International Journal of Data Science and Analytics manuscript No.
(will be inserted by the editor)

Predicting Individual Socioeconomic Status from Mobile Phone Data: A Semi-supervised Hypergraph-based Factor Graph Approach

Tao Zhao* · Hong Huang · Xiaoming Yao · Jar-der Luo · Xiaoming Fu

Received: date / Accepted: date

Abstract Socioeconomic Status (SES) is an important economic and social aspect widely concerned. Assessing individual SES can assist related organizations in making a variety of policy decisions. Traditional approach suffers from the extremely high cost in collecting large-scale SES-related survey data. With the ubiquity of smart phones, mobile phone data has become a novel data source for predicting individual SES with low cost. However, the task of predicting individual SES on mobile phone data also proposes some new challenges, including sparse individual records, scarce explicit relationships and limited labeled samples, unconcerned in prior work restricted to regional or household-oriented SES prediction. To address these issues, we propose a semi-supervised Hypergraph-based Factor Graph Model (HyperFGM) for individual SES prediction. HyperFGM is able to efficiently capture the associations between SES and individual mobile phone records to handle the individual record sparsity. For the scarce explicit relationships, HyperFGM models implicit high-order relationships among users on the hypergraph structure. Besides, HyperFGM explores the limited labeled data and unlabeled data in a semi-supervised way. Experimental results show that HyperFGM greatly outperforms the baseline methods on a set of anonymized real mobile phone data for individual SES prediction.

Keywords Socioeconomic status · Mobile phone data · Hypergraph · Factor graph

1 Introduction

Socioeconomic Status (SES) is an indicator that measures an individual, a household or a region's economic and social position in relation to others, which is typically divided into three levels (high, middle, and low)[1]. The rich information carried by SES not only helps governments and research institutes study and make public policies, but also assists in meeting the needs of target clients by evaluating their purchasing power from a commercial perspective. Furthermore, SES can benefit a wide range of other fields, such as health [17,29], education [21] and public transportation [5]. National statistical offices measure socioeconomic information typically by a large number of personal/household interviews. However, assessing SES for a whole country/region's population by this method is extremely expensive and time-consuming (e.g., usually once every 5 to 10 years). It is critical to develop a low-cost means for timely capturing and assessing individual SES in a population.

Due to the worldwide ubiquity of smart phones, mobile phone data captures abundant information regarding personal social attributes, relation networks and mobility patterns in

This work is partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 824019, by the National Natural Science Foundation of China (No. 61802140), and by the Hubei Provincial Natural Science Foundation (No. 2018CFB200).

Tao Zhao*
Institute of Computer Science, University of Goettingen, Germany
E-mail: tao.zhao@cs.uni-goettingen.de

Hong Huang
School of Computer Science, Huazhong University of Science and Technology, China
E-mail: honghuang@hust.edu.cn

Xiaoming Yao
China TelecomCo., Ltd., China
E-mail: yaoxm@chinatelecom.cn

Jar-Der Luo
Department of Sociology, Tsinghua University, China
E-mail: jdluo@tsinghua.edu.cn

Xiaoming Fu
Institute of Computer Science, University of Goettingen, Germany
E-mail: fu@cs.uni-goettingen.de

[1] https://en.wikipedia.org/wiki/Socioeconomic_status

a large-scale population, which to some extent reflects SES. Hence, mobile phone data has been used as a novel data source for efficiently inferring SES with low cost. Some efforts have been made to infer regional or household SES from mobile phone data by directly applying classic supervised machine learning methods [4,10,23]. Different from prior work that concentrates on aggregated records of a region/household, we are motivated to study the SES prediction on mobile phone data at an individual level, the first trial in the community as far as we know. Intuitively, even living in the same household, several individuals probably share different SES levels. Inferring the individual SES provides the finest level of evidence and indication to improve the quality of corresponding public policies-making. Furthermore, it can enable numerous fine-grained applications at an individual level, such as precision marketing, fine service and assessment. However, individual SES prediction on mobile phone data proposes several main challenges:

- **Sparse individual records.** Compared with aggregated records of a region/household, a large portion of individual mobile phone users actually generate sparse valid usage records every day. With the ubiquity of WiFi, individual records that telco service providers can identify are becoming rarer. For example, 71.9% users generate less than two valid daily records in the data provided by an ISP in China. It is difficult to explore enough information from sparse individual records for revealing personal SES as done in the existing SES prediction work, thus causing poor prediction performance.
- **Scarce explicit relationships.** Due to the increasing popularity of mobile communication applications like WhatsApp and Wechat, an increasing number of mobile phone users are giving up traditional voice calling and SMS services[2]. Subsequently, the communication relationships built in these mobile applications are disconnected from ISP-provided mobile phone data. Therefore, explicit relationships among users extracted from mobile phone records become scarce, which makes the methods based on such relationships failed to work.
- **Limited labeled samples.** Since the cost of assessing individual SES by existing methods is extremely high, it is rather difficult to obtain enough SES-labeled samples for learning models. To the best of our knowledge, prior work on the SES prediction only employ supervised learning methods to predict SES, which does not work well on data with limited labeled samples.

To simultaneously address the above challenges for enabling individual SES prediction from mobile phone data, we propose a novel semi-supervised probabilistic model called Hypergraph-based Factor Graph Model (HyperFGM). First, to reduce the performance loss caused by the individual record sparsity, leveraging the idea of factor graph model, HyperFGM utilizes customized factor functions to efficiently capture the correlations between SES and numerous attributes of users extracted from individual mobile phone records, which significantly exploits the power of sparse records compared with the prior methods on SES prediction. Second, to address the explicit relationship scarcity problem, HyperFGM leverages the advantage of hypergraph on high-order relationship modeling to model implicit high-order relationships among users based on the hypergraph structure, which avoids the performance loss caused by ignoring the implicit high-order relationships. Third, for handling the limited labeled samples, HyperFGM explores both labeled and unlabeled data on a hypergraph network in a semi-supervised way, thereby achieving better performance than supervised learning methods in prior SES prediction work.

Furthermore, compared with our proposed hypergraph-based factor graph model, traditional hypergraph-based models [7,20,33], focusing on the relationships among objects, need to convert the numerous attributes of objects into various relationships among objects, causing conversion loss. Traditional factor graph models [25,27,30] only consider objects' attributes and explicit pair-wise relationships between objects in a simple graph, which ignore implicit and high-order relationships among objects. However, there actually exist many complex high-order relationships among objects [33]. Therefore, in order to solve the disadvantages of these two traditional methods, HyperFGM, combining hypergraph-based model and factor graph model into one model, predicts individual SES by not only directly considering the SES-related attributes of users but also modeling the implicit high-order mobility pattern-based relationships among users in the hypergraph structure.

