Minimally Supervised Contextual Inference from Human Mobility: An Iterative Collaborative Distillation Framework

Jiayun Zhang[1] , **Xinyang Zhang**[2] , **Dezhi Hong**[3] , **Rajesh K. Gupta**[1] and **Jingbo Shang**[1]
[1]University of California, San Diego
[2]University of Illinois at Urbana-Champaign
[3]Amazon*
{jiz069,rgupta,jshang}@ucsd.edu, xz43@illinois.edu, hondezhi@amazon.com

Abstract

The context about trips and users from mobility data is valuable for mobile service providers to understand their customers and improve their services. Existing inference methods require a large number of labels for training, which is hard to meet in practice. In this paper, we study a more practical yet challenging setting—contextual inference using mobility data with minimal supervision (i.e., a few labels per class and massive unlabeled data). A typical solution is to apply semi-supervised methods that follow a self-training framework to bootstrap a model based on all features. However, using a limited labeled set brings high risk of overfitting to self-training, leading to unsatisfactory performance. We propose a novel collaborative distillation framework STCOLAB. It sequentially trains spatial and temporal modules at each iteration following the supervision of ground-truth labels. In addition, it distills knowledge to the module being trained using the logits produced by the latest trained module of the other modality, thereby mutually calibrating the two modules and combining the knowledge from both modalities. Extensive experiments on two real-world datasets show STCOLAB achieves significantly more accurate contextual inference than various baselines.

1 Introduction

The prevalence of location-based mobile services offers new opportunities for businesses to better understand the context of trips (e.g., transportation mode and purpose) and their customers (e.g., ethnicity, disability, and socioeconomic status). Such information can facilitate a wide spectrum of mobile applications, including human mobility recovery [Fang *et al.*, 2021], urban planning [Liu *et al.*, 2017], and personalized location recommendation [Wang *et al.*, 2020b]. In practice, given the sensitive nature, very few users would share the contextual information, especially at the beginning of the business [Schein *et al.*, 2002; Quercia *et al.*, 2010; Shen *et al.*, 2018]. As such, we study the problem of contextual inference from mobility data *under minimal supervision*, which is an extreme case of semi-supervised learning when using a few labels (e.g., 10) per class. Specifically, we study the inference of people's demographic attributes.

Related work on demographics inference from human mobility [Wang *et al.*, 2017; Zhong *et al.*, 2015; Xu *et al.*, 2020; Wang *et al.*, 2020a] requires a large number (e.g., tens of thousands) of users to share labels for training and is prone to overfitting in the minimally-supervised setting. To mitigate the label scarcity problem, existing semi-supervised methods [Cascante-Bonilla *et al.*, 2021; Lee and others, 2013; Li *et al.*, 2019] typically follow a self-training framework that bootstraps a single model using all features at once. However, for mobility data, simple concatenation of spatial and temporal features does not always guarantee improvement in predictions. Due to different network sizes for different modalities, the model inference might lean heavily on one of the modalities [Gat *et al.*, 2020]. Especially when training data is limited, the model does not have enough supervision to find the optimal combination of different modalities, leading to unsatisfactory performance and poor generalization.

To better learn with limited supervision, we propose to alternatingly train two separate modules—one for spatial and one for temporal information—and then let them iteratively distill knowledge from each other following a novel collaborative distillation framework STCOLAB[1] (see Figure 1).

Instead of building one large model that fuses the spatial and temporal modules, we separate the training of the two modules to sufficiently learn features of both modalities with supervision and avoid one modality dominating the learning. Both modules in STCOLAB are supervised by the limited labeled data for contextual inference in an alternating manner—the spatial module first learns geographic features from maps describing where a user has visited; the temporal module then utilizes the features extracted by the trained spatial module and learns cyclic temporal patterns.

In addition to training each module using labeled data, we propose a novel collaborative distillation framework that combines the knowledge of spatial and temporal modules to improve model generalization in an iterative manner. We use the latest trained spatial and temporal modules as the teacher

*Work unrelated to Amazon.

[1]Code is available at https://github.com/jiayunz/STColab

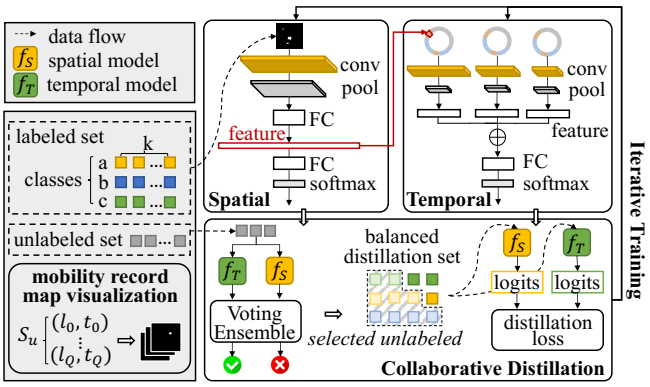


Figure 1: An overview of STCOLAB. It consists of two modules: the spatial module takes a map and learns spatial features; the temporal module then utilizes the spatial features and learns temporal patterns in trajectories to make predictions. A collaborative distillation process distills knowledge from the latest trained modules to guide the training of both. The process iterates in self-training cycles.

model to guide the training of the current spatial/temporal module. Specifically, we construct guidance based on the unlabeled data for which the latest trained modules have consistent and confident predictions. We regulate the learning of the module being trained at each iteration by forcing it to approximate the logits produced by the teacher model on such selected unlabeled data. The two modules give complementary supervision from different views to each other and calibrate the predictions. This way, we combine spatial and temporal information to improve model generalization.

