DemoMotif: Demographic Inference from Sparse Records of Shopping Transactions based on Motif Patterns

Jiayun Zhang[1], Xinyang Zhang[2], Dezhi Hong[1], Rajesh K. Gupta[1], Jingbo Shang[1]

[1]University of California, San Diego, La Jolla, CA, USA

[2]University of Illinois at Urbana-Champaign, Champaign, IL, USA

{jiz069,dehong,rgupta,jshang}@ucsd.edu,xz43@illinois.edu

ABSTRACT

Inferring contextual information such as demographics from historical transactions is valuable to public agencies and businesses. Existing methods are data-hungry and do not work well when the available records of transactions are sparse. We consider here specifically inference of demographic information using limited historical grocery transactions from a few random trips that a typical business or public service organization may see. We propose a novel method called DemoMotif to build a network model from heterogeneous data and identify subgraph patterns (i.e., motifs) that enable us to infer demographic attributes. We then design a novel motif context selection algorithm to find specific node combinations significant to certain demographic groups. Finally, we learn representations of households using these selected motif instances as context, and employ a standard classifier (e.g., SVM) for inference. For evaluation purposes, we use three real-world consumer datasets, spanning different regions and time periods in the U.S.. We evaluate the framework for predicting three attributes: ethnicity, seniority of household heads, and presence of children. Extensive experiments and case studies demonstrate that DemoMotif is capable of inferring household demographics using only a small number (e.g., fewer than 10) of random grocery trips, significantly outperforming the state-of-the-art.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Information systems → Data mining**.

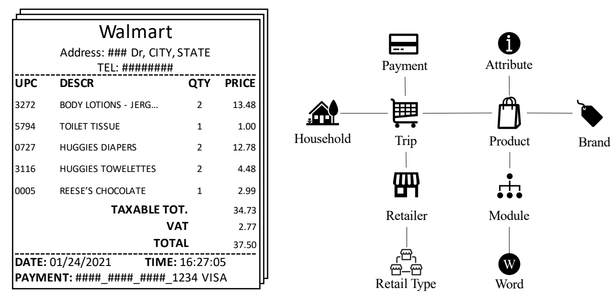
KEYWORDS

demographic information, transaction data, heterogeneous network, motif

1 INTRODUCTION

Contextual information is critical to accurate assessments in healthcare, disbursement of services by public agencies, and appropriate marketing in online and in-store business. This information is often not collected to ensure privacy or for practical reasons of convenience, even when the service provider is explicitly authorized to collect such information. Instead, such information must be inferred, often after anonymizing the records to ensure the removal of personally identifiable information. A range of services depend upon such information including user profiling [1], recommendation [27, 38], and market analysis [2, 15]. We focus on consumer marketing information for a demonstration of inference capabilities.

Prior work has explored contextual inference from user transaction records [11, 16, 34]. These typically require a large amount of data from each household (such as receipts) in order to generate meaningful bag-of-item [34] or bag-of-retailer [11, 16] features for the household. This requirement is hard to meet in practice as there exist a large number of occasional customers of offline retailers who make only a limited number of transactions, and the transactions are usually random in nature. E-commerce faces the same problem since many new users have only limited shopping records.



(a) An example shopping receipt. (b) Schema of the heterogeneous trip network constructed from the receipts.

Figure 1: We organize scattered purchase information into a heterogeneous shopping trip network, from which meaningful shopping patterns can be mined.

In this paper, we specifically study the demographic inference problem using *only a few* grocery transactions (e.g., 1 to 10) *per household*, and we propose a novel inference framework DemoMotif for this challenging problem that can be applied to many other application domains. Data sparsity makes it hard to detect simple recurring patterns. That is, we cannot tell if a particular purchase is part of a recurring purchase habit reflecting a household's preference or if it is only one-time and coincidental. The lack of sufficient purchase records significantly hinders modeling shopping patterns over time. Existing methods built upon bag-of-items, a representation that accumulates records over time, may fail in our setting.

Instead, we explore another perspective in modeling limited transaction data—leveraging combinatorial purchase patterns (e.g., the combination of item categories and their brands). Intuitively, combining different types of data gives us more expressive power than modeling them alone. For example, rice is a typical daily essential for many households, which on its own is barely an indicative predictor. However, by combining it with brand information, we may find that households in different ethnic groups have different preferences for the brands of rice.

We, therefore, first organize user transaction records into a heterogeneous trip network (as shown in Figure 1). This large heterogeneous network contains the information recorded in the receipts (e.g., retailer chain, method of payment, purchased items) from shopping trips and their co-occurrence relationships. The network schema is shown in Figure 1(b). To extract meaningful purchase patterns, we then leverage *motif patterns* to capture combinations of typed entities in the graph, such as brands, attributes, and modules. Figure 2 shows two example motif patterns and concrete purchase patterns that match these two motif patterns. We refer to such concrete purchase patterns as *motif contexts*.

In addition to informative combinations in motif patterns, there are noisy and distracting purchase patterns as well. In other words, not all the purchase patterns matching the same motif pattern are equally useful; only a subset of them would be significant to the target demographic group. We propose a novel motif selection algorithm that sifts through the motif contexts to identify only those indicative ones.

In summary, DemoMotif is a graph-based representation learning framework. The workflow is sketched in Figure 3. We mine purchase patterns predictive of household contextual information with the motif selection algorithm. Then, we embed the households and purchase patterns into the same feature space. Finally, we employ a standard classification model on the representations of the households to infer their demographics.

We show that with a small number of receipts from each household, DemoMotif can infer important demographic attributes about the household with reasonable accuracy. According to our experiments, with 10 random receipts per household, DemoMotif can predict the presence of children, seniority of household heads, and ethnicity with macro-F1 scores of around 0.78, 0.72, 0.48 on all datasets, significantly outperforming the state-of-the-art methods. We conducted a stress test for DemoMotif by varying the number of available receipts per household. It shows that when there is *only 3* receipt per household, DemoMotif can still achieve macro-F1 of around 0.728, 0.665, 0.372 on the three tasks with the country-wide dataset. These results also point to the need for new means to ensure privacy, which is beyond the scope of this paper. Our contributions are summarized as follows:

- We address a practical yet challenging contextual inference problem: inferring household demographics from only limited random transaction records that cannot be addressed by existing models.
- We propose a new framework DemoMotif using a novel motif selection algorithm to effectively mine user purchase patterns that are strongly linked to the target demographic group.
- We conduct extensive experiments on three real-world datasets collected and publicly available in the U.S.. We evaluate DemoMotif on inferring three demographic measures: presence of children, seniority of household heads, and ethnicity of the household. The choice of these parameters is purely arbitrary and driven by the available datasets. Experimental results show that DemoMotif could predict the information with reasonable accuracy even when there are only 3 shopping receipts per household.

