2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Online Social Community Sub-Location Classification

1st Jiarui Wang
University of California, Davis
jrwwang@ucdavis.edu

2nd Xiaoyun Wang
Nvidia
xiaoyunw@nvidia.com

3rd Chun-Ming Lai
Tunghai University
cmlai@thu.edu.tw

4th S. Felix Wu
University of California, Davis
sfwu@ucdavis.edu

***Abstract*—Facebook public pages are a popular form of online social network (OSN) communities. The "like" connections between public pages create a graph of pages on Facebook. Geographic location is a crucial piece of metadata for pages, but it is often omitted by page managers. We propose a classification algorithm to restore the missing subdivision location of Facebook public pages. We propose neighborhood state distribution vectors as features for graph neural networks to classify the state of the pages. Then, we define intrastate and interstate Facebook public pages based on the high-probability state label outputted by the classification model. Finally, we profile states with different influences over the online communities by analyzing the classification confusion matrix, interstate page percentages, and interstate pages across state borders. Our method achieves better accuracy (87.52%) and F1 score (0.8756) than previous studies (66.2% and 73.08%).**

I. Introduction

The social relationship has long been a topic of academic interest. The social network is a representation of 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 world to the digital world. Not only did personal social networks move online, but also social communities. People are often part of multiple communities and participate in conversations and activities within the communities offline. These communities could be neighborhoods, workplaces, or groups with shared interests. Many groups or communities have online information pages or discussion groups on Facebook or other forums online.

Facebook is the most popular online community platform and has been the subject of many research projects. Our research focuses on public Facebook pages. The Facebook public page is a platform for information announcements, user discussion, news dispersion, public relations promotion, and business promotion. One important attribute of Facebook pages is location, which indicates where the majority of page activities and users are located. Many meaningful research activities and promotions can be conducted based on page location, such as targeting highly influential pages in a specific area.

However, not all pages have their location information filled out by their page managers. In our data set of public Facebook pages, only 30.8% (18,895,994 pages) out of a total of 61,263,729 pages listed their location. Predicting the missing location of pages is essential for conducting other research related to geographic location. Sub-location classification is even more challenging and important, such as classifying the state of pages in the United States. In this research, we aim to make the following contributions:

- Propose neighborhood state distribution vectors as features and graph neural networks to classify the state of the pages. This method outperforms previous algorithms by improving the classification accuracy from 66.2% and 73.08% to 87.52% accuracy and 0.88 F1 score.
- Define intrastate and interstate Facebook public pages based on the high-probability state label outputted by the classification model.
- Profile states with different influences over online communities by analyzing the classification confusion matrix, interstate page percentages, and interstate pages across state borders.

II. Data Description and Cleaning

In the metadata for each Facebook public page, page managers can fill in the city location. However, this is often omitted. Pages can also like other pages, just as Facebook users can. This is managed by the page's managers.

We built a page-likes graph using only ground truth data. This data consists of 6,194,277 pages with city locations inside the United States and edges between any two of these pages. We ignored 55,069,452 pages and their connecting edges that have city locations outside the United States or no city locations at all. These pages were ignored because our focus is on sub-location classification in the United States, and the algorithms would not be able to handle all the pages and edges. There was no need to process them.

The generated subgraph of the ground truth U.S. pages has disconnected components because some pages that connect the U.S. pages are ignored. The largest connected component has 5,873,395 pages. We focus on the largest connected

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
http://dx.doi.org/10.1145/3625007.3627504

component because other components are too small. The state location classification problem for Facebook public pages can be translated into a more practical problem. Given a directed graph, where each node is a Facebook public page labeled with a state. Each edge in the graph starts from one page and ends at other pages which are liked by the starting page. The goal is to achieve good state classification accuracy on this page-likes graph, which is the largest connected component of the ground truth U.S. pages.

There are two kinds of U.S. pages in the subgraph:

- **Deterministic Pages**: 2,147,399 pages whose cities have unique names among all the states inside the United States.
- **Non-deterministic Pages**: 3,725,996 pages whose cities share the same names with other cities in different states inside the United States.

We can use deterministic pages as ground truth pages directly because the states of the pages are determined. However, we cannot use non-deterministic pages, because the states of those pages are not determined. We include non-deterministic pages in the graph, but only count the state labels of the deterministic page neighbors when computing the neighborhood state distribution feature vectors. Non-deterministic Pages are only connecting nodes.

