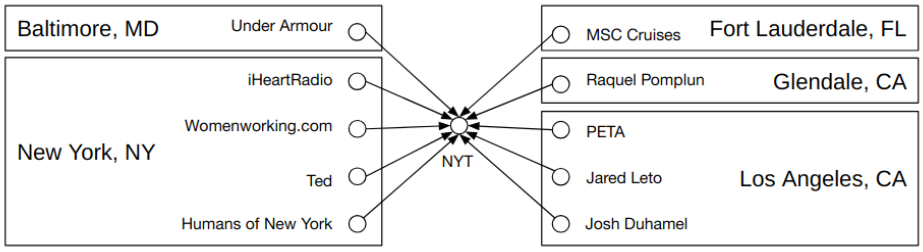


Figure 2.6: Example of inward one-hop neighborhood state distribution for the New York Times

consideration to three-hop neighbors, as this would significantly enlarge the total neighbor count, potentially reaching millions. Such an extensive neighbor set would render the state distribution vectors too homogeneous across nodes, due to an excessively broad receptive field.

2.6.2 Graph Neural Network Model Selection

The Graph Convolutional Network (GCN) model employs a neighborhood aggregation scheme, where the feature vector of each node is updated through message passing and aggregation from its neighbors' feature vectors[7][8]. Initially, we considered using the GCN as our baseline GNN model. However, GCN encounters scalability issues due to its requirement to update the feature vectors for all nodes simultaneously and its reliance on the entire graph's adjacency matrix for computing aggregated messages at each layer, resulting in substantial GPU memory usage[8]. Given the considerable size of our graph data, which exceeds GPU memory capacity, we decided against adopting the GCN model.

GraphSAGE[9], in contrast, updates feature vectors following a similar propagation rule to GCN but differs significantly in its approach. Unlike GCN, which updates all nodes' feature vectors in each iteration, GraphSAGE updates a subset of nodes per iteration by uniformly sampling a fixed number of neighboring nodes for each node in the batch[9][33]. This method significantly reduces both memory and computational demands, enabling GraphSAGE to efficiently process large-scale graphs like ours. Consequently, we selected GraphSAGE as our baseline model. Figure 2.7 illustrates the process of sampling neighbors for the New York Times in a two-layer GraphSAGE model, with the sampled pages

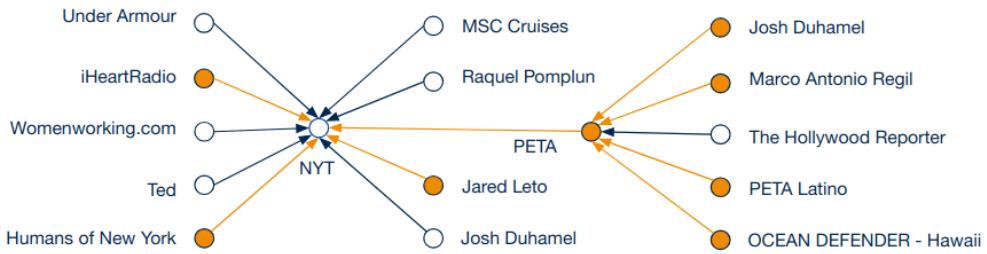


Figure 2.7: Example of a two-layer GraphSAGE neighbor sampling

highlighted in orange.

We also incorporated the GraphSAINT[10] model into our analysis. GraphSAINT diverges from GraphSAGE's neighborhood sampling technique by employing graph sampling. It runs a complete GCN-like model on a subgraph sampled from the original graph for each batch. This approach of reducing the original graph to manageable subgraphs allows GraphSAINT to accommodate large graphs while offering enhanced training efficiency and speed.

2.7 Evaluating Page State Classification

2.7.1 Model Setup

Our implementations of the GraphSAGE and GraphSAINT models comprise two layers each, with the number of output channels set to 51. This figure corresponds to the total number of states in the United States, including Washington D.C. The output represents the probabilities for each of the 51 classes, indicating the likelihood of a page belonging to a specific state. Both the number of input channels and hidden channels are configured to 306, aligning with the number of features in the neighborhood state distribution feature vectors. These models are developed using the PyTorch Geometric (PyG) framework for Graph Neural Networks[34].

2.7.1.1 GraphSAGE

GraphSAGE serves as our baseline model due to its efficiency in updating feature vectors by loading only the sampled neighboring nodes for each node in a batch, significantly reducing memory usage compared to the GCN approach of loading the entire graph into

Table 2.2: Accuracy for state-known pages (A)

Algorithm	Precision	Recall	F1-score	Overall Accuracy
Majority Voting	-	-	-	0.7308
BFS-based ML	0.7019	0.6620	0.6718	0.6620
GraphSAGE	0.8715±0.0004	0.8684±0.0002	0.8678±0.0003	0.8682±0.0006
GraphSAINT	0.8770±0.0004	0.8752±0.0003	0.8756±0.0003	0.8752±0.0002

GPU memory. However, uniformly sampling a fixed number of neighboring nodes introduces random bias and may slow the model's convergence.

To mitigate the random bias, we opted to aggregate messages from all neighboring nodes for each node, foregoing the sampling of neighboring nodes. This approach, while eliminating random bias, introduces the "neighbor explosion" issue for nodes with a high degree of connections in the batch[10]. This phenomenon refers to the computational challenge of aggregating messages from a vast number of neighbors for high-degree nodes.

By adjusting the batch size downward, we can manage this issue at the cost of increased training time. A reduced batch size decreases the likelihood of including multiple high-degree nodes within the same batch, thereby lessening the computational load. Consequently, our GraphSAGE implementation resembles the GCN model in its operation but updates feature vectors in batches rather than processing the entire graph in each iteration.

2.7.1.2 GraphSAINT

GraphSAINT differentiates itself by sampling a subgraph from the original graph for each batch during every iteration. GraphSAINT[10] provides multiple choices for the sampling schemes, such as random node sampling, random edge sampling, and random walks sampling that samples the nodes by their importance intuitively. We opted for the random walk sampler, which typically yields superior performance.

2.7.2 Accuracy for State-known Pages

Our experimental dataset comprises the largest connected subgraph of ground truth U.S. pages, as detailed in Section 2.3.3. This dataset includes both state-known and state-unknown pages, described in Section 2.3.4. State-known pages, whose state locations are unambiguous, serve as our direct ground truth. In contrast, state-unknown pages, with uncertain state locations, cannot be directly utilized for this purpose. To circumvent this limitation, we incorporate state-unknown pages within the graph while exclusively considering the state labels of neighboring state-known pages for the computation of neighborhood state distribution feature vectors. The state labels of state-unknown page neighbors are disregarded due to their uncertainty. Consequently, our models are trained and tested solely on state-known pages, constituting Dataset A.

Previous efforts to address the sub-location classification challenge, namely the majority voting algorithm[4] and the BFS-based machine learning algorithm[5], are discussed in Section 2.4. These methods serve as benchmarks for comparison. All experiments are conducted using the same dataset of state-known U.S. pages. Our GNN algorithms, GraphSAINT and GraphSAGE, significantly outperform both the majority voting and BFS-based machine learning algorithms, as evidenced by the accuracy results presented in Table 2.2. The reported accuracy for GraphSAINT and GraphSAGE reflects the mean and 95% confidence intervals over three experimental runs.

2.7.2.1 Node Sampling

The accuracy of the GraphSAINT model surpasses that of GraphSAGE by a small margin, as detailed in Table 2.2. This superior performance may be attributed to GraphSAINT's strategy of prioritizing 'important' neighbors through random walk sampling, in contrast to GraphSAGE, which considers all neighboring nodes in its message propagation without employing a sampling mechanism. The divergence in their sampling approaches results in the models engaging with different subsets of nodes from the dataset.

Table 2.3 illustrates the variance in how these models sample nodes of varying degrees, with the percentages in each row representing the average from five samples. Notably, both samplers exhibit comparable effectiveness when sampling nodes with degrees ex-

Table 2.3: GraphSAINT and GraphSAGE sampled node degree distribution

node degree	percentage in data	GraphSAINT sampler	GraphSAGE sampler
1	13.1%	15.5%	0.3%
2 - 4	29.7%	12.3%	5.7%
5 - 9	16.8%	12.9%	11.5%
10 - 99	34.2%	42%	65.3%
100 - 999	6.07%	16.72%	16.83%
1000 - 9999	0.13%	0.12%	0.18%
10000 and above	0.000834%	0.46%	0.19%
total	100%	100%	100%

ceeding 100. However, the random walk sampler utilized by GraphSAINT demonstrates a pronounced efficiency in engaging nodes with degrees below 100. Consequently, the random walk sampler achieves a node degree distribution that more closely mirrors the actual distribution within the graph. This fidelity to the real degree distribution potentially enhances model performance by offering a more accurate reflection of the diverse relationships existing among nodes of different degrees.

2.7.3 Data Expansion

State-unknown pages comprise 63.4% of our total dataset, prompting an investigation into whether their inclusion could enhance model performance. The challenge lies in accurately determining the states for these pages, whose city names are shared across different states. To address this, we propose two heuristic methods for assigning state labels to state-unknown pages:

1. **Population Labeling:** For each state-unknown page, we select the city with the highest population among those sharing the same name across different states, and assign the page to that city's state. This approach is based on the observed positive

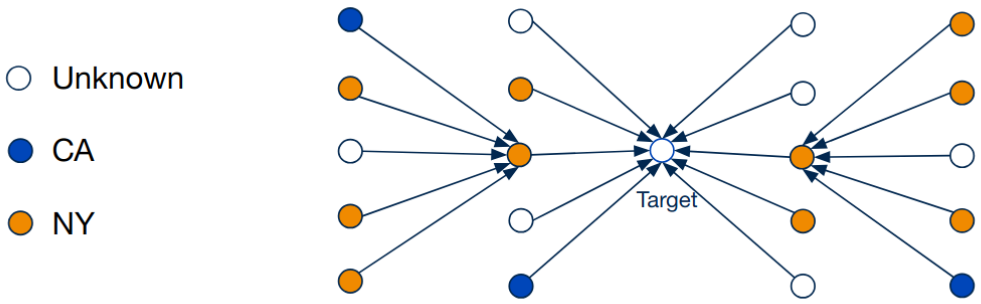


Figure 2.8: Example of Most-Neighbors labeling

Table 2.4: Accuracy for labeling methods

A: State-known Pages		B: State-unknown Pages		
Labeling Method for B	Training Data	Test Data	GraphSAGE Accuracy	GraphSAINT Accuracy
-	A	A	0.8682	0.8752
Most-neighbors labeling	A + B	A	0.8256	0.8433
Population labeling	A + B	A	0.8285	0.8394
Most-neighbors labeling	B	A	0.8015	0.8129
Population labeling	B	A	0.8017	0.8103

correlation between a state's population and its number of pages[5]; states with larger populations are likely to have a higher number of pages.

2. **Most-Neighbors Labeling:** Each state-unknown page is labeled with the state of the majority of its neighbors within a two-hop distance. This method leverages the principle that similar nodes (in this context, pages) tend to cluster together or have direct connections, making the most common neighboring state the probable location for the page. Figure 2.8 illustrated the example of labeling the target page as NY.

Table 2.4 reveals that models trained exclusively on Dataset A (state-known pages)

achieve the highest accuracy when evaluated against the ground truth test set, also comprised of Dataset A. Conversely, models trained solely on Dataset B (state-unknown pages), which were labeled using heuristic methods, exhibit the lowest accuracy. This outcome underscores the limitations of both population labeling and most-neighbors labeling in accurately reconstructing state labels to the fidelity of ground truth data, with both methods yielding comparable accuracy levels.

Interestingly, models trained on a combination of Datasets A and B perform better than those trained only on Dataset B but do not match the accuracy of models trained solely on Dataset A. Among the models, GraphSAINT consistently outperforms GraphSAGE across different training scenarios, highlighting its superiority in handling this classification task. Given these findings, the GraphSAINT model, trained on Dataset A, emerges as the optimal approach for subsequent experiments.

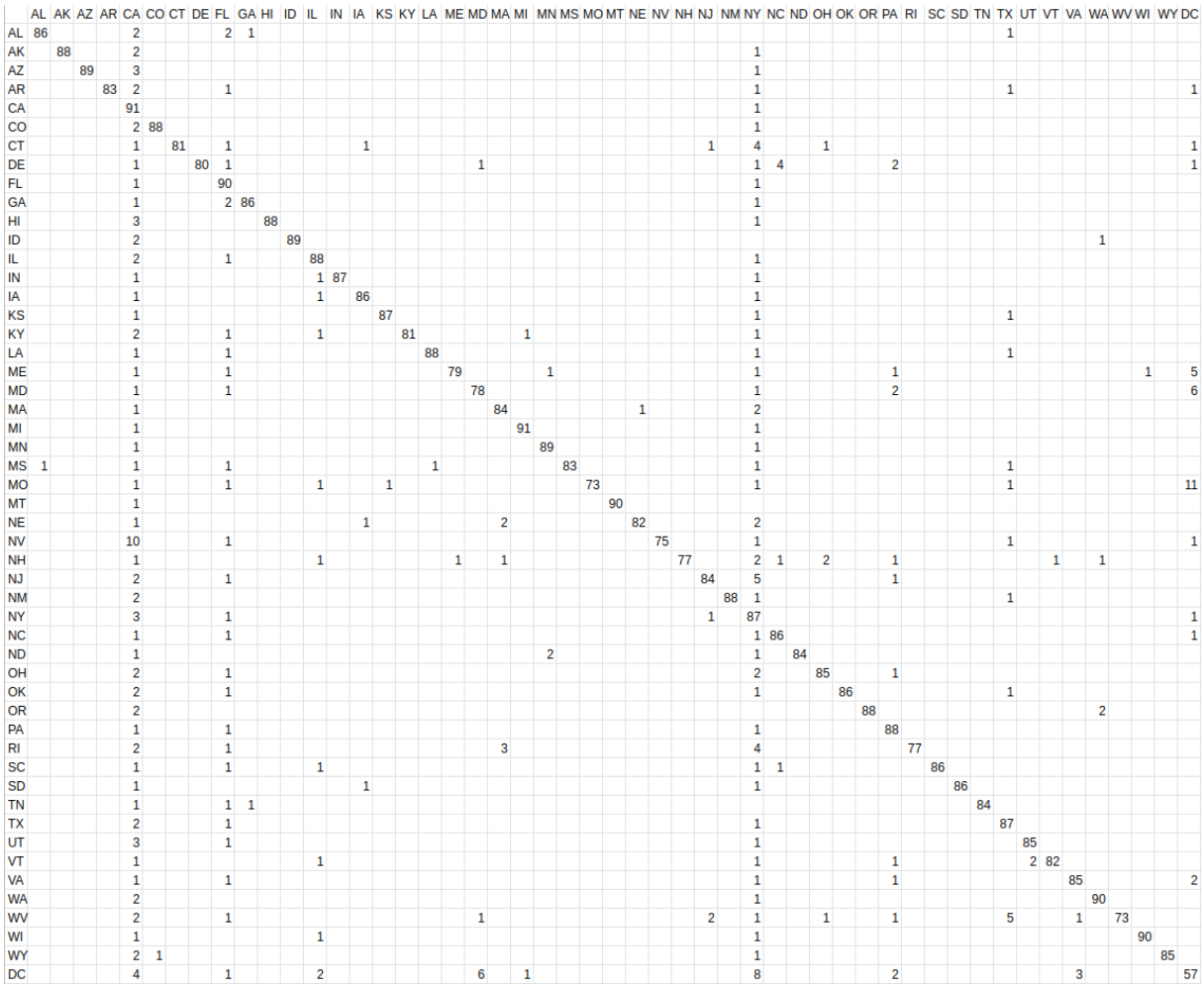
2.7.4 Confusion Matrix

The confusion matrix shows the mismatch between each class pair, revealing interesting findings hidden in the data. The confusion matrix in Figure 2.9, is computed from the ground truth label and the classification result of running GraphSAINT on state-known pages. Each state row represents how the ground truth data is classified into each state column in the matrix. The confusion matrix is normalized by ground truth data, meaning each row adds up to 100%. Every number in the matrix is a percentage number, blank cells mean the number is less than 1%.

The confusion matrix, depicted in Figure 2.9, is derived from comparing the ground truth labels with the classification outcomes of the GraphSAINT model applied to state-known pages. Each row in the matrix corresponds to a state's ground truth data, showing its distribution across the predicted states (columns). The matrix is normalized by the ground truth for each state, ensuring that the sum of each row equates to 100%. Values within the matrix represent percentages, with blank cells indicating values less than 1%.

