Birds of a Feather Purchase Together: Accurate Social Network Inference using Transaction Data

Jiaxing Shen*, Yulin He†, Yunfei Long‡, Jiaqi Wen§, Yanwen Wang¶, Yu Yang‖
*Lingnan University; ‡University of Essex; §The University of Auckland; ¶Hunan University;
†Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ); ‖The Hong Kong Polytechnic University

***Abstract*—Social networks play a crucial role in providing valuable contextual information across disciplines and applications. However, the unobservable nature of physical-world social connections has led to the development of social network inference. Existing approaches rely on co-occurrences and universal thresholds to infer social networks from spatiotemporal data. Yet, these methods suffer from two limitations: disregarding individual social preferences and failing to address "familiar strangers". Our analysis reveals that relying solely on common spatiotemporal data is inadequate for accurate social network inference. Fortunately, the availability of extensive transaction data, encompassing spatiotemporal and consumption information, presents an opportunity. Our approach involves integrating individuals' lifestyles with co-occurrences, driven by the fact that different lifestyles entail distinct social preferences and that true friends share similar lifestyles. However, we face two significant challenges: flexible extraction of lifestyle features and personalized threshold setting. To overcome these challenges, we propose nonparametric methods applicable to various scenarios and leverage domain knowledge for threshold determination. Evaluation on a real dataset of over $2,000$ individuals demonstrates an impressive improvement of over 20% in F1-score compared to the baselines.**

***Index Terms*—social network inference, physical social networks, transaction data, feature extraction**

I. Introduction

Social networks serve as the foundation for various disciplines, from epidemiology to sociology [1], [2]. They can be categorized into physical social networks (PSN) in real-life and online social networks (OSN) in cyberspace. PSN primarily have unobservable social ties, while OSN platforms like Facebook allow direct observation of social links [3]. Research indicates that only 22% of individuals reported complete overlap between their PSN and OSN connections, highlighting significant structural differences [4], [5].

Although social ties in PSN are unobservable, they can be inferred from various behavioral indicators, a process known as *PSN inference* [6], [7]. PSN inference benefits real-world applications by extending their scope and improving performance. It enables group-level applications such as group recommendation and collective behavior analysis [8], [9]. Additionally, it facilitates a more comprehensive understanding of human behaviors and more effective modeling of preferences [8], [10]. For instance, knowledge of the network structure of a target society aids accurate modeling of infectious disease diffusion processes [11].

To infer PSNs, researchers have explored various spatiotemporal data sources, including check-in histories [12], geo-tagged photos [13], and smartphone proximity information [7]. These approaches rely on the assumption that individuals appearing in the same place simultaneously (co-occurrence) may have a latent social relationship. Co-occurrence networks are used to approximate social networks in two steps. First, links between individuals are established based on the significance of their co-occurrences, with larger weights indicating closer relationships. Second, strangers who sporadically encounter each other are filtered out by comparing their link weight with a predefined threshold.

However, existing approaches are often primitive and inaccurate in addressing individual differences and familiar strangers. Individuals have varying social preferences, with some favoring a large number of ordinary friends while others prefer a few close friends [14]. Consequently, it is impossible to filter out strangers using a universal threshold since the same weight carries different meanings for different individuals. "Familiar strangers," characterized by regular and frequent co-occurrences due to shared daily routines, pose another challenge [15], [16]. Examples include students living in the same dormitory or employees working in the same office building. Familiar strangers are undesired and detrimental in marketing and epidemiological scenarios as they lack intentional contact or information exchange, unlike friends [15]. Although the phenomenon has been recognized for decades, it is largely neglected in the literature. Our analysis reveals that identifying familiar strangers solely based on co-occurrences is intractable, rendering most existing approaches ineffective.

The proliferation of e-payment has resulted in the accumulation of substantial *smart card transaction data (SD)*, presenting an opportunity for accurate PSN inference. SD comprises digital records of consumption events in the physical world, such as credit card and campus card transactions. In addition to spatiotemporal information, SD contains details about the total amount and content of the consumption event. This work proposes an accurate PSN inference framework based on SD. To address individual differences and familiar strangers, our key idea is to identify lifestyles from SD, such as chronotypes. This approach is motivated by two factors. Firstly, individuals with different chronotypes exhibit different social preferences [14], and this information can help personalize the threshold setting, partially mitigating individual differences. Secondly, real friends not only frequently co-appear but also have similar lifestyles [6], [17], [18]. Familiar strangers can be addressed by appropriately fusing proximity and lifestyle similarities.

However, applying our vision to real-world conditions poses several challenges. First, ground truth availability is a challenge, as many existing works use OSN data from platforms like Gowalla [12] and Flickr [13] as ground truth. Although OSN and PSN overlap to some extent, they exhibit considerable differences, as explained earlier. Second, extracting effective lifestyle features with complicated parameter settings may be infeasible in different scenarios or for different individuals. For example, calculating the regularities of dietary behaviors requires specifying meal times, which may not be clearly defined for certain individuals. Lastly, finding an appropriate threshold for filtering strangers, particularly personalized thresholds for different individuals, is challenging.

To overcome these challenges, we adopt and revise the approach of treating inferred networks as data representations for specific tasks and evaluate network inference based on task performance, as practiced in Reference [3]. Our work consists of an inference component for network construction and a task component for evaluation. Second, we develop nonparametric methods to extract lifestyle features like regularity and chronotypes that are flexible enough to be applied to any scenario. The effectiveness of these extracted features is implicitly validated through quantitative analysis, which aligns with findings in the literature. Lastly, we leverage domain knowledge to measure the goodness of threshold settings and propose a search-based scheme to find optimal personalized thresholds for different individuals.

Through evaluation on a campus card dataset of over 2,274 students spanning three months and five predictive tasks, our approach significantly outperforms baseline methods, achieving an average F1-score improvement of over 20%.