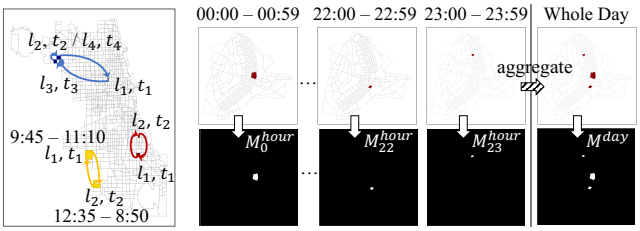
We conduct extensive experiments on two real-world mobility datasets collected from two metropolitan cities in different countries: Chicago in the United States and Brasilia in Brazil. We show that with a small number of labeled samples per class (e.g., 10), STCOLAB can infer important demographic attributes about users with reasonable accuracy, significantly outperforming the state-of-the-art methods. To the best of our knowledge, we are the first to address the contextual inference problem using mobility data with minimal supervision. We make the following contributions:

- We study the problem of contextual inference from mobility data under the challenging yet practically important minimally supervised setting, where only a few annotated samples are available per class.
- We propose a novel framework called STCOLAB, which learns from spatial and temporal modalities iteratively and distills knowledge from both modalities collaboratively to improve generalization using unlabeled data.
- We conduct extensive experiments on two real-world mobility datasets to predict demographic attributes. Results show STCOLAB can predict such information with reasonable accuracy, improving upon state-of-the-art methods.

2 Preliminaries

2.1 Concepts

Definition 2.1 (Region)**.** *A region is a geographic unit with a corresponding polygon in the coordinates, denoted as l.*



(a) Mobility Records (b) Visualized Maps

Figure 2: Examples of mobility records and visualized maps: (a) mobility records of three random people in Chicago; (b) visualized hour maps and day map generated from a person's mobility records.

The two datasets we used in the experiments divide cities into regions according to certain criteria. In the *Chicago* dataset, the city is divided into 866 *census tracts*, each of which is defined for the purpose of taking a census. In the *Brasilia* dataset, the city is divided into 233 *micro zones*, which is defined by the government for statistical purposes. The divisions are shown as grey grids in Figure 2(a). Note that STCOLAB can also deal with location data of other forms such as fine-grained GPS coordinates. The format of location data is flexible and driven by the available dataset.

Definition 2.2 (Daily Mobility Record)**.** *An entry of mobility records is a triplet (u, l, t), which denotes user u visits region l during time period t in the day. By sorting the records of user u by time, we get a sequence of time-location pairs:*

$$\mathcal{S}_u = [(l_0, t_0), (l_1, t_1), \ldots, (l_Q, t_Q)],$$

where Q is the total number of records of user u and l_q and t_q are the region and time period of the q-th record. The time range of each mobility record from a user is one day.

Figure 2(a) shows the daily mobility records of three random users from the Chicago dataset in different colors. For example, the yellow trajectory shows the user stays at l_1 during t_1 (9:45 a.m. - 11:10 a.m.) and stays at l_2 during t_2 (12:35 p.m. - 8:50 a.m.).

2.2 Problem Definition

We aim to predict demographic attributes from people's mobile data under minimal supervision. This is an extreme case of semi-supervised learning, where the number of labeled data is very limited. For each class of a demographic attribute, only a few (e.g. 10) ground truth labels are available.

The number of samples known in each class is denoted as k. The training dataset $\mathcal{D}$ consists of two parts: the labeled set $\mathcal{D}_{lb} = \{(x, y)|x \in \mathcal{X}, y \in \mathcal{C}\}$ and the unlabeled set $\mathcal{D}_{ul} = \{x|x \in \mathcal{X}\}$, where x and y denote the input features and the label of a sample, $\mathcal{X}$ is the set of mobility data and $\mathcal{C}$ is the set of classes. The total number of labeled samples $|\mathcal{D}_{lb}| = k * |\mathcal{C}|$. $|\mathcal{D}_{ul}| \gg |\mathcal{D}_{lb}|$. We aim to learn a model for each prediction task to assign an attribute class label a to each person u given the daily mobility records.

3 Our STCOLAB Framework

As shown in Figure 1, STCOLAB consists of two modules—a spatial module and a temporal module—to learn from differ-

Algorithm 1: Iterative Collaborative Distillation

Require: labeled set $\mathcal{D}_{lb}$, unlabeled set $\mathcal{D}_{ul}$

Initialize iteration $t \leftarrow 1$;
while $t \leq$ *MaxIteration* **do**
  Train spatial model $f_S^{(t)}$ according to Eq. 3;
  if $t > 1$ **then**
    Construct distillation dataset $\tilde{\mathcal{D}}_S^{(t)}$ using $f_S^{(t-1)}$ and $f_T^{(t-1)}$ based on voting ensemble;
    Distill knowledge to $f_S^{(t)}$ using $f_T^{(t-1)}$ as teacher according to Eq. 2.
  end
  Train temporal model $f_T^{(t)}$ according to Eq. 4;
  if $t > 1$ **then**
    Construct distillation dataset $\tilde{\mathcal{D}}_T^{(t)}$ using $f_S^{(t)}$ and $f_T^{(t-1)}$ based on voting ensemble;
    Distill knowledge to $f_T^{(t)}$ using $f_S^{(t)}$ as teacher according to Eq. 2.
  end
  $t \leftarrow t + 1$;
end

ent modalities. We design a collaborative distillation process to combine knowledge from both modules in a self-training manner. The pseudo code is presented in Algorithm 1.

3.1 Iterative Collaborative Distillation

Conventional methods for integrating spatial and temporal information involve fusing features from two sources within one large model, or ensembling the predictions of the spatial and temporal modules. However, training a model with data from spatial and temporal sources in a single pass is likely to bias the model to one modality [Gat *et al.*, 2020]. To sufficiently learn features from both modalities, we propose an iterative learning process that alternates the training of the spatial and temporal modules in several iterations.