We demonstrate the feasibility and power of HyperFGM on individual SES prediction using a set of anonymized real mobile phone data collected from a major ISP in China. Experimental results indicate that HyperFGM outperforms previous work on SES prediction by 5-22% w.r.t. the F1-score and provides a considerable improvement (2-9%) compared with the state-of-the-art hypergraph-based methods and factor graph methods. It is worth to note that the proposed HyperFGM is a general semi-supervised classification method, which can be applied not only to the SES prediction problem but also to other similar tasks.

Our major contributions in this paper are summarized as follows.

- We first identify the issue of predicting individual SES from mobile phone data. To our knowledge, no previous work has extensively studied this issue.
- We propose a semi-supervised probabilistic hypergraph model, HyperFGM, to solve the SES prediction problem, which jointly considers user attributes and high-

[2] http://uk.businessinsider.com/

order relationships among users based on the hypergraph structure.

- We apply our model on a collection of anonymized real mobile phone data. Experimental results show that HyperFGM outperforms the baseline models.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 shows the data collection. Section 4 describes the proposed HyperFGM model. Section 5 evaluates the prediction performance of HyperFGM with extensive experiments. Finally, Section 6 concludes the paper.

2 Related Work

SES prediction on mobile phone data emerges as a very recent application of artificial intelligence (AI) for social and economic good. One direction is to investigate the relation between regional economic development and mobile phone usage. [22] analyzed cell phone data from two countries to extract a set of important features correlated with poverty indexes. [15] defined several indicators of mobile phone usage to analyze their correlations with economic status indicators. Other efforts are on applying classic supervised machine learning techniques to predict regional or household SES. [23] applied SVM and Random Forest on the aggregated cell phone records to predict regional SES. [10] introduced a supervised LDA-based topic modeling approach to inferring regional SES from large-scale spatio-temporal calling data. [4] developed a deterministic finite automation (DFA)-based method to generate a large number of features and relied on a linear regression method (elastic net) to predict the SES of each household in Rwanda on mobile phone data. However, these classic supervised learning methods cannot solve three challenges mentioned above, would lead to poor performance in predicting individual SES from mobile phone data.

Factor graph based models, as a specific type of graphical models, have been widely applied in many areas, such as social network modeling, disease forecasting and medical informatics. [25] proposed a partially-labeled pairwise factor graph model by considering pairwise relations and attribute factors to infer social tie. [30] proposed a sparse factor graph model to forecast potential diabetes complications. [27] presented a probabilistic-graphical model to infer characteristics of instantaneous brain activities by jointly analyzing spatial, temporal and observational relationships in electroencephalograms. However, these traditional factor graph models are unable to exploit high-order relationships among objects. To formulate the complex relationships among objects beyond pairwise relationship, hypergraph learning has obtained some interest recently. [33] extended spectral clustering methods from graphs to hypergraphs and further proposed a transductive learning model. [12] proposed to employ the hypergraph structure to formulate the relationship among images. [7] proposed to employ the weighted multiple hypergraphs to formulate the higher order relationships among objects. [20] modeled multi-way relations as hypergraphs and extended the discriminative random walk (DRW) framework, originally proposed for transductive inference on single graphs, to the case of multiple hypergraphs. These hypergraph learning methods focus on the relationship called hyperedge and need to convert the attributes of objects into various relationships among objects, which causes some conversion loss. To our knowledge, there is no effort on directly considering the attributes of objects to well exploit the dependencies between a large number of real-valued attributes of objects and label.

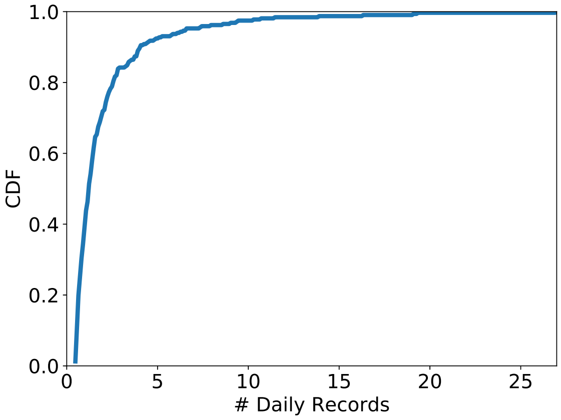

Fig. 1: Distribution of average daily record count of each user.

3 Data Collection

Before presenting our proposed model, we first describe the data used in our work. The data was provided by our collaborator, a major ISP in China. We got the anonymized mobile phone's Internet records of 317 active mobile phone users (each user generated more than 50 records in a given period) who agree to provide their personal SES-related information, including occupation, education, income for this research. The data contains these users' Internet records from October 31, 2016 to February 13, 2017 from the city of Shanghai, one of the largest cities in China. For the user privacy and ISP's privacy agreement, the data can be only used for our research. Each user averagely generated about 219 valid records in the period. As shown in Figure 1, most of users generate very sparse daily records. The key data statistics are summarized in Table 1.

In our dataset, each user has an Internet record sequence generated from his/her mobile phone during the given period. Each record provided by the ISP contains the anonymized

Table 1: Description of dataset

Location	Shanghai, China
Time duration	Oct.31, 2016-Feb.13, 2017
Number of users	317
Number of records	69621

userID, the occurred time and the Uniform Resource Locator (URL). A URL indicates the address of a resource on the Internet, specifically, the HTTP and HTTPS requests issued from user to the cellular towers. As shown in the previous studies [11, 14], a user's spatio-temporal mobility pattern is correlated with his/her socioeconomic status. Besides, in order to make our work applicable to traditional call detail records (CDRs) and other location-based data, our work mainly focus on spatio-temporal information, i.e., latitude-longitude pair and timestamp, to exploit the power of mobile phone records for predicting individual SES. Through analyzing the content of URL, we find that the URLs generated from location-based mobile applications mostly contain GPS location information. For example, if a URL is "http://www. example.com/...lat=i&lng=j...", then we can extract a latitude-longitude pair (i, j) from it. Through this extraction, we can obtain a set of spatio-temporal data from the raw data, which will be used for the SES prediction in our work.

In order to obtain the SES label for each user, a sociologist mapped users into three SES levels, which are high (level A), middle (level B), or low (level C) level, according to their personal SES-related information [2,9,19]. Finally, in our data, the resultant user distribution across classes is 70 users with Level A, 160 users with Level B and 87 users with Level C. Consequently, like most previous work [10, 13,23] on SES level prediction, our work regards the SES prediction as a classification problem. To be more specific, the aim of our work is to predict a SES label (high, middle or low) for each individual user as accurately as possible.

4 The HyperFGM Model

The purpose of our work is to predict individual SES based on mobile phone user's records, which proposes three main challenges. To address these challenges, in this section, we propose and introduce the proposed HyperFGM for the individual SES prediction.