Our main contributions can be summarized as follows:

- We propose an accurate PSN inference framework based on transaction data, addressing the challenges of familiar strangers and individual differences by extracting lifestyle features and fusing them with proximity features.
- We develop novel methods for extracting lifestyle features that are applicable to various scenarios, and their effectiveness is implicitly validated through quantitative analysis.
- We evaluate the proposed inference framework on a large-scale real dataset and conduct various analyses, including the verification of familiar strangers and the examination of social homophily in lifestyles.

The remainder of this paper is organized as follows. Section II elaborates on the feature extraction, while Section III describes the feature fusion process. Section IV presents the experimental evaluation results, followed by a discussion of related works and a conclusion in the final sections.

II. Extracting Proximity Features and Lifestyle Features

A. Proximity Features

Co-occurrence is usually defined as two individuals appearing close to each other within a very short period of time. The definitions of how close in both space and time vary across different scenarios. In this work, we empirically define *co-occurrence* as having transactions in merchants of the same building within 1 minute. *Co-occurrence network (CN)* is a graph with nodes representing individuals and links between nodes showing the frequency of co-occurrences. CN is extensively adopted to represent PSN for both humans and animals [12], [19]–[21]. The main focus of CN construction is to separate real co-occurred friends from strangers that occasionally encountered. To this end, there are two essential steps: measuring the importance of the co-occurrences and finding the appropriate threshold.

Algorithm 1: CN construction

Input : SD - a list of smart card transaction data
Assume: Within SD, there are a sorted list of students $\mathbf{S}$, a list of merchants $\mathbf{M}$, a list of locations $\mathbf{L}$.
SD have K days, each day is split into a list of time slots $\mathbf{T}$.
Output : $\mathcal{A}$ - a adjacent matrix of the co-occurrence network

Initialize $\mathcal{A}$ with a zero matrix.
// LC is a list of transaction clusters. A cluster represents a co-occurrence event.
LC $\leftarrow$ group SD by days, locations, and time slots
// $C \in LC, C \subset SD$. Each C is associated with a day k, a location $l \in \mathbf{L}$, a time slot $t \in \mathbf{T}$, and a subset of students $\mathbf{S}_C \subseteq \mathbf{S}$, i.e., $C = LC[k, l, t]$.
foreach *cluster $C \in LC$* **do**
  // Find the largest size of cluster across different locations.
  $\text{msc} \leftarrow \max_{C' \in \{LC[k,l',t] \mid l' \in \mathbf{L}\}} |\mathbf{S}_{C'}|$
  // $\mathcal{A}_{ij}$ measures the importance of co-occurrence of individuals i and j.
  $\mathcal{A}_{ij} \leftarrow \mathcal{A}_{ij} + 1 - |\mathbf{S}_C|/\text{msc}, \quad i, j \in \mathbf{S}_C$
end
$\mathcal{A} \leftarrow \mathcal{A}/\mathbf{K}$; *// Average by the number of days*

Measure the importance of co-occurrences: To measure the importance, frequency-based approaches were initially proposed which are direct yet naive. The methods simply count the number of unique locations individuals co-occurred or sum up the number of co-occurrences across different locations. However, both measures consider different locations equally important leading to overestimated social strength, especially for co-occurred strangers. For example, 5 random encounters at a crowded student restaurant are considered 5 times more important than a private dinner at an unpopular bar. If the frequency (number of co-occurrences) is considered only, we may derive a wrong conclusion about relation strength.

The drawback of the frequency-based approach could be attributed to the fact that some restaurants are popular that most people will go during dinner time. The observation of co-appearance in popular restaurants may thus not be a strong indication of social connection. On the contrary, a few co-occurrences in unpopular places are perhaps a better indication of friendship. According to this, an entropy-based approach has been proposed to measure the diversity of co-occurrences [12]. It quantifies each social strength by considering how diverse the distribution of the co-appearance is in the context of locations.

We borrow a similar idea from the entropy-based approach to measure the importance of co-occurrences by considering the probability in both space and time. When a small-

probability event occurs like two individuals co-appeared in an unpopular place in midnight, it bears more information and should be treated with more importance. Algorithm 1 illustrates how CN is constructed with SD. The output is an adjacent matrix $\mathcal{A}$. We first find out all co-occurrence events by grouping SD into many transaction clusters according to the day, the location, and the time slot (1-minute time window). For example, the transaction generated between $12{:}00$ and $12{:}01$ of day k in location l is regarded as a co-occurrence event. Students within an event are assumed to have a latent social relationship. The relation strength is calculated using a simple yet effective rule: the more students co-appeared in an event, the smaller strength is accumulated to their relationships. This rule takes the popularity of both time and location into consideration.

We constructed a CN based on the campus card dataset (refer to Section IV-A for more details) using Algorithm 1. We also analyzed the distribution of edge weights which turns out to obey a power-law distribution. Power-law distributions are commonly identified in many human dynamics research [22], [23]. It implies the majority of weights are within a very small range with a long tail of high weights. Given these results, it remains challenging to address individual differences and familiar strangers.

Find appropriate thresholds: Although we have measured the importance of co-occurrences, noises (junk links) are still inevitable. It is necessary to apply an appropriate threshold to filter out those junk links. Given a constructed CN in an adjacent matrix format $\mathcal{A}$, it could also be formulated as an undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$ where $\mathbf{V}$ is a set of vertices, $\mathbf{E}$ is a set of links between vertices, and $\mathbf{W}$ indicates the weight of those links. A predefined threshold θ alters the network into $\widetilde{\mathbf{G}} = (\mathbf{V}, \mathbf{E}', \mathbf{W}')$ (corresponds to $\widetilde{\mathcal{A}}$) where $\mathbf{W}'$ and $\mathbf{E}'$ are defined in Equation 1.

$$w'_{ij} = \begin{cases} w_{ij}, & \text{if } w_{ij} \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{1}$$
$$\mathbf{W}' = \{w'_{ij}\}, \quad \mathbf{E}' = \{(i,j) | w'_{ij} \neq 0\}$$

Given different thresholds, the network structure would change dramatically. Small thresholds are unable to filter out normal strangers let alone familiar strangers. When large filtering thresholds are applied, CN naturally transforms into many small groups. To conclude, proximity features are difficult to filter strangers out let alone familiar strangers. The reasons are as follows. First, proximity features are incapable of capturing fine-grained social relationships. Second, it is hard to find the appropriate threshold. However, it might be a good option to study social networks of close relationships using co-occurrence networks with large filtering thresholds.