2 PRELIMINARIES

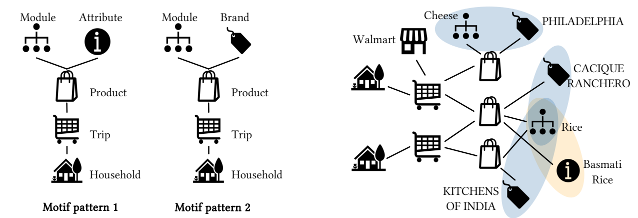
We first introduce the structure of the receipt data and some preliminary concepts. Then, we formulate the problem by specifying the input and output.

2.1 Receipt Data

Figure 1(a) gives an example shopping receipt. It contains information about which shop the household visited (e.g., Walmart), the type of retailer (e.g., grocery), the method of payment (e.g., Visa), and a list of purchased items with details. Each purchased item has its UPC code (unique identity ID) and comes with metadata information, as shown in Table 1. Modules tell the category of the items; attributes tell additional information about the items such as flavor, scent, etc. Such metadata can either be directly extracted from the description fields in the receipt or be retrieved from publicly available knowledge bases[1] by matching the UPC codes.

Table 1: Product Metadata Available in the Nielsen Data.

Information	Example
UPC	2100000727
Module	cheese - processed - cream cheese
Brand	Philadelphia
Attribute	strawberry (flavor), soft (style), ...



(a) Motif patterns combining "Module-Attribute" and "Module-Brand".

(b) Motif contexts extracted with motif patterns given in (a).

Figure 2: Examples of motif pattern and motif context

2.2 Heterogeneous Trip Network

Given a collection of shopping receipts, we organize them into a heterogeneous trip network to model the relationship among multi-type objects. The network $G = (V, E, \phi, \psi)$ consists of a set of nodes V, a set of edges E, a node type mapping $\phi : V \rightarrow T_V$, and an edge type mapping $\psi : E \rightarrow T_E$.

The schema of our heterogeneous trip network is shown in Figure 1(b). There are $|T_V| = 10$ different types of nodes in total, including households, shopping trips, trip metadata (method of payment, chain stores, retail types), products, and product metadata (brands, attributes, modules, and the words describing the modules). Households are connected to their shopping trips, and trips are connected to all the products purchased. The metadata nodes, if available, are connected to the respective trips and products.

2.3 Motif Pattern and Motif Context

Motif patterns are sub-graph patterns at the meta level (i.e., nodes are abstracted by their types). They are able to capture the higher-order inter-connectivity among nodes.

A *motif instance* is an instantiation of a motif pattern by replacing the node types with concrete values. We define "open" nodes in a motif instance to be those single-degree nodes except for the household nodes. The "open" nodes define a *motif context*.

Figure 2 shows some examples. Given the two motif patterns in Figure 2(a), the combination of module and brand and the combination of module and attribute, there are four motif contexts in

[1]https://www.upcitemdb.com

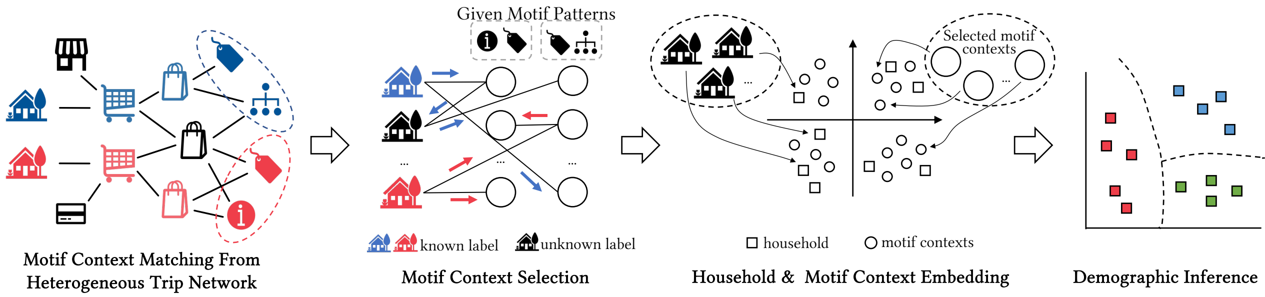


Figure 3: An overview of DEMOMOTIF: It learns embeddings for households with motif contexts and performs a motif context selection process to identify contexts that are strongly linked to specific user demographics.

the trip network: three "module-brand" contexts (marked in blue in Figure 2(b)) being "cheese-Philadelphia", "rice-Cacique Ranchero" and "rice-Kitchens of India", and one "module-attribute" context (marked in yellow) being "rice-basmati rice". Considering "module" as another motif pattern, the two households share the same motif context "rice".

2.4 Problem Definition

In this paper, we aim to predict household demographics, a specific form of contextual information, given a small number of shopping receipts per household. The input user transaction data contain a set of households U and a set of shopping receipts $R = \bigcup_{u \in U} R_u$ collected from these households. The number of receipts given for each household is a small, fixed number $N_R = |R_u|, \forall u \in U$, e.g., 10. We also require the input of k motif patterns $M = \{M_1, M_2, ..., M_k\}$ to extract useful information from the heterogeneous trip network constructed from receipts as described earlier. Our objective is to learn a demographic inference function $f : U \mapsto Y$, which predicts the demographic attributes of each household. For training, we know the demographic labels of a subset of the households $(U_{\text{train}}, Y_{\text{train}})$. We test our model on the unlabeled households $U_{\text{test}} = U \setminus U_{\text{train}}$.

3 THE DEMOMOTIF FRAMEWORK

The key principle underlying DEMOMOTIF is to assign similar representations to households that share purchase patterns predictive of a demographic group. Purchase patterns critical to the representation learning process are captured by our novel motif context matching and selection methods.