Key insights from the confusion matrix include:

- **National Centers**: California, New York, and Florida emerge as national centers for Facebook pages in the U.S., with significant misclassification scores observed



	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY	DC
AL	86				2				2	1																						1												1							
AK		88			2																										1																					
AZ			89		3																										1																					
AR				83	2				1																						1												1									1
CA					91																										1																					
CO					2	88																									1																					
CT					1		81		1						1															1	4			1																		1
DE					1			80	1											1											1	4				2														1
FL					1				90																						1																					
GA					1				2	86																					1																					
HI					3						88																				1																					
ID					2							89																																									
IL					2				1				88																		1																					
IN					1								1	87																	1																					
IA					1								1		86																1																					
KS					1											87															1																					
KY					2				1				1				81					1									1																					
LA					1				1									88													1																					
ME					1				1										79				1								1																					1
MD					1				1											78											1																					
MA					1																84						1				2																					
MI					1																	91										1																					
MN					1																		89										1																					
MS	1				1				1									1						83								1																					
MO					1				1				1			1									73							1																					
MT					1																					90																										
NE					1										1						2						82					2																					
NV					10				1																			75				1																					1
NH					1								1						1		1								77			2	1		2		1							1		1					
NJ					2				1																					84		5					1													
NM					2																										88	1																			
NY					3				1																					1		87																				1
NC					1				1																							1	86																		1
ND					1																		2									1		84																	
OH					2				1																							2			85		1														
OK					2				1																							1				86															
OR					2																															88														
PA					1				1																							1				88														
RI					2				1													3										4					77													
SC					1				1				1																			1	1					86												
SD					1										1																	1							86											
TN					1				1	1																												84											
TX					2				1																							1							87										
UT					3				1																							1								85									
VT					1								1																			1					1				2	82							
VA					1				1																							1					1						85						2
WA					2																											1									90						
WV					2				1											1										2		1			1		1					5		1		73		
WI					1								1																			1																90		
WY					2	1																										1																85	
DC					4				1				2							6		1										8					2									3					57

Figure 2.9: Confusion Matrix

across nearly all states. This trend suggests a higher likelihood of pages from various states being connected to pages from these three dominant states, attributed to their having the largest number of pages.

- **Regional Centers**: Texas, Pennsylvania, and Illinois function as regional centers. Pages from states adjacent to these regions are more prone to being incorrectly classified as belonging to one of these three states, highlighting their influence as regional centers.

- **Neighboring States**: Notably, states sharing borders, such as Nevada and California, New Jersey and New York, Connecticut and New York, Rhode Island and New York, Washington D.C. and Maryland, Washington D.C. and Virginia, Rhode

Island and Massachusetts, and Oregon and Washington, exhibit elevated misclassification scores. This pattern underscores the tendency for pages from neighboring states to be more interconnected than those from distant, non-center states.

To further explore the center states' influence, we conducted a control experiment by excluding all center state labels from Dataset A. The GraphSAINT model was then trained and evaluated exclusively on pages from non-center states. Upon training completion, we applied the model to classify pages in Dataset A from both center and non-center states, but restricted the classification to non-center state labels only. This approach notably increased the mislabeled pages among neighboring states to the center states. However, this adjustment inadvertently skewed the dataset, with DC, Washington, and New Jersey emerging as prominent new center states. These states exhibited a marked increase in mislabeled pages with nearly every other state.

2.7.5 Intrastate Page and Interstate Page

2.7.5.1 Definition

In our multi-class classification framework, we calculate the cross-entropy loss by comparing the ground truth labels with the predictions from our GNN models. The model outputs, representing unnormalized scores for each class, do not necessarily have to be positive nor sum to one. By applying the softmax function[35] to the outputs generated from state-known Dataset A, we obtain the probabilities of a page being associated with each state. These probabilities range from 0 to 1, with their total summing to 1.

Rather than picking one state with the highest probability as the prediction for the page in classification, we are interested in all the states with relatively high probabilities for one page. We define intrastate pages and interstate pages as follows:

- **Intrastate page**: A page associated with only one state having significantly high probabilities.
- **Interstate page**: A page associated with more than one state having significantly high probabilities.

Table 2.5: Interstate and Intrastate page example

Page id	ground truth	Cut off probability	High probability states
5606629547	FL	0.08	FL 0.37, IL 0.17, DC 0.20
5479739307	NY	1.23e-07	NY 0.54, DC 0.45
4846711747	CA	5.94e-15	CA 1.0
5602549475	NJ	0.04	NJ 0.76

To effectively categorize the 51 probabilities into two distinct groups—those with higher probabilities and those with lower—we employ the Jenks natural breaks algorithm [36]. This technique aims to minimize the variance within each group, ensuring that the probabilities grouped together are as similar as possible. The group with the higher probabilities includes a specified number of probabilities associated with different states. If this group comprises probabilities for more than one state, the page is classified as an interstate page; if it contains probabilities for only one state, the page is considered an intrastate page. Table 2.5 presents examples of both interstate and intrastate pages, illustrating the criteria for their classification.

2.7.5.2 Interstate Page Distribution

Table 2.6 details the distribution of pages across varying numbers of states with high probabilities within Dataset A. Among the total, 213,035 pages are classified as interstate, constituting 9.92% of the combined count of 2,147,399 interstate and intrastate pages. The proportions of interstate and intrastate pages for each state are further enumerated in Table 2.8.

2.7.5.3 State Interstate Page Percentage

Figure 2.10 illustrates the percentage of interstate pages for each state across the U.S. map. The interstate page percentage is defined as the ratio of interstate pages to the total number of pages within a state. The map reveals that Nevada, Missouri, West Virginia, Virginia, and Washington D.C. exhibit the highest percentages of interstate pages. To delve deeper into the dynamics of interstate pages, it is necessary to examine the number of interstate pages between each pair of neighboring states.

Table 2.6: Page distribution with different numbers of high-probability states in state-known page data A

High-probability States	Page Numbers
1	1934364
2	120582
3	37357
4	18959
5	10758
6	7319
7	4652
8	3437
9	2426
10	1691
11-20	5690
21-31	164

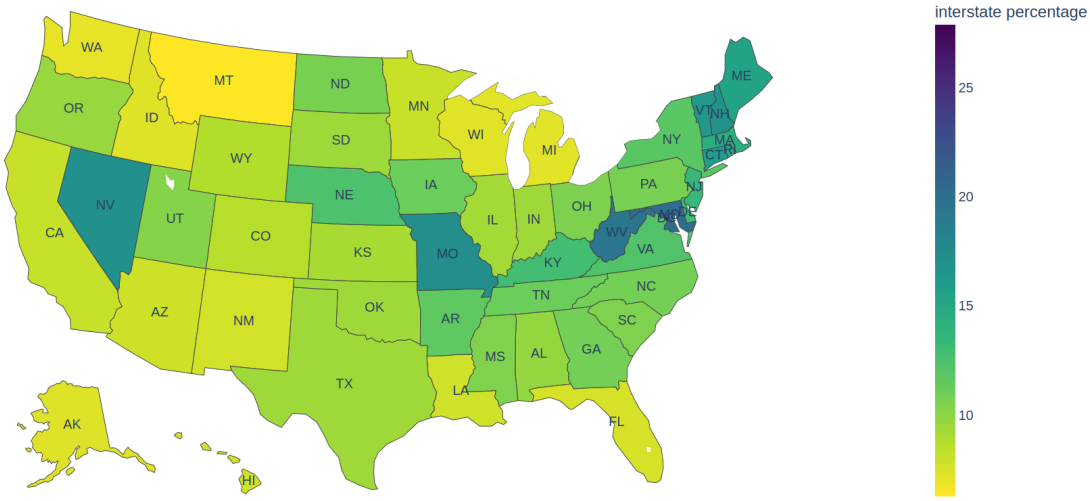


Figure 2.10: Interstate page percentage map

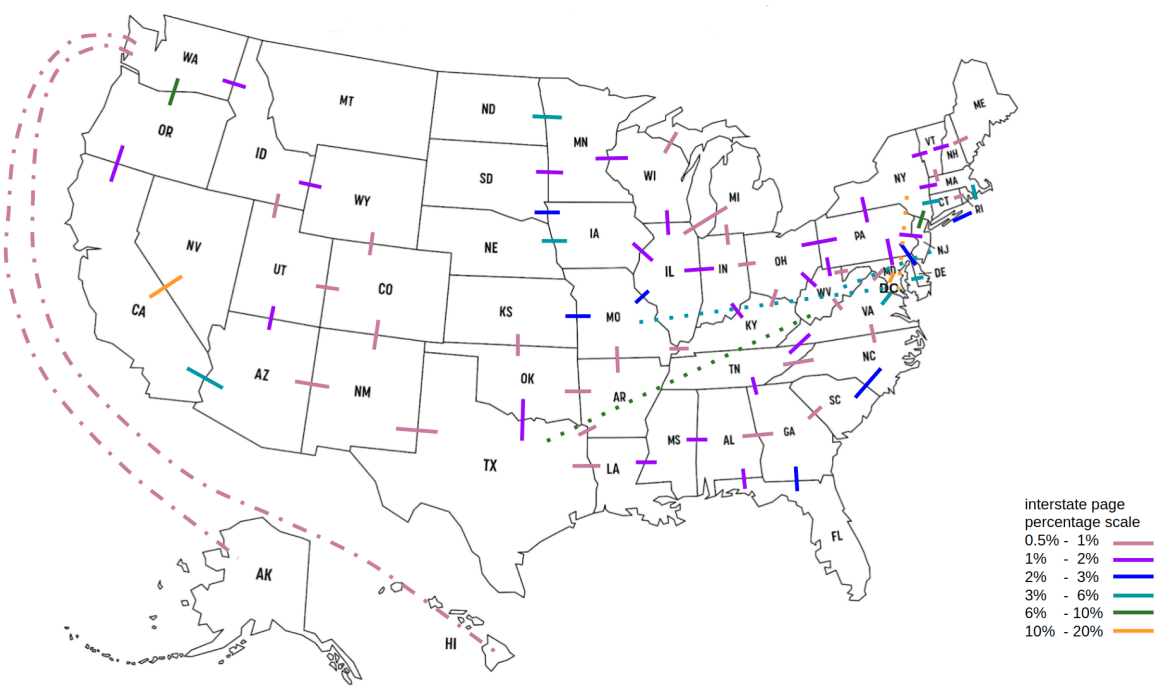


Figure 2.11: Numbers of interstate page across state borders

2.7.5.4 Interstate Pages Across Borders

In Figure 2.11, we depict the distribution of interstate pages across borders for each pair of neighboring states, normalizing the number of interstate pages by the total number of pages of the state with fewer pages in each pair. For clarity, interstate pages constituting less than 0.5% on a border are not included. Although Alaska and Hawaii do not share borders with any states, they both have the highest number of interstate pages with Washington, reflecting geographic proximity over direct borders. We can see that the high interstate page percentage states, Nevada, Missouri, West Virginia, and Virginia, have more interstate pages shared with their neighboring states, and some center states. Specifically, Nevada shares a significant number of pages with California; Missouri with Illinois, Kansas, and the District of Columbia (DC); West Virginia with Texas, New Jersey, Pennsylvania, and Ohio; and Maryland with DC and Delaware. Interestingly, DC emerges as a sub-regional center not evident in the confusion matrix, showing substantial interstate page sharing with Maryland, New York, Virginia, California, Missouri, North Carolina, and Pennsylvania.

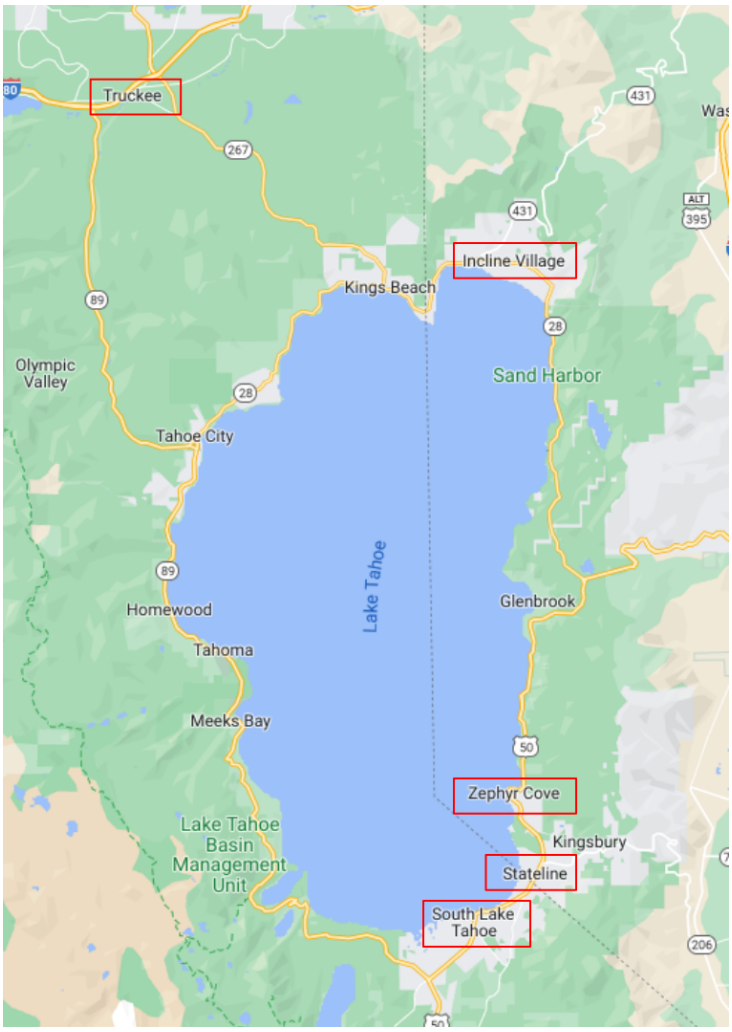


Figure 2.12: Interstate pages from border cities of CA and NV

2.7.5.5 Interstate Pages Fact Check

We conducted a detailed examination of interstate pages between pairs of states. Notably, interstate pages between California (CA) and Nevada (NV) predominantly originate from cities near Lake Tahoe, straddling the CA-NV border—such as Stateline, Zephyr Cove, Incline Village in Nevada, and South Lake Tahoe and Truckee in California—as shown in Figure 2.12. A similar pattern of interstate pages is observed between Maryland (MD) and Washington D.C. (DC), with pages from Washington D.C. closely connected to Maryland cities like Gaithersburg, Silver Spring, and Hyattsville, among others, as depicted in Figure 2.13.

Initially, Washington D.C. was not included in the ground truth Dataset A. However,