B. Lifestyle Features

Due to the limitations of proximity features, we explored lifestyle features to address individual differences in social interactions and familiar strangers. Certain lifestyles are associated with particular social interaction patterns [14]. For example, evening-active people tend to make more friends than morning-active people. Therefore, identify lifestyles could help to address individual differences. Evidence has shown that friendships are also reflected in lifestyles. Synchronized circadian rhythm indicates latent social relationships [18]. Distances of individuals in the behavior space, consumption behaviors included, can be used to estimate friendships [6]. A combination of lifestyle and proximity could alleviate the negative impact of familiar strangers.

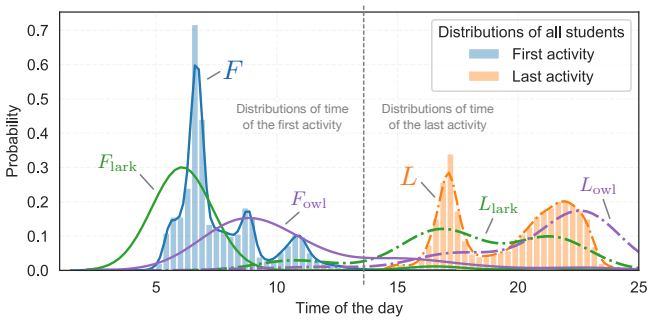


Figure 1: Illustration of chronotype detection using distributions of the first activity and the last activity.

Consumption: The consumption information is inherent with SD. It contains rich features related to consumption habits, especially food preferences. The following features include three aspects that are mostly based on a daily frequency. 1) For consumption, we consider the total amount of money, the frequency of transactions. 2) For locations, we focus on the number of visited locations. More specifically, the diversity of location visitation patterns is examined with Shannon entropy. 3) For the time, the number of active hours and the diversity of time patterns are considered. 4) For Points of Interest (PoIs), the total number of PoIs and the diversity are taken into consideration. For most of the features, we further calculate their mean values and standard variations over the whole period of three months. Besides, there are large gaps in behaviors between working days and free days. Therefore, the basic consumption features consist of "work/free" $\times$ "mean/std" $\times$ "aspects". For example, the feature "work_mean_consumption" refers to the average amount of consumption in working days.

Chronotype: Life on Earth follows a circadian rhythm, a 24-hour internal clock, reflected at the physiological, biochemical, and psychological levels. Even though human beings are endogenously controlled by an internal circadian clock, there are individual differences in how the clock is synced with the environment's daily rhythm [14]. These differences can be classified into three chronotypes. At the two extremes are the morning-active people ("larks") and the evening-active people ("owls"), and the rest is intermediate people whose rhythms do not deviate too much from the population average. Larks wake up and sleep early while owls wake up and sleep late. Although a person's chronotype can change, it is relatively stable within a few years.

Conventional ways of measuring chronotype are ques-

tionnaires like the Morningness-Eveningness Questionnaire (MEQ) [24]. However, there are no absolute criteria for any given chronotype [14]. Recently, emerging modalities like smartphones are used as a proxy to infer chronotype. Since SD keeps recording our daily activities, it could also be an indicator of chronotype. The first activities happened early in the morning means the cardholder got up early. While the last activities that happened late in the night indicate the cardholder sleep late.

The proposed chronotype detection aims to find people who go to bed early and get up early or both late. It mainly consists of three steps. First, find two distributions of the time of the first activity F and the last activity L for all individuals. Second, for individual i, find the two distributions for him: F_i and L_i. Third, compare the corresponding distributions using Mann–Whitney U test and get the chronotype as shown in Equation 2.

$$\begin{cases} \text{Morningness (Lark)} : (F_i < F) \wedge (L_i < L) \\ \text{Eveningness (Owl)} : (F_i > F) \wedge (L_i > L) \\ \text{Intermediate} : \text{other situations} \end{cases} \tag{2}$$

Mann–Whitney U test is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. The advantages are two folds. The first advantage is parameter-free which automatically adapts to different situations. Second, unlike the t-test, it does not require the assumption of normal distributions. Figure 1 shows the distributions F and L from all students and two examples of a lark and an owl. Notably, only SD from Monday to Thursday was used in the analysis. It is suggested that individuals may have different behaviors during weekdays and weekends, and the extent of these differences may vary from one chronotype to another [14].

The detection result using the campus card dataset indicates there are 15.3% larks and 14.6% owls among all students, which is close to the reported percentages (20%, 20%) in the literature [14], [25]. Among intermediate students, the percentages of males (50.7%) and females 49.3% are close to each other. However, 37.3% of larks are males and 62.7% of larks are females. While in owls, the numbers are 74.3% and 25.7% for males and females, respectively. This finding shows that male students are much more likely to be evening-active than female students. We also identified chronotypes have certain relationships with GPA as depicted in Figure 2. First, there is a high correlation between the academic performance of the two semesters. Second, for both semesters, the academic performance follows the order: Morningness $>$ Intermediate $>$ Eveningness. One of the reasons is that most exams happened in the morning which is not good for owls [14].

Regularity: Regularity refers to the predictability of biological and behavioral patterns. It is an important aspect of the internal circadian clock. Previous work derived regularity by calculating the entropy of activities like having breakfast and taking a shower [26]. However, it is nontrivial to clearly identify the time of certain activities since individuals have varied lifestyles. Even for the same person his behavior changes from time to time. For example, some people may have brunch instead of breakfast or even do not have the habit of breakfast. This makes identifying the time of specific activities difficult and inaccurate.