As shown in Figure 3, DEMOMOTIF is a graph-based representation learning framework. DEMOMOTIF takes the following steps. We first build a heterogeneous trip network based on users' transaction records and extract motif contexts by matching the user-provided motif patterns to the network. We then link each household to its motif contexts and construct a bipartite motif context graph. On the motif context graph, we perform motif context selection with authority ranking to find label-indicative contexts for different demographic groups. Once the high-quality motif contexts are selected, we embed the households and the motif contexts into a shared feature space. Finally, we use the learned embeddings to predict the corresponding demographic attributes.

We organize the rest of the section as follows. We first present our embedding learning method in Section 3.1, assuming that the rest of the motif-based mining part of the framework has already been done. Then, we present the novel motif context matching and selection methods for mining label-indicative motif contexts in Section 3.2. In Section 3.3, we describe how the learned embeddings are used for classification.

3.1 Household and Motif Context Embedding

In DEMOMOTIF, each household will be associated with a set of label-indicative purchase patterns, i.e., its *motif context*. As the motif contexts (purchase patterns) in a motif pattern are carefully chosen and are highly predictive of the demographic attributes, characterizing the households with motif contexts would allow us to estimate their demographic information. The principle of the embedding learning is to enforce households with similar motif contexts to have similar representations.

Formally, we denote the set of households as H and the associated motif contexts of household h as $C(h)$. We use each household to predict its motif contexts. The predicted probability of a motif context given a household is measured by the similarity between the embeddings of the motif context and the household. We train the embeddings to approximate the probabilities of the occurrence of motif contexts in the household's shopping records. The number of times each context c gets sampled is proportional to its occurrences in $C(h)$. The goal is to minimize the following loss term:

$$\mathcal{L}_{emb} = \sum_{h \in V_H} \mathbb{E}_{c \sim C(h)} - \log P(c|h)$$

The probabilities are approximated with negative sampling [22]:

$$\log P(c|h) = \log \sigma(u_c^T v_h) - \mathbb{E}_{c' \sim P_{neg}(c)} \log \sigma(u_{c'}^T v_h)$$

where u_c and v_h are the embedding vectors of motif context c and household h, respectively, $P_{neg}(c)$ is the negative sampling distribution. We set $P_{neg}(c)$ as the distribution of context raised to the 3/4rd power [22], i.e., the probability of choosing c_i as a negative sample is $P_{neg}(c_i) = f(c_i)^{3/4} / \sum_{j=1}^{|C|} f(c_j)^{3/4}$, where $f(c_i)$ is the number of occurrences of context c_i in all households.

3.2 Motif Context Matching and Selection

The advantage of our embedding lies in mining purchase patterns that are strongly linked to specific demographic groups. We achieve this goal by modeling typed entity combinations from transaction data with motif patterns. Given k user-defined motif patterns M, we extract a set of initial motif contexts $\tilde{C}$ from the heterogeneous trip network G by matching the network structures around households and the motif patterns M.

Simply using all motif contexts for training embeddings gives sub-optimal performance based on our observation. Not all motif contexts are indicative of the demographics of households. Some common products, such as tissue paper, are daily purchases for almost all households. These contexts are not helpful for characterizing households. In addition, some contexts are indicative of one demographic attribute but are less so for others. For example, the frequency of buying oriental food could tell something about the ethnicity, but is almost meaningless for judging if the household has children. Carefully selecting indicative motif contexts can help learn the embeddings for specific tasks.

A label-indicative motif context should be important to the households and be able to distinguish different household groups. We apply authority ranking [31] on the *motif context graph* to refine the motif contexts used for embedding learning.

We construct a bipartite motif context graph $G^B = \langle V, E\rangle$. The node set V consists of two types of objects, households V_H and motif contexts V_C. There are two types of edges, household-context edges E_{HC} and context-context edges E_{CC}. We denote the adjacency matrix of G^B as $\mathbf{W} \in R^{|V_H|\times|V_C|}$. The weight $w_{h,c}$ in the adjacency matrix describes the number of occurrences of context c in household h. We decompose the adjacency matrix $\mathbf{W}$ into four blocks: $\mathbf{W}_{HH}$, $\mathbf{W}_{HC}$, $\mathbf{W}_{CH}$, and $\mathbf{W}_{CC}$, each denoting a sub-graph of objects between two different types of objects. Since there is no edge between any two households, sub-matrix $\mathbf{W}_{HH}$ is a zero matrix. Then the motif context graph can be written as $G^B = \langle\{V_H \cup V_C\}, \mathbf{W}\rangle$, where $\mathbf{W}$ could be represented as:

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{W}_{HC} \\ \mathbf{W}_{CH} & \mathbf{W}_{CC} \end{bmatrix}$$

The principles of authority ranking are: 1) highly ranked households (which means they are representative) with label l are frequently connected with motif contexts that are highly label-indicative for label l; 2) the ranking of a context is increased if it has many highly ranked neighbor contexts. Based on these two principles, the ranking scores will be first propagated from the household nodes to the context nodes, then spread among contexts, and finally propagated back from context nodes to the households.

The authority ranking is an iterative propagation process. Denote $\mathrm{S}_H^{(t)} \in R^{|H|\times|L|}$ and $\mathrm{S}_C^{(t)} \in R^{|C|\times|L|}$ as the ranking scores of households and contexts at the t-th iteration, where L is the set of classes. We initialize the ranking scores of the households as $\mathrm{S}^{(0)}(h,l) = 1/N_l$ for all the households with known label l and zero for the other households, where N_l is the number of households known to be samples of class l. On each iteration, the ranking scores are updated with the following equations:

$$\mathrm{S}_H^{(t)} \leftarrow \tilde{\mathbf{W}}_{HC}\mathrm{S}_C^{(t-1)}$$
$$\mathrm{S}_C^{(t)} \leftarrow \alpha\tilde{\mathbf{W}}_{HC}^T\mathrm{S}_H^{(t)} + (1-\alpha)\tilde{\mathbf{W}}_{CC}\mathrm{S}_C^{(t-1)}$$

where α is a hyper-parameter balancing the influence of neighbor households and contexts. $\tilde{\mathbf{W}}_{HC}$ and $\tilde{\mathbf{W}}_{CC}$ are the normalized adjacency matrices of household-context sub-matrix and context-context sub-matrix. The weight matrices are normalized by $\tilde{\mathbf{W}}_* = \mathbf{D}_{*r}^{-1/2}\mathbf{W}_*\mathbf{D}_{*c}^{-1/2}$, where $\mathbf{D}_{*r}$ and $\mathbf{D}_{*c}$ are the row degree and column degree matrices of the corresponding sub-matrix, respectively.