- For the sparse individual records, we leverage factor graph model to efficiently capture the correlations between SES and mobile phone records, which can greatly enhance the performance compared with classic machine learning methods [30]. To this end, we first extract SES-related user attributes from sparse mobile phone data through employing a DFA-based method and a relief-based feature selection method. Then in the HyperFGM, we define attribute factor functions to represent the correlations between SES and each attributes.
- For the scarce explicit relationships, we first extract semantic mobility pattern similarity between users as implicit relationships, and then construct a hypergraph network structure among users based on the mobility pattern similarity to capture more implicit high-order relationships among users. Through defining hyperedge factor function in HyperFGM, we utilize these implicit high-order relationships among users to enhance the prediction performance.
- For the limited labeled samples, HyperFGM can explore both labeled and unlabeled data in a semi-supervised way. Specifically, the input data to our model is partially labeled so that the prediction model is learned by leveraging the labeled data and unlabeled data on the hypergraph network to infer the unknown label.

In this section, we first present the SES-related user attribute extraction from mobile phone's Internet records, and then propose a hypergraph construction method based on their semantic mobility patterns for exploring the implicit high-order relationships among users, as shown in Figure 2. Lastly, HyperFGM is conducted based on the user attributes and the hypergraph structure to infer the SES of each mobile phone user.

4.1 SES-related User Attribute Extraction

We first preprocess the obtained records for the user attribute extraction. The first step utilizes a stay point estimating method proposed by Ye et al. [31] to obtain the stay points of each user. A stay point represents a geographic region in which the user stays for a while, which carries its semantic meaning, such as home, working place and the spot the user traveled. The second step employs the Baidu Map API and a land price crawler to obtain semantic location information, including POI type, city level and land price.

The user attribute extraction transforms each user's mobile phone's Internet records into a set of SES-correlated attribute metrics. To this end, we first employ a DFA-based method [18] to generate a large number of potentially correlated attributes. In our work, the structured and combinatorial method automatically generates more than 400 attributes from the records. These user attributes are generated from different attribute spaces such as the record volume, movement distance, POI type, city level and land price. For each attribute space, we compute all possible attributes such as mean, maximum, minimum, standard deviation, sum, radius of gyration and count/fraction of unique values over time.

To eliminate irrelevant attributes, we utilize a relief-based feature selection method, MultiSURF* [8,26] to select SES-related user attributes according to the importance score ranking. In our work, top 20% attributes are selected as the final attribute input for the best prediction performance. As a result, for each user v_i, there is an associated attribute vector $\mathbf{x}_i$, in which each element denotes a user attribute.

4.2 Mobility Pattern-based Hypergraph Construction

In this part, we aim at generating the implicit high-order mobility pattern relationships among users, namely, high-order relationships among users on a hypergraph structure based on users' semantic mobility patterns. As mentioned above, users with the same SES are more likely to have similar mobility patterns. For instance, persons, who typically stay in office during the daytime of a workday and visit entertainment places on the weekend, might belong to the same SES. Inspired by this intuition, we first extract the semantic mobility pattern of each user by leveraging POI types and occurred time. A user's semantic mobility motifs can be defined as follows.

Definition 1. **Semantic Mobility Motifs.** A user v_i has a set of semantic spatio-temporal records $\{s_{i1}, s_{i2}, ..., s_{im}\}$. Each record is a tuple of $s = (t, p)$, which means that the user visited the POI type p at time t. Our work divides the time into workday/weekend and day/night so that a semantic mobility motif is defined as $smm = (w, d, p)$ if a user was at the POI p at time (w, d) where $w = 1$ if the time is in a workday otherwise 0; $d = 1$ if it is daytime otherwise 0. Hence, a user v_i's semantic mobility sequence is represented as $\mathbf{sm}_i = \{smm_{i1}, smm_{i2}, ..., smm_{im}\}$.

Given the defined semantic mobility motifs, we employ Latent Dirichlet Allocation (LDA) [3], a topic modeling method, to extract individual's semantic mobility patterns from mobility motifs. Each semantic mobility motif is regarded as a word and a user's semantic mobility sequence is treated as a document. As a result, each user's mobility pattern is represented as a topic distribution vector. Given two users' topic distribution vectors $\mathbf{m}_i, \mathbf{m}_j$, using a distance metric for probability distribution called *Jensen–Shannon Divergence* [6], the mobility pattern distance between each pair of users can be calculated as:

$$Mdistance(i, j) = \frac{1}{2}D_{KL}(\mathbf{m}_i||M)+\frac{1}{2}D_{KL}(\mathbf{m}_j||M) \quad (1)$$

where $M = \frac{1}{2}(\mathbf{m}_i + \mathbf{m}_j)$, D_{KL} is the *Kullback-Leibler Divergence* which defines the divergence from distribution $\mathbf{p}$ to $\mathbf{q}$ as: $D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_i \mathbf{p}(i) \log \frac{\mathbf{p}(i)}{\mathbf{q}(i)}$.

Based on the mobility pattern distance, we build a hypergraph structure $G = (V, E)$, where V represents a set of vertices (users), E is the hyperedge set such that for any hyperedge $e_i \in E, e_i \subseteq V$. Different from a simple graph that only contains pair-wise edges, the hypergraph is a graph where an edge called hyperedge can connect more than two vertices. Accordingly, to build the hypergraph, by using the star expansion strategy [12], we take each vertex as a centroid and generate a hyperedge for this vertex by connecting this centroid and its k-1 nearest neighbors. The strength of connectivity is determined by the mobility pattern distance between the centroid vertex and the other vertices. That is, each hyperedge connects k vertices. Following this construction method, we can choose different k (e.g., $k = 2, 3, 4, 5$) to generate different hyperedges in a hypergraph. Finally, the hypergraph can be represented by a $|V| \times |E|$ incidence matrix H:

$$h(v_i, e_j) = \begin{cases} 1, & \text{if } v_i \in e_j \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

We note that the employed methods of user attribute extraction and hypergraph construction are flexible and can be expanded/replaced by other methods.

4.3 Model Description

Our work focuses on investigating the prediction of individual SES through combining traditional hypergraph model and a probabilistic factor graph method into one model. Given the above constructed hypergraph, we define the input of our problem as a partially labeled hypergraph network. The hypergraph network is denoted as $G = (V^L, V^U, E, Y^L, \mathbf{X})$, where V^L is a set of labeled users (vertices) and V^U is a set of unlabeled users with $V^L \cup V^U = V$; E is a set of hyperedges; Y^L is a set of SES labels corresponding to the users in V^L. Let an attribute matrix $\mathbf{X} = \{\mathbf{x}_i\}$ which means each user v_i is associated with an attribute vector $\mathbf{x}_i$. Given the partially labeled hypergraph network, the goal of our work is to predict the labels (SES) of all SES-unknown users in the network, which is formulated as the following prediction problem.