III. Facebook Page States Classification

Since the pages are connected to each other in a graph, where edges represent the "likes" relationship between pages, it is natural to apply graph-based algorithms to the page graph. Graph neural networks are a good fit for this page sub-location classification within the country border, as they can learn the relationships between pages and use this information to classify pages.

A. Graph Neural Network Model Selection

The GraphSAGE [6] model updates the feature vectors with the same propagation rule as the GCN model [3]. The difference is that while GCN updates the feature vectors of all nodes in the graph in each iteration, GraphSAGE only updates a batch of nodes in each iteration by uniformly sampling a fixed number of neighboring nodes for each node in the batch [6] [11]. This reduces the memory and computation footprints and allows GraphSAGE to work on large graphs like ours, compared to the GCN model [14]. We choose GraphSAGE as our baseline model.

Another GNN model we used is GraphSAINT [16]. Unlike GraphSAGE uses neighborhood sampling, GraphSaint leverages graph sampling. For each batch, a full GCN-like model runs on a subgraph of the original graph. By downsizing the original graph to a subgraph, GraphSAINT can handle large graphs in a superior training time.

B. Machine Learning Feature Selection

Machine learning algorithms need features associated with the pages to perform training and classification. We propose neighborhood state distribution vectors as the features instead. This is because every page in the connected graph will have a non-zero number of neighbors, which means non-zero neighborhood state distribution vectors. The neighborhood state distribution is the ratio of the number of neighbors from each state, over the total number of neighbors.

We choose both neighborhood state distributions within one hop and within two hops as the feature vector. For one-hop and two-hop neighborhood state distributions, we consider inward edge direction neighbors, outward edge direction neighbors, and undirected edge neighbors. For each direction, the neighborhood state distribution is the percentage of neighbors from every state over the total neighbors from all states. Following is the definition of neighborhood state distributions (NSD) vector:

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

Where:

- $INSD_1(State_i)$/$ONSD_1(State_i)$/$UNSD_1(State_i)$ represent the inward/outward/undirected neighbor state distribution for $State_i$ within one-hop distance from the $Page$.
- $INSD_2(State_i)$/$ONSD_2(State_i)$/$UNSD_2(State_i)$ represent the inward/outward/undirected neighbor state distribution for $State_i$ within two-hop distance from the $Page$.

We define the element of neighborhood state distribution for each page, $INSD$, $ONSD$, $UNSD$ as the following:

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

Where:

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

For some pages with a small number of one-hop neighbors, the state distribution could be biased. This is because the state distribution could be heavily influenced by the dominant number of neighbors from a single state. To address this issue, we consider two-hop neighbors as well. This increases the number of neighbors for each page, which helps to reduce the bias in the state distribution. We do not consider three-hop neighbors because the number of total neighbors would easily reach millions. This would make the state distribution vectors indistinguishable for every node, as the receptive field would be too large.

IV. Evaluating Page Location Classification

A. Model Setup

Both our GraphSAGE and GraphSAINT models have two layers. The number of output channels is 51, the same as the total number of states in the United States, including Washington D.C. The output is the probabilities for each of the 51 classes, representing the probability that the page belongs to a particular state. The number of input channels and the number of hidden channels are both 306, the same as the number of features in the neighborhood state distribution feature vectors. Models are implemented in the Graph Neural Networks framework PyG (PyTorch Geometric) [5].

Our GraphSAGE Model aggregates messages from all neighboring nodes, instead of sampling the neighboring nodes, which introduces random bias and makes the model slow to converge in our experiments. GraphSAINT samples a subgraph of the original graph for every batch in each iteration, using random walk sampling that samples the nodes by their importance intuitively, which usually has better performance than random node sampling and random edge sampling.

TABLE I
Accuracy for deterministic pages

Algorithm	Precision	Recall	F1	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

B. Experiment Result

The majority voting algorithm [8] and the BFS-based machine learning algorithm [10] are the previous research that tried to solve the sub-location classification problem, which are introduced in Section V-B. We use them as the comparison algorithm. All the experiments use the same data set described in Section II. Both our GNN algorithms perform much better than the majority voting and the BFS-based machine learning algorithms in Table I. The accuracy results of GraphSAINT are GraphSAGE are means and 95% confidence intervals of 3 runs.