At the end of each iteration, we normalize the ranking scores by

$$\mathrm{S}_*^{(t)} \leftarrow \mathrm{S}_*^{(t)}\mathbf{D}_{*c}^{(t)^{-1}}$$

where $\mathbf{D}_{*c}^{(t)}$ is the corresponding column degree matrix of $\mathrm{S}_*^{(t)}$.

The propagation process is repeated until convergence or reaching a maximum iteration number. In practice, we observe the numerical convergence at around 8 iterations. After the iterations, each motif context has the ranking scores for the $|L|$ classes. Then, we consider the following two aspects in order to combine the ranking scores for the $|L|$ classes:

- **Importance.** The selected motif contexts should have high ranking scores on at least one of the $|L|$ classes in order to provide useful information for representation learning. As a result, for each motif context c, we take the maximum of its ranking score across the $|L|$ classes as its importance score.

$$\text{Importance}(c) = max(\mathrm{S}_C(c,\cdot))$$

- **Concentration.** The selected motif contexts should be important to only a subset of the classes so as to be discriminative for different groups of households. We normalize the ranking score of motif contexts as $\tilde{\mathrm{S}}_C = \mathrm{S}_C(c,l)/\sum_{l_i\in L}\mathrm{S}_C(c,l_i)$ and calculate the concentration score based on entropy:

$$\text{Concentration}(c) = 1 + \frac{1}{\log|L|}\sum_{l_i\in L}\tilde{\mathrm{S}}_C(c,l_i)\log\tilde{\mathrm{S}}_C(c,l_i)$$

We follow previous work [9, 29] and use a geometric mean to combine the two scores:

$$\text{Score}(c) = (\text{Importance}(c) \times \text{Concentration}(c))^{\frac{1}{2}}$$

The initial motif contexts $\tilde{C}$ are ranked according to the final scores. The top K_c percent of the contexts are selected, which we denote as C. Then, we can train household embeddings with the method described in Section 3.1 using these selected motif contexts.

3.3 Demographic Inference

With the learned embeddings of all the households, we can apply a classifier to perform demographic inference. We note that DemoMotif is compatible with common classification models, such as logistic regression, support vector machine (SVM), and random forest. Without loss of generality, we adopt the widely used SVM to perform the classification.

Given the embeddings and the labels of N training samples $\{(v_i, y_i)\}_{n=1}^N$, where v_i and y_i are the embedding and the label of the i-th training household, respectively, we train an SVM classifier following the objective to find the hyperplanes to separate data points to their potential classes with maximum margin. During testing, given the embedding of the testing households, the classifier is able to assign an attribute class to each household.

4 EXPERIMENTS

We describe datasets and the experimental setup before discussing quantitative results. We then perform sensitivity analysis of the performance of DemoMotif under limited data. Finally, we present case studies to illustrate the learned motif contexts for different demographic prediction tasks and some interesting findings.

Table 2: Number of Nodes in Our Constructed Motif Context Graph on the Datasets After Pre-processing

Dataset	Household	Product	Retailer	Retail Type	Payment	Module	Brand	Attribute	Word	Module & Brand	Module & Attribute	Attribute & Attribute
US-2015	30,144	25,126	386	56	9	7,397	991	9,597	1,097	9,688	15,633	34,749
CA-2016	2,556	52,783	273	57	9	973	9,552	5,849	1,016	13,681	13,460	32,034
NY-2014	1,538	37,296	243	59	9	913	6,587	3,819	965	9,367	9,169	20,225

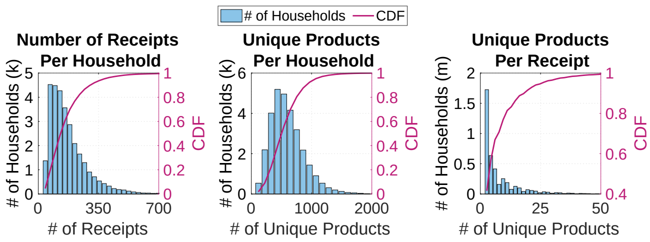


Figure 4: Statistics about the US-2015 dataset. We process and sample from the dataset to simulate the scenarios where each household only has limited, random receipts.

4.1 Dataset

We use the **Nielsen data**[2] for our study. It collects purchase records from households in the United States in a longitudinal study through voluntary participation. Participating households provided their demographic information, including household composition, presence of children, seniority of household heads, ethnicity, etc. The households uploaded receipts from their shopping trips to any outlets using in-home scanners throughout the year. They were also asked to specify when and where they made purchases. The metadata information about the retailer chains and purchased items is included in the Nielsen data. Note that the datasets used by the previous papers [11, 16, 34] are not suitable for evaluation since they do not have the information about the items purchased by the customers, or the items are renamed by digital IDs so we cannot extract the metadata.

The Nielsen data is a large-scale nationwide dataset, collecting transactions across the U.S. for over 16 years. To evaluate the generalizability of DemoMotif, we form three datasets by selecting transaction records from the Nielsen data that differ in regions and data collection time: 1) **US-2015** includes all the transaction records in Nielsen data collected from Dec. 29, 2014 to Dec. 27, 2015; 2) **CA-2016** includes the transactions in California State from Dec. 28, 2015 to Dec. 25, 2016; 3) **NY-2014** includes the transactions in New York State from Dec. 30, 2013 to Dec. 28, 2014.

In order to make the datasets better represent the census estimates of demographic composition, the Nielsen data provide a sampling weight for each household for re-weighting the raw panel. We sample the households with respect to the given sampling weights. We filter out those motif contexts present in fewer than 20 households in US-2015 and those motif contexts present in fewer than 5 households in CA-2016 and NY-2014. The number of nodes in the motif context graph after pre-processing is shown in Table 2.

Figure 4 shows statistics about the US-2015 dataset. In total, there are 4,797,444 reported receipts. The average number of receipts per household is 159.15. Every household purchases an average of

Table 3: Demographic Distributions in the Datasets

Attribute	Child		Head Age		Ethnicity			
Value	Yes	No	< 65	≥ 65	Caucasian	African American	Asian	Other
US-2015	28%	72%	75%	25%	78%	12%	4%	6%
CA-2016	29%	71%	73%	27%	64%	8%	14%	14%
NY-2014	23%	77%	74%	26%	71%	16%	5%	8%

544.82 unique products in a year. The average number of unique products per receipt is 7.01. In our study, we randomly sample $N_R = 10$ receipts per household from the datasets. According to the statistics described above, it is a small subset (about 6%) of the shopping records that the households would have in a year.