Problem 1. **Individual Socioeconomic Status Prediction.** Given a hypergraph network $G = (V^L, V^U, E, Y^L, \mathbf{X})$, the objective is to learn a classification function:

$$f : G = (V^L, V^U, E, Y^L, \mathbf{X}) \to Y \quad (3)$$

As defined above, the input data is partially labeled. Therefore, to solve this problem, the HyperFGM model is learned in a semi-supervised way, i.e., exploring the labeled data as well as the unlabeled data on the hypergraph network to infer the unknown labels. Figure 2 shows the graphical representation of the HyperFGM model, where each user has a corresponding attribute vector $\mathbf{x}_i$ while the implicit complex relationships among users are exploited and represented on the hypergraph G. For example, y_1, y_2 and y_3 are connected by the hyperedge e_2. Furthermore, to efficiently model the power of the user attributes and the implicit high-order relationships among users, we define the following two kinds of factor functions respectively:

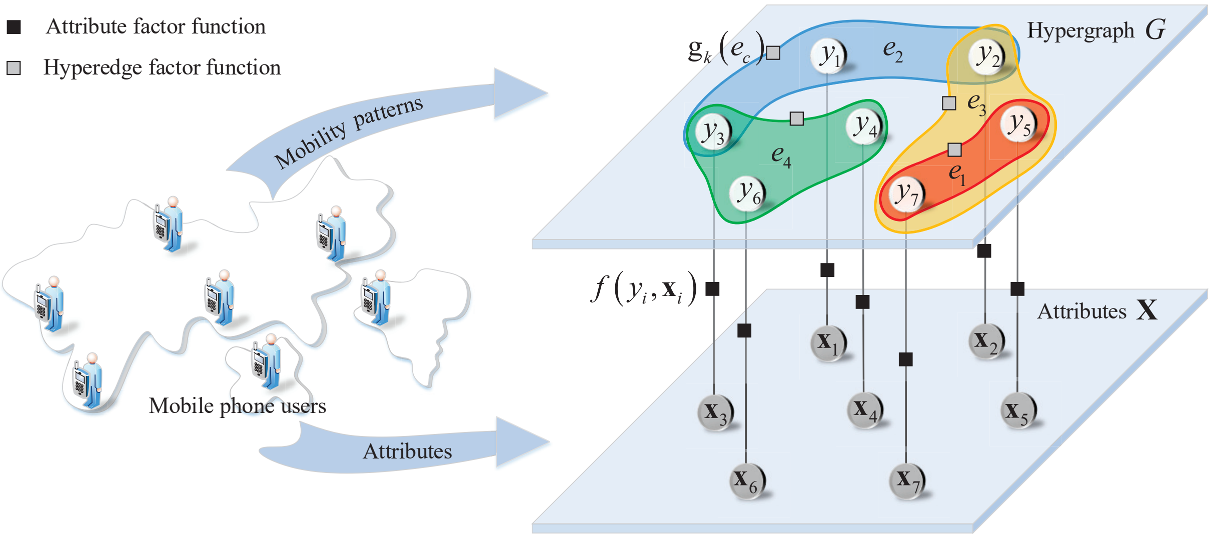

Fig. 2: Graphical representation of the HyperFGM model

- **Attribute factor:** $f(y_i, \mathbf{x}_i)$ (denoted as black rectangles in Figure 2) represents the correlation between y_i and its attribute vector $\mathbf{x}_i$.
- **Hyperedge factor:** $g_k(e_c)$ (denoted as gray rectangles in Figure 2) represents the complex correlation among users, where e_c denotes the c-th hyperedge in the hypergraph and k denotes the vertex number of the hyperedge.

According to the proposed model, given a partially labeled hypergraph network $G = (V^L, V^U, E, Y^L, \mathbf{X})$, we first define the posterior probability of $P(Y|\mathbf{X}, G)$ according to Bayes' theorem as follows:

$$\begin{aligned} P(Y|\mathbf{X}, G) &= \frac{P(\mathbf{X}, G|Y)P(Y)}{P(\mathbf{X}, G)} \\ &\propto P(\mathbf{X}|Y)P(Y|G) \\ &\propto (\prod_i P(\mathbf{x}_i|y_i))P(Y|G) \end{aligned} \tag{4}$$

We assume that the generative probability of user attributes given each user's label is conditionally independent, and the attributes and the network structure G are conditionally independent given labels Y. In Equation 4, $P(\mathbf{X}|Y)$ denotes the probability of generating the attributes $\mathbf{X}$ given their labels Y and $P(\mathbf{x}_i|y_i)$ is the probability of generating attributes $\mathbf{x}_i$ given the label y_i; $P(Y|G)$ indicates the labels' probability in a given hypergraph network structure G.

These two kinds of factors can be instantiated in different ways. In this work, we use exponential-linear functions. Accordingly, the probability of generating attributes $\mathbf{x}_i$ given the label y_i is instantiated as:

$$P(\mathbf{x}_i|y_i) = \frac{1}{Z_\alpha} exp\{\sum_{j=1}^{m} \alpha_j f_j(y_i, x_{ij})\} \tag{5}$$

where $f_j(y_j, x_{ij})$ denotes the attribute factor function of an attribute x_{ij} associated with user v_i; α_j is the weight of the attribute function f_j, and Z_α is a normalization factor. $f_j(y_i, x_{ij})$ can be defined as either a binary function or a real-valued function. Without losing generality, we define it as a real-valued function, e.g., the land price of the place that user v_i visited most frequently.

For the hyperedge factor function, we define it as a binary function based on the hypergraph network. For instance, if there is a 3-node hyperedge $e_4 = \{y_3, y_4, y_6\}$ among three users in Figure 2, then the value of the corresponding hyperedge factor function $g_3(e_4) = 1$; otherwise 0. Hyperedges in the network can be obtained from the incidence matrix H. We accumulate all hyperedge factor functions and obtain the probability of labels given the hypergraph as follows:

$$P(Y|G) = \frac{1}{Z_\beta} exp\{\sum_{e_c \in E} \sum_k \beta_k g_k(e_c)\} \tag{6}$$

where $g_k(e_c)$ denotes a hyperedge factor function of a hyperedge e_c which connects k nodes (vertices), and β_k is the weight of the k-node hyperedge factor function.

According to Equations 4-6, a hypergraph-based factor graph model is constructed as follows:

$$\begin{aligned} P(Y|\mathbf{X}, G) = & \frac{1}{Z} exp\{\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_j f_j(y_i, x_{ij}) \\ & + \sum_{e_c \in E} \sum_k \beta_k g_k(e_c)\} \end{aligned} \tag{7}$$

where $Z = Z_\alpha Z_\beta$ is a normalization factor; m denotes the length of the attribute vector $\mathbf{x}_i$; $n = |V|$ is the number of users.