The small training set shows very limited information about the data distribution so the model is prone to overfitting. While labels are hard to collect, the unlabeled data itself provides valuable information for model generalization. Moreover, the spatial and temporal modules learn from two different views of the data. By utilizing the knowledge learned by both modules, they can reduce confirmation bias [Arazo *et al.*, 2020] and mutually calibrate each other. Thus, we design a novel collaborative distillation process to distill knowledge between the two modules by using unlabeled samples.

In the conventional form of knowledge distillation, knowledge is distilled to the new model by training it to approximate the output of a teacher model on a distillation dataset [Hinton *et al.*, 2015]. However, the model trained on a very small training set is likely to make random predictions on unseen data, which causes knowledge given by the teacher model to be noisy. Iterative learning from such noisy knowledge can act like a negative feedback loop and degrade performance [Arazo *et al.*, 2020]. To increase the chance that the teacher model gives correct predictions, we design a voting ensemble method by evaluating the consistency and confidence of its predictions to select samples.

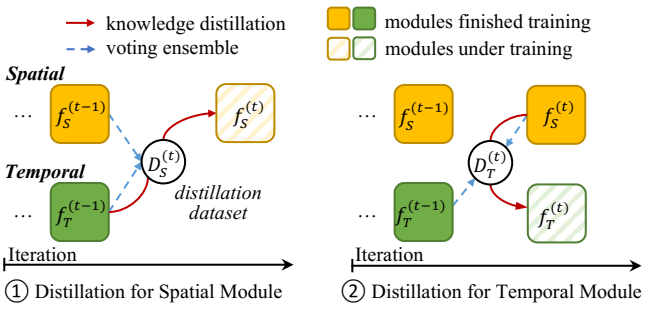


Figure 3: Collaborative distillation in a self-training cycle. At iteration t ($t \geq 2$), ① for spatial module $f_S^{(t)}$, we use $f_S^{(t-1)}$ and $f_T^{(t-1)}$ to form distillation dataset via voting ensemble and use $f_T^{(t-1)}$ as teacher for distillation. Then, ② for temporal module $f_T^{(t)}$, we use $f_T^{(t-1)}$ and $f_S^{(t)}$ for voting ensemble and use $f_S^{(t)}$ as teacher.

Figure 3 explains the iterative collaborative distillation process. In the first iteration, both modules are trained using only the labeled training data. Starting from the second iteration, in addition to training with ground-truth data, we further use the latest trained models from both modalities to conduct a voting ensemble for constructing a distillation dataset and use the latest trained model as the teacher model for distilling knowledge to the current module. Denote the spatial model and temporal model at iteration t as $f_S^{(t)}$ and $f_T^{(t)}$ respectively. Without loss of generality, we illustrate the process of distillation for the temporal model $f_T^{(t)}$.

Voting ensemble. There is a higher chance that the modules give a correct prediction if both modules give the same prediction to one sample, compared to the case when the two modules disagree on the prediction. Thus, we use $f_T^{(t-1)}$ and $f_S^{(t)}$ to make predictions on all unlabeled samples and only select those for which the two modules give the same predictions. In addition, the predicted probability can be regarded as the prediction confidence which indicates how certain the model thinks the prediction is correct. We further use percentile scores and choose a subset of the unlabeled samples whose prediction probabilities given by $f_S^{(t)}$ are above the r-th percentile. The threshold of the prediction confidence of class a is $\mathbb{T}_a^{(t)} = \text{percentile}(f_{S,a}^{(t)}(*), r)$, where $f_{S,a}^{(t)}(*)$ are the prediction confidence of all unlabeled samples with respect to class a given by $f_S^{(t)}$. Combined with the voting condition, the selected unlabeled set at the t-th iteration is:

$$\tilde{\mathcal{D}}_{T_{ul}}^{(t)} \leftarrow \{x_i | f_{S,a}^{(t)}(u_i) \geq \mathbb{T}_a^{(t)} \text{ and } \hat{y}_{i,a}^S = \hat{y}_{i,a}^T = 1\}_{a \in \mathcal{C}}, \quad (1)$$

where x_i is the input of user u_i. We do upsampling to get a balanced distillation dataset. Denote the number of samples in $\tilde{\mathcal{D}}_{T_{ul}}^{(t)}$ that are predicted to be class a (i.e., $\hat{y}_{i,a}^S = 1$) as $\hat{N}_a$. For each class a, we randomly sample $max\{\hat{N}_m\}_{m \in \mathcal{C}} - \hat{N}_a$ samples from the original training set, making all classes have the same amount of samples. Denoted the labeled dataset sampled from the original training set as $\tilde{\mathcal{D}}_{T_{lb}}^{(t)}$. The resulting

balanced distillation dataset $\tilde{\mathcal{D}}_T^{(t)} = \tilde{\mathcal{D}}_{T_{ul}}^{(t)} \cup \tilde{\mathcal{D}}_{T_{lb}}^{(t)}$.

Knowledge distillation. We use the latest trained model $f_S^{(t)}$ as the teacher to make predictions on the unlabeled set $\tilde{\mathcal{D}}_{T_{ul}}^{(t)}$. Denote $f_S^{(t)}(u_i)$ as the predicted probability of user u_i given by $f_S^{(t)}$. We let $f_T^{(t)}$ approximate the predicted probabilities of samples in $\tilde{\mathcal{D}}_{T_{ul}}^{(t)}$ and the ground-truth of samples in $\tilde{\mathcal{D}}_{T_{lb}}^{(t)}$. The distillation loss is:

$$\tilde{\mathcal{L}}_T^{(t)} = -\frac{1}{|\tilde{\mathcal{D}}_{T_{ul}}^{(t)}|} \sum_i f_S^{(t)}(u_i) \log p_i^{(t)} - \frac{1}{|\tilde{\mathcal{D}}_{T_{lb}}^{(t)}|} \sum_j y_j \log p_j^{(t)}, \quad (2)$$

where $p_i^{(t)}$ is the predicted probability of user u_i given by $f_T^{(t)}$. In this way, the knowledge from the latest spatial model $f_S^{(t)}$ is distilled to the temporal model $f_T^{(t)}$. A similar process applies to the distillation for spatial model $f_S^{(t+1)}$, which uses $f_T^{(t)}$ and $f_S^{(t)}$ for voting ensemble and $f_T^{(t)}$ as the teacher for distilling knowledge.