4.2 Experimental Setup

Implementation Details. We randomly partition each dataset into 5 equal-sized subsamples and conduct 5-fold cross-validation (CV) for all experiments. The dimension of the embeddings is set to be 128. For motif context selection, we keep top $K_c = 1 - 1/|L|$ motif contexts since if $|L|-1$ classes can be distinguished with the selected contexts, the rest class would be easily classified. Specifically, we set $K_c = 0.5$ for predicting the presence of children and seniority of household heads, and $K_c = 0.75$ for predicting ethnicity. We perform sensitivity analysis on this parameter in Section 4.5.1.
Motif Patterns. We define 11 motif patterns for the trip network. All open nodes except for "household" and "trip" in the network schema shown in Figure 1 are included (i.e., retailer, retail type, method of payment, product, module, attribute, brand, word). In addition, three combinatorial patterns: "module-brand", "module-attribute", "attribute-attribute", are added to the motif patterns.
Prediction Tasks. We conduct two prediction tasks on all datasets: the presence of children (whether there is any child under 18, denoted as "child" in the tables) and the seniority of household heads (whether the male or female head is older than 65, denoted as "age"), and ethnicity (denoted as "ethnic"). The distributions of attributes are shown in Table 3.
Metrics. Due to label imbalance, we adopt macro-F1 to evaluate the performance:

$$\text{macro-F1} = \frac{1}{|L|} \sum_{l \in L} \frac{2 * \text{Precision}_l * \text{Recall}_l}{\text{Precision}_l + \text{Recall}_l}$$

where $\text{Precision}_l = \text{TP}_l/(\text{TP}_l + \text{FP}_l)$ and $\text{Recall}_l = \text{TP}_l/(\text{TP}_l + \text{FN}_l)$. TP_l, FP_l and FN_l are the number of true positives, false positives and false negatives of class l respectively.

4.3 Compared Methods

To evaluate our framework, we compare it with 11 baseline models. We group them into three categories: 1) deep neural networks-based

[2] https://www.chicagobooth.edu/research/kilts/datasets/nielsen

Table 4: Results (%) of all compared methods using 10 receipts per household based on 5-fold CV. The macro-F1 of DemoMotif is significantly greater than that of any baseline method (p-value < 0.05, paired t-test).

Method	US-2015			CA-2016			NY-2014		
	Child	Age	Ethnic	Child	Age	Ethnic	Child	Age	Ethnic
bag-of-item [13]	68.9	61.7	36.4	63.7	56.8	26.7	53.1	54.7	31.9
bag-of-item++	73.8	67.6	46.9	57.8	56.8	25.8	63.4	56.8	30.8
random walk [24]	73.5	58.8	26.6	69.1	59.4	26.7	63.7	48.8	30.8
GCN [17]	74.8	67.1	37.4	68.3	61.2	35.2	64.9	58.0	38.7
GAT [33]	74.1	67.5	39.5	66.7	62.1	36.2	58.0	59.2	40.1
metapath2vec [6]	77.1	70.1	45.2	73.6	66.3	41.8	66.3	58.2	38.4
ESim [28]	74.7	66.3	40.0	72.8	63.9	40.2	67.5	59.0	40.6
Transformer [32]	75.9	69.8	47.3	72.3	63.2	28.9	65.7	50.6	21.0
SNE [34]	48.4	51.9	25.0	33.9	44.7	22.0	47.8	50.4	27.2
ETNA [16]	71.8	65.7	40.5	59.1	55.0	29.2	53.2	56.4	32.3
HURA [35]	75.9	67.2	41.0	60.9	50.2	23.9	63.6	54.0	33.9
DemoMotif w/o S	78.6	72.4	48.4	76.3	69.8	45.2	71.7	64.5	46.3
DemoMotif	**78.9**	**72.8**	**48.5**	**79.5**	**72.7**	**47.7**	**75.3**	**71.0**	**48.2**

methods (SNE, ETNA, HURA, Transformer), among which there are three state-of-the-art methods (SNE, ETNA, HURA) designed for demographic inference, 2) feature engineering-based methods (bag-of-item, bag-of-item++), 3) graph-based methods (random walk, GCN, GAT, metapath2vec, ESim). The baselines are described below:

- **SNE** [34] is the first model designed for demographic inference using transaction records. It maps items into latent vectors and applies average pooling to the vectors of the items purchased by each household. Then, a linear layer is used to make predictions.
- **ETNA** [16] is a state-of-the-art neural model for demographic inference based on transactions. It first learns global user embeddings with transaction histories. Then, a linear transformation operation maps the global embeddings into a task-oriented vector space. An attention layer is adopted to assign weights to each transaction and a linear layer is used to make predictions.
- **HURA** [35] is a state-of-the-art model for demographic inference from search queries. To adapt this model, we treat items as words and shopping trips as sentences to accommodate transactional data. The model learns representations in a hierarchical way: It first learns representations of shopping trips from the products and then learns user representations from the trip representations. The model consists of CNN and attention layers which select informative items and trips.
- **Transformer** [32] is a neural network model that adopts multi-head attention to weigh the significance of each part of the input. Each household's input data includes the items it bought and the metadata about the items and trips.
- **bag-of-item** [13] is a traditional representation method that uses the list of items with their counts that a user has bought in N_R shopping trips to represent the user.
- **bag-of-item++** is an advanced version of *bag-of-item*. In addition to items, it also extracts textual and categorical features from the metadata of items and trips. The elements in the representation are normalized into TFIDF values.

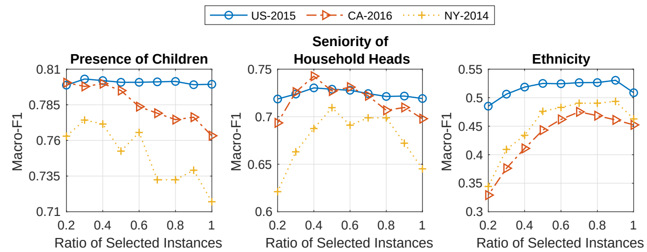


Figure 5: Performance of DemoMotif w.r.t. Ratios of Selected Motif Contexts Based on 5-Fold CV.