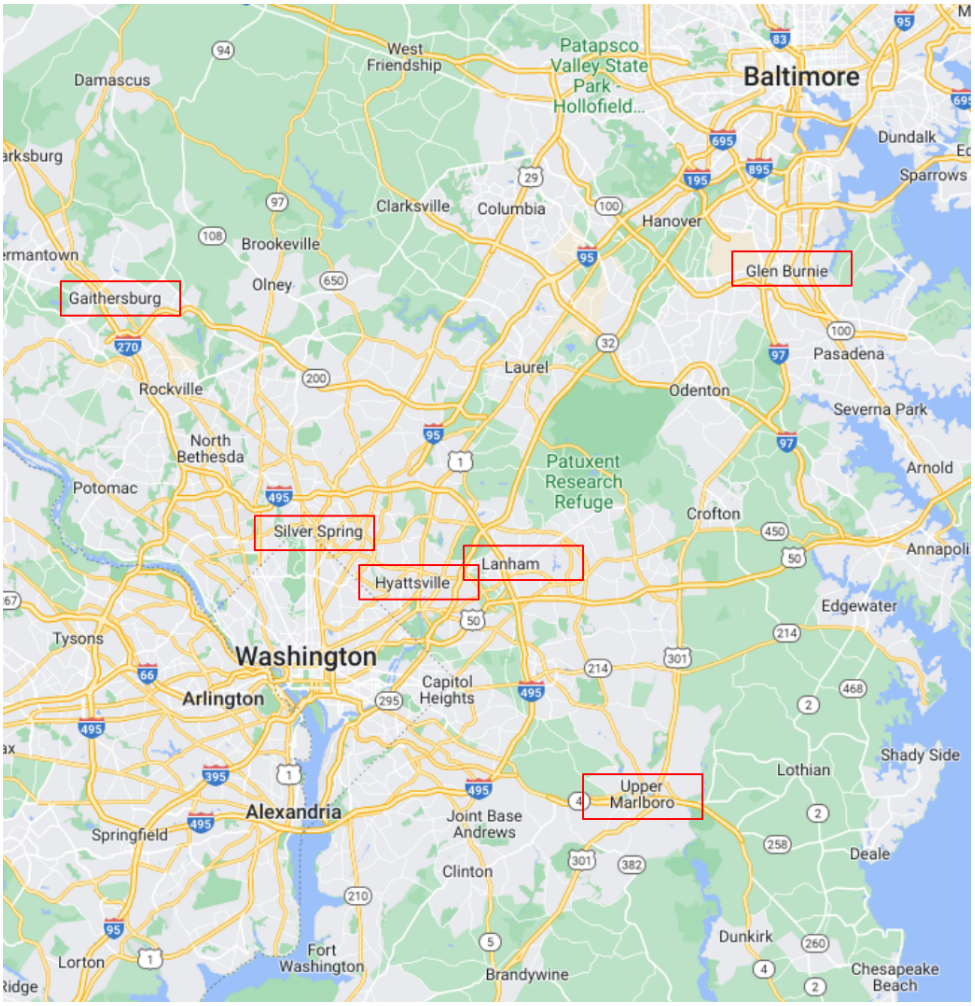


Figure 2.13: Interstate pages from border cities of MD and DC

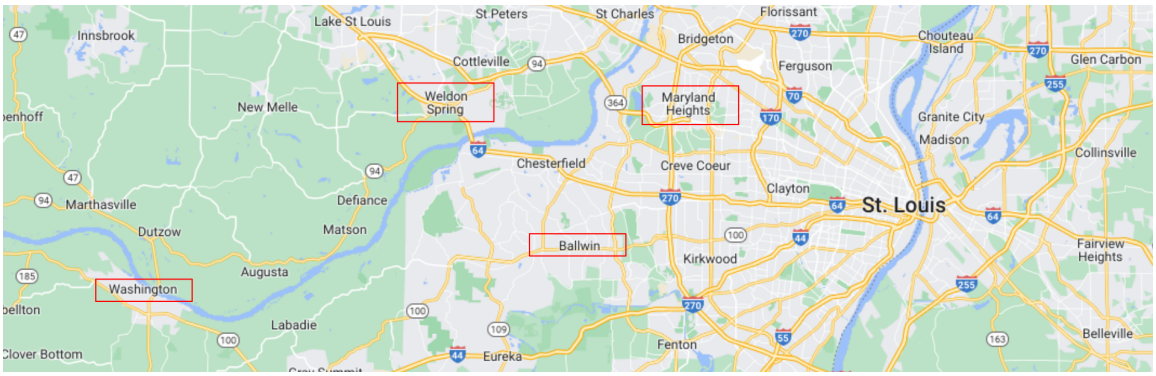


Figure 2.14: Interstate pages from border cities of MO and DC

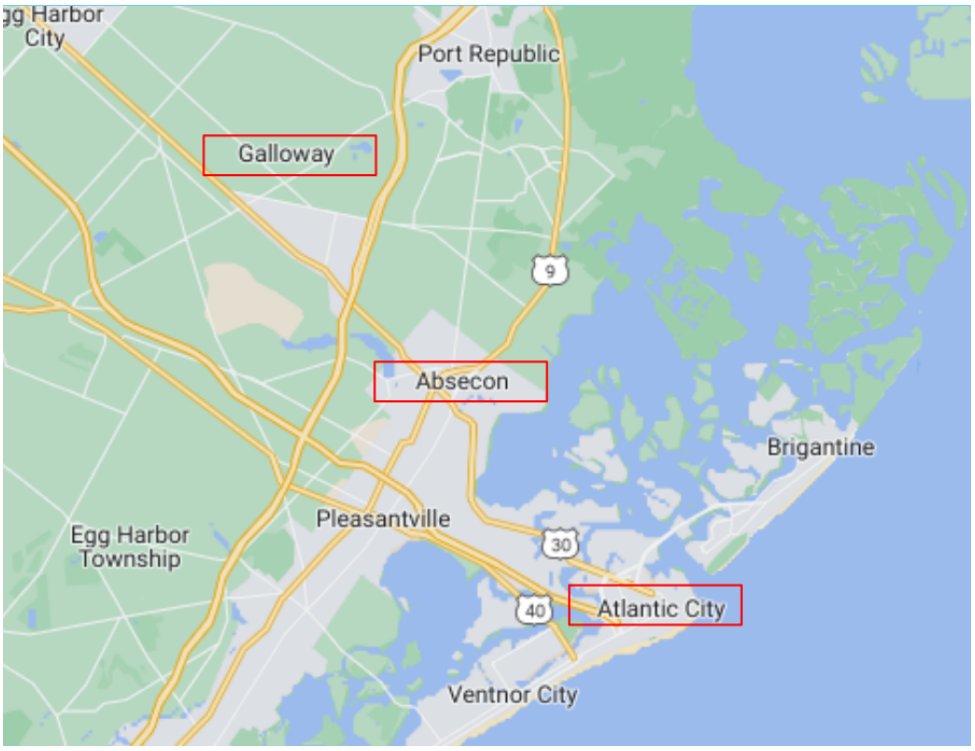


Figure 2.15: Interstate pages from border cities of NJ and WV

its significant population and location adjoining Maryland and Virginia warranted its inclusion as a noteworthy case. Figure 2.7 displays populations for various 'Washington' cities across states. Due to the smaller populations of other cities named 'Washington', we label pages from all cities of Washington as 'DC.' Consequently, many pages from Missouri (MO) and Maine (ME) were inaccurately labeled as 'DC,' likely due to mislabeling of their 'Washington' cities as 'DC' in the ground truth data. We confirmed this hypothesis by analyzing the interstate pages between Missouri (MO) and Washington D.C. (DC). As illustrated in Figure 2.14, the interstate pages originate from Missouri cities such as St. Louis, Nixa, Wentzville, and Kirksville, all of which are in proximity to Washington, MO.

Further investigation into the interstate pages between New Jersey (NJ) and West Virginia (WV) revealed that pages from Galloway, WV, were connected to Atlantic City and Absecon in NJ. The issue stems from the absence of Galloway, NJ in our U.S. city dataset, with only Galloway, WV being listed. Consequently, what should have been identified as page connections between Galloway, NJ, and the nearby cities of Absecon and Atlantic City in NJ, were incorrectly recognized as connections between Galloway,

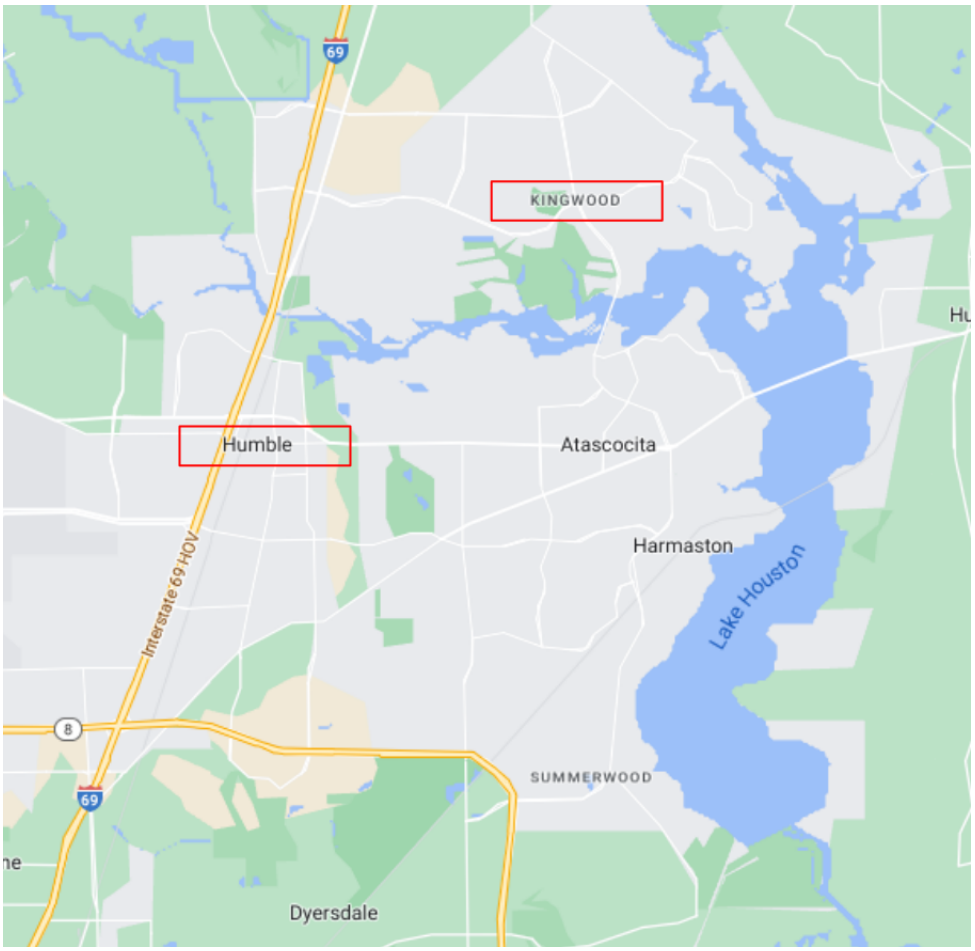


Figure 2.16: Interstate pages from border cities of TX and WV

WV, and these New Jersey cities, in Figure 2.15. This error leads to the misclassification of these connections as interstate pages between New Jersey (NJ) and West Virginia (WV).

A similar data issue was identified between WV and Texas (TX), where the absence of Kingwood, TX, in our dataset resulted in pages from Kingwood, TX, being incorrectly assigned to Kingwood, WV. This mislabeling falsely suggests interstate connections between TX and WV, as both Kingwood, TX, and nearby Humble, TX, share page connections, as shown in Figure 2.16.

This analysis underscores the utility of interstate pages in validating social ties between neighboring cities across state borders. It also highlights gaps in U.S. city data and suggests that closer cities tend to establish social connections, as reflected in our page

graph.

Table 2.7: City populations for Washington across states

State	Population	State	Population
DC	5,066,973	IN	12,514
UT	28,192	NC	9,555
IL	16,555	IA	7,318
MO	14,052	NJ	6,475
PA	13,404	GA	3,946
WV	1,303	KS	993
AR	134	NE	124
LA	860	VA	77
OK	687	CA	137

2.8 Conclusion

In this chapter, we explored the challenge of subdivision location classification for Facebook public pages within the United States. We critically analyzed the limitations of previous studies in sub-location classification and introduced a novel approach leveraging the GraphSAINT model. This model utilizes neighborhood state distribution vectors to accurately classify pages. Our evaluation on a dataset of U.S. Facebook public pages demonstrated a notable improvement in classification accuracy compared to prior methods.

Furthermore, we applied our model to distinguish between intrastate and interstate Facebook public pages. Our findings indicate that intrastate pages tend to garner "likes" from pages within the same state, whereas interstate pages are more commonly liked by pages from different states. Through an analysis of the state classification confusion matrix, the percentages of interstate pages by state, and the distribution of interstate pages across state borders, we conclude that geographic location plays a crucial role in the formation of online community networks and the accuracy of sub-location classification for Facebook public pages.

Table 2.8: Accuracy of GraphSAINT on state-known page data A

Begin of Table 2.8

State	Precision	Recall	F1 score	Intrastate Pages %	Interstate Pages %	Number of Pages
Alabama(AL)	0.90	0.86	0.88	90.2	9.8	27907
Alaska(AK)	0.94	0.88	0.91	92.7	7.3	8324
Arizona(AZ)	0.91	0.89	0.90	92.1	7.9	56288
Arkansas(AR)	0.84	0.83	0.83	88.3	11.7	10869
California(CA)	0.85	0.91	0.88	91.9	8.1	252922
Colorado(CO)	0.91	0.89	0.90	91.4	8.6	45043
Connecticut(CT)	0.81	0.81	0.81	84.1	15.9	4796
Delaware(DE)	0.86	0.80	0.83	86.9	13.1	3760
Florida(FL)	0.89	0.91	0.90	92.4	7.6	203544
Georgia(GA)	0.87	0.86	0.86	89.1	10.9	44859
Hawaii(HI)	0.94	0.88	0.91	92.3	7.7	25969
Idaho(ID)	0.93	0.89	0.91	92.7	7.3	25502
Illinois(IL)	0.88	0.88	0.88	90.7	9.3	120819
Indiana(IN)	0.89	0.88	0.89	90.6	9.4	50521
Iowa(IA)	0.90	0.86	0.88	88.8	11.2	28872
Kansas(KS)	0.91	0.88	0.89	90.9	9.1	23206
Kentucky(KY)	0.88	0.81	0.84	87.1	12.9	12856
Louisiana(LA)	0.93	0.89	0.91	92.2	7.8	47353
Maine(ME)	0.86	0.79	0.82	84.6	15.4	4787
Maryland(MD)	0.83	0.79	0.81	79.8	20.2	30884
Massachusetts(MA)	0.85	0.84	0.85	85.7	14.3	13084
Michigan(MI)	0.92	0.91	0.92	92.8	7.2	69652
Minnesota(MN)	0.88	0.89	0.89	92.0	8.0	30654

Continuation of Table 2.8

State	Precision	Recall	F1 score	Intrastate Pages %	Interstate Pages %	Number of Pages
Mississippi(MS)	0.92	0.84	0.88	89.5	10.5	10642
Missouri(MO)	0.87	0.73	0.79	82.8	17.2	20231
Montana(MT)	0.93	0.91	0.92	93.7	6.3	17009
Nebraska(NE)	0.88	0.83	0.85	87.5	12.5	8893
Nevada(NV)	0.84	0.76	0.80	83.0	17.0	4256
New Hampshire(NH)	0.86	0.77	0.81	83.1	16.9	1388
New Jersey(NJ)	0.86	0.84	0.85	86.6	13.4	46598
New Mexico(NM)	0.94	0.88	0.91	92.3	7.7	21613
New York(NY)	0.82	0.87	0.85	88.1	11.9	175611
North Carolina(NC)	0.89	0.87	0.88	89.1	10.9	54721
North Dakota(ND)	0.91	0.85	0.88	89.3	10.7	6255
Ohio(OH)	0.87	0.85	0.86	89.4	10.6	40344
Oklahoma(OK)	0.90	0.86	0.88	90.6	9.4	38480
Oregon (OR)	0.90	0.89	0.89	90.3	9.7	24246
Pennsylvania(PA)	0.87	0.89	0.88	89.3	10.7	82160
Rhode Island(RI)	0.88	0.78	0.83	85.0	15.0	3361
South Carolina(SC)	0.91	0.87	0.89	89.5	10.5	30984
South Dakota(SD)	0.92	0.86	0.89	90.5	9.5	10389
Tennessee(TN)	0.89	0.85	0.87	88.8	11.2	16198
Texas(TX)	0.89	0.88	0.88	90.6	9.4	127645
Utah(UT)	0.88	0.86	0.87	89.7	10.3	27648
Vermont(VT)	0.82	0.82	0.82	83.6	16.4	7521
Virginia(VA)	0.87	0.86	0.86	87.7	12.3	40355
Washington(WA)	0.92	0.91	0.91	93.0	7.0	97228

Continuation of Table 2.8

State	Precision	Recall	F1 score	Intrastate Pages %	Interstate Pages %	Number of Pages
West Virginia(WV)	0.86	0.73	0.79	80.6	19.4	7163
Wisconsin(WI)	0.93	0.90	0.92	92.8	7.2	45665
Wyoming(WY)	0.92	0.86	0.89	91.2	8.8	5058
Washington D.C.(DC)	0.52	0.57	0.54	72.1	27.9	33266
macro avg	0.88	0.85	0.86	90.1	9.9	2147399
weighted avg	0.88	0.88	0.88	90.1	9.9	2147399

End of Table 2.8

Chapter 3

Online Social Community Neighborhood Formation

3.1 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. As of May 2023, 33 online social platforms boast over 100 million monthly active users (MAU)[37], highlighting the vast reach of these networks. Facebook, in particular, leads as the most popular platform, with nearly 2.99 billion monthly active users. Figure 3.1 presents the top 20 online social platforms ranked by their number of monthly active users, illustrating the scale and diversity of OSNs in facilitating digital social interactions.

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

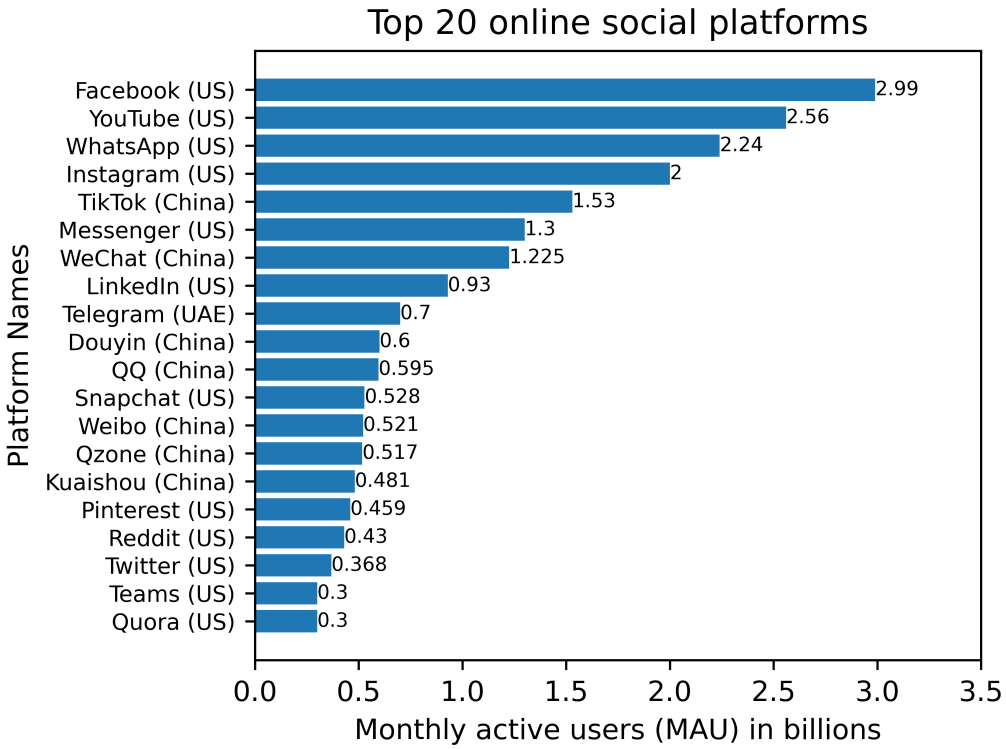


Figure 3.1: Top 20 Online Social Platforms

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.