The goal of learning the model is to estimate a parameter configuration $\theta = (\alpha, \beta)$, based on the input hypergraph structure and the attributes, to maximize the log-likelihood objective function $\mathcal{L}(\theta) = \log P_\theta((Y|\mathbf{X}, G)$, i.e.,

$$\begin{aligned}\theta^* &= \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_j f_j(y_i, x_{ij}) + \sum_{e_c \in E}\sum_{k} \beta_k g_k(e_c) \\ &\quad - \log Z \end{aligned} \tag{8}$$

Solution. We use a gradient descent method (or a Newton-Raphson method) to solve the objective function. The gradient for each parameter θ is calculated as:

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \alpha} &= \mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{m} f_j(y_i, x_{ij})] - \mathbb{E}_{P_\alpha(Y)}[\sum_{i=1}^{n}\sum_{j=1}^{m} f_j(y_i, x_{ij})] \\ \frac{\partial \mathcal{L}(\theta)}{\partial \beta} &= \mathbb{E}[\sum_{e_c \in E}\sum_{k} h_k(e_c)] - \mathbb{E}_{P_\beta(Y)}[\sum_{e_c \in E}\sum_{k} g_k(e_c)] \end{aligned} \tag{9}$$

where $\mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{m} f_j(y_i, x_{ij})]$ is the expectation of factor function $f_j(y_i, x_{ij})$ given the data distribution in the training data, and $\mathbb{E}_{P_\alpha(\mathbf{y})}[\sum_{i=1}^{n}\sum_{j=1}^{m} f_j(y_i, x_{ij})]$ is the expectation of factor function $f_j(y_i, x_{ij})$ under the distribution $P_\alpha(Y)$ (i.e., $P_\alpha(Y|\mathbf{X}, G)$) given by the estimated model. For the other equation, the expectation has the similar notations.

Algorithm 1: Learning algorithm for HyperFGM

```
Input: attribute matrix $\mathbf{X}$, hypergraph $G$, learning rate $\eta$
Output: estimated parameters $\theta$
Initialize $\theta \leftarrow 0$;
repeat
    Call LBP to calculate $\mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{m} f_j(y_i, x_{ij})]$ and
        $\mathbb{E}_{P_\alpha(Y)}[\sum_{i=1}^{n}\sum_{j=1}^{m} f_j(y_i, x_{ij})]$;
    Call LBP to calculate $\mathbb{E}[\sum_{e_c \in E}\sum_{k} g_k(e_c)]$ and
        $\mathbb{E}_{P_\beta(Y)}[\sum_{e_c \in E}\sum_{k} g_k(e_c)]$;
    Compute $\frac{\partial \mathcal{L}(\theta)}{\partial \alpha}$ and $\frac{\partial \mathcal{L}(\theta)}{\partial \beta}$ according to Equation 9;
    Update the parameter $\theta$ with the learning rate $\eta$:
        $\alpha_{new} = \alpha_{old} - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \alpha}$
        $\beta_{new} = \beta_{old} - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \beta}$
until Convergence;;
```

As shown in Algorithm 1, to solve the intractable problem of calculating the marginal distributions (e.g., $P_\alpha(Y)$), which is caused by the arbitrariness and the possible cycles of the graphical structure in the HyperFGM, we adopt Loopy Belief Propagation (LBP) [16] to calculate the marginal probability of Y and all hyperedges E such that the gradient for each parameter can be calculated. Then, with the gradient, we update α and β with a learning rate η. With the learned parameters, we can predict the label of unknown users Y^U by finding a label configuration which maximizes the objective function, i.e., $Y^* = \operatorname{argmax} P(Y|\mathbf{X}, G)$. We need to utilize LBP to compute the marginal probability of each user $P(y_i|\mathbf{x}_i, G)$ again and then assign each user the label with the maximal marginal probability. Please notice that the proposed HyperFGM is a general framework, which can be utilized to other similar tasks with appropriate definitions of factor functions and their hypergraphs.

Finally, we present a case study to further demonstrate the proposed model. As shown in Figure 2, each user v_i has an attribute vector $\mathbf{x}_i$, containing SES-related attributes, and has its own mobility pattern $\mathbf{m}_i$ extracted from its mobility motifs. With LDA, each user's mobility pattern is represented as a probability distribution over some latent topics, while each topic is represented as a probability distribution over a number of mobility motifs. Then, a hypergraph is constructed based on each user's mobility pattern. For example, user v_1 has an attribute vector $\mathbf{x}_1$ and has a hyperedge e_2 to connect with v_2 and v_3, which means they have similar mobility patterns. The SES label y_1 of the user may be known or unknown according to the actual case. Next, the attribute factor and hyperedge factor are used to capture the correlations between SES and attributes and the mobility pattern relationships among users respectively. Based on Algorithm 1, the labeled and unlabeled users can be used to infer these unknown label on the hypergraph network.

5 Evaluation

In this section, we apply the proposed HyperFGM to a real-life data for predicting individual SES levels. We first describe the experimental setup, and then report the experimental results to demonstrate the efficiency of HyperFGM compared with the baseline methods.

5.1 Experimental Setup

To evaluate the performance of our model, all the previous related work on SES prediction, traditional hypergraph-based methods and traditional factor graph methods are considered below for comparison.

Logistic Regression (LR): [4] relied on the elastic net model for SES prediction. We choose LR with the elastic net regularization as a baseline model for SES prediction.

SVM & Random Forest (RF): [23] utilized SVM and RF for SES prediction.

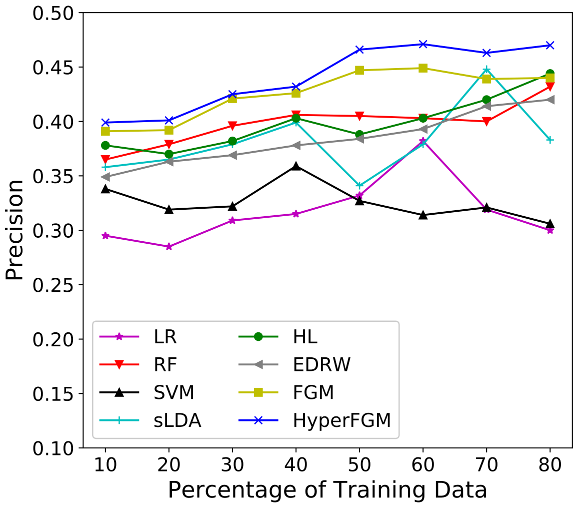

Fig. 3: Performance (Precision) comparison of different methods with different percentages of training data.