- **random walk** [24] computes personalized PageRank for each class by starting the random walk from households of the corresponding class in the training set. Households are assigned to the class for which they have the largest PageRank score.
- **GCN** [17] builds a multi-layer neural network that operates on the heterogeneous trip network to learn node embeddings based on the properties of each node's neighbors. Nodes are initialized as one-hot representations.
- **GAT** [33] is an advanced graph neural network structure equipped with self-attention. It operates on the heterogeneous trip network and computes the hidden representations of each node in the graph by attending over its neighbors.
- **metapath2vec** [6] is a node embedding algorithm for HIN, where the node sequence is guided by metapath-based random walks. Since it only accepts one meta-path scheme, we merge the metadata information of trips and products as two individual node types and test all possible meta-paths. We choose "household–trip–product–brand/module/attribute–product–trip–household" as the final one as it works the best.
- **ESim** [28] is another node embedding algorithm for HIN, which accepts multiple meta-path schema as guidance and assigns different weights for meta-paths. We include all the path patterns used in DemoMotif and fine-tune the path weights to get the best performance.

We also examine an ablation of DemoMotif without motif context selection as DemoMotif $^{\textbf{w/o S}}$.

For a fair comparison, we set the input graph the same for all the graph-based methods, including all products and all types of metadata of products and trips, and also make the dimension as 128. We use SVM to perform the predictions based on these learned embeddings. We have tried Logistic Regression on all the representation learning-based baselines. It gives the same relative performance order as using SVM. The Macro-F1 of DemoMotif with Logistic Regression are 0.785, 0.727, 0.523, respectively for the two tasks on the US-2015 dataset, 0.761, 0.697, 0.456 on the CA-2016 dataset, and 0.713, 0.654, 0.497 on the NY-2014 dataset.

4.4 Main Results and Analysis

Table 4 summarizes the main results. Recall that we use only 10 receipts per household to obtain these results. We conduct paired t-test on the results of 5-fold CV and find the macro-F1 of DemoMotif is significantly greater than that of any baseline (p-value < 0.05).

The deep neural networks-based methods, SNE, ETNA, HURA, and Transformer, show limited performance when only a small amount of information is available from each household since deep

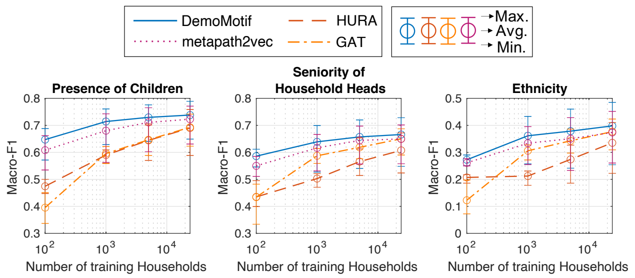


Figure 6: Model performance w.r.t. the number of training households based on 5-fold CV on US-2015. The "o" markers show the *average* scores obtained by using different numbers of receipts per household. The top and bottom lines of the bars across the markers show the *max* and *min*.

learning models usually need more data for training. Among the three state-of-the-art methods designed for demographic inference, HURA performs the best. The two feature engineering-based methods, bag-of-item and bag-of-item++, also show poor inference performance. The representations of the households based on these two methods are very high-dimensional due to the large set of items and metadata, which degrade the classification performance of machine learning classifiers. The three network embedding-based methods, metapath2vec, ESim, and DemoMotif, outperform feature engineering-based methods and deep neural networks-based methods in most of the tasks, especially with the two state-based datasets (CA-2016 and NY-2014), which demonstrates the advantage of these semi-supervised network embedding methods over conventional supervised machine learning when the training data is sparse. The superiority of DemoMotif over the two meta-path guided methods, metapath2vec and ESim, shows the power of combinatorial purchase patterns. Incorporating network motifs could capture richer network semantics, which helps to produce better embeddings.

By comparing the results of DemoMotif and those of DemoMotif w/o selection, we see the effect of the motif context selection procedure. DemoMotif can more effectively characterize households by selecting label-indicative motif contexts and using them to learn the embeddings for households, thus achieving better performance.

4.5 Sensitivity Analysis

4.5.1 Number of Selected Motif Contexts. We evaluate the system performance with different ratios of selected motif contexts K_c. The results are shown in Figure 5. For predicting the presence of children and the seniority of household heads, selecting the top 30% or more motif contexts is effective, compared to the results w/o selection. For ethnicity, selecting the top 60% or more contexts can achieve comparable or better results than using all motif contexts. According to the results, DemoMotif could achieve better results than DemoMotif w/o selection after K_c is increased to a certain value. The performance could be further improved by tuning K_c.

4.5.2 Number of Training Households and Receipts. We also examine model performance under even less transaction information. To do so, we decrease the number of training households in US-2015. In addition, for each setting, we vary the number of receipts available from each household from 1, 3, 5, 7 to 10. We compare

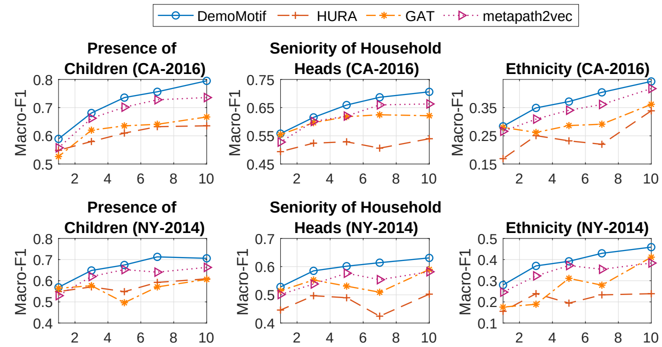


Figure 7: Performance of DemoMotif on CA-2016 and NY-2014 w.r.t Different Numbers of Receipts Based on 5-Fold CV

DemoMotif with HURA, which performs the best among the three state-of-the-art models, and GAT and metapath2vec, which are the two best performing baselines.

The results on US-2015 are shown in Figure 6. For all three tasks, having more training households or more receipts from each household helps improve models. DemoMotif always performs the best and the advantage is more significant when the number of training samples is smaller. When there are only 100 training households, DemoMotif increases the average macro-F1 by 0.026, 0.014, 0.014 compared with metapath2vec, 0.278, 0.117, 0.147 compared with GAT and 0.241, 0.200, 0.073 compared with HURA. Although the differences decrease as the number of training households increases, the average macro-F1 of DemoMotif are still 0.015, 0.016, 0.024 higher than metapath2vec, the second-best baseline, when there are 24,115 (80%) training samples.