Facebook stands out as the most popular platform for online communities, attracting

considerable attention from researchers. This study specifically focuses on public Facebook pages, which serve as a platform for disseminating information, facilitating user discussions, spreading news, and promoting businesses or public relations activities. Like individual users, these pages can like or follow other Facebook 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.

This chapter is organized as follows: Section 3.2 introduces related research on user geographic location analysis, link prediction, page geographic location analysis, and graph neural networks we used for classification. Section 3.3 describes the data used for this study and the ground truth data for verification. Section 3.4 introduced the methodology we used and some proposed node features to perform link prediction. Section 3.5 shows experiment setups, results of the experiments, and analysis of the results. Finally, sSection 3.6 offers a summary of this Chapter.

3.2 Related Work

3.2.1 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[2]. 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].

3.2.2 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 [38]. 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.

3.2.3 Online Social Community Location Classification

Facebook 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.[4] explored the Facebook 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. In Chapter 1, we introduced the concept of neighborhood state distribution vectors and applied Graph Neural Networks for the classification of Facebook pages' subdivision locations, achieving notable accuracy. This methodology offers insights into a page's influence across different states.

3.2.4 Graph Neural Networks

Graph neural networks (GNNs) are a subset of artificial neural networks designed for processing graph-structured data[6]. 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[7][8].

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].

3.3 Data Description

3.3.1 Data Acquisition

In this chapter, we use the same Facebook public page data as Chapter 1, sourced via Facebook Graph API 2.8. 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 Facebook public pages and progressively encompassing pages liked by these initial nodes. This approach organically constructs a directed graph representation of the Facebook page network.

3.3.2 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 Facebook 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 state-known pages. Conversely, the remaining 63.4% of pages are linked to cities with names that duplicate across multiple states; we classify these as state-unknown pages. Our study concentrates on the state-known pages.

3.4 Page-Likes Link Prediction

3.4.1 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[39]. 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[40].

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[41]. Our use of link prediction algorithms aims to identify key factors influencing the formation of page neighborhoods.

3.4.2 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 in Chapter 1 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.

3.4.3 Feature Selection

In our study, we delve into the dynamics behind the "likes" relationships among Facebook 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 Facebook 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.

3.4.3.1 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[42]. 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.

3.4.3.2 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 3.2, with linear and logarithmic scales used in Figure 3.2a and Figure 3.2b, respectively. 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[43], indicating adherence to a log-normal distribution.

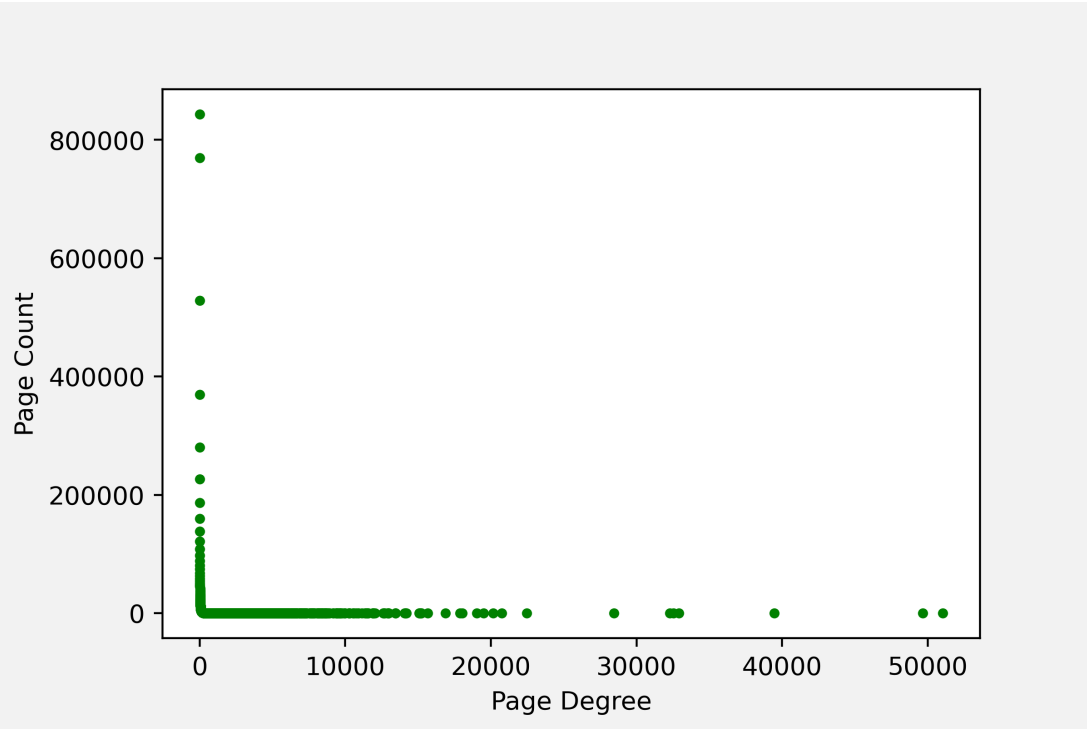
Node degree is an often used feature in network analysis. Therefore, we propose the degree of the page node as one candidate feature. The page graph is a directed graph. Hence, we use both normalized inward degree and normalized outward degree as features.

3.4.3.3 Page's Category

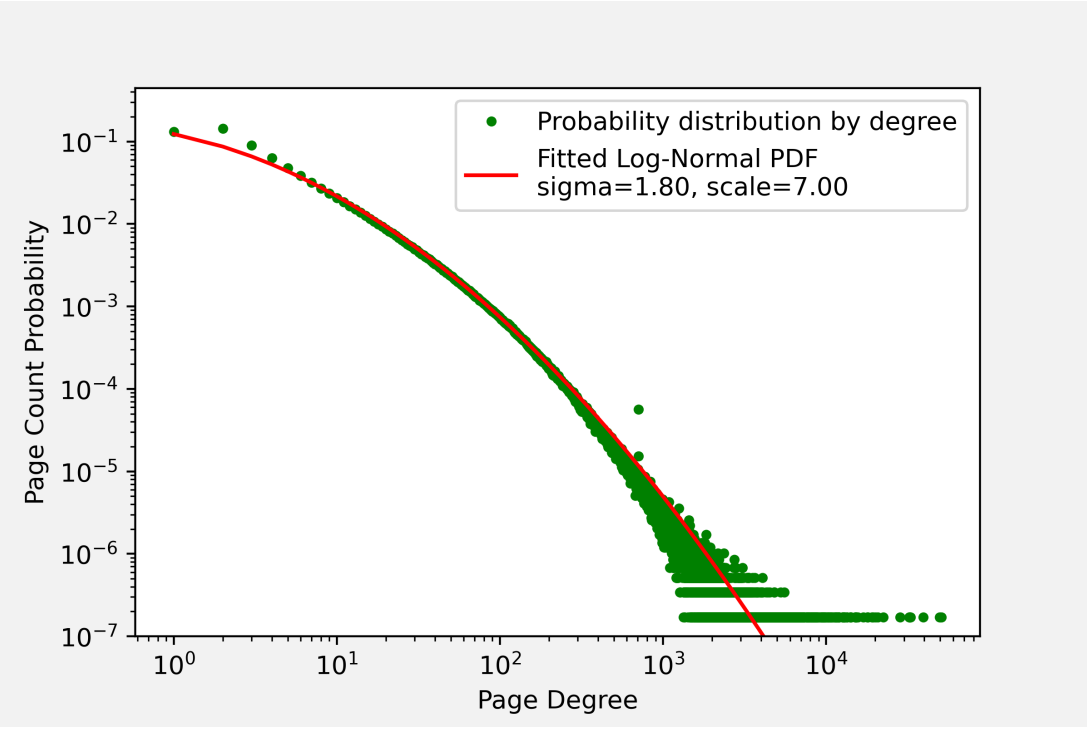
Facebook public pages categorize their topics as assigned by their managers, encompassing over a thousand distinct categories within the page-likes graph. For instance, the top 20 categories are enumerated in Table 3.1, extracted directly from the page metadata without modification. Despite the presence of duplicate categories, their impact on prediction accuracy is minimal. According to the theory of homophily[42], 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[44].

3.4.3.4 Page's State Label

In previous analyses, notably in Chapter 1, neighborhood location information has emerged as a pivotal feature for classifying pages within the Facebook page-likes graph, aligning with the principles of homophily theory[42]. 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 state-known



(a) Page Degree Distribution



(b) Page Degree Distribution on Log Scale

Figure 3.2: Page Degree Distributions

Table 3.1: Top 20 Categories of Facebook 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

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.

3.4.3.5 Page's State Neighborhood Distribution

In Chapter 1, we employed state neighborhood distribution vectors as node features for classifying the states of Facebook 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.

3.4.3.6 Page's City Label

Inspired by the insights gained from analyzing page state neighborhood distributions and aligned with the principles of homophily theory[42], our investigation extends into more granular location data of Facebook 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 state-known pages, as their city affiliations are unequivocally determined, in contrast to state-unknown 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.

3.4.3.7 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.

3.4.3.8 Page's City Geographic Coordinate

In Chapter 1, 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 state-known pages—as features to explore the extent to which geographic location factors into the formation of these online community networks.

3.5 Evaluating Page Features

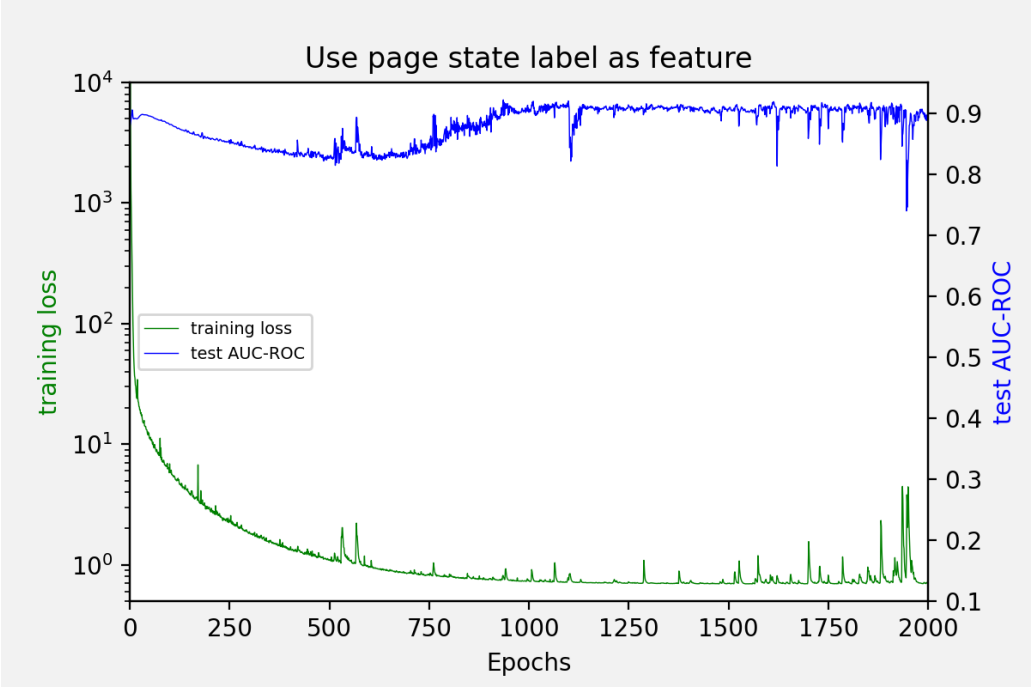
3.5.1 Experimental Setup

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.

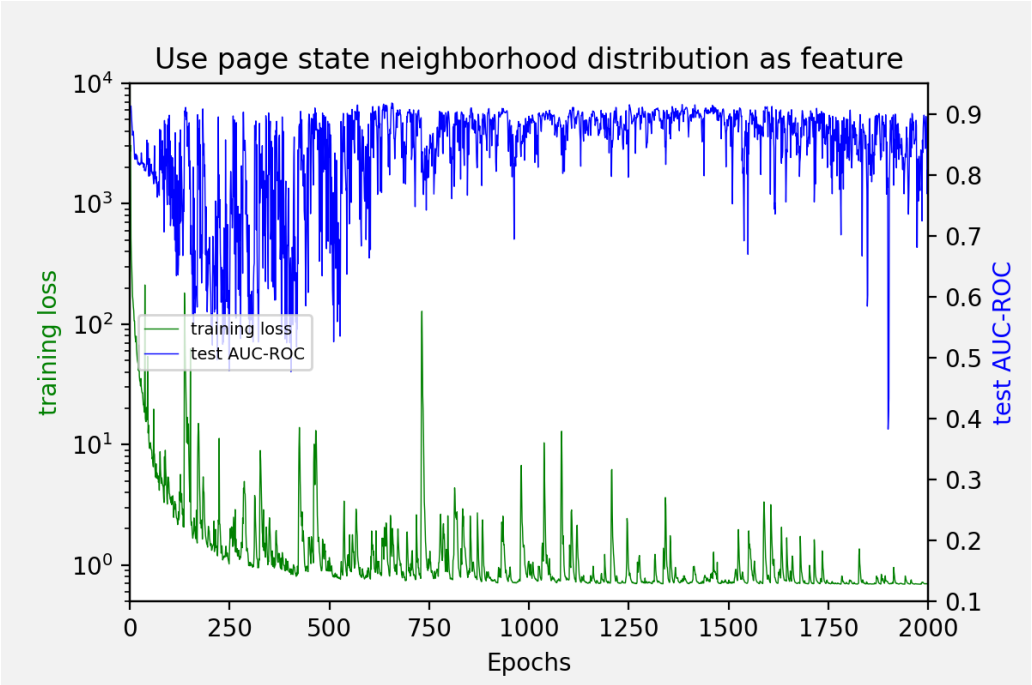
Table 3.2: Highest AUC-ROC on test set on different Positive/Negative edge ratio with page state label as feature

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

In this study, 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, contributes to a GNN layer[45]. The outputs from both layers are