3.2 Spatial Module

To utilize spatial information, we visualize the mobility records as *visualized maps*. The maps contain rich information about the spatial structure. Even with minimal supervision, locations that are not present in the labeled set may still have a similar geographical distribution on a map. This enables the model to better generalize to unseen locations. A visualized map is a one-channel image M showing the regions that the user has been to during a time period. Given the set of regions $\mathcal{L}$ that a user has been to during time period t, we visualize the regions on the map of the city by highlighting the regions in $\mathcal{L}$ in white and marking the others in black. The images are rendered using the GeoPandas Library [Jordahl *et al.*, 2020]. The maps of two time periods t_1 and t_2 can be aggregated by taking the maximum in each entry of the two maps, describing the regions that the user has been to during t_1 and t_2. In STCOLAB, we leverage maps of two different temporal granularities: *hour map* and *day map*. The hour map includes the locations visited in each hour and the day map is the aggregated map of the 24 hours. Figure 2(b) gives examples of the visualized maps generated from a random person's daily mobility records in the Brasilia dataset.

The spatial module learns spatial features, such as the geographic location and distance of people's daily trajectories, from the visualized map M. The spatial module performs convolution operation upon M as $c = \text{conv}(W_c, M) + b_c$, where W_c is the weight matrix and b_c is the bias term. We use Parametric ReLU (PReLU) as the activation function and use max pooling after the convolution operations. Then, a Fully-Connected (FC) Layer is applied to generate vector $z \in \mathcal{R}^d$ representing the spatial features learned from the map image.

To train the spatial module, we aggregate the visualized maps of each hour to get the day map M_i^{day}. The spatial module extracts spatial features z_i^{day} from M_i^{day} and applies an FC layer to make predictions. We denote p_i^{spatial} as the spatial module's predicted probability and y_i as the ground truth. The learning objective is to minimize the cross entropy loss:

$$\mathcal{L}_S = -1/|\mathcal{D}_{lb}| \sum_i y_i \log p_i^{\text{spatial}}. \quad (3)$$

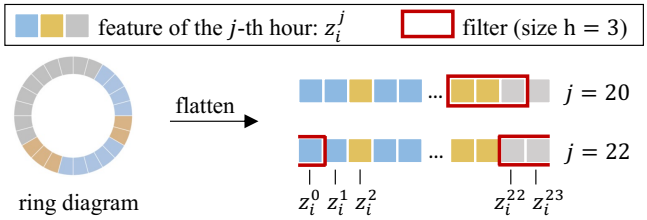


Figure 4: Illustration of periodic convolution

3.3 Temporal Module

The temporal module takes the sequence of spatial features in each hour of the day as inputs and is equipped with convolution layers with different filters sliding over the sequence to extract temporal features within different time periods.

Let $z_i^j \in \mathcal{R}^d$ be the spatial representation vector corresponding to the j-th hour of user u_i. The input to the temporal module is a sequence of the spatial features of each hour. We denote the concatenation of the spatial features from the j-th to the $(j+h)$-th hours of user u_i as $z_i^{j:j+h}$. A convolution filter of size $h \times d$ moves along the time dimension and is applied to a window of h hours to produce a new feature $e^{j:j+h}$ each time. For example, a feature e_j is generated from a window (of size h) of spatial features $z_i^{j:j+h}$ by $e_j = W_e \cdot z_i^{j:j+h} + b_e$, where W_e is the weight matrix of filter kernel and $b_e \in R$ is the bias term. This filter is applied to each possible window of spatial features in the sequence.

A person's temporal mobile pattern can be represented as a cyclic ring, where the beginning and the end of a day are continuous and connected to each other. To capture the cyclic patterns, we employ periodic convolution operations through circular padding. Specifically, the start of the sequence is padded with features from the end of the sequence, and vice versa. Figure 4 illustrates how periodic convolution works.

We use PReLU as the activation function and apply a max-pooling operation over the spatial features to get the features corresponding to a particular filter. The module uses N_f filters (with varying window sizes) to obtain multiple features and then concatenates them together to get the temporal features. An FC Layer is applied to make predictions. We denote p_i^{temporal} as the temporal module's predicted probability. The learning objective is to minimize the cross-entropy loss:

$$\mathcal{L}_T = -1/|\mathcal{D}_{lb}| \sum_i y_i \log p_i^{\text{temporal}}. \quad (4)$$

4 Experiments

4.1 Experimental Setup

Datasets and prediction tasks. We conduct experiments on *Chicago Dataset*[2] and *Brasilia Dataset*[3]. They collect mobility data in two cities in the United States and Brazil. We remove visits outside of the designated city and filter out users whose mobility records span less than 12 hours. For

[2]https://datahub.cmap.illinois.gov/dataset/traveltracker0708
[3]https://metro.df.gov.br/?page_id=47685

the Chicago dataset, we predict employment status (i.e., employed or not) and ethnicity (i.e., Caucasian, African American, or others). For the Brasilia dataset, we predict employment status, education level (i.e., whether the person has a college degree), and age group (i.e., ≤ 17, 18-59, or ≥ 60).