We shall note that, even when there is only *one receipt* per household, DemoMotif can still predict the presence of children, seniority of household heads, and ethnicity with macro-F1 of 0.649, 0.548, 0.254 respectively. When each household has 3 receipts available, the f1 reaches 0.728, 0.655, 0.372, respectively. When there are 5 receipts per household, the f1 reaches 0.753, 0.694, 0.423.

We also examine the performance with respect to different numbers of available receipts on CA-2016 and NY-2014. Figure 7 shows the comparison of the performance among DemoMotif, HURA, GAT, and metapath2vec. On both datasets, DemoMotif always performs the best among the compared methods for the three tasks.

4.6 Case Studies

We next present some case studies for a closer look at the selected motif contexts by DemoMotif. Figure 8 lists the top selected motif contexts for the three tasks.

For predicting the presence of children, the most indicative contexts include words that directly imply the presence of children like "baby" and "children", as well as kids' products such as "diapers" and "combination lunch". For predicting the seniority of household heads, "adult-incontinence" and "denture" suggest that the household head is likely to be an elderly, while "tampon" is a strong indicator of having females under 65. In addition, "paying by check" is another indicative context since elderly people are more likely to use checks than young people. For ethnicity, we find certain chain stores are important, which indicates that different ethnic groups have preferences for different chain stores for shopping.

Word | Module | Brand | Attribute | Payment
Retailer | Module & Brand | Module & Attribute | Attribute & Attribute

Presence of Children	Head Age	Ethnicity
· Diapers	· Adult-incontinence & Pad	· *** (a grocery store)
· Disposable diapers	· Adult-incontinence	· Credit Card/American Express
· Baby	· Tampons	· Hair preparations-ethnic
· Lunchables	& Non-deodorant	· House Foods
· Combination lunches	· Non-deodorant & Tube	· *** (a warehouse club)
& Lunchables	· Tampons	· No USDA organic seal & Tofu
· Aseptic pouch	· Tampons & Unscented	· Oriental foods-misc.
· Children	· Tube & Plastic	& Hinoichi
· Cat food - wet	· Non-deodorant & Plastic	· Hinoichi
· Aseptic pouch & Regular	· Tampons & Tube	· Hapi
· Combination lunches	· Long & Pad	· Tofu
· Lunches	· Denture	· TruRoots
· Combination	· Unscented & Tube	· Oriental foods-misc. & Tofu
· Wet	· Check	· *** (a grocery store)

Figure 8: Case Study: Top Selected Motif Contexts by DemoMotif for All Three Prediction Tasks With US-2015.

5 RELATED WORK

We review two important lines of related work: demographic inference and heterogeneous information network mining.

5.1 Demographic Inference

Inferring demographic attributes from behavioral data has been studied in different scenarios. Researchers have proposed methods to infer user demographics from online behaviors, such as webpage browsing [8, 18], search queries [3, 14, 35], social media activities [3, 5, 20] and mobile app usage [7, 20]. For example, Hu et al. [8] proposed a method to predict the age and gender of the Internet users with the webpage click-through log. Bi et al. [3] inferred users' demographics based on their search queries extracted from social media. Similarly, behavior data such as travel trajectories [39, 40] and automotive driving data [30] have been used to infer demographic attributes. Zhong et al. [39] predict users' demographics attributes based on mobile data, such as users' usage logs, physical activities and environmental context. Zhong et al. [40] proposed a framework to infer various demographic attributes with location check-in data. Stachl and Bühner [30] studied the identification of drivers' gender using automotive driving data.

Shopping data has also been used to predict customers' demographics for companies to make marketing strategies. As the first paper to predict demographics from purchase data, Wang et al. [34] proposed a Structured Neural Embedding (SNE) model that trains embeddings from bag-of-items and aggregates them as user representations for demographic inference. Resheff et al. [25] proposed a similar approach, which integrates sequence embedding of transaction data with structured relational data (i.e., type and number of financial accounts held by the user) to form user representation. Kim et al. [16] proposed Embedding Transformation Network with Attention (ETNA) model, which learns global user embeddings with the transaction histories at the bottom of the model and transforms them into task-specific embeddings using a linear transformation. Jiang et al. [11] proposed a Knowledge-Aware Embedding (KAE) method that leverages convolutional networks to fuse embeddings learned from transactions and external knowledge graphs.

These studies design neural networks-based frameworks to model the long, continuous purchase history of customers, establishing the potential of transaction data to reveal the demographics. By contrast, our work seeks to construct an effective method that learns from very limited shopping transaction records.

5.2 Heterogeneous Information Network

A *heterogeneous information network* (HIN) can capture the connectivities among nodes of multiple types in diverse semantic relationships. HINs have been applied to solve many real-world applications. For example, bibliographic data has been organized into HINs that consist of multi-typed objects, such as papers, authors, venues, and terms, to classify or cluster the objects into different research communities [9, 10, 31]. In the recommender system domain, user-item interactions have been formed into HINs to help the decision of making recommendations [4, 12, 19, 36].

More recently, network motifs [23] have been adopted to capture higher-order connectivity in complex real-world networks and provide richer network semantics. Recently studies have shown that incorporating network motifs into the learning of node embeddings yields better performance, compared to the widely-used meta-path patterns [21, 26, 29, 37]. Shang et al. [29] integrated motifs into text-rich networks for topic taxonomy construction. Mekala et al. [21] leveraged motifs for text classification in a weak supervision way.

In our work, we enrich the contexts of households by extracting combinatorial purchase patterns from the heterogeneous trip network with network motifs. We further design a novel motif context selection algorithm. This novel framework enables us to refine the contexts captured by network motifs, generating embeddings with label-indicative contexts.

6 CONCLUSIONS AND FUTURE WORK

We studied the problem of demographic inference from sparse grocery shopping records. We proposed DemoMotif, a graph-based demographic inference framework, and demonstrated its use in the inference of demographic information from purchase receipts. DemoMotif uses network motifs to enrich the context for households and incorporates a novel motif context selection algorithm to select label-indicative contexts for improved embedding learning. We have shown that DemoMotif can predict demographic attributes of households with satisfying accuracy, which provides indicative context for distinguishing different groups of households. We note that the choice of specific demographic information was strictly a function of the availability of data and ground truth to assess effectiveness. Such demographic information can surely raise legitimate privacy concerns in practice. This raises an important ethical issue that is not considered in this paper: under what circumstances is an inference appropriate in a legal or moral context? A methodology that can satisfactorily answer such imperative questions is essential to building necessary *awareness* that accompanies such work. As we leave such questions for future exploration, on a technical level, we are exploring means to further reduce the amount of data required to derive an accurate model, and privacy-preserving methods for mining users' transaction records.