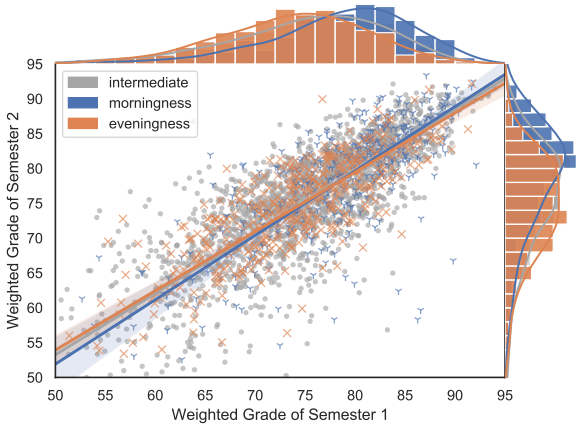


Figure 2: Relationship between chronotypes and academic performance: Morningness $>$ Intermediate $>$ Eveningness.

To address the concern, we propose a nonparametric method that looks at the regularity of activities as a whole using Singular Value Decomposition (SVD). It consists of three steps. First, we extract a consumption matrix M^k for student k where M^k_{ij} represents the total amount of money consumed in the i-th hour of the j-th day. In a similar way, we extract a consumption matrix from dietary transactions only. Then, the dietary consumption matrix is transformed into a binary dietary matrix N^k, where $N^k_{ij} = 1$ means student k had dietary-related transactions during the i-th hour of the j-th day. Lastly, we apply SVD to decompose $N \in \mathbb{R}^{m\times n}$ into three submatrices $U \in \mathbb{R}^{m\times k}, S \in \mathbb{R}^{k\times k}, V \in \mathbb{R}^{n\times k}$. The rationale is that the higher regularity an individual's dietary behaviors, the less nonzero eigenvalues ($S' = \{s_i | s_i > 0.001, s_i \in S\}$). To avoid the negative effect of sparse dietary matrices which also have few nonzero eigenvalues due to the sparsity, we use the difference in ratios of the first and the last nonzero eigenvalues as illustrated in Equation 3. The advances of the proposed approach are robustness and interpretability without identifying the exact time of certain activities.

$$\text{eig\_diff} = (\max(S') - \min(S'))/\sum S' \tag{3}$$

According to the analysis of the campus card dataset, the distributions of regularity for both genders are quite similar and resemble a normal distribution $N(0.25, 0.06)$. However, there are significant differences between students in different grades. Figure 3 reveals an interesting effect that the regularity of meals from high to low follows the order: seniors (admission year: 2014) $>$ juniors $>$ sophomores $>$ freshmen indicating students' dietary routines become more stable over the years on the campus.

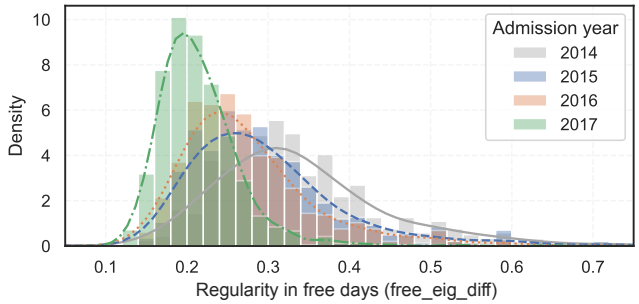


Figure 3: Regularity of students from different grades.

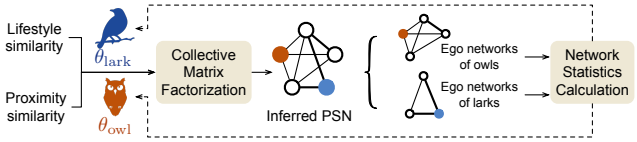


Figure 4: An illustration of the fusion process.

To sum up, we devise and extract three types of lifestyle features including consumption behavior, chronotype, and regularity. The common advantage of our extraction methods is parameter-free which could be applied to different scenarios. Although the effectiveness of those extracted features is not directly measured, the quantitative results of feature analytics align with the findings in the literature.

III. Fusing Proximity and Lifestyle

In this section, we fuse proximity and lifestyles to accurately infer PSN. The fusion process is illustrated in Figure 4. First, we measure similarity using proximity and lifestyle respectively. Then junk links are filtered out in both similarity matrices using different thresholds for people of different chronotypes. Collective matrix factorization is used to fuse both filtered similarities to derive an inferred PSN. We identify ego networks of larks and owls respectively to calculate their network statistics. The discrepancy between statistics of the inferred PSN and that of real networks indicates the goodness of the thresholds. The goodness serves as a criterion to search for the optimal threshold setting.

Measure similarity and filter junk links: As introduced in Section II, $\mathcal{A}$ is already a similarity matrix of proximity. We also have a list of lifestyle features. Directly combining them will spoil the structure of CN and add more noises since individuals with similar lifestyles could also be strangers. A more appropriate way is to use the proximity-based CN as the main structure and refine the social strength with lifestyle features. To this end, we constructed a similarity matrix of lifestyles $\mathcal{F}$ where $\mathcal{F}_{ij}$ is the cosine similarity of lifestyles between i and j.