Implementation details. The dimension of spatial features $d = 64$. The filter sizes in the temporal module are $[3, 5, 12]$. The convolution layers in the spatial module output 32 channels. The maximum number of iterations is 5. All experiments are repeated 5 times with a fixed set of random seeds.

Metrics. Due to label imbalance, we adopt macro- and micro-F1 scores to evaluate the performance. For all compared methods, we rank the prediction probabilities of the test data and assign classes according to label distribution [Meng *et al.*, 2017; Yuan *et al.*, 2018]. The label distributions are estimated by randomly sampling $N_{est} = 100$ pieces of data from the training set and calculating the ratio of each class. We show STCOLAB is robust to N_{est} via sensitivity analysis.

4.2 Compared Methods

We compare STCOLAB with the state-of-the-art methods designed for demographic inference from mobility data and for general-purpose spatio-temporal tasks.

- **L2P** [Zhong *et al.*, 2015] is a tensor factorization-based method for demographic inference from location check-ins. It extracts spatial and temporal semantics from check-ins and mines location knowledge from social networks and customer reviews. User representations are obtained by tensor factorization and are used to train classifiers.
- **SUME** [Xu *et al.*, 2020] is an embedding-based method that learns mobility patterns by modeling a heterogeneous network that describes relations among users and locations. SVM is adopted for classification with learned embeddings.
- **Transformer** [Vaswani *et al.*, 2017] is a neural network model which learns temporal patterns from sequential data and utilizes multi-head attention mechanism to select important inputs. We organize the data of each user into a sequence of location IDs, showing the main locations where the person stays during each hour of the day.
- **ConvLSTM** [Shi *et al.*, 2015] is a recurrent neural network for spatio-temporal prediction. It receives a sequence of visualized maps as input, uses the convolutional networks to extract features from the maps, and feeds the features into the LSTM networks in chronological order.

We also craft two strong baselines by using some modules in STCOLAB. For a fair comparison with the state-of-the-art methods, the modules in both baselines are combined through late fusion by taking the average of the outputs as the final prediction and are updated together in back-propagation.

- **CNN+Transformer** uses convolutional networks for the spatial module (same as STCOLAB) and Transformer for the temporal module to learn patterns from location IDs.
- **CNN+PeriodicCNN (our ablation)** is equipped with the same spatial and temporal modules in STCOLAB.

Additionally, we use the same spatial and temporal modules in STCOLAB and compare different ways to combine them.

- **STFC** adopts intermediate fusion: the last hidden outputs of the two modules are concatenated and an FC layer is applied toward the concatenated features to make predictions.
- **S+T** adopts the late fusion which takes the average of the outputs given by the two modules as the final prediction.
- **ST (our ablation)** follows the alternating training in STCOLAB and is trained for only one iteration.

We denote our proposed method as **STCOLAB**. We compare STCOLAB with a state-of-the-art self-training method by pairing it with the same base model as in STCOLAB.

- **ST w/ CL** adopts curriculum labeling (CL) [Cascante-Bonilla *et al.*, 2021]. It uses a self-paced curriculum and re-initializes the model at each round to avoid concept drift.

4.3 Main Experimental Results and Analysis

The results are shown in Table 1. Overall, STCOLAB performs the best compared to all the baseline models. The existing methods for demographic inference from mobility data and for general-purpose spatial-temporal tasks show inferior performance in the minimally-supervised setting. The performance of L2P, SUME, and Transformer indicates the limitation of these methods in capturing the geographic distribution. By comparison, the models which learn from visualized maps (i.e., ConvLSTM, CNN+Transformer, and CNN+PeriodicCNN) show better results.

The comparison among STFC, S+T and ST shows the advantage of alternating training over using conventional ways of modal fusion. Simply combining the predictions or intermediate hidden features of the two modules together does not fully leverage their respective strengths and may even yield worse results than using either module alone.

Finally, the comparison among ST, ST w/ CL, and STCOLAB shows the effectiveness of iterative collaborative distillation. Applying CL does not guarantee improvement and even causes performance degradation on some tasks. This suggests that, in the minimally-supervised setting, iteratively guiding the model with pseudo labels generated by the same model, hence the same view, may harm model performance. By comparison, STCOLAB provides guidance from two different views, which lets the two modules give complementary supervision for each other and mutually enhance themselves.

4.4 Ablation Studies and Sensitivity Analysis

More comparisons with state-of-the-art self-training method. We further compare STCOLAB with ST w/ CL by applying them to three different model architectures. (1) **CNN & Transformer** uses convolutional networks as the spatial module (same as STCOLAB) and uses Transformer as the temporal module to learn temporal patterns from location IDs. The average of the predictions given by the two modules is taken as the final results. (2) **CNN & LSTM** uses convolutional networks to process the visualized maps and uses LSTM as the temporal module to process the spatial features of each hour generated by the convolutional networks in chronological order. (3) **CNN & PeriodicCNN** is the architecture in STCOLAB. As shown in Figure 5, STCOLAB is robust to different model architectures and always brings improvement to the vanilla models, while applying CL to the models may lead to worse performance.