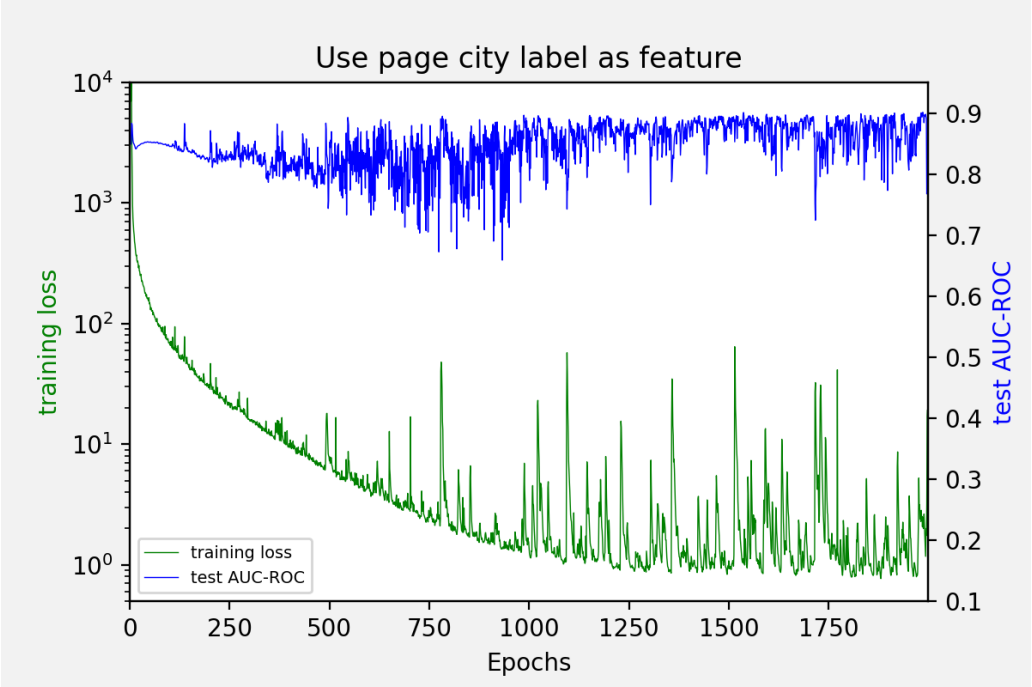


(a) page state label as feature

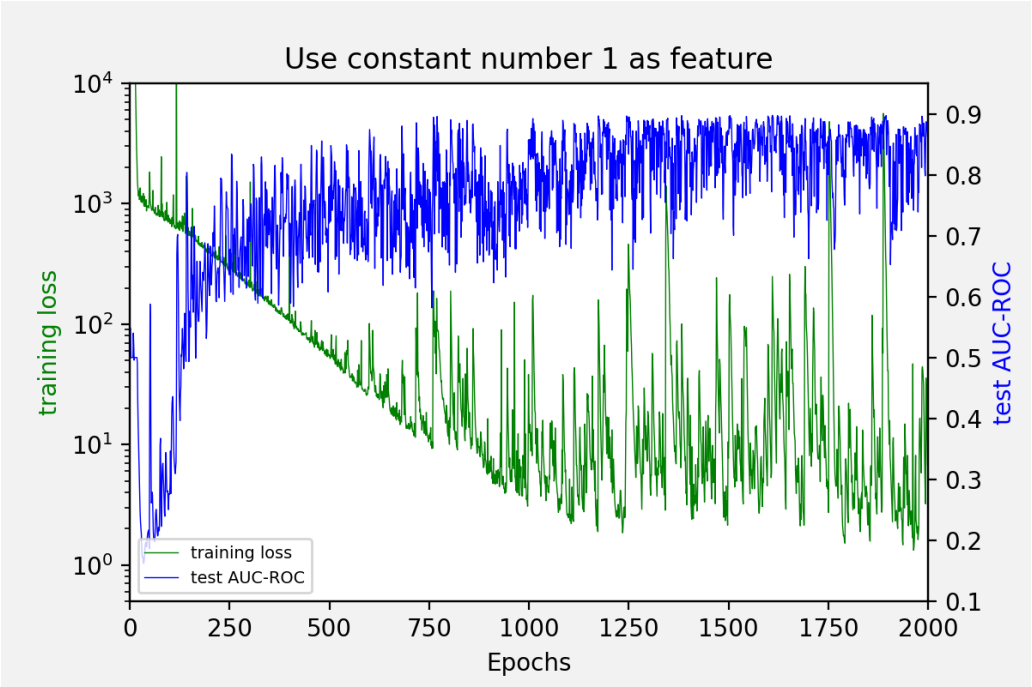


(b) page state neighborhood distribution as feature

Figure 3.3: Loss and AUC-ROC curves of different features (Part 1)

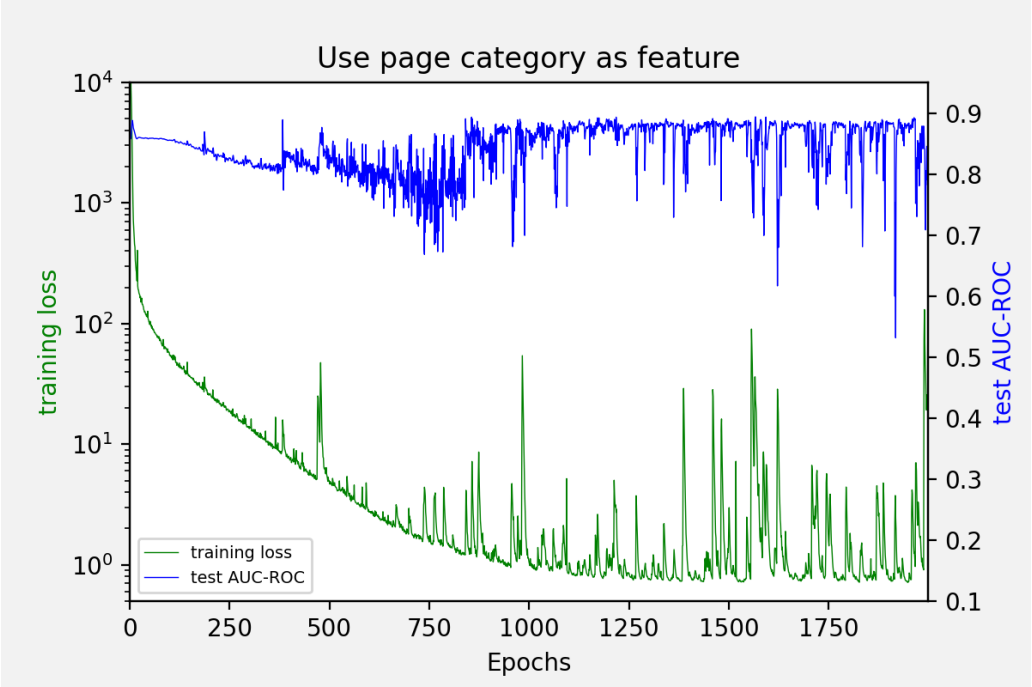


(c) page city label as feature

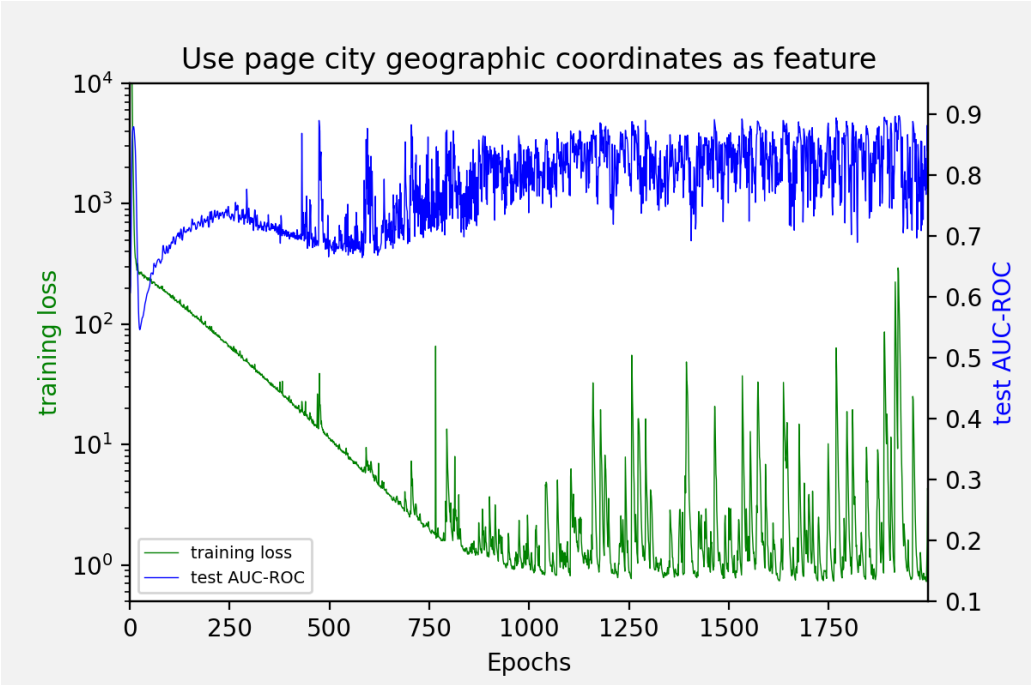


(d) constant number 1 as feature

Figure 3.3: Loss and AUC-ROC curves of different features (Part 2)

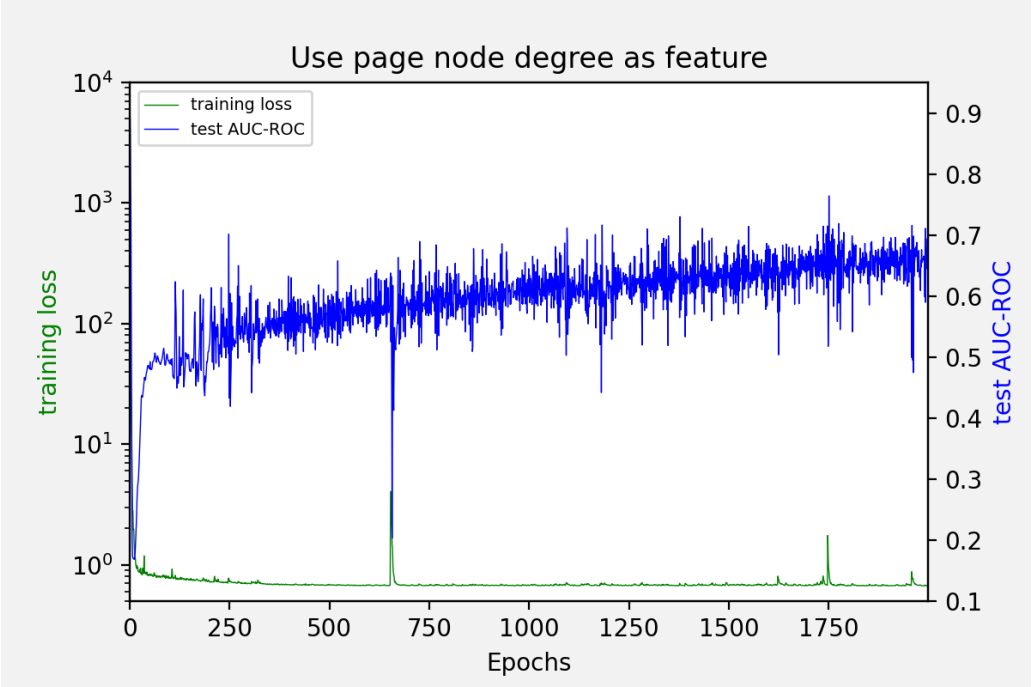


(e) page category as feature

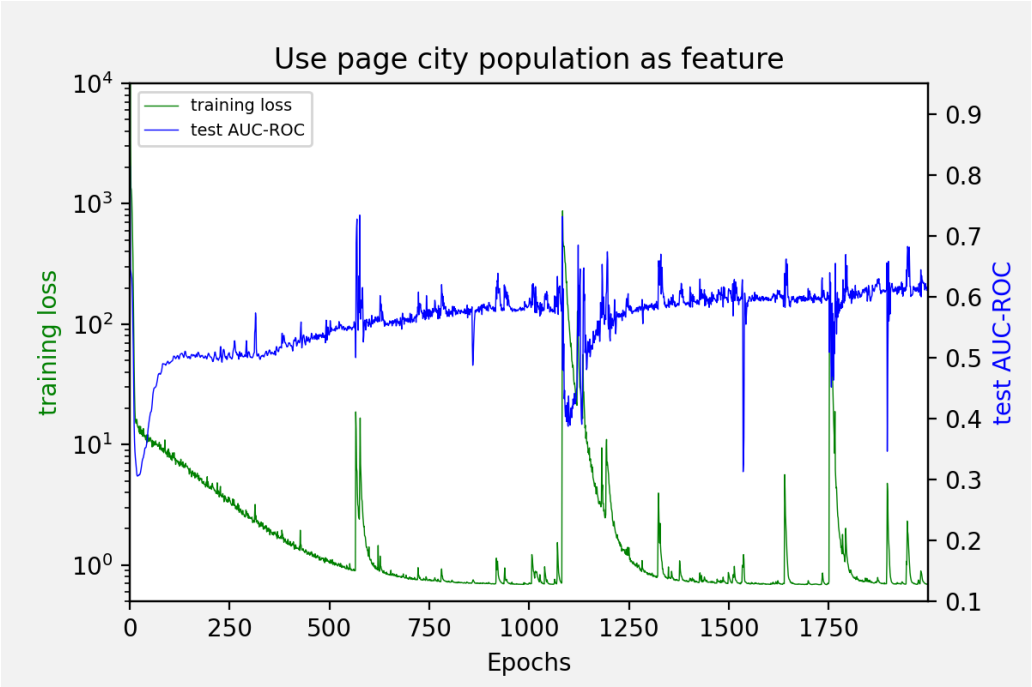


(f) page city geographic coordinates as feature

Figure 3.3: Loss and AUC-ROC curves of different features (Part 3)



(g) page node degree as feature



(h) page city population as feature

Figure 3.3: Loss and AUC-ROC curves of different features (Part 4)

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, the actual edges are significantly outnumbered by the 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[41]. 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 3.2. The training process has 2000 epochs, necessitating approximately 20 hours to complete.

3.5.2 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 3.3, the most stable training loss curve, and the most consistent testing AUC-ROC score curve in Figure 3.3a.

3.5.2.1 Performance

Table 3.3: Summary of feature analysis across the entire dataset, ordered by average AUC-ROC

Feature	Avg. AUC-ROC	Avg. TPR	Avg. TNR
Page State Label	0.9308	0.8485	0.8832
Page State Neighborhood Distribution	0.9306	0.8481	0.8729
Page 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
Page City Geographic Coordinates	0.8964	0.7844	0.8609
Page Node Degree	0.7632	0.5832	0.9672
Page City Population	0.6849	0.5140	0.9043

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 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 3.3 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 3.4 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

Table 3.4: 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

label exhibits the best performance. Page location label shows a strong effect on pages connecting with their neighbors.

3.5.2.2 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 3.3, 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 3.3a, 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 3.3b), page city label (Figure 3.3c), constant number 1 (Figure 3.3d), page category (Figure 3.3e), and page city geographic coordinates (Figure 3.3f) 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 3.5: Combine one feature with page state label feature analysis across the entire dataset

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

3.5.3 Combined Feature

We have determined that the page state label feature outperforms other single features for link prediction within the page graph. Given the relative underperformance of other features, we proceed to enhance link prediction accuracy by investigating feature combinations.

3.5.3.1 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 3.5 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 com-

Table 3.6: Combine more features with page state label feature analysis across the entire dataset

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

bined 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 3.6. The table also illustrates that merging the page category, page city geographic coordinates, constant number 1, and page city label features with the page state label feature resulted in suboptimal performance. While exhaustive combinations

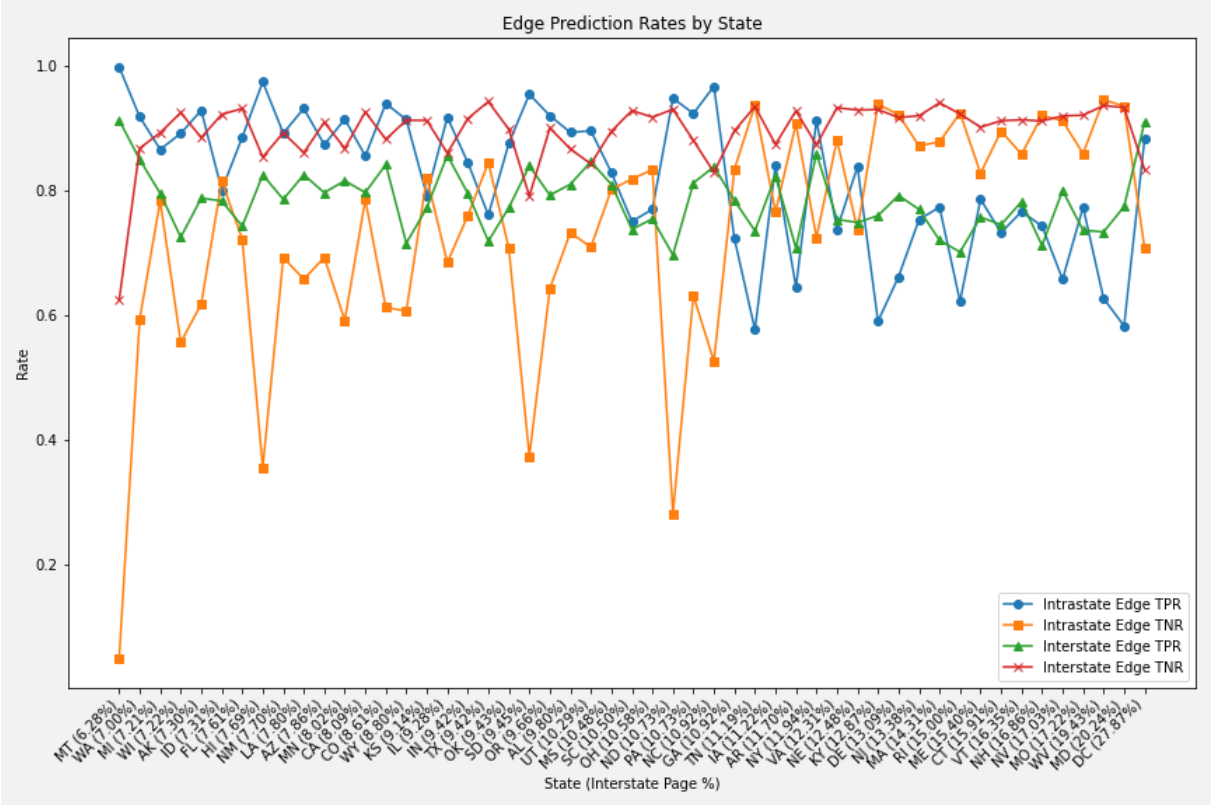


Figure 3.4: Edge prediction rates by state

of these less effective features were not explored, a few examples are provided in Table 3.6 for illustrative purposes.