Then to filter out noises or junk links, we updated $\mathcal{A}$ and $\mathcal{F}$ according to personalized thresholds. Equation 4 shows how the thresholds are applied for $\mathcal{A}$. The conditions are applicable for $\mathcal{F}$ as well.

$$\widetilde{\mathcal{A}}_{ij} = \begin{cases} \mathcal{A}_{ij}, \text{ if } \mathcal{A}_{ij} \geq \min(\Theta(i), \Theta(j)) \\ 0, \text{ otherwise} \end{cases}$$
$$\Theta(k) = \begin{cases} \theta_{\text{owl}}, \text{ if } k \in \mathbf{V}_{\text{owl}} \\ \theta_{\text{lark}}, \text{ if } k \in \mathbf{V}_{\text{lark}} \\ (\theta_{\text{lark}} + \theta_{\text{owl}})/2, \text{ otherwise} \end{cases} \tag{4}$$

There are three threshold settings: θ_{lark} for morningness people, θ_{owl} for eveningness people, and $(\theta_{\text{lark}} + \theta_{\text{owl}})/2$ for intermediate people. Set $\mathbf{V}_{\text{owl}}$ and set $\mathbf{V}_{\text{lark}}$ refer to individuals that are identified as owls and larks, respectively.

Collective matrix factorization: Given two matrices $\widetilde{\mathcal{A}}, \widetilde{\mathcal{F}} \in \mathbb{R}^{n \times n}$, we resort to nonnegative collective matrix factorization to derive three submatrices $\mathcal{W}, \mathcal{H}, \mathcal{P} \in \mathbb{R}^{n \times k}$ so that $\widetilde{\mathcal{A}} \approx \mathcal{W}\mathcal{H}^\intercal$ and $\widetilde{\mathcal{F}} \approx \mathcal{P}\mathcal{H}^\intercal$ with the objective:

$$\min_{\mathcal{W},\mathcal{H},\mathcal{P}} \frac{1}{2}\Big[\gamma\Big\|\widetilde{\mathcal{A}} - \mathcal{W}\mathcal{H}^\intercal\Big\|_F^2 + (1-\gamma)\Big\|\widetilde{\mathcal{F}} - \mathcal{P}\mathcal{H}^\intercal\Big\|_F^2 + \eta(\|\mathcal{W}\|_F^2 + \|\mathcal{P}\|_F^2)\Big] \text{ s.t. } \mathcal{W}, \mathcal{H}, \mathcal{P} \geq 0, \tag{5}$$

where parameter $\gamma \in [0, 1]$ weighs the relative importance of two input matrices $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{F}}$. Parameter $\eta > 0$ controls the size of the elements of $\mathcal{W}$ and $\mathcal{P}$. It is usually determined by the largest element of input matrices. Although the function is a non-convex, it is convex separately in each factor, i.e., finding the optimal factor $\mathcal{W}$ corresponding to fixed factors $\mathcal{H}$ and $\mathcal{P}$ reduce to a convex optimization problem. Algorithms based alternating nonnegative least squares are often used for this purpose. More details about solving the optimization and finding parameters and η can be found in references [27], [28].

Once $\mathcal{H}$ is derived, we construct the final adjacent matrix like the way we get $\mathcal{F}$. The non-negativity constraint is to ensure the values in the decomposed matrix make sense. Besides, the main advantage of collective matrix factorization is the capacity of considering the latent association between proximity and lifestyle matrices by factorizing them simultaneously.

Measure the goodness of thresholds: Different threshold settings affect the network structure dramatically. To ensure the thresholds are appropriate, we resort to domain knowledge of real ego networks of larks and owls in the literature [14]. Ego networks (or personal networks) are local networks with one central node, known as the ego. The network is based on the ego and all other nodes directly connected to the ego are called alters. For real social networks, the ratios of owls over larks in ego network size and social tie strength are around $7:5$ and $4:5$, respectively.

For each setting of thresholds, we could calculate the average network size and the average link weight for both larks and owls. The statistics of real social networks serve as criteria for searching the optimal threshold setting. It is worth noting that, multiple settings of threshold might be close to the criteria. We facilitate the search by imposing additional conditions like $\theta_{\text{owl}} < \theta_{\text{lark}}$.

IV. Experimental Evaluation

After finishing network inference component, then we focus on the task evaluation component. As explained, we use attribute prediction to evaluate the effectiveness of the inferred PSN. The inferred PSN is regarded as a data representations for different predictive tasks including major, admission year, class, gender, and GPA levels. The results of the tasks are utilized to demonstrate the performance of those inferred PSN.

This component consists of two steps: extract numerical features from PSN and train machine learning models to predict different attributes. For the first step, instead of using conventional graph features, we apply a popular network embedding method (node2vec [29]) to learn low-dimensional representations for nodes in the graph by optimizing a neighborhood preserving objective. There are two hyperparameters p and q controlling the transition probability of the walk. We also use the same default parameter setting ($p = 0.25$ and $q = 0.25$) for different PSN. For the second step, we adopt classic machine learning models including Random Forest, Neural Networks, and Support Vector Machine (SVM) which are implemented with Python Library scikit-learn[1] (Version 0.22.1). All the models use the same default parameter setting for all approaches and tasks. For example, the parameter $n\_neighbors$ of Nearest Neighbors is 5.

In this section, the experimental settings are firstly introduced including the dataset, baseline approaches, and evaluation metrics. Then we show the evaluation results of the proposed approach and conduct some further analysis.

A. Experimental Settings

The campus card dataset: In this work, we use campus card transaction data to validate the proposed research framework. It is contributed by a Chinese university involving 2,274 students from 6 faculties ranging from freshmen to seniors. For each faculty, students are attached to different classes of different majors. Their campus cards could be used to cover the majority of daily expenses within campus including three meals a day, snacks and drinks, and payment in the hospital and library.

During 3 months, starting from 1 Oct 2017 to 31 Dec 2017, a total number of 633,180 transaction records have been accumulated. Besides, there are 86 points of consumption in the merchant table including different kinds of shops and university facilities like the library and the hospital. Over half of the transactions are food-related as indicated in Table I. The student table has both demographic information (age, gender) and academic information (major, class, and performances of two semesters) of the students. We use both information as ground truth to verify the results of our approaches. Students' identity information has been carefully processed with proper anonymization to reduce the risk of privacy leakage.