Method	Comment	Chicago				Brasilia					
		Employment		Ethnicity		Employment		Education		Age	
		Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1
L2P [Zhong *et al.*, 2015]	baseline models	49.70	61.40	25.80	54.50	49.70	69.10	49.20	50.30	31.12	55.54
SUME [Xu *et al.*, 2020]	for demographic	50.40	59.30	37.20	46.10	41.00	70.00	51.20	52.50	30.85	54.71
Transformer [Vaswani *et al.*, 2017]	inference and	50.42	62.51	30.20	40.89	50.36	67.79	50.05	51.11	28.14	51.43
ConvLSTM [Shi *et al.*, 2015]	general-purpose	53.10	64.10	48.37	59.54	52.87	69.45	50.72	52.16	29.06	53.55
CNN+Transformer	spatio-temporal	53.30	64.71	53.67	65.28	50.56	67.97	50.39	51.48	29.91	50.53
CNN+PeriodicCNN*	tasks	54.82	65.82	54.46	66.71	52.54	69.24	48.83	49.93	29.20	52.14
S+T	late fusion	54.82	65.82	54.46	66.71	52.54	69.24	48.83	49.93	29.20	52.14
STFC	intermediate fusion	53.34	64.67	54.06	65.93	51.76	68.73	51.48	52.54	30.48	52.10
ST*	alternating training	55.68	65.61	56.08	68.97	52.43	69.13	52.97	54.06	30.75	54.52
ST w/ CL [Cascante-Bonilla *et al.*, 2021]	w/ self-training	54.85	65.79	54.66	65.92	54.76	70.62	49.98	51.07	33.25	55.46
STCOLAB		**56.47**	**67.02**	**58.87**	**75.27**	**54.77**	**70.67**	**59.34**	**60.21**	**35.64**	**59.37**

Table 1: Experimental results averaged over 5 runs. The first section of the table compares different neural architectures. The second section focuses on different fusion solutions. The third section shows different self-training methods. We use * to mark the ablations of STCOLAB.

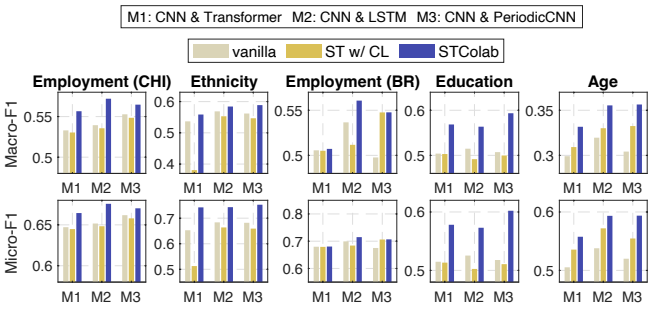


Figure 5: Performance of STCOLAB and ST w/ CL applied to different model architectures. Base models without applying self-training strategies are denoted as vanilla. (CHI: Chicago, BR: Brasilia)

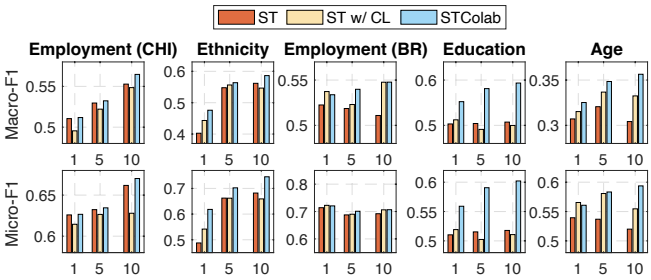


Figure 6: Performance of ST, ST w/ CL and STCOLAB w.r.t the number of labeled samples per class. (CHI: Chicago, BR: Brasilia)

Key designs in iterative collaborative distillation. We examine three ablations of STCOLAB. (1) **STCOLAB w/o vote** removes the voting ensemble strategy and uses all unlabeled samples as the distillation set. (2) **STCOLAB w/ union** uses the union of the prediction probabilities from both the latest trained spatial and temporal models to distill knowledge at every distillation process. (3) **STCOLAB w/o balance** removes the upsampling for getting a balanced distillation dataset and uses the dataset selected by the voting ensemble directly. The results are shown in Table 2. We notice performance degradation after replacing the key designs with the ablations, which indicates the importance of these designs. The voting ensemble helps the model choose the appropriate distillation set, giving better guidance during distillation. Moreover, by letting two modules alternate as teachers, the student model acquires complementary knowledge from the other modality and avoids confirmation bias. Furthermore, the balancing strategy ensures the number of samples for each class is the same and does not affect learning.

Contributions of spatial and temporal modules. We conduct experiments on the model with spatial module only (denoted as **Spatial**), and the model with temporal module only which is trained from scratch including the spatial feature extractor (denoted as **Temporal**). The models are trained for only one pass without knowledge distillation. We also examine the single modules with knowledge distillation. We use the latest trained model at the previous iteration to construct the distillation dataset based on percentile score and to distill knowledge at the current iteration. **SCOLAB** and **TCOLAB** are the ablations of STCOLAB with only the temporal module and only the spatial module respectively. As shown in Table 2, removing either module will cause performance degradation. The performance of ST is slightly better than that of Temporal. This indicates the spatial features extracted by the pretrained spatial model are label-indicative, which improves the inference ability. By comparing SCOLAB, TCOLAB, and STCOLAB, we observe that knowledge distillation with a single module does not guarantee better performance. This again demonstrates the importance of the proposed collaborative distillation strategies for combining the modalities.

Number of labeled samples. We evaluate the system performance with even less label information. To do so, we decrease the number of training samples per class k and compare the performance of ST, ST w/ CL, and STCOLAB. The results are shown in Figure 6. In general, STCOLAB outperforms the vanilla model and the model applying CL.

Robustness of class distribution estimation. The parameter N_{est} is used when randomly selecting training samples for

Method	Chicago				Brasilia					
	Employment		Ethnicity		Employment		Education		Age	
	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1
Spatial	54.56	65.62	55.33	68.50	50.42	67.86	49.85	50.93	29.69	52.58
Temporal	53.00	64.51	53.03	64.92	51.64	68.76	52.95	53.98	30.13	53.03
ST	55.68	65.61	56.08	68.97	52.43	69.13	52.97	54.06	30.75	54.52
SCOLAB	54.10	65.31	56.28	70.47	50.57	67.92	56.39	57.32	33.63	56.43
TCOLAB	53.75	65.07	52.38	68.11	52.62	68.54	58.93	59.85	35.00	57.61
STColab w/o vote	56.08	66.77	57.35	70.21	51.67	68.63	59.18	60.05	32.99	55.02
STCOLAB w/ union	55.92	66.64	57.84	71.74	51.70	68.71	58.61	59.50	33.50	56.92
STCOLAB w/o balance	56.00	66.68	56.06	73.01	52.67	69.32	50.43	51.50	28.94	51.91
STCOLAB	**56.47**	**67.02**	**58.87**	**75.27**	**54.77**	**70.67**	**59.34**	**60.21**	**35.64**	**59.37**

Table 2: Ablation studies on the contributions of the key designs in iterative collaborative distillation and the spatial and temporal modules.