3.5.3.2 Prediction Analysis

In Chapter 1, we identified two types of edges in the page graph: interstate edges and intrastate edges. 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 3.8.

In this examination, 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 3.8. The "Start" and "End" columns denote the originating and terminating states of the edges, respectively. 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 3.7a. 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 3.7b. 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 3.8 using a line chart in Figure 3.4 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.

3.6 Conclusion

In this chapter, we delve into the identification of 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 topological algorithm GNN to perform the link prediction. Through meticulous

Table 3.7: Percentage distribution of edges

(a) Positive/Negative edges percentage distribution

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

(b) Intrastate/Interstate edges percentage distribution

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

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 concept introduced in Chapter 1, underscoring the nuanced dynamics of link formation within the page graph.

Table 3.8: Link Prediction Performance by State

Intrastate Edge				Interstate Edge			
Start	End	TPR	TNR	Start	End	TPR	TNR
AL	AL	0.8939	0.7331	AL	other states	0.8107	0.8672
AK	AK	0.9293	0.6189	AK	other states	0.7883	0.8856
AZ	AZ	0.8742	0.6932	AZ	other states	0.7972	0.9101
AR	AR	0.6455	0.9074	AR	other states	0.7083	0.9284
CA	CA	0.8560	0.7862	CA	other states	0.7974	0.9266

Continued on next page

Table 3.8 – continued from previous page

Intrastate Edge				Interstate Edge			
Start	End	TPR	TNR	Start	End	TPR	TNR
CO	CO	0.9402	0.6131	CO	other states	0.8423	0.8828
CT	CT	0.7332	0.8960	CT	other states	0.7457	0.9125
DE	DE	0.6622	0.9225	DE	other states	0.7927	0.9178
FL	FL	0.8862	0.7226	FL	other states	0.7434	0.9322
GA	GA	0.7234	0.8347	GA	other states	0.7847	0.8964
HI	HI	0.9750	0.3560	HI	other states	0.8258	0.8542
ID	ID	0.8014	0.8158	ID	other states	0.7839	0.9227
IL	IL	0.9182	0.6853	IL	other states	0.8571	0.8606
IN	IN	0.8454	0.7603	IN	other states	0.7961	0.9160
IA	IA	0.8410	0.7661	IA	other states	0.8238	0.8747
KS	KS	0.7926	0.8218	KS	other states	0.7730	0.9130
KY	KY	0.5909	0.9390	KY	other states	0.7604	0.9304
LA	LA	0.9322	0.6585	LA	other states	0.8258	0.8613
ME	ME	0.7867	0.8281	ME	other states	0.7580	0.9023
MD	MD	0.5835	0.9360	MD	other states	0.7749	0.9338
MA	MA	0.7740	0.8788	MA	other states	0.7210	0.9413
MI	MI	0.8666	0.7839	MI	other states	0.7960	0.8934
MN	MN	0.9151	0.5924	MN	other states	0.8163	0.8683
MS	MS	0.8294	0.8026	MS	other states	0.8104	0.8958
MO	MO	0.7741	0.8600	MO	other states	0.7369	0.9215
MT	MT	0.9986	0.0511	MT	other states	0.9133	0.6254
NE	NE	0.8387	0.7380	NE	other states	0.7493	0.9298
NV	NV	0.6582	0.9123	NV	other states	0.8005	0.9203
NH	NH	0.7436	0.9211	NH	other states	0.7125	0.9121

Continued on next page

Table 3.8 – continued from previous page

Intrastate Edge				Interstate Edge			
Start	End	TPR	TNR	Start	End	TPR	TNR
NJ	NJ	0.7538	0.8722	NJ	other states	0.7703	0.9204
NM	NM	0.8925	0.6930	NM	other states	0.7865	0.8933
NY	NY	0.9129	0.7251	NY	other states	0.8600	0.8739
NC	NC	0.9677	0.5267	NC	other states	0.8389	0.8298
ND	ND	0.9486	0.2824	ND	other states	0.6974	0.9308
OH	OH	0.7705	0.8342	OH	other states	0.7552	0.9183
OK	OK	0.8764	0.7080	OK	other states	0.7732	0.8985
OR	OR	0.9206	0.6436	OR	other states	0.7933	0.9008
PA	PA	0.9242	0.6318	PA	other states	0.8127	0.8808
RI	RI	0.6227	0.9249	RI	other states	0.7019	0.9234
SC	SC	0.7511	0.8189	SC	other states	0.7383	0.9288
SD	SD	0.9551	0.3744	SD	other states	0.8402	0.7917
TN	TN	0.5773	0.9375	TN	other states	0.7359	0.9354
TX	TX	0.7614	0.8455	TX	other states	0.7194	0.9433
UT	UT	0.8965	0.7098	UT	other states	0.8469	0.8433
VT	VT	0.7668	0.8589	VT	other states	0.7828	0.9140
VA	VA	0.7386	0.8817	VA	other states	0.7542	0.9333
WA	WA	0.9186	0.5946	WA	other states	0.8503	0.8684
WV	WV	0.6275	0.9462	WV	other states	0.7342	0.9368
WI	WI	0.8921	0.5569	WI	other states	0.7257	0.9254
WY	WY	0.9143	0.6074	WY	other states	0.7161	0.9134
DC	DC	0.8838	0.7079	DC	other states	0.9104	0.8340

Chapter 4

Online Social Community City Classification

4.1 Introduction

In the digital realm of online spaces, people's behaviors remain closely linked to location. Individuals tend to show greater interest in local news, are more likely to connect with nearby friends, and have preferences for local dining options, among other location-centric activities. Location information plays a crucial role in both economic activities and public services, including targeted news dissemination, product and service recommendations, and emergency event notifications.

Since the inception of online social network platforms, automatically identifying users' geographic locations has gained popularity. A substantial body of research has explored various methods for geolocating users. Some studies predict location based on content analysis, including words in posts, comments, and tweets. Other research examines user networks, such as friendships and following relationships, to predict locations based on the tendency of users to interact with geographically close individuals.

The geolocation of online social communities, such as Facebook pages and Reddit, which serve as digital town halls for information dissemination and user discussion, has not been extensively explored. The task of predicting the geolocation of geographically unlabeled Facebook public pages has been approached with varying levels of granularity. Hong introduced the Majority Voting method to categorize Facebook public pages by

country. In Chapter 1, we furthered this research by utilizing the GraphSAINT model in conjunction with neighborhood state distribution (NSD) feature vectors. This approach facilitated the more challenging task of classifying pages into specific sublocations, such as States within the U.S.

Classifying Facebook pages by cities presents a greater challenge because a city is a much smaller and more fragmented area than a country or state. For example, our dataset includes 630 California cities, which complicates classification.

In this chapter, we introduced a virtual geographic structure of cities, which are city clusters resembling counties, to enhance classification performance. This virtual geographic structure of cities is not represented in the data explicitly. The composition of cities in each cluster results from our clustering algorithm, which is based on the confusion matrix of the flat city classification. Based on the results of the clustering, we implemented a two-stage hierarchical classification method that classifies pages by city clusters first and then by cities within each cluster. These innovations have significantly improved our city classification performance.

This chapter is structured as follows: Section 4.2 delves into related research on user geographic location analysis, studies related to online community locations, and hierarchical classification. Section 4.3 details the data used in this study, including the ground truth data for validation purposes. Section 4.4 outlines the baseline models for city and county page classification, serving as benchmarks for subsequent experiments. Section 4.5 explores feature engineering with neighborhood distribution vectors to enhance performance. Section 4.6 discusses hierarchical classification leveraging the natural taxonomy structure of counties and cities. Section 4.7 presents our clustering method designed to construct a city hierarchical structure to improve classification performance. The chapter concludes with Section 4.8, providing a summary of the discussions and findings.

4.2 Related Work

4.2.1 User Location Prediction

Twitter user location prediction has been extensively studied, with research efforts focusing on both user home location prediction[46, 47, 48, 49, 50] and tweet location prediction [51, 52]. Our interest primarily lies in user home location prediction, which aligns more closely with our objectives. There are two predominant approaches to predicting user home location. The first relies on content analysis, identifying local vernacular or place-specific words, such as "howdy" and "Phillies," which are frequently used in certain regions[46, 47]. The second approach analyzes user networks, focusing on friendships, interactions, or other relational ties to infer location[48, 49].

4.2.2 Online Community Location Studies

Online communities have been primarily studied in the context of user engagement[53, 54] and information consumption[55]. Facebook, known for its emphasis on location, prioritizes local recommendations and advertising[56]. Several studies have explored the geographical aspects of Facebook communities. For instance, [57] analyzed the location data of businesses' Facebook pages to provide geolocation recommendations for new businesses. Another study[58] found that Facebook pages belonging to news providers tend to interact more with other pages within the same geographical confines, such as continents and countries. The study [59] detects the geolocation of Twitter user communities by extracting and summarizing users' location data within each community.

4.2.3 Hierachical Classification

Hierarchical classification is widely utilized in various real-world classification challenges, as documented in the literature [60]. This method is particularly beneficial in domains where classes or categories inherently form hierarchical structures, including bioinformatics, text mining [61, 62], among others [63]. Typically, hierarchical classification involves the initial classification of meta-classes, followed by a more detailed classification within each meta-class. This approach leverages the advantage of model specialization in a multi-stage classification process [64], where employing distinct models for different data

subsets or specific types of information can lead to superior performance compared to using a single, flat classifier that may not capture all nuances effectively.

4.3 Data Description

4.3.1 Data Acquisition

In this chapter, we utilize the same dataset of Facebook public pages as presented in Chapter 1. This data was obtained via the Facebook Graph API 2.8, courtesy of Facebook. The metadata for each page includes its ID, name, description, category, country, city, the other pages it likes, and the number of fans, among other details. Importantly, these datasets do not contain any private user information. To collect the data, we employed snowball sampling[11], initiating the crawl from several popular Facebook public pages and progressively moving to the pages they like. This process naturally constructs a directed graph of Facebook pages.

4.3.2 Data Cleaning

In Section 2.3.4, we differentiated between state-known and state-unknown pages. State-known pages correspond to cities with unique names across the United States, while state-unknown pages are linked to cities whose names are shared by cities in different states. For the purposes of city classification, we focus exclusively on state-known pages.

Facebook pages from California, having the highest number of state-known pages, serve as our dataset for city classification experiments. The California page graph comprises 324,887 pages and 2,378,881 edges, encompassing 58 counties and 630 cities.

However, we encountered cities listed as ground truth that were absent from our city database, including some rural and small community areas within larger cities or counties. To retain as much data as possible, we manually relabeled these pages to their nearest recognized city. This process involved identifying these cities individually, relocating urban communities to their larger parent cities, and rural communities to the nearest cities in our database. As a result, we modified the city labels for 26,371 pages.

4.4 Classification Baseline

4.4.1 GraphSAINT Model

As discussed in Chapter 1, the GraphSAINT model addresses the challenge of processing large graphs by reducing them into smaller, sampled subgraphs through random walk sampling. This method significantly decreases memory usage compared to the GCN strategy, which involves loading the entire graph into GPU memory. By transforming the original graph into manageable subgraphs, GraphSAINT enhances the model's ability to process large datasets, thereby improving training efficiency and speed. Evidence of GraphSAINT's superior performance over GraphSAGE is presented in Table 2.2. Consequently, we employ two-layer GraphSAINT models for all experiments in this chapter, utilizing the PyTorch Geometric (PyG) framework, a specialized tool for Graph Neural Networks[34].

4.4.2 City Neighborhood Distribution Vector

To effectively classify pages by city, we require not only the GraphSAINT model to understand the page graph's topology but also node features that offer additional information to enhance performance.

Building on the findings from Chapters 1 and 2, we have demonstrated that neighborhood state distribution serves as an effective node feature for both page state classification and link prediction. Extending this approach to city classification, we introduce the city neighborhood distribution vector ($City - ND$) as a novel node feature.

The $City - ND$ vectors represent the ratio of a page's neighbors from each city to its total number of neighbors. Since every page in the California page graph is connected, these vectors are guaranteed to be non-zero, offering a reliable feature for machine learning-based classification. To ensure a thorough understanding of a page's local network, we calculate $City - ND$ for both one-hop and two-hop distances, considering neighbors connected through inward, outward, and undirected edges. This comprehensive, multi-faceted strategy enriches the representation of page associations, thereby enhancing the accuracy of city classification. The formulation of the $City - ND$ vector is as follows:

$$City - ND(Page) = [$$
$$[City - IND_1(Page, City_i), City - OND_1(Page, City_i), City - UND_1(Page, City_i),$$
$$City - IND_2(Page, City_i), City - OND_2(Page, City_i), City - UND_2(Page, City_i)] :$$
$$i \in 1, ..., N_{number\ of\ cities}]$$

where:

- $City - IND_k(Page, City_i)$ denotes the inward city neighborhood distribution for $City_i$, calculated within a k-hop distance from the $Page$.
- $City - OND_k(Page, City_i)$ denotes the outward city neighborhood distribution for $City_i$, calculated within a k-hop distance from the $Page$.
- $City - UND_k(Page, City_i)$ denotes the undirected city neighborhood distribution for $City_i$, calculated within a k-hop distance from the $Page$.

Furthermore, each element of the $City - ND$ for a page, whether $City - IND$, $City - OND$, or $City - UND$, is defined as the ratio of neighbors from city i within a j-hop distance, normalized by the total number of neighbors across all cities within the same hop distance:

$$City - XND_j(Page, City_i) = \frac{XNeighbor_{ij}}{\sum_{i=1}^{N_{number\ of\ cities}} XNeighbor_{ij}},$$
$$i \in \{1, ..., N_{number\ of\ cities}\},$$
$$j \in \{1, 2\},$$
$$X \in \{I, O, U\},$$

Where:

- i denotes the ith city.

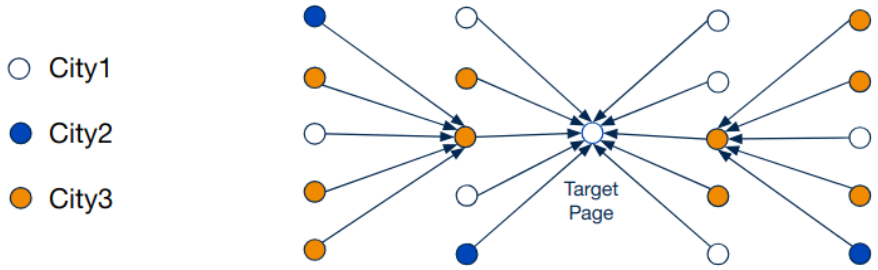


Figure 4.1: Example of a two-hop inward neighborhood for a target page within a page graph covering three cities

- j denotes the one-hop or two-hop distance.
- X denotes one of three edge directions, inward I, outward O, or undirected U.
- $XNeighbor_{ij}$: the total number of neighbors from City i within j hop distance for inward I, outward O, or undirected U edge direction.

For example, Figure 4.1 illustrates the two-hop inward neighborhood of a target page within a page graph that includes three cities. For the target page, the one-hop inward city neighborhood distribution, $City - IND_1(Target)$, is [0.5, 0.1, 0.4], and the two-hop inward city neighborhood distribution, $City - IND_2(Target)$, is [0.35, 0.5, 0.15]. Given the dataset encompasses 630 cities, the $City - ND$ vector features 3780 dimensions for each page. This characteristic is leveraged for our baseline experiment as it seamlessly extends the concept of state neighborhood distribution discussed in Section 2.6.

Using the GraphSAINT model and the city neighborhood distribution vectors as node features, we achieved a page city classification accuracy of 0.6928, as shown in Table 4.1. This accuracy is lower than the page state classification accuracy of 0.8752, reported in both Table 2.2 and Table 4.1.

4.4.3 County Neighborhood Distribution Vector

Page city serves as the ground truth data. In Chapter 1, we utilized the cities of state-known pages to determine their respective states. Similarly, we can derive the page county from the page city. Given that the city classification accuracy in Table 4.1 falls short of

Table 4.1: Baseline accuracy for Pages in California Page Graph

Baseline	Overall Accuracy
Page City Classification	0.6928
Page County Classification	0.8869
Page State Classification	0.8752

the state classification accuracy, we also undertake county classification as an additional reference point. This effort aims to explore avenues for enhancing the performance of page city classification.

We introduce the county neighborhood distribution ($County - ND$) vectors as node features for the classification of page counties. The definition is as follows:

$$
\begin{aligned}
&County{-}ND(Page) = [\\
&\quad [County - IND_1(Page, County_i),\ County - IND_2(Page, County_i),\\
&\quad County - OND_1(Page, County_i),\ County - OND_2(Page, County_i),\\
&\quad County - UND_1(Page, County_i),\ County - UND_2(Page, County_i)] :\\
&\quad i \in 1, ..., N_{number\ of\ counties}]
\end{aligned}
$$

where:

- $County - IND_k(Page, County_i)$ denotes the inward county neighborhood distribution for $County_i$, calculated within a k-hop distance from the $Page$.
- $County - OND_k(Page, County_i)$ denotes the outward county neighborhood distribution for $County_i$, calculated within a k-hop distance from the $Page$.
- $County - UND_k(Page, County_i)$ denotes the undirected county neighborhood distribution for $County_i$, calculated within a k-hop distance from the $Page$.