Baseline approaches: To recap, two of our contributions are personalized thresholds to address individual differences and the fusion of lifestyle features to address familiar strangers. To demonstrate their effectiveness, two baseline approaches have been proposed as depicted in Figure 5. P_A is a conventional way of PSN inference based on proximity information using a single threshold. PL_A combines proximity and lifestyle information while only uses a single threshold. PL_M is our proposed approach that uses personalized thresholds and fuses proximity and lifestyles. By comparing P_A and PL_A, the effectiveness of lifestyle features could be revealed. By comparing PL_A and PL_M, we verify the effectiveness of personalized thresholds.

[1]Scikit-learn, https://scikit-learn.org/

Food	Beverages	Hot water	Hospital	Library †	Top-up	Electricity
51.29%	31.17%	17%	0.27%	0.15%	0.07%	0.05%

† The transactions in library refer to payments for losing borrowed books.

Table I: Percentage of transaction data from different PoIs.

Approach	Features	Parameter Setting	
P_A	Proximity	A single threshold	§
PL_A	Proximity + Lifestyles	A single threshold	§
PL_M	Proximity + Lifestyles	Multiple thresholds	†

§ : Evaluate lifestyle features † : Evaluate parameter setting

Figure 5: Configurations of baseline approaches.

Evaluation metrics: Since all the prediction tasks are essentially multi-label classification problems, we use F1-score to evaluate the performance.

$$\begin{cases} \text{precision}(p) = \frac{\mathbf{tp}}{\mathbf{tp}+\mathbf{fp}} \\ \text{recall}(r) = \frac{\mathbf{tp}}{\mathbf{tp}+\mathbf{fn}} \\ \text{F1-score} = 2 \cdot \frac{p \cdot r}{p+r} \end{cases}$$

Prediction \ Truth	X	$\bar{X}$
X	tp	fp
$\bar{X}$	fn	tn

X: Target label
$\bar{X}$: Non-target label

Parameter selection: The proposed approach only requires setting γ that controls the relative importance of proximity and lifestyles which is scenario-dependent. Since γ is also require in baseline approaches, we experimentally set $\gamma = 0.4$ that is a mild setting for all approaches.

Although the parameter θ is not required in our approach, it is necessary for P_A and PL_A. It controls the minimum link weight in the inferred PSN. In many existing works, this parameter is empirically set. Here are the rules of thumb. If θ is set too large, we will derive a sparse graph that reveals strong homophily. The prediction of attributes is thus quite extreme under this situation [3]. For nodes that are connected to others, the prediction is of high accuracy. On the contrary, there are more isolated nodes that are disconnected from others. The prediction of attributes of those nodes is like a random guess which results in poor performance. If θ is too small, it will lose its original meaning which is to filter out junk links generated by co-occurrences of strangers. Under these considerations, we experimentally set $\theta = 0.1$ which can achieve relatively better performance for P_A and PL_A.

B. Evaluation Results

All the experimental results are derived from 10-fold cross-validation and are summarized in Table II, including 5 attribute

Method\Task	Admission Year			Gender			Major			GPA Level			Class		
	P_A	PL_A	PL_M	P_A	PL_A	PL_M	P_A	PL_A	PL_M	P_A	PL_A	PL_M	P_A	PL_A	PL_M
AdaBoost	0.403	0.69	**0.723**	0.651	0.745	**0.773**	0.37	**0.526**	0.515	0.268	**0.331**	0.328	0.019	0.064	**0.069**
Decision Tree	0.319	0.587	**0.642**	0.587	0.661	**0.698**	0.294	**0.52**	0.513	0.247	0.305	**0.32**	0.071	**0.231**	0.217
Linear SVM	0.524	0.816	**0.864**	0.682	0.762	**0.862**	0.455	**0.706**	0.686	0.282	0.328	**0.356**	0.286	0.647	**0.684**
Naive Bayes	0.514	**0.726**	0.699	0.676	0.672	**0.731**	0.442	**0.668**	0.642	0.29	0.327	**0.336**	0.271	**0.611**	0.605
Nearest Neighbors	0.388	0.805	**0.86**	0.631	0.792	**0.854**	0.378	0.737	**0.769**	0.264	0.315	**0.345**	0.16	0.568	**0.649**
Neural Net	0.498	0.845	**0.905**	0.635	0.815	**0.864**	0.422	0.738	**0.748**	0.276	0.335	**0.341**	0.241	0.63	**0.643**
Random Forest	0.318	0.682	**0.761**	0.675	0.707	**0.772**	0.323	0.592	**0.606**	0.253	0.317	**0.342**	0.077	0.267	**0.305**
RBF SVM	0.565	0.888	**0.912**	0.613	0.807	**0.89**	0.459	0.755	**0.77**	0.279	0.315	**0.36**	0.307	0.67	**0.697**

Table II: F1-score of five predictive tasks of all approaches on different machine learning models. Bold text represents the best of three approaches on a certain learning model. Underlined text highlights the best performance among all learning models.

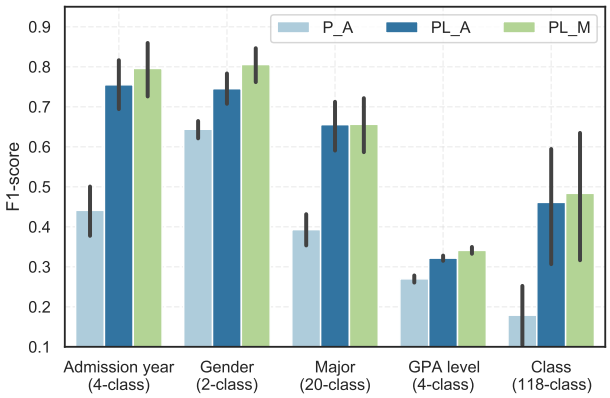
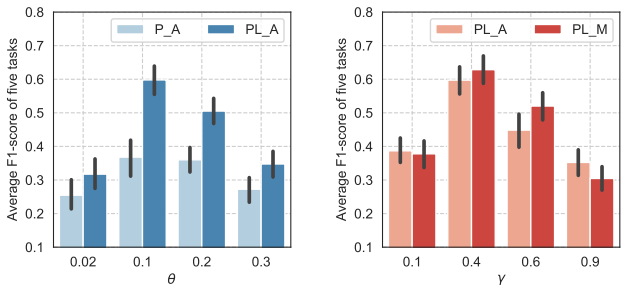


Figure 6: Average F1-score of five prediction tasks.