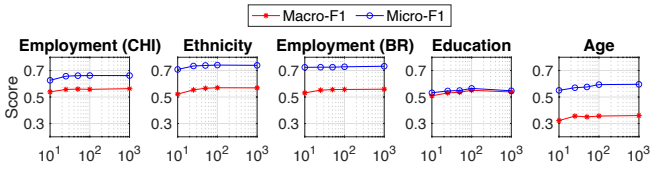


Figure 7: Performance of STCOLAB w.r.t number of sampled data N_{est} for class distribution estimation. (CHI: Chicago, BR: Brasilia)

estimating the class ratio. To test the robustness of STCOLAB with respect to this parameter, we change N_{est} to different values. As shown in Figure 7, the differences in performance are small when increasing or decreasing N_{est}, which shows that STCOLAB framework is robust to N_{est}.

5 Related Work

Contextual inference from human mobility. Our work studies one of the important lines in contextual inference—inferring the demographic attributes of mobile users. The demographic inference problem has been studied with the support of abundant behavioral data from various fields, such as web and social media activities [Bi *et al.*, 2013; Culotta *et al.*, 2015; Wang *et al.*, 2019], transactions [Wang *et al.*, 2016; Kim *et al.*, 2019] and ratings [Shang *et al.*, 2018]. Mobile data, which is ubiquitous in life, has been proven to have correlations with people's demographics [Luo *et al.*, 2016; Zhang *et al.*, 2016]. Several methods have been proposed for demographic inference from human mobility including tensor factorization-based methods, [Zhong *et al.*, 2015; Montasser and Kifer, 2017] and network embedding-based method [Xu *et al.*, 2020]. These studies typically require a large number (e.g., thousands) of users to share labels for model training. By contrast, we seek to develop a data-efficient method that can achieve meaningful results with a very small amount of annotated data. To the best of our knowledge, we are the first to address the contextual inference problem using mobility data with minimal supervision.

Minimally supervised classification. Minimally supervised classification is an extreme case of semi-supervised learning. It is a challenging problem due to the scarcity of labeled training data. The problem has recently received much attention [Zhang *et al.*, 2020; Zhang *et al.*, 2021]. A similar setting is few-shot learning which focuses on learning from limited data samples. The key difference between the two settings is that few-shot learning does not consider the availability of unlabeled data. Common solutions for few-shot classification such as metric learning [Koch *et al.*, 2015] and meta-learning [Finn *et al.*, 2017] are not ideal for our scenario, as they do not utilize underlying data distribution from massive unlabeled data. Self-training is arguably the most popular semi-supervised method for mitigating label scarcity. Self-training strategies [Cascante-Bonilla *et al.*, 2021; Li *et al.*, 2019; Shi *et al.*, 2018] achieve remarkable results by iteratively utilizing the predictions of unlabeled data from previous rounds to augment the training set and support subsequent rounds of training. Mobility data usually consists of information from two different modalities. Traditional self-training solutions that bootstrap a single model using all features at once would incur a high risk of overfitting. Different from traditional methods, we propose collaborative distillation between the spatial and temporal modules in self-training cycles, which fuses information and improves generalization.

6 Conclusions and Future Work

We studied the problem of contextual inference in the context of minimal supervision. We proposed STCOLAB framework and demonstrated its capability in predicting demographic attributes from human mobility. STCOLAB learns mobility features from spatial and temporal information alternatingly and adopts iterative collaborative distillation to enhance model generalization in self-training cycles. STCOLAB achieves reasonable accuracy with only a small amount of annotated data (i.e., 10 samples per class), outperforming the state-of-the-art methods. Our future work is on extending our framework to mobility data of different granularities, such as fine-grained GPS trajectories, to support more applications. Specifically, we plan to incorporate additional spatial features according to data specificities, such as points of interest, to enrich the location contexts. Additionally, the network architecture of each modality can be adapted accordingly to accommodate different formats of mobility data.

Ethical Statement

We strictly follow the Term-of-Use statements of the datasets when conducting analysis and experiments. The identity information in the datasets is properly anonymized by the data owners before they make the data public. We recognize the potential privacy concerns when inferring demographic information. We believe exploring the feasibility of inferring demographics with a limited number of labels is a first step towards privacy-preserving data mining. The good inference performance motivates us to develop privacy-preserving approaches to user modeling with spatio-temporal data in future work. We are committed to preventing misuse of the model to predict sensitive information. To this end, we release the code for research use with an ethical license[4] and will not share the pretrained model to prevent misuse.

Acknowledgments

This work was sponsored in part by NSF Convergence Accelerator under award OIA-2040727, NIH Bridge2AI Center Program under award 1U54HG012510-01, as well as generous gifts from Google, Adobe, and Teradata. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes not withstanding any copyright annotation hereon.

References

[Arazo *et al.*, 2020] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *2020 International Joint Conference on Neural Networks*, pages 1–8, 2020.

[Bi *et al.*, 2013] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. Inferring the Demographics of Search Users: Social Data Meets Search Queries. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 131–140, 2013.

[Cascante-Bonilla *et al.*, 2021] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6912–6920, 2021.