C. Confusion Matrix

The confusion matrix shows the mismatch between each class pair, revealing interesting findings hidden in the data. The confusion matrix in Figure 1, is computed from the ground truth label and the classification result of running GraphSAINT on deterministic 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%.

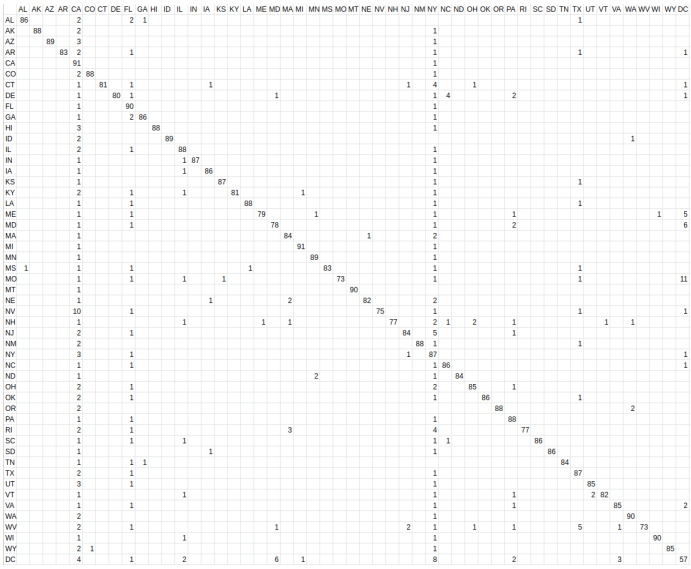

Fig. 1. Confusion Matrix

Here are some observations from the confusion matrix:

- CA, NY, and FL are the top-level center states for Facebook pages in the United States. Almost all states have noticeable mismatching scores with them. This is because pages from other states have a higher possibility to be connected with pages from these 3 top-level center states, which have the most pages in the U.S.
- TX, PA, and IL are regional center states. Pages from their neighboring states have a higher probability of being mislabeled to these 3 regional center states.
- There are pairs of states which have higher mismatching scores and are sharing borders, such as NV and CA, NJ and NY, CT and NY, RI and NY, DC and MD, DC and VA, RI and MA, and OR and WA. This shows that pages from adjacent states have a higher possibility to connect, compared to pages from two non-top-level states that are far away from each other.

Washington D.C. was not initially included in the ground truth data set A. However, it has a large population and is located on the border of Maryland and Virginia, making it an ideal example. Since there are dozens of other cities called "Washington" in states with small populations, we labeled all of these cities as "DC." Missouri and Maine have a significant number of pages that have been mislabeled as "DC." This could be due to their own "Washington" cities being mislabeled as "DC" in the ground truth data, which then led to more pages from these states being mislabeled as "DC".

D. Interstate Page and Intrastate Page

For our multi-class classification problem, we compute the cross-entropy loss between the ground truth labels and the outputs from the GNN models. The outputs are expected to be

TABLE II
INTERSTATE AND INTRASTATE PAGE EXAMPLE

Page ID	Truth	Threshold	High probability states
5XXXX547	FL	0.08	FL 0.37, IL 0.17, DC 0.20
5XXXX307	NY	1.23e-07	NY 0.54, DC 0.45
4XXXX747	CA	5.94e-15	CA 1.0
5XXXX475	NJ	0.04	NJ 0.76

TABLE III
PAGE DISTRIBUTION WITH DIFFERENT NUMBERS OF HIGH-PROBABILITY STATES IN DETERMINISTIC PAGE DATA

States #	1	2	3	4	5
Page #	1934364	120582	37357	18959	10758
States #	6	7	8	9-20	21-31
Page #	7319	4652	3437	9807	164

unnormalized for each class, which do not need to be positive or sum to 1. We apply the trained model on deterministic page data A to compute outputs for each page, then input the outputs into a softmax function [2] to compute the probability that the page belongs to each state. The probabilities that one page belongs to each state are between 0 and 1, and the sum is 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**: There is only one state with a high probability for the page.
- **Interstate page**: There is more than one state with high probabilities for the page.

To properly group 51 probabilities into two groups, higher and lower probability groups, we use Jenks natural breaks algorithm [4]. This method minimizes the variation of the probabilities within each group, so the probabilities within each group are as close as possible in value to each other. The higher probability group contains a certain number of probabilities for states. If the number of states is greater than one, the page is an interstate page, otherwise, it is an intrastate page.