(a) F1-score vs θ with $\gamma = 0.4$. (b) F1-score vs γ with $\theta = 0.1$.

Figure 7: F1-score under different parameter settings.

prediction tasks, 8 learning models, and 3 approaches. Tasks are separated by different background colors and each of them consists of three columns of different approaches. For every machine learning model of a certain task, we marked the best performance out of three approaches with bold text. For all tasks, we marked the best performance among all machine learning models with an underline. Take admission year prediction as an example, the best performance on Radial Basis Function (RBF) kernel SVM is 0.912 achieved by PL_M which is also the best among all learning models.

Figure 6 gives a more intuitive view of the results, the performance follows the order: PL_M > PL_A > P_A. On average, PL_M outperforms PL_A by 4.9% and PL_A outperforms P_A by 66.1%. This proves the effectiveness of our contributions, especially lifestyle features since PL_A significantly outperforms P_A in prediction tasks of admission year (71.1% performance gain), major (66.8%), and class (157.5%).

P_A assumes real friends will co-appear regularly while strangers only have occasional encounters. It identifies social relationships based on the strength of co-occurrences. However, as explained, familiar strangers also have regular and frequent co-occurrences due to similar daily routines. P_A suffers from the negative impacts of familiar strangers. While combining with lifestyle features, the situation becomes much better since lifestyles will not be affected by familiar strangers. Therefore, fusing proximity and lifestyles could significantly improve the task performance indicating a more accurate underlying PSN. Besides, except for major prediction, PL_M has an average improvement of 6.1% over PL_A. This indicates the superiority of personalized thresholds.

Different difficulty levels of prediction tasks: According to Table II, we found the tasks are of different levels of difficulty. It is not only related to the number of classes but also the strength of the association between input features and the task labels. The most difficult task is GPA level prediction with 4 classes rather than the class prediction which has over 100 types of labels. The underlying reason is academic performance is neither strongly associated with social relationships nor consumption behaviors. Despite the varying difficulty levels, our proposed approach (PL_M) still achieves the best performance.

The impact of θ and γ: Figure 7 (a) and (b) illustrate the F1-score of different approaches under different θ and γ respectively. With γ fixed to 0.4, the average F1-score of five predictive tasks peaks when $\theta = 0.1$ for p_A and PL_A. The parameter θ is supposed to filter out junk links. Large values of θ mostly retain strong relationships of proximity. People connected under this situation are of high homogeneity which contributes to the high performance of prediction. However, it also results in more isolated nodes which are difficult to predict their attributes. While small values of θ contain various familiar strangers which also degrades the performance.

When $\gamma = 0.4$ the performance of both approaches peak (with θ of PL_A fixed to 0.1). The parameter γ controls the relative importance of proximity and lifestyle. Since the lifestyle features have been updated with the proximity information as shown in Equation 4, it bears relatively more information than

the proximity information. The performance under extreme situations ($\gamma = 0.1$ and $\gamma = 0.9$) are significantly worse. It indicates the single source of information can hardly achieve the best performance.

Homophily and familiar strangers: Homophily is one of the fundamental organizational principles of human societies. It has a number of important social implications such as the origin of segregation or the perpetuation of economic inequality and social immobility [30]. We examine the homophily of different aspects of students using the assortativity coefficient [31]. The coefficient r could tell in a concise fashion how vertices of different types are preferentially connected among themselves. The value of 1 of the coefficient indicates the strongest homophily. Equation 6 shows the definition of r where e_{ij} is the fraction of edges from a vertex of type i to a vertex of type j.

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}, \quad a_i = \sum_j e_{ij}, \quad b_j = \sum_i e_{ij} \tag{6}$$

The homophily of grade, gender, class, age, GPA level, and node degree under different thresholds are illustrated in Figure 8 using PL_A. The shadow of each line represents variance which is created via Jackknife resampling. There is an increasing trend of homophily for all aspects with the increase of thresholds, especially for grade, class, and gender. This indicates close friends are more similar in those aspects which are also reported in the literature [30].

On the contrary, GPA and age remain relatively stable and consistently low under different thresholds. According to our dataset, there may not exist such a strong association between academic performance and social relationships. Interestingly, although age and grade have a strong positive correlation, their assortativity coefficients are quite different. A latent reason could be the varied regulations on the time of enrollment in different regions of China. In other words, students of the same grade may have different ages.

From Figure 8 we could find that under large thresholds, the connected individuals are quite similar in different aspects revealing strong homophily. Those connected individuals could be regarded as "close friends" who frequently interact with each other and thus could be easily identified with large thresholds. The order of co-occurrence frequency generally follows the rules: close friends $>$ normal friends, familiar strangers $>$ normal strangers. However, it is infeasible to differentiate normal friends from familiar strangers since they co-occur from time to time in both situations.

Effectiveness of lifestyle: Although the comparison of P˙A and PL_A has already demonstrated the effectiveness of lifestyle features, we show more details about the homophily of some of those features in CN in Figure 9. It is clear that chronotype is the strongest homophily indicator than other features. We also use two random features of uniform and normal distributions as baselines. The assortativity coefficients of both random variables approximate to 0 which validates the homophily of the devised features. Besides, the features extracted from working days have stronger homophily than

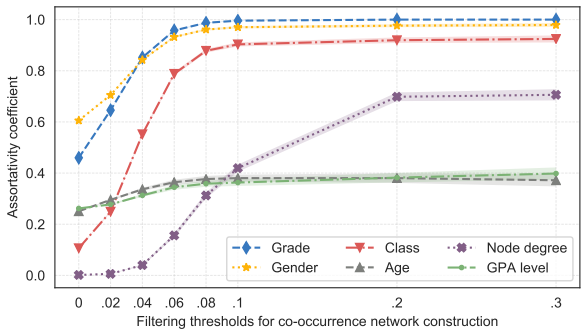


Figure 8: Homophily of different attributes in CN under different thresholds.

that of free days. A latent reason is the number of data samples in working days is double that of free days.