[Culotta *et al.*, 2015] Aron Culotta, Nirmal Kumar, and Jennifer Cutler. Predicting the Demographics of Twitter Users from Website Traffic Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[Fang *et al.*, 2021] Zhihan Fang, Yu Yang, Guang Yang, Yikuan Xian, Fan Zhang, and Desheng Zhang. CellSense: Human Mobility Recovery via Cellular Network Data Enhancement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–22, 2021.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*, pages 1126–1135, 2017.

[Gat *et al.*, 2020] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies. *Advances in Neural Information Processing Systems*, 33:3197–3208, 2020.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[Jordahl *et al.*, 2020] Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, Matthew Perry, Adrian Garcia Badaracco, Carson Farmer, et al. geopandas/geopandas: v0. 8.1. *Zenodo*, 2020.

[Kim *et al.*, 2019] Raehyun Kim, Hyunjae Kim, Janghyuk Lee, and Jaewoo Kang. Predicting Multiple Demographic Attributes with Task Specific Embedding Transformation and Attention Network. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 765–773, 2019.

[Koch *et al.*, 2015] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese Neural Networks For One-Shot Image Recognition. In *ICML deep learning workshop*, volume 2, 2015.

[Lee and others, 2013] Dong-Hyun Lee et al. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 896, 2013.

[Li *et al.*, 2019] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to Self-Train for Semi-Supervised Few-Shot Classification. *Advances in Neural Information Processing Systems*, 32:10276–10286, 2019.

[Liu *et al.*, 2017] Yanchi Liu, Chuanren Liu, Xinjiang Lu, Mingfei Teng, Hengshu Zhu, and Hui Xiong. Point-of-Interest Demand Modeling with Human Mobility Patterns. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 947–955, 2017.

[Luo *et al.*, 2016] Feixiong Luo, Guofeng Cao, Kevin Mulligan, and Xiang Li. Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography*, 70:11–25, 2016.

[Meng *et al.*, 2017] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. Deep Keyphrase Generation. *arXiv preprint arXiv:1704.06879*, 2017.

[4] https://firstdonoharm.dev

[Montasser and Kifer, 2017] Omar Montasser and Daniel Kifer. Predicting Demographics of High-Resolution Geographies with Geotagged Tweets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Quercia *et al.*, 2010] Daniele Quercia, Neal Lathia, Francesco Calabrese, Giusy Di Lorenzo, and Jon Crowcroft. Recommending Social Events from Mobile Phone Location Data. In *2010 IEEE International Conference on Data Mining*, pages 971–976, 2010.

[Schein *et al.*, 2002] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and Metrics for Cold-Start Recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, 2002.

[Shang *et al.*, 2018] Jin Shang, Mingxuan Sun, and Kevyn Collins-Thompson. Demographic Inference Via Knowledge Transfer in Cross-Domain Recommender Systems. In *2018 IEEE International Conference on Data Mining*, pages 1218–1223, 2018.

[Shen *et al.*, 2018] Ting Shen, Haiquan Chen, and Wei-Shinn Ku. Time-aware Location Sequence Recommendation for Cold-start Mobile Users. In *Proceedings of 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 484–487, 2018.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, 28, 2015.

[Shi *et al.*, 2018] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive Semi-Supervised Deep Learning using Min-Max Features. In *Proceedings of the European Conference on Computer Vision*, pages 299–315, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2016] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. Your Cart tells You: Inferring Demographic Attributes from Purchase Data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 173–182, 2016.

[Wang *et al.*, 2017] Pinghui Wang, Feiyang Sun, Di Wang, Jing Tao, Xiaohong Guan, and Albert Bifet. Inferring Demographics and Social Networks of Mobile Device Users on Campus From AP-Trajectories. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 139–147, 2017.

[Wang *et al.*, 2019] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. In *Proceedings of the 2019 World Wide Web Conference*, pages 2056–2067, 2019.

[Wang *et al.*, 2020a] Daheng Wang, Meng Jiang, Munira Syed, Oliver Conway, Vishal Juneja, Sriram Subramanian, and Nitesh V. Chawla. Calendar graph neural networks for modeling time structures in spatiotemporal user behaviors. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2581–2589, 2020.

[Wang *et al.*, 2020b] Qinyong Wang, Hongzhi Yin, Tong Chen, Zi Huang, Hao Wang, Yanchang Zhao, and Nguyen Quoc Viet Hung. Next Point-of-Interest Recommendation on Resource-Constrained Mobile Devices. In *Proceedings of the Web Conference*, pages 906–916, 2020.

[Xu *et al.*, 2020] Fengli Xu, Zongyu Lin, Tong Xia, Diansheng Guo, and Yong Li. SUME: Semantic-enhanced Urban Mobility Network Embedding for User Demographic Inference. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–25, 2020.

[Yuan *et al.*, 2018] Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases. *arXiv preprint arXiv:1810.05241*, 2018.

[Zhang *et al.*, 2016] Ke Zhang, Yu-Ru Lin, and Konstantinos Pelechrinis. EigenTransitions with Hypothesis Testing: The Anatomy of Urban Mobility. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, 2016.

[Zhang *et al.*, 2020] Yu Zhang, Yu Meng, Jiaxin Huang, Frank F Xu, Xuan Wang, and Jiawei Han. Minimally Supervised Categorization of Text with Metadata. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1231–1240, 2020.

[Zhang *et al.*, 2021] Xinyang Zhang, Chenwei Zhang, Xin Luna Dong, Jingbo Shang, and Jiawei Han. Minimally-Supervised Structure-Rich Text Categorization via Learning on Text-Rich Networks. In *Proceedings of the Web Conference*, pages 3258–3268, 2021.

[Zhong *et al.*, 2015] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 295–304, 2015.