Table II shows the examples for interstate pages and intrastate pages, page ID is hidden for privacy. Table III shows the page distribution over different numbers of high-probability states in data. The total number of interstate pages is 213,035, 9.92% of 2,147,399 both interstate and intrastate pages together in the data set.

Figure 2 shows the interstate page percentages of each state plotted on the U.S. map. Interstate page percentage is the ratio of the number of interstate pages to the number of total pages in one state. From the map, we can see that Nevada, Missouri, West Virginia, Virginia, and Washington D.C. have the highest interstate page percentages. To further investigate the interstate pages of these states, we need to know how many pages are interstate between each pair of neighboring states.

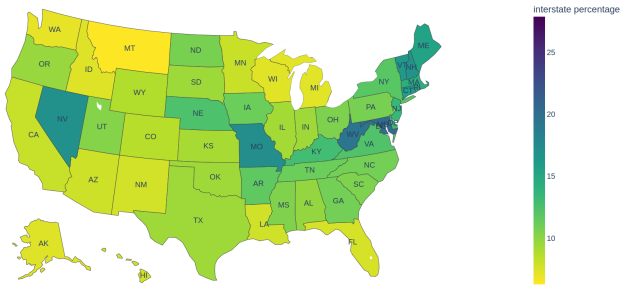

Fig. 2. Interstate page percentage map

We plot the number of interstate pages at different scales across the borders of every pair of neighboring states in Figure 3. The interstate page number is normalized by the total page number of the state with fewer pages in the state pairs. Interstate pages that are less than 0.5% on the border are omitted for easy reading. State Alaska and Hawaii don't share borders with any states, but they both share the most interstate pages with Washington, which is closer to them than other states. 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. Nevada heavily shares pages with California. Missouri shares pages with Illinois, Kansas, and the District of Columbia (DC). West Virginia shares pages with Texas, New Jersey, Pennsylvania, and Ohio. Maryland shares pages with DC and Delaware. DC is a newly found sub-region center that does not show on the confusion matrix. It heavily shares pages with states, Maryland, New York, Virginia, California, Missouri, North Carolina, and Pennsylvania.

To further explore the center states' influence, we performed a control experiment in which we remove all the center state labels from data set. We then trained and tested the GraphSAINT model only on the pages from non-center states. After obtaining the trained model, we classify the pages in data A from both center states and non-center states, but only to the labels of non-center states. This experiment indeed increased the number of interstate pages between neighboring states of the center states. However, it also skews the data, as DC, Washington, and New Jersey become the top center states, sharing a significant number of interstate pages with almost every state.

V. RELATED WORK

A. *Facebook User Graph Analysis*

The Facebook user graph is a network of users who are connected to each other by friendship ties. This graph has been the subject of much research by social scientists and computer scientists. Ugander et al. characterized the global structure of the Facebook user graph and computed numerous network properties [13]. Barnett and Benefield investigated the determinants of the Facebook user network. They found that proximity and cultural homophily are two important factors

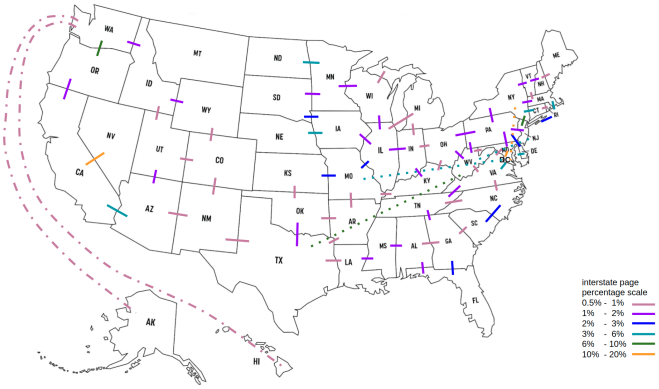

Fig. 3. Numbers of interstate page across state borders

that influence who are friends with whom on Facebook. They also found that countries with international Facebook friendship ties tended to share borders, language, civilization, and migration [1].