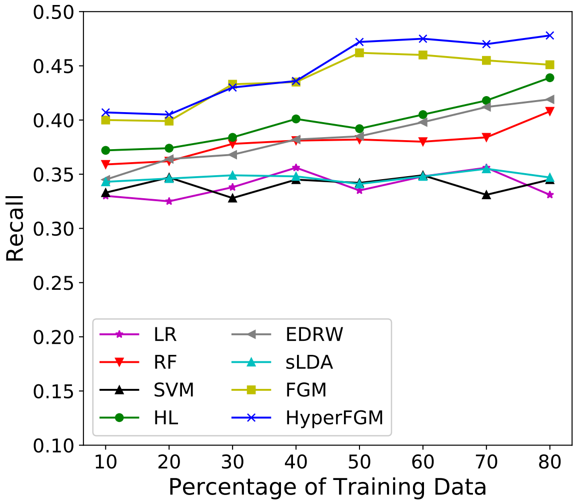

Fig. 4: Performance (Recall) comparison of different methods with different percentages of training data.

Supervised LDA (sLDA): [10] employed the supervised topic model to infer SES. We use mobility pattern vectors as the input.

Hypergraph Learning (HL): A classic hypergraph learning model [7].

Extended Discriminative Random Walk (EDRW): A hypergraph-based model[20] that extends the discriminative random walk framework.

Factor Graph Model (FGM): A traditional factor graph model [27] that does not consider the implicit relationship factors.

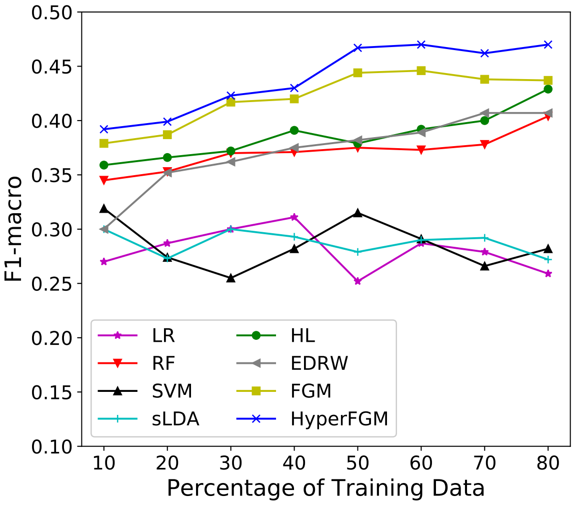

Fig. 5: Performance (F1-macro) comparison of different methods with different percentages of training data.

LR, SVM and RF use the same user attributes and mobility pattern vectors as their inputs. For the hypergraph-based models HL and EDRW, two kinds of hyperedges that are respectively based on the user attributes and mobility pattern vectors are considered.

In our experiments, in order to evaluate the performance of our model with different percentages of training data (i.e., labeled data), 10% to 80% samples for each SES level are randomly selected as the labeled training data and the rest as the unlabeled testing data. More specifically, we consider several kinds of data splitting, i.e., we randomly select $k\%$ samples for each SES level as the labeled training data and the rest samples for the unlabeled testing data. In our work, we set $k = [10, 20, 30, 40, 50, 60, 70, 80]$. In order to ensure the soundness and robustness of experimental results, like the traditional evaluation method of semi-supervised method [24], this procedure with different percentages of training data repeats 10 times and we report the averaged prediction performance as final results. The prediction performance is evaluated in terms of precision, recall, and macro F1-score (F1-macro). In the presence of class-imbalance, the F1-macro that balances precision and recall is deemed to be better than other measures such as accuracy [20].

5.2 Prediction Performance

Figure 3, 4 and 5 compare the prediction performance of different methods with different training data percentage (10%-80%) in terms of precision, recall and F1-macro respectively. The proposed HyperFGM achieves the highest performance under any percentages in terms of all metrics. Specifically, on the sparse individual records, HyperFGM significantly outperforms previous models for SES prediction, i.e., LR, SVM, RF and sLDA, by 11-22%, 7-19%, 5-9% and 9-20% respectively in terms of F1-macro. There is a similar improvement in terms of precision and recall. This is because HyperFGM, taking advantage of factor graph model, can effectively capture the relations between SES and numerous SES-related attributes by the customized factor functions. In addition, thanks to the implicit high-order mobility pattern relationships among users represented on the hypergraph structure, HyperFGM outperforms FGM (with a about 2-3% higher F1-macro score). Meanwhile, the recall and precision of HyperFGM also increase with the similar improvement. Furthermore, compared with the traditional hypergraph-based methods HL and EDRW, HyperFGM also increases 2-8%, 3-9% and 3-9% in terms of precision, recall and F1-macro respectively. This is because traditional hypergraph-based methods is unable to directly represent the relations between various attributes of users and SES by the hyperedges, and they then convert numerous attributes into relationships among users, which leads to some performance loss.

Performance of Each SES Level. Table 2 shows the prediction performance of different methods on the prediction task for each SES level. Due to the space limitation, here we only present the results in the context of taking 50% of users as training data and the rest for test. We observe that LR, SVM, RF, sLDA, HL and EDRW have much low performance on the prediction tasks of Level A and Level C while achieving relatively high performance on the Level B prediction task, which indicates that these methods may suffer from the label bias problem. On the contrary, FGM and HyperFGM have significantly higher performance with about 9-36% and 4-27% higher F1-macro scores in terms of Levels A and C, which shows that factor graph models handle the label imbalance problem much better. Furthermore, HyperFGM considers the attributes of users and exploits the implicit high-order relationships among users, thus achieving better performance than FGM in each SES level prediction.

Mobility Pattern Relationship Contribution Analysis. Figure 6 demonstrates the contribution of mobility pattern relationships in the graph-based models. Generally, the models considering the mobility pattern relationships among users mostly increase the prediction performance compared with their counterparts, i.e., HL-M, EDRW-M and FGM, which do not consider the mobility pattern relationships. Intuitively, from the social science perspective, the mobility pattern relationship factor improves the performance by bringing the prior knowledge that "the mobility patterns of users with a similar socioeconomic status tend to be similar". For example, users with similar SES have similar life style, i.e., they would work and live at the similar place areas during the similar time period. As a result, the results further prove this social science phenomenon.

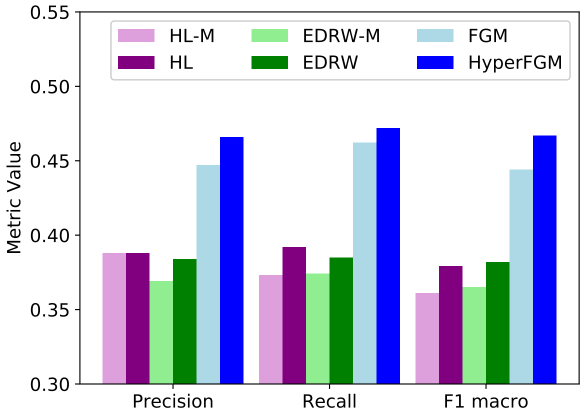

Fig. 6: Mobility pattern relationship contribution analysis.