In the California page graph, page cities are associated with 58 distinct counties, resulting in 364 dimensions in the county neighborhood distribution ($County - ND$)

vectors. Employing the GraphSAINT model with county neighborhood distribution vectors as node features, we achieved a county classification accuracy of 0.8869, detailed in Table 4.1. This accuracy surpasses the city classification accuracy of 0.6928 by a large margin. This outcome suggests potential avenues for enhancing the performance of city classification.

4.5 City Classification Feature Engineering

4.5.1 Integrate Predicted County Information

To improve the page city classification performance, our initial strategy focused on analyzing and adjusting the expressiveness of features. Specifically, the $City - ND$ vectors comprise six subvectors: $City - IND_1$, $City - OND_1$, $City - UND_1$, $City - IND_2$, $City - OND_2$, and $City - UND_2$. Each subvector is a 630-dimensional vector, corresponding to the 630 cities. For a target page associated with city A, if city A's distribution in any of these subvectors is the highest and unequivocal, then there is a strong likelihood that the page will be correctly classified to city A. For instance, as shown in Table 4.2, within the subvector $City - IND_1$, the highest city distribution that matches the city label accounts for 70.42% of the pages. Among these pages, 73.99% have the matching city label uniquely, without ties to other city labels.

Table 4.2: Percentage of highest distribution match and no tie for city label

Subvector	Highest Distribution Match City Label %	no Tie %
$City - IND_1$	70.42	73.99
$City - OND_1$	82.83	38.09
$City - UND_1$	68.85	84.99
$City - IND_2$	63.65	82.51
$City - OND_2$	78.34	38.69
$City - UND_2$	57.97	98.40

Given the superior performance of page county classification over city classification in Table 4.1, we explored the integration of county information into the $City - ND$ vectors

to potentially enhance accuracy. Initially, we predicted each page's county label through county classification, then amplified the city distribution values within the $City - ND$ vectors for cities corresponding to the predicted county. This approach aimed to highlight the cities within the predicted counties. However, this modification did not significantly improve accuracy, suggesting that amplifying the city distribution for all cities in the predicted county may not be the correct way to integrate the county information. We need to explore other options.

4.6 City Classification within Counties

4.6.1 Derived County Classification

We compare the baseline city classification and baseline county classification by analyzing the county classification accuracy derived from the baseline city classification results, as discussed in Section 4.4.2. Consider a scenario where $City - A$ and $City - C$ belong to $County - A$, and $City - B$ belongs to $County - B$; these represent ground truth one-to-one relationships. If $Page - A$, with a ground truth label of $City - A$ and consequently belonging to $County - A$, is correctly classified as $City - A$, it implies that $Page - A$ is also correctly classified as belonging to $County - A$. Conversely, if $Page - A$ is misclassified as $City - C$, it is incorrectly classified at the city level but correctly at the county level ($County - A$). However, if $Page - A$ is misclassified as $City - B$, it indicates an incorrect classification at both the city and county levels, as it would be incorrectly assigned to $County - B$. This derivation approach allows for an assessment of how the baseline city classification performs at the county level.

The derived county classification accuracy stands at 0.8425, lower than the baseline county classification accuracy of 0.8869 but significantly surpassing the baseline city classification accuracy of 0.6928. This observation suggests the potential benefit of first classifying pages by county using a county classifier, which demonstrates superior performance at the county level, followed by classifying pages into specific cities within those counties. This method necessitates a two-step approach: initially classifying pages by county, then further classifying pages into cities within those counties. This process requires two

distinct types of classifiers: a county classifier and fifty-eight city classifiers, one for each county.

4.6.2 County Classifier

The county classifier employs a two-layer GraphSAINT model to predict the classification of pages across 58 counties within the California page graph dataset. Utilizing county neighborhood distribution ($County-ND$) vectors, as introduced in Section 4.4.3, as node feature inputs, and the California (C.A.) Page graph as the graph input, this classifier achieves a high prediction performance, with an accuracy of 0.8869.

4.6.3 City Classifier For Each County

All city classifiers for each county utilize a two-layer GraphSAINT model, along with city neighborhood distribution ($City - ND$) vectors, as introduced in Section 4.4.2, as node features to predict page classifications within each county. The key differences include the graph input, which is the specific county page graph for each classifier, and the ($City - ND$) vectors for each page, calculated based on the cities within the respective county. The total number of cities within each county ranges from 1 to 60, leading to significantly fewer dimensions in the ($City - ND$) vectors compared to the baseline ($City - ND$) vectors for 630 cities. However, this approach requires the training and inference process to be executed fifty-eight times.

4.6.4 City Classification Accuracy

After training two kinds of classifiers—a county classifier and city classifiers for each county—we perform inference for all California pages in two steps:

1. Classify all California pages into different counties using the county classifier based on the California page graph and $County - ND$ node features. We disregard the pages misclassified at the county level, retaining only those correctly classified for the subsequent step.

2. For each county, we take the pages correctly classified to the respective county from step one and classify these pages into different cities within the county using the

city classifier specific to that county. This classification is based on the page graph and $City - ND$ node features specific to the respective county. We record the pages correctly classified at the city level in this step to calculate the overall city classification accuracy later. This step is repeated for every county.

Table 4.3: Accuracy for City Classification within Counties

Algorithms	City Level Accuracy	County Level Accuracy	Total Pages
City Classification within Counties	0.7494	0.8869	324887
Baseline City Classification	0.6928	0.8425	324887
Improvement	0.0566	0.0444	324887

The overall accuracy of city classification within all counties is 0.7494, significantly higher than the baseline city classification accuracy of 0.6928, as shown in Table 4.3. The improvement in city level accuracy is greater than the improvement in county level accuracy between these two methods, as illustrated in Table 4.3. This indicates that the higher county level accuracy achieved by the county classifier in step one not only improves performance at the county level but also enhances city level classification performance within each county.

4.6.5 Hierarchical Classification

Page city classification within counties essentially adopts a hierarchical classification approach, commonly employed in real-world classification problems [60][63][64][65]. By contrast, the baseline city classification represents a flat classification model. Here, page cities serve as the target classes, while page counties act as meta-classes for these cities, forming a natural taxonomy based on ground truth. This method, which involves classifying pages into a meta-class followed by classification within that meta-class, exemplifies hierarchical classification. Such an approach benefits from model specialization in multi-stage classification[64], where training distinct models for data subsets or specific information leads to improved performance compared to a singular, flat classifier that struggles to

encompass all information effectively. As evidenced in Table 4.3, ensembled specialized models demonstrate superior performance at every class level compared to a single model.

4.7 City Classification within Clusters

4.7.1 Building Hierarchical Structure

In hierarchical classification, two primary types of meta-classes are identified: the first type comprises pre-existing taxonomies related to the target classes, such as counties for cities in the context of city classification within counties. The second type involves meta-classes that are newly created based on the similarity among target classes. A common methodology entails initially conducting a flat classification of the target classes, followed by the generation of a confusion matrix for this classification. The confusion matrix serves to reveal class similarities, which are then used to construct meta-classes. Subsequently, based on the hierarchical structure of these meta-classes, hierarchical classifiers are developed to execute the hierarchical classification process.

The critical step in forming meta-classes involves determining the optimal number of these groups by analyzing the classification confusion matrix. Attempting to explore all possible clustering configurations is computationally equivalent to identifying all possible partitions of n samples. The complexity of this task is quantified by Bell numbers, which increase rapidly in a manner known as combinatorial explosion. For example, the number of possible partitions for just 10 items is 115,975. Considering the challenge involves 630 city classes, the task of assessing all potential clustering options for these classes is practically unfeasible due to the exponential growth in the number of possible partitions.

4.7.2 Affinity Clustering

A common strategy for constructing meta-classes involves adopting a systematic approach that leverages clustering algorithms. These algorithms cluster classes based on their similarity, as indicated within the confusion matrix. This method is documented in various studies, including those focused on the use of confusion matrices for hierarchical classification construction and others that explore semantic and probability-based approaches to understanding class similarities[66][67][68][69]. In this research, given the unknown opti-

mal number of meta-classes, we reference the affinity clustering algorithm, which organizes classes into a suitable number of meta-classes based on their similarity distances.

Affinity clustering is particularly beneficial in scenarios where the number of clusters is not predetermined. By adjusting its configuration, we managed to group the city classes into two distinct sets of city clusters: one comprising 3 city clusters and another encompassing 75 city clusters. Subsequent hierarchical classifications were conducted based on these two varying hierarchical structures to facilitate a comparative analysis. Each hierarchical classification setup requires a dedicated city cluster classifier along with multiple city classifiers. Table 4.4 illustrates how the adoption of 75 city-cluster and 3 city-cluster hierarchical structures enhances the accuracy of city classification.

Table 4.4: Accuracy for Hierarchical Classification

Hierarchical Structure	City Level Accuracy	Cluster Level Accuracy
Baseline City Classification (1 Cluster)	0.6928	1
75 City Clusters (Affinity Clustering)	0.7512	0.8573
3 City Clusters (Affinity Clustering)	0.7744	0.9375
17 City Clusters (Our Clustering Method)	0.8014	0.9778

4.7.3 Our Clustering Method

4.7.3.1 Intuition of Confusion Matrix

The confusion matrix of the flat page city classification reveals the extent to which pages from each city are incorrectly labeled as belonging to other cities. A high number of misclassified pages between city A and city B suggests that the flat city classifier struggles to distinguish pages between these two cities, indicating a certain level of similarity or closeness between them in the context of 630 cities. This challenge arises because the flat city classifier must differentiate among pages from all 630 cities, making it difficult to capture the subtle distinctions between any two specific cities, such as city A and city B. For instance, as depicted in Figure 4.2, if cities A and B have a high misclassification

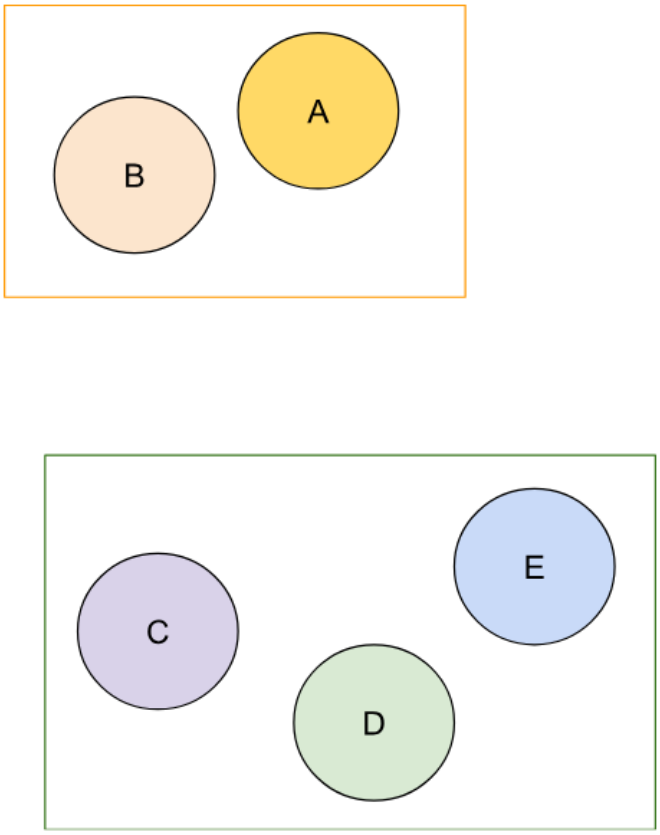


Figure 4.2: City Cluster Example

rate, and cities C, D, and E also share high misclassification rates among themselves, it implies that cities A and B are close to each other, and cities C, D, and E form another close group. We could interpret these findings as indicating two clusters, with the flat classifier being more capable of distinguishing between these two clusters, since cities A and B have lower misclassification rates with cities C, D, and E.

4.7.3.2 City Clustering Based on Misclassification Rates

Based on an intuitive analysis of the confusion matrix, we propose an algorithm for clustering cities according to their misclassification rates. For each city, we sort the misclassification rates from its respective row in the confusion matrix (normalized by row) in descending order. By connecting a city to its neighbor with the highest misclassification rate via an edge, we cluster these two cities together in the city graph. Upon adding edges for all cities, the resulting city clusters are identified as disconnected components within

Algorithm 1 City Clustering Based on Misclassification Rates

Require: $desc\_ordered\_rates$, max_edges, $thresholds$

Ensure: Edge list E with tuples (city, neighbor, rate)

$E \leftarrow []$ ▷ Initialize edge list
for $city \leftarrow 0$ to $number\_of\_cities - 1$ **do**
    $edges\_added \leftarrow 0$
    for $j \leftarrow 0$ to $number\_of\_cities - 1$ **do**
        $(rate, neighbor) \leftarrow desc\_ordered\_rates[city][j]$
        if $neighbor \neq city$ **and** $thresholds[edges\_added] \leq rate$ **then**
            E.append($(city, neighbor, rate)$)
            $edges\_added \leftarrow edges\_added + 1$
        end if
        if $edges\_added = max\_edges$ **then**
            break
        end if
    end for
end for

the graph.

The configuration of clusters can be adjusted by two parameters: the number of edges to add based on the highest misclassification rates for each city, and the misclassification rate threshold for including an edge. Increasing the number of edges enhances the connectivity of the city graph and decreases the number of clusters. Conversely, raising the threshold for edge inclusion filters out connections with lower misclassification rates, leading to reduced connectivity and an increased number of city clusters. The detailed methodology is outlined in Algorithm 1.

By setting the number of edges to 1 and the threshold rate to 0, we exclusively link each city to its most frequently misclassified neighboring city, resulting in the formation of 17 city clusters. Subsequent hierarchical classification leverages this cluster configuration. Altering the threshold rate to 0.05 for the edge leads to an increased number of city

clusters. Introducing a second edge with a threshold rate of 0.5 slightly consolidates the clusters.

4.7.4 Cluster Neighborhood Distribution Vector

Based on the city cluster configurations, we calculate node features for cluster classification, introducing the cluster neighborhood distribution ($Cluster - ND$) vector as the cluster-level node feature. The definition is as follows:

$$
\begin{aligned}
Cluster{-}ND(Page) = [\ & \\
& [Cluster - IND_1(Page, Cluster_i),\ Cluster - IND_2(Page, Cluster_i), \\
& Cluster - OND_1(Page, Cluster_i),\ Cluster - OND_2(Page, Cluster_i), \\
& Cluster - UND_1(Page, Cluster_i),\ Cluster - UND_2(Page, Cluster_i)] : \\
& i \in 1, ..., N_{number\ of\ clusters}]
\end{aligned}
$$

where:

- $Cluster - IND_k(Page, Cluster_i)$ denotes the inward cluster neighborhood distribution for $Cluster_i$, calculated within a k-hop distance from the $Page$.
- $Cluster - OND_k(Page, Cluster_i)$ denotes the outward cluster neighborhood distribution for $Cluster_i$, calculated within a k-hop distance from the $Page$.
- $Cluster - UND_k(Page, Cluster_i)$ denotes the undirected cluster neighborhood distribution for $Cluster_i$, calculated within a k-hop distance from the $Page$.

In the California page graph, with pages associated with 17 city clusters, the $Cluster - ND$ vectors result in 102 dimensions. Using the GraphSAINT model with $Cluster - ND$ vectors as node features, we achieved a cluster classification accuracy of 0.9778, as detailed in Table 4.4.

4.7.5 City Classifier within Each Cluster

City classifiers for each cluster utilize a two-layer GraphSAINT model, along with city neighborhood distribution ($City - ND$) vectors, as introduced in Section 4.4.2, as node

features to predict page classifications within each county. The key differences include the graph input, which is the page graph for each cluster, and the $(City - ND)$ vectors for each page, calculated based on the cities within the respective cluster.

4.7.6 City Classification Accuracy

After training two kinds of classifiers—a cluster classifier and city classifiers for each cluster—we perform inference for all California pages in two steps:

1. Classify all California pages into different clusters using the cluster classifier based on the California page graph and $Cluster - ND$ node features. We disregard the pages misclassified at the cluster level, retaining only those correctly classified for the subsequent step.

2. For each cluster, we take the pages correctly classified to the respective cluster from step one and classify these pages into different cities within the cluster using the city classifier specific to that cluster. This classification is based on the page graph and $City - ND$ node features specific to the respective cluster. We record the pages correctly classified at the city level in this step to calculate the overall city classification accuracy later. This step is repeated for every cluster.

The overall accuracy of city classification within all clusters reaches 0.8014, which is significantly higher than the baseline city classification accuracy of 0.6928 (with no hierarchy, implying a single city cluster) and the accuracies achieved using other hierarchical structures, as depicted in Table 4.4. The accuracy for cluster classification stands at 0.9778. This performance underscores the effectiveness of our clustering algorithm in grouping similar cities within our dataset. Such grouping facilitates the task of the cluster classifier in differentiating pages across clusters, thereby contributing to the improvement in overall city classification accuracy.