B. Facebook Page Location Classification

The Facebook page graph is a network of pages where the edges represent the relationship that one page likes another page. Hong et al. [7] [8] studied and characterized this graph, and proposed a majority voting algorithm to classify the missing country location information of Facebook pages. This algorithm performs well on the country location classification problem as most pages that are connected by the edges are within one country due to the same cultural, language, and social context. However, the majority voting method is not effective for subdivision location classification, such as state labeling within the United States.

To resolve this issue, Lin et al. proposed a Breadth-First Search(BFS) based machine learning algorithm that uses handpicked anchor pages as seeds to start the Breadth-First Search from [10]. However, this algorithm has major issues. First, it does not have full coverage of data, because there always are pages that not reachable from any seed anchor page. Unreachable pages have all-zero distance feature vectors. The anchor pages of each state are likely not the centroid of the clusters for each state as claimed. There are some arbitrary thresholds handpicked.

VI. Conclusion

In this paper, we introduced our study on subdivision location classification of Facebook public pages from the United States. First, We investigated the drawbacks of the previous research activities for sub-location classification. Then, we proposed a new method that uses GNN models to learn from the neighborhood state distribution(NSD) vectors of each page. We then use these models to classify the pages into their respective states. Our method was able to significantly improve the classification accuracy to 87.52% and F1 score to 0.8756, evaluated on the data set of Facebook public pages in the United States.

We also used our method to define intrastate and interstate Facebook public pages. We found that intrastate pages were more likely to be liked by other pages from the same state, and interstate pages were more likely to be liked by pages from other states. Finally, we profiled states with different influences over the online communities by analyzing the state classification confusion matrix, state interstate page percentages, and interstate pages across state borders. We conclude that the geographic location of the Facebook public pages is an important factor in both the formation of the "likes" relationship between pages and the sub-location classification for the pages.

References

[1] George A Barnett and Grace A Benefield, "Predicting international Facebook ties through cultural homophily and other factors", New Media & Society, vol. 19(2), pp. 217-239, 2017.

[2] John Bridle, "Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters", Advances in Neural Information Processing Systems, vol. 2, pp. 217-239, 1989.

[3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun, "Spectral networks and locally connected networks on graphs", Proc. Int. Conf. Learn. Representations, pp. 1-14, 2014.

[4] JENKS G. F. "The data model concept in statistical mapping", International Yearbook of Cartography, vol. 7, pp. 186-190, 1967.

[5] Matthias Fey and Jan Eric Lenssen, "Fast graph representation learning with pytorch geometri", arXiv preprint arXiv:1903.02428, 2019.

[6] Will Hamilton, Zhitao Ying, and Jure Leskovec, "Inductive representation learning on large graphs", Advances in Neural Information Processing Systems, vol. 30, 2017.

[7] Yunfeng Hong, "The Application of the Concept of Abstraction in Program Analysis and Social Network", University of California, Davis, 2017.

[8] Yunfeng Hong, Yu-Cheng Lin, Chun-Ming Lai, S. Felix Wu, and George A. Barnett, "Profiling facebook public page graph", 2018 International Conference on Computing, Networking and Communications (ICNC), pp. 161–165, 2018.

[9] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong, "Statistical properties of sampled networks", Phys. Rev. E, vol. 73, pp. 016102, Jan 2006.

[10] Yu-Cheng Lin, Chun-Ming Lai, Jon William Chapman, S. Felix Wu, and George A. Barnett, "Geo-location identification of facebook pages", 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 441–446, 2018.

[11] Seung Won Min and Kun Wu and Sitao Huang and Mert Hidayetoğlu and Jinjun Xiong and Eiman Ebrahimi and Deming Chen and Wen-mei Hwu, "Large graph convolutional network training with gpu-oriented data communication architecture", arXiv preprint arXiv:2103.03330, 2021.

[12] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini, "The graph neural network model", IEEE Transactions on Neural Networks, vol. 20(1), pp. 61-80, 2009.

[13] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow, "The anatomy of the facebook social graph", arXiv preprint arXiv:1111.4503, 2011.

[14] Mingyu Yan, Zhaodong Chen, Lei Deng, Xiaochun Ye, Zhimin Zhang, Dongrui Fan, and Yuan Xie, "Characterizing and understanding gcns on gpu", IEEE Computer Architecture Letters, vol. 19(1), pp. 22–25, 2020.

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

[16] Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna, "Graphsaint: Graph sampling based inductive learning method", arXiv preprint arXiv:1907.04931, 2019.