We believe exploring the capability of inferring demographics with a limited number of context data (e.g., receipts) is a stepping stone for further investigation of the privacy-leakage problem in data mining. We are committed to preventing misuse of the model to predict sensitive user information. To this end, we will release the code for research use with an ethical license[3]. The pre-trained

[3]https://firstdonoharm.dev

embeddings will not be shared to prevent misuse. We will not release our datasets according to the term-of-use of the Nielsen data. However, access to the data can be applied through its official procedures.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 1999. User profiling in personalization applications through rule discovery and validation. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 377–381.

[2] Rick L Andrews and Imran S Currim. 2002. Identifying segments with identical choice behaviors across product categories: An intercategory logit mixture model. *International Journal of Research in Marketing* 19, 1 (2002), 65–79.

[3] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22nd international conference on World Wide Web*. 131–140.

[4] Bin Cao, Nathan Nan Liu, and Qiang Yang. 2010. Transfer learning for collective link prediction in multiple heterogenous domains. In *ICML*. Citeseer.

[5] Aron Culotta, Nirmal Kumar, and Jennifer Cutler. 2015. Predicting the demographics of twitter users from website traffic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[6] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 135–144.

[7] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla. 2014. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 15–24.

[8] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. 2007. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*. 151–160.

[9] Ming Ji, Jiawei Han, and Marina Danilevsky. 2011. Ranking-based classification of heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1298–1306.

[10] Ming Ji, Yizhou Sun, Marina Danilevsky, Jiawei Han, and Jing Gao. 2010. Graph regularized transductive classification on heterogeneous information networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 570–586.

[11] Yiwen Jiang, Wei Tang, Neng Gao, Ji Xiang, and Yijun Su. 2019. Demographic Prediction from Purchase Data Based on Knowledge-Aware Embedding. In *International Conference on Neural Information Processing*. Springer, 297–308.

[12] Jiarui Jin, Jiarui Qin, Yuchen Fang, Kounianhua Du, Weinan Zhang, Yong Yu, Zheng Zhang, and Alexander J Smola. 2020. An Efficient Neighborhood-based Interaction Model for Recommendation on Heterogeneous Graph. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 75–84.

[13] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, 137–142.

[14] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. " I know what you did last summer" query logs and user privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 909–914.

[15] Kirthi Kalyanam and Daniel S Putler. 1997. Incorporating demographic variables in brand choice models: An indivisible alternatives framework. *Marketing Science* 16, 2 (1997), 166–181.

[16] Raehyun Kim, Hyunjae Kim, Janghyuk Lee, and Jaewoo Kang. 2019. Predicting multiple demographic attributes with task specific embedding transformation and attention network. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 765–773.

[17] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference for Learning Representations (ICLR)*.

[18] Huaxin Li, Zheyu Xu, Haojin Zhu, Di Ma, Shuai Li, and Kai Xing. 2016. Demographics inference through Wi-Fi network traffic analysis. In *The 35th Annual IEEE International Conference on Computer Communications*. 1–9.

[19] Yuanfu Lu, Yuan Fang, and Chuan Shi. 2020. Meta-learning on heterogeneous information networks for cold-start recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1563–1573.

[20] Eric Malmi and Ingmar Weber. 2016. You are what apps you use: Demographic prediction based on user's apps. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 10.

[21] Dheeraj Mekala, Xinyang Zhang, and Jingbo Shang. 2020. META: Metadata-Empowered Weak Supervision for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8351–8361.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013), 3111–3119.

[23] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.

[24] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[25] Yehezkel S Resheff and Moni Shahar. 2017. Fusing multifaceted transaction data for user modeling and demographic prediction. *arXiv preprint arXiv:1712.07230* (2017).

[26] Aravind Sankar, Xinyang Zhang, and Kevin Chen-Chuan Chang. 2019. Meta-GNN: Metagraph neural network for semi-supervised learning in attributed heterogeneous information networks. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 137–144.

[27] Suvash Sedhain, Scott Sanner, Darius Braziunas, Lexing Xie, and Jordan Christensen. 2014. Social collaborative filtering for cold-start recommendations. In *Proceedings of the 8th ACM Conference on Recommender systems*. 345–348.

[28] Jingbo Shang, Meng Qu, Jialu Liu, Lance M Kaplan, Jiawei Han, and Jian Peng. 2016. Meta-path guided embedding for similarity search in large-scale heterogeneous information networks. *arXiv preprint arXiv:1610.09769* (2016).

[29] Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. 2020. Nettaxo: Automated topic taxonomy construction from text-rich network. In *Proceedings of The Web Conference 2020*. 1908–1919.

[30] Clemens Stachl and Markus Bühner. 2015. Show me how you drive and I'll tell you who you are recognizing gender using automotive driving parameters. *Procedia Manufacturing* 3 (2015), 5587–5594.

[31] Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu. 2009. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. 565–576.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[34] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2016. Your cart tells you: Inferring demographic attributes from purchase data. In *Proceedings of the ninth ACM international conference on web search and data mining*. 173–182.

[35] Chuhan Wu, Fangzhao Wu, Junxin Liu, Shaojian He, Yongfeng Huang, and Xing Xie. 2019. Neural demographic prediction using search query. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. 654–662.

[36] Jixing Xu, Zhenlong Zhu, Jianxin Zhao, Xuanye Liu, Minghui Shan, and Jiecheng Guo. 2020. Gemini: A Novel and Universal Heterogeneous Graph Information Fusing Framework for Online Recommendations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3356–3365.

[37] Carl Yang, Yichen Feng, Pan Li, Yu Shi, and Jiawei Han. 2018. Meta-graph based hin spectral embedding: Methods, analyses, and insights. In *2018 IEEE International Conference on Data Mining (ICDM)*. 657–666.

[38] Xin Wayne Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. 2014. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1935–1944.

[39] Erheng Zhong, Ben Tan, Kaixiang Mo, and Qiang Yang. 2013. User demographics prediction based on mobile data. *Pervasive and mobile computing* 9, 6 (2013), 823–837.

[40] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. 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*. 295–304.