Hyperedge Contribution Analysis. In this part, we evaluate the contribution of hyperedges in HyperFGM model. We implement four HyperFGM models, denoted as HyperFGM+k: HyperFGM+2 only considers pairwise (2-node) relationships; HyperFGM+3 considers 2-node and 3-node hyperedge relationships; HyperFGM+4 considers 2-node, 3-node and 4-node hyperedges; HyperFGM+5 considers more 5-node hyperedges than HyperFGM+4. We evaluate their prediction performance with the same experimental settings. We plot Figure 7 as an example to show the performance comparison of different versions of HyperFGM. The results show that HyperFGM+3 achieves the best performance. When considering higher-order hyperedges (i.e., $k = 4, 5$), the performance decreases; This may be because the discriminative ability of this hypergraph would be limited and even the hypergraph may confuse the correlations when each hyperedge connects to many vertices. Some work [32] has proven that the optimal k is data-dependent. This result shows the optimal k is 3 on our data. Therefore, when applying HyperFGM on other datasets, we first need to select the optimal k through grid search and use the model in other similar tasks. Compared with some other machine learning methods or deep learning methods which have many hyperparameters, our model only need to be tuned for searching one optimal hyperparameter, which simplifies the tuning procedure and decreases the tuning cost. Besides, according to previous hypergraph-based work [32], $k = 3$ always results in a good performance. Therefore, we could set $k = 3$ as default. In our future work, we plan to apply HyperFGM on different kinds of datasets to further investigate the influence of k and demonstrate the power of HyperFGM in other classification tasks.

Table 2: Performance of the prediction task for each SES level.

Models		LR	SVM	RF	sLDA	HL	EDRW	FGM	HyperFGM
Precision	A	0.188	0.153	0.264	0.146	0.270	0.243	0.357	0.396
	B	0.495	0.513	0.534	0.513	0.417	0.559	0.597	0.600
	C	0.316	0.316	0.419	0.369	0.325	0.352	0.388	0.401
	Avg	0.332	0.327	0.405	0.342	0.388	0.384	0.447	**0.466**
Recall	A	0.151	0.143	0.154	0.034	0.417	0.306	0.525	0.469
	B	0.528	0.546	0.738	0.918	0.450	0.483	0.446	0.548
	C	0.327	0.339	0.257	0.075	0.311	0.368	0.415	0.400
	Avg	0.335	0.342	0.382	0.342	0.392	0.385	0.462	**0.472**
F1-macro	A	0.115	0.127	0.192	0.054	0.324	0.210	0.423	0.428
	B	0.437	0.518	0.619	0.654	0.498	0.517	0.509	0.572
	C	0.202	0.299	0.315	0.122	0.315	0.359	0.400	0.400
	Avg	0.252	0.315	0.375	0.276	0.379	0.382	0.444	**0.467**

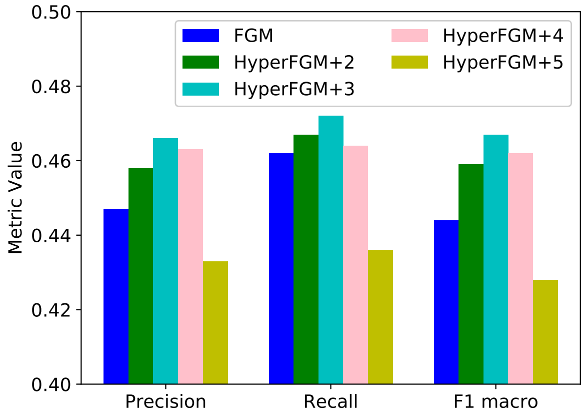

Fig. 7: Hyperedge contribution analysis.

5.3 Case study

Compared with traditional method, e.g., demographic census, estimating individual socioeconomic status based on their own real-time mobile phone usage data provides a much more real-time and cheaper method, which can benefit a wide range of applications. In order to further demonstrate the social and economic impact of this work, we take several specific case studies to show the practical value of our work.

From a commercial perspective, estimating users' SES in real time can assist in capturing each user's social and economic factors, such as income, wealth, education, health, which can improve many business applications. For example, [28] has shown that consumer's perceptions of food safety vary with socio-economic status and consumer may concern more about ingredient, ecology and food culture when purchasing food. Thus, food businesses can estimate the perception degrees of food safety of potential consumers according to related sociological achievements [28] and then recommend different kinds of food to different groups of persons by advertising. Another example may be that assessing individual SES can help banks and finance companies estimate users' credit risk index. In a word, companies can more efficiently recommend different levels of services and products to consumers with different SES. Furthermore, obtaining the personal SES distribution of each area or community can help companies select more suitable sites to start their business.

From a social and economic perspective, previous sociological articles have investigated the social and economic value of predicting SES. Measuring SES can not only help capture and understand changes to the structure of a society, but also assist in investigating the relationship between other important social variables[3]. In addition, predicting SES can assist in studying and making public polices in many fields, such as economic, education, health. For instance, regarding strong relationship between SES and health [1], assessing SES can help make sound policy decisions for health care.

5.4 Discussion and Future Work

The proposed HyperFGM, as a general semi-supervised classification method, can be applied not only to the SES prediction problem but also to other similar tasks. For example, based on similar mobile phone data like CDR, with extracting related attributes and relationships this model can be utilized for mobile phone user profiling, such as occupation, income, gender, etc. Another typical use case is to infer user demographics based on their social media data. For instance, besides social media users' attributes, with customized factor functions HyperFGM can take into account various high-order relationships based on online behavior pattern similarity, e.g., following the similar users or mentioning the similar topics. Consequently, the proposed HyperFGM can be used in a classification problem, where each object has attributes while there exist explicit or implicit relationships among objects.

The general problem of predicting individual SES based on mobile phone data represents a interesting and promis-

[3] http://www.esourceresearch.org/portals/0/uploads/documents/public/oakes_fullchapter.pdf

ing research direction in social computing field. There are many potential future directions of this work. First, in order to predict finer grained SES value of each user, some other methods can be further explored and utilized such as ranking method and regression method. For example, this work could be regarded as a ranking problem. The goal of the new ranking problem is to optimally sort the users in terms of SES, which would be a more challenging and interesting problem. Next, it is interesting to study how to further explore more implicit relationships, e.g., involving mobile Internet behavior of each user. Another potential issue is to further validate the feasibility and efficiency of the proposed model on other similar tasks. In addition, to further verify the feasibility and reusability of the proposed model, we plan to apply HyperFGM on different kinds of datasets to demonstrate the power of HyperFGM in other classification tasks. Furthermore, to verify the practical benefits of our work, the quantification of the business or social impact of the identified SES would be a very interesting and promising research direction in the future work.