This method classifies pages into a meta-class followed by classification within each meta-class. This approach benefits from model specialization in multi-stage or hierarchical classification [70, 71, 64], where training distinct models for data subsets or specific information leads to improved performance compared to a singular, flat classifier that struggles

to encompass all information effectively. The performance improvement in this two-step hierarchical classification primarily results from reducing the classification complexity for each classifier involved. Handling 630 classes is overwhelming for a single machine learning classifier, as it can easily confuse similar classes. In contrast, dividing these into 17 clusters makes the task more manageable, as each cluster contains a significantly smaller number of classes. These clusters are distinctly different from each other and can be easily distinguished by a classifier, as they have been clustered based on the confusion matrix of the 630 cities. Moreover, each cluster does not contain too many cities, facilitating better performance by the classifiers. In the second step, the city classifiers focus on learning the nuanced differences between similar cities within the same cluster, free from interference from cities in other clusters.

4.8 Conclusion

In this chapter, we looked into the task of Facebook public page classification by cities within California. We introduced city, county, and cluster neighborhood distribution vectors as distinctive features for page classifications. With an initial city classification accuracy of 0.6928, the complexity of distinguishing among 630 cities presents a significant challenge. We introduced a virtual geographic city structure resembling counties to improve the classification performance. We developed a clustering algorithm that leverages the confusion matrix from the flat city classification to construct a virtual geographic city structure. Based on this virtual structure, we implemented a two-stage hierarchical classification method, first classifying pages by virtual city clusters and then within clusters by city. This implementation of a cluster-city hierarchical classification achieves a notable improvement in city classification accuracy to 0.8014.

Chapter 5

Concluding Remarks

5.1 Summary

In this dissertation, we focus on the geographic characterization of online social communities within social network platforms. Our objective is to predict missing geographic locations at various levels of granularity, including state, county, and city levels, and to demonstrate the significant impact of geo-location on the formation of links and neighborhoods within online community graphs. We outline the primary contributions of our research in the sections below, highlighting our achievements in enhancing the understanding and predictive capabilities related to the geographic dimensions of online social communities.

5.1.1 Online Social Community Sub-Location Classification

In Chapter 2, we addressed the challenge of classifying subdivision locations for Facebook public pages, focusing on pages within the United States. This section critically evaluated the limitations of previous approaches to sub-location classification and introduced a new method using the GraphSAINT model, which leverages neighborhood state distribution vectors for more accurate classification. Our method demonstrated significant improvement in classification accuracy over previous techniques when applied to a dataset of U.S. Facebook public pages.

Additionally, our model was used to differentiate between intrastate and interstate Facebook public pages, revealing that intrastate pages typically receive "likes" from within

the same state, while interstate pages attract likes from pages across different states. Through an analysis of the state classification confusion matrix, examination of interstate page percentages by state, and exploration of interstate pages across state borders, we demonstrated the significant influence of geographic location on the formation of online community networks and the precision of sub-location classification for Facebook public pages.

5.1.2 Online Social Community Neighborhood Formation

In Chapter 3, our investigation targeted key factors contributing to link formation and neighborhood structuring within the Facebook page graph. We examined an array of features—graph-based and content-based—potentially influencing link formation. We then introduced our method, leveraging node similarity and the topological capabilities of Graph Neural Networks (GNN) for link prediction.

Through detailed experiments with both individual and combined features, we identified the page state label as the most effective single feature for predicting link formation. Additionally, enhancing the page state label with page node degree and page city population features improved link prediction accuracy. Our findings also highlighted a relationship between the true positive rate for intrastate positive edges and the interstate page percentage, introduced in Chapter 1. This correlation further illuminates the complex dynamics of link formation within the page graph.

5.1.3 Online Social Community City Classification

In Chapter 4, we tackled city classification for Facebook public pages in California, introducing innovative features such as city, county, and cluster neighborhood distribution vectors for classification. The task posed a considerable challenge, given the necessity to discriminate among 630 cities with an initial accuracy of only 0.6928. By implementing a two-stage hierarchical classification method—initially classifying pages by county, then further classifying within each county by city—we managed to improve accuracy to 0.7494.

Building on this method, we devised a clustering algorithm based on the confusion

matrix from city classification, enabling the construction of a hierarchical city structure. This approach facilitates a more nuanced cluster-city hierarchical classification, markedly enhancing city classification accuracy from the baseline of 0.6928 to 0.8014.

5.2 Future Works

5.2.1 Confusion Matrix Clustering and Hierarchical Classification Accuracy

In Chapter 4, we explored a novel clustering algorithm applied to the confusion matrix from flat city classification, aiming to construct a hierarchical city structure that enhances the accuracy of hierarchical city classification. Additionally, we compared this method with affinity clustering. However, exploring all possible clustering algorithms on the confusion matrix was impractical. This limitation arises because each clustering approach requires training a cluster classifier and multiple city classifiers for city clusters, a process that could extend over one or two days given our large dataset. The possible number of city clustering combinations also grows exponentially.

Each city clustering modifies the flat city confusion matrix into a matrix representing city clusters, alongside multiple city-specific confusion matrices within those clusters. We sought to understand how these transformed confusion matrices correlate with the accuracy of hierarchical classification for each city clustering. Identifying a relationship that allows these transformed confusion matrices to serve as indicators of final hierarchical classification accuracy—without undergoing the actual classification process—could significantly conserve computational resources by guiding our efforts based on the insights gained from the transformed confusion matrices.

5.2.2 Applications of Geolocation Characterized Networks

In this dissertation, we utilize location information to accurately classify Facebook public pages by city and state and to predict links between pages. The high accuracy achieved underscores the effectiveness of location information in characterizing online community networks. Based on the geographic characteristics identified, we can extend our research to other areas related to geolocation.

One such area is the analysis of the political spectrum within online community networks. The political landscape of each state, typically categorized as blue, red, or purple, has remained relatively stable over the past decades. By examining the content of these pages, we can explore how political pages across different parts of the political spectrum interact with one another.

Another area of interest is state economics. Economic powerhouses such as California, Texas, and New York likely exhibit more independent economic activities and exert influence on other states rather than being influenced by them. For instance, Nevada experiences significant influence from California. This presents an opportunity to scrutinize the economic dynamics for national and regional economic centers, as identified in Section 2.7, and their interactions with peripheral states, providing insights into the inter-state economic landscape and influence patterns.

References

[1] J. Wang, X. Wang, C.-M. Lai, and S. F. Wu, "Online social community sub-location classification," in *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '23. New York, NY, USA: Association for Computing Machinery, 2024, p. 276–280. [Online]. Available: https://doi.org/10.1145/3625007.3627504

[2] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," 2011. [Online]. Available: https://arxiv.org/abs/1111.4503

[3] G. A. Barnett and G. A. Benefield, "Predicting international facebook ties through cultural homophily and other factors," *New Media & Society*, vol. 19, no. 2, pp. 217–239, 2017. [Online]. Available: https://doi.org/10.1177/1461444815604421

[4] Y. Hong, Y.-C. Lin, C.-M. Lai, S. Felix Wu, and G. A. Barnett, "Profiling facebook public page graph," in *2018 International Conference on Computing, Networking and Communications (ICNC)*, 2018, pp. 161–165.

[5] Y.-C. Lin, C.-M. Lai, J. W. Chapman, S. F. Wu, and G. A. Barnett, "Geo-location identification of facebook pages," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018, pp. 441–446.

[6] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[7] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013. [Online]. Available: https://arxiv.org/abs/1312.6203

[8] M. Yan, Z. Chen, L. Deng, X. Ye, Z. Zhang, D. Fan, and Y. Xie, "Characterizing and understanding gcns on gpu," *IEEE Computer Architecture Letters*, vol. 19, no. 1, pp. 22–25, 2020.

[9] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf

[10] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. Prasanna, "Graphsaint: Graph sampling based inductive learning method," 2019. [Online]. Available: https://arxiv.org/abs/1907.04931

[11] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Phys. Rev. E*, vol. 73, p. 016102, Jan 2006. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.73.016102

[12] W. W. Zachary, "An information flow model for conflict and fission in small groups," in *Journal of anthropological research*, 1977, p. 452–473.

[13] Y. Hong, *The Application of the Concept of Abstraction in Program Analysis and Social Network*. University of California, Davis, 2017.

[14] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[15] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *Journal of school psychology*, vol. 48, no. 1, pp. 5–37, 2010.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[17] D. B. Rubin, "Multiple imputation," in *Flexible Imputation of Missing Data, Second Edition*. Chapman and Hall/CRC, 2018, pp. 29–62.

[18] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang, "Deep learning versus conventional methods for missing data imputation: A review and comparative study," *Expert Systems with Applications*, p. 120201, 2023.

[19] W.-C. Lin, C.-F. Tsai, and J. R. Zhong, "Deep learning for missing value imputation of continuous data and the effect of data discretization," *Knowledge-Based Systems*, vol. 239, p. 108079, 2022.

[20] S. J. Choudhury and N. R. Pal, "Imputation of missing data with neural networks for classification," *Knowledge-Based Systems*, vol. 182, p. 104838, 2019.

[21] I. Spinelli, S. Scardapane, and A. Uncini, "Missing data imputation with adversarially-trained graph convolutional networks," *Neural Networks*, vol. 129, pp. 249–260, 2020.

[22] G. E. Batista, M. C. Monard *et al.*, "A study of k-nearest neighbour as an imputation method." *His*, vol. 87, no. 251-260, p. 48, 2002.

[23] E.-L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, and M.-D. Cubiles-de-la Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons," *Neural Networks*, vol. 24, no. 1, pp. 121–129, 2011.

[24] J. Yoon, J. Jordon, and M. Schaar, "Gain: Missing data imputation using generative adversarial nets," in *International conference on machine learning*. PMLR, 2018, pp. 5689–5698.

[25] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[26] L. Gondara and K. Wang, "Mida: Multiple imputation using denoising autoencoders," in *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*. Springer, 2018, pp. 260–272.

[27] J. You, X. Ma, Y. Ding, M. J. Kochenderfer, and J. Leskovec, "Handling missing data with graph representation learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 075–19 087, 2020.

[28] A. Cini, I. Marisca, and C. Alippi, "Filling the g_ap_s: Multivariate time series imputation by graph neural networks," *arXiv preprint arXiv:2108.00298*, 2021.

[29] G. Kossinets, "Effects of missing data in social networks," *Social networks*, vol. 28, no. 3, pp. 247–268, 2006.

[30] M. Huisman, "Imputation of missing network data: Some simple procedures," *Journal of Social Structure*, vol. 10, no. 1, pp. 1–29, 2009.

[31] P. S. Bearman, J. Moody, and K. Stovel, "Chains of affection: The structure of adolescent romantic and sexual networks," *American journal of sociology*, vol. 110, no. 1, pp. 44–91, 2004.

[32] J. A. Smith, J. H. Morgan, and J. Moody, "Network sampling coverage iii: Imputation of missing network data under different network and missing data conditions," *Social Networks*, vol. 68, pp. 148–178, 2022.

[33] S. W. Min, K. Wu, S. Huang, M. Hidayetoğlu, J. Xiong, E. Ebrahimi, D. Chen, and W.-m. Hwu, "Large graph convolutional network training with gpu-oriented data communication architecture," 2021. [Online]. Available: https://arxiv.org/abs/2103.03330

[34] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," 2019. [Online]. Available: https://arxiv.org/abs/1903.02428

[35] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 2. Morgan-Kaufmann, 1989. [Online]. Available: https://proceedings.neurips.cc/paper/1989/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf

[36] J. G. F., "The data model concept in statistical mapping," *International Yearbook of Cartography*, vol. 7, pp. 186–190, 1967. [Online]. Available: https://cir.nii.ac.jp/crid/1573668925394541312

[37] Wikipedia contributors, "List of social platforms with at least 100 million active users — Wikipedia, the free encyclopedia," 2023, [Online; accessed 30-May-2023]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=List_of_social_platforms_with_at_least_100_million_active_users&oldid=1155463228

[38] 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

[39] 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

[40] 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.

[41] 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

[42] 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

[43] J. Leskovec and E. Horvitz, "Planetary-scale views on an instant-messaging network," 2008.

[44] 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.

[45] 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.

[46] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 759–768. [Online]. Available: https://doi.org/10.1145/1871437.1871535

[47] Z. heng, J. Caverlee, and K. Lee, "A content-driven framework for geolocating microblog users," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, feb 2013. [Online]. Available: https://doi.org/10.1145/2414425.2414427

[48] C. A. Davis Jr., G. L. Pappa, D. R. R. de Oliveira, and F. de L. Arcanjo, "Inferring the location of twitter messages based on user relationships," *Transactions in GIS*, vol. 15, no. 6, pp. 735–751, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2011.01297.x

[49] D. Rout, K. Bontcheva, D. Preoţiuc-Pietro, and T. Cohn, "Where's @wally? a classification approach to geolocating users based on their social ties," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, ser. HT '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 11–20. [Online]. Available: https://doi.org/10.1145/2481492.2481494

[50] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "Supervised text-based geolocation using language models on an adaptive grid," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing*

and Computational Natural Language Learning, ser. EMNLP-CoNLL '12. USA: Association for Computational Linguistics, 2012, p. 1500–1510.

[51] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Who, where, when and what: discover spatio-temporal topics for twitter users," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 605–613. [Online]. Available: https://doi.org/10.1145/2487575.2487576

[52] M. Hulden, M. Silfverberg, and J. Francom, "Kernel density estimation for text-based geolocation," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI Press, 2015, p. 145–150.

[53] E. Newell, D. Jurgens, H. Saleem, H. Vala, J. Sassine, C. Armstrong, and D. Ruths, "User migration in online social networks: A case study on reddit during a period of community unrest," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 1, pp. 279–288, Aug. 2021. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14750

[54] K. K. Aldous, J. An, and B. J. Jansen, "View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, no. 01, pp. 47–57, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/3208

[55] S. Rahman, "Tourism destination marketing using facebook as a promotional tool," *IOSR Journal of Humanities and Social Science*, vol. 22, pp. 87–90, 02 2017.

[56] R. Wilken, "Places nearby: Facebook as a location-based social media platform," *New Media & Society*, vol. 16, pp. 1087–1103, 10 2014.

[57] J. Lin, R. Oentaryo, E.-P. Lim, C. Vu, A. Vu, and A. Kwee, "Where is the goldmine? finding promising business locations through facebook data analytics,"

in *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, ser. HT '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 93–102. [Online]. Available: https://doi.org/10.1145/2914586.2914588

[58] A. L. Schmidt, F. Zollo, M. D. Vicario, A. Bessi, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "Anatomy of news consumption on facebook," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. 3035–3039, 2017. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1617052114

[59] J. Ruiz, J. D. Featherstone, and G. A. Barnett, "Identifying vaccine hesitant communities on twitter and their geolocations: a network approach," *Proceedings of Hawaii International Conferences on System Science (HICSS-54)*, 2021.

[60] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data mining and knowledge discovery*, vol. 22, pp. 31–72, 2011.

[61] L. Cai and T. Hofmann, "Hierarchical document categorization with support vector machines," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, ser. CIKM '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 78–87. [Online]. Available: https://doi.org/10.1145/1031171.1031186

[62] H. Peng, J. Li, Y. He, Y. Liu, M. Bao, L. Wang, Y. Song, and Q. Yang, "Large-scale hierarchical text classification with recursively regularized deep graph-cnn," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 1063–1072. [Online]. Available: https://doi.org/10.1145/3178876.3186005

[63] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2740–2748.

[64] T. Senator, "Multi-stage classification," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005, pp. 8 pp.–.

[65] Y. Xiong, "Building text hierarchical structure by using confusion matrix," in *2012 5th International Conference on BioMedical Engineering and Informatics*, 2012, pp. 1250–1254.

[66] P. Cavalin and L. Oliveira, "Confusion matrix-based building of hierarchical classification," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, R. Vera-Rodriguez, J. Fierrez, and A. Morales, Eds. Cham: Springer International Publishing, 2019, pp. 271–278.

[67] A. Temko and C. Nadeu, "Svm-based-clustering-schemes," *Pattern Recognition*, vol. 39, no. 4, pp. 682–694, 2006, graph-based Representations. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132030500419X

[68] D. E. Zomahoun, "A semantic collaborative clustering approach based on confusion matrix," in *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2019, pp. 688–692.

[69] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Information Sciences*, vol. 340-341, pp. 250–261, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002002551600044X

[70] S. Dumais and H. Chen, "Hierarchical classification of web content," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 256–263.

[71] R. Babbar, I. Partalas, E. Gaussier, and M. R. Amini, "On flat versus hierarchical classification in large-scale taxonomies," *Advances in neural information processing systems*, vol. 26, 2013.