6 Conclusions

In social science and public services, precisely assessing individual SES is very critical for informing public policy-making, which is yet very costly and challenging. With the advancement of AI techniques and availability of mobile phone data, existing work studied region/household-level SES assessment using mobile phone data. Compared with previous work, this paper takes a new attempt to predict individual SES on mobile phone data, which aims to provide richer insight about the relations between SES and personal attributes and networking while also address the issues in existing work on SES prediction and direct applications of existing analytic methods. A semi-supervised Hypergraph-based Factor Graph Model (HyperFGM) is introduced to leverage customized factor functions on a hypergraph structure. It effectively captures the influence of user attributes and the implicit high-order mobility pattern relationships among users on SES. HyperFGM handles both labeled and unlabeled data in a semi-supervised way. HyperFGM is tested on a set of anonymized real-life mobile phone data and sociological domain knowledge for SES labeling. The extensive experiments demonstrate that HyperFGM provides more reasonable individual SES prediction results than all existing work on SES prediction, and also achieves better performance than the state-of-the-art hypergraph-based methods and factor graph methods.

References

1. Adler, N.E., Boyce, T., Chesney, M.A., Cohen, S., Folkman, S., Kahn, R.L., Syme, S.L.: Socioeconomic status and health: the challenge of the gradient. American psychologist **49**(1), 15 (1994)
2. Blau, P.M., Duncan, O.D.: The american occupational structure. (1967)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
4. Blumenstock, J., Cadamuro, G., On, R.: Predicting poverty and wealth from mobile phone metadata. Science **350**(6264), 1073–1076 (2015)
5. Carlsson-Kanyama, A., Linden, A.L.: Travel patterns and environmental effects now and in the future:: implications of differences in energy consumption among socio-economic groups. Ecological Economics **30**(3), 405–417 (1999)
6. Dagan, I., Lee, L., Pereira, F.: Similarity-based methods for word sense disambiguation. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 56–63. Association for Computational Linguistics (1997)
7. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3-d object retrieval and recognition with hypergraph analysis. IEEE Transactions on Image Processing **21**(9), 4290–4303 (2012)
8. Granizo-Mackenzie, D., Moore, J.H.: Multiple threshold spatially uniform relieff for the genetic analysis of complex human diseases. In: EvoBIO, pp. 1–10. Springer (2013)
9. Hauser, R.M., Warren, J.R.: Socioeconomic indexes for occupations: A review, update, and critique. Sociological methodology **27**(1), 177–298 (1997)
10. Hong, L., Frias-Martinez, E., Frias-Martinez, V.: Topic models to infer socio-economic maps. In: AAAI, pp. 3835–3841 (2016)
11. Huang, Q., Wong, D.W.: Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? International Journal of Geographical Information Science **30**(9), 1873–1898 (2016)
12. Huang, Y., Liu, Q., Zhang, S., Metaxas, D.N.: Image retrieval via probabilistic hypergraph ranking. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 3376–3383. IEEE (2010)
13. Lampos, V., Aletras, N., Geyti, J.K., Zou, B., Cox, I.J.: Inferring the socioeconomic status of social media users based on behaviour and language. In: European Conference on Information Retrieval, pp. 689–695. Springer (2016)
14. Lotero, L., Hurtado, R.G., Floría, L.M., Gómez-Gardeñes, J.: Rich do not rise early: spatio-temporal patterns in the mobility networks of different socio-economic classes. Royal Society open science **3**(10), 150654 (2016)
15. Mao, H., Shuai, X., Ahn, Y.Y., Bollen, J.: Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to côte d'ivoire. EPJ Data Science **4**(1), 15 (2015)
16. Murphy, K.P., Weiss, Y., Jordan, M.I.: Loopy belief propagation for approximate inference: An empirical study. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 467–475. Morgan Kaufmann Publishers Inc. (1999)
17. Propper, C., Damiani, M., Leckie, G., Dixon, J.: Impact of patients' socioeconomic status on the distance travelled for hospital admission in the english national health service. Journal of Health Services Research & Policy **12**(3), 153–159 (2007)
18. Rabin, M.O., Scott, D.: Finite automata and their decision problems. IBM journal of research and development **3**(2), 114–125 (1959)
19. Rose, D., Pevalin, D.: Re-basing the ns-sec on soc2010 (2010)
20. Satchidanand, S.N., Ananthapadmanaban, H., Ravindran, B.: Extended discriminative random walk: A hypergraph approach to multi-view multi-relational transductive learning. In: IJCAI, pp. 3791–3797 (2015)
21. Sirin, S.R.: Socioeconomic status and academic achievement: A meta-analytic review of research. Review of educational research **75**(3), 417–453 (2005)

22. Smith-Clarke, C., Mashhadi, A., Capra, L.: Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 511–520. ACM (2014)
23. Soto, V., Frias-Martinez, V., Virseda, J., Frias-Martinez, E.: Prediction of socioeconomic levels using cell phone records. User modeling, adaption and personalization pp. 377–388 (2011)
24. Su, L., Gao, Y., Zhao, X., Wan, H., Gu, M., Sun, J.: Vertex-weighted hypergraph learning for multi-view object classification. In: IJCAI, pp. 2779–2785 (2017)
25. Tang, W., Zhuang, H., Tang, J.: Learning to infer social ties in large networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 381–397. Springer (2011)
26. Urbanowicz, R.J., Olson, R.S., Schmitt, P., Meeker, M., Moore, J.H.: Benchmarking relief-based feature selection methods. arXiv e-print. https://arxiv.org/abs/1711.08477 (2017)
27. Varatharajah, Y., Chong, M.J., Saboo, K., Berry, B., Brinkmann, B., Worrell, G., Iyer, R.: Eeg-graph: A factor-graph-based model for capturing spatial, temporal, and observational relationships in electroencephalograms. In: Advances in Neural Information Processing Systems, pp. 5377–5386 (2017)
28. Wilcock, A., Pun, M., Khanona, J., Aung, M.: Consumer attitudes, knowledge and behaviour: a review of food safety issues. Trends in Food Science & Technology **15**(2), 56–66 (2004)
29. Winkleby, M.A., Jatulis, D.E., Frank, E., Fortmann, S.P.: Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. American journal of public health **82**(6), 816–820 (1992)
30. Yang, Y., Luyten, W., Liu, L., Moens, M.F., Tang, J., Li, J.: Forecasting potential diabetes complications. In: AAAI, pp. 313–319 (2014)
31. Ye, Y., Zheng, Y., Chen, Y., Feng, J., Xie, X.: Mining individual life pattern based on location history. In: Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on, pp. 1–10. IEEE (2009)
32. Yu, J., Tao, D., Wang, M.: Adaptive hypergraph learning and its application in image classification. IEEE Transactions on Image Processing **21**(7), 3262–3272 (2012)
33. Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. In: Advances in neural information processing systems, pp. 1601–1608 (2007)