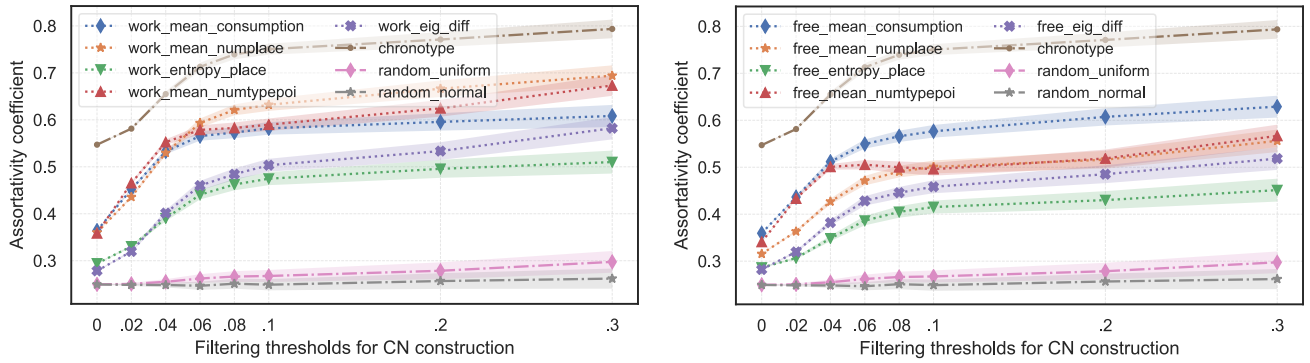
V. Related Works

A. Network Structure Inference (NSI)

As introduced in Preliminaries, inferring the structure of many networks requires a network model and a task model. Network models are used to construct a network based on the input data. There are two categories: parametric models and non-parametric models. Although both categories may have some tuning parameters, the key difference is parametric network models made assumptions on the distributions of edges in the network [3]. For example, in epidemiology, Bernoulli random graph is used to describe potential contacts among a population of individuals, including transmission and interaction rates [32]. Based on the assumption, different methods have been adopted to infer the parameters of parametric models like Markov Chain Monte Carlo [32] and maximum likelihood methods [33]. Non-parametric models measure interactions between nodes directly and determine appropriate edge weights via statistical tests [3].

Task models are often used to evaluate the fitness of the inferred networks, which consist of predictive analysis and descriptive analysis. Predictive analysis describes the prediction of the original data and edges including higher-order attributes like change-point and rank. Descriptive analysis is a qualitative evaluation of the inferred networks based on analysis of different levels [3]. Node-oriented analysis is primary exploratory like studying distributions of simple node statistics. Role-oriented analysis aims to characterize nodes using network features by the structural roles they play in the system such as bridges between social communities. Other high-order analysis examines communities and larger subgraph structures beyond node and edge-based descriptive statistics.

In certain scientific areas, parametric models are widely used. In Neuroscience, it could model relationships between brain regions, physiological structures, and functions [34]. In Epidemiology, it is used to model hidden networks from observed infections [35]. However, for other areas like Ecology and Sociology, finding appropriate models for hidden networks



(a) Working days (b) Free days

Figure 9: Homophily of the proposed lifestyle features in CN with varied thresholds. (a) Features from SD in working days; (b) Features from SD in free days.

is usually intractable and difficult. The underlying reason might be the lack of knowledge about animal and human behaviors. Therefore, the majority of works in NSI are based on non-parametric models.

Besides methodology, different data modalities have been explored for NSI, including online and offline measurements. Online measurements represent unambiguous and countable interactions among entities like emails, phone calls, and instant messages [36], [37]. Offline measurements are mostly the proximity information collected by wearable sensors [6], [19], [20], [38]. The proximity information could be proxies for face-to-face interaction which is the richest communication modality available to humans [37], [39].

B. Smart Card Transaction Data Analysis

Smart card transactions capture rich information of human mobility and urban dynamics, therefore are of particular interest to urban planners and location-based service providers [40]. Based on these data, tremendous analysis and applications have been explored. First of all, security and privacy issues within transaction data are investigated [41], [42]. Understanding the privacy of these data sets is key to their broad use and, ultimately, their impact. Researchers studied 3 months of credit card records for 1.1 million people and found that 4 spatiotemporal points are enough to uniquely reidentify 90% of individuals. Another line of research is the recommendation of goods and services. Researchers have experimented with new information from transaction data to improve the performance of recommendations. They tried to mine frequent patterns [43], cluster individual transaction records [44] and recently incorporated transactional context [45], customer preference, and price sensitivity [46]. The rest are behavior analysis and understanding include traveling and shopping behaviors [46], [47], urban lifestyles [17], [48], and physical social networks [49], [50]. These analyses are further applied to predict visitation patterns of merchants [51], academic performance [21], [26], and macro-socioeconomic indicators [10].

VI. Conclusions

In this study, we leveraged smart card transaction data to infer physical social networks. To tackle the challenges posed by individual differences in social behavior and the identification of familiar strangers, we devised an approach that combines lifestyle features with proximity information. Our contributions can be summarized in three key aspects. Firstly, we present the first accurate inference framework for physical social networks (PSNs) utilizing transaction data. Secondly, we introduce nonparametric methods for lifestyle feature extraction that offer flexibility and applicability across diverse scenarios. Lastly, we evaluate our inference framework using real-world datasets. Through the extraction and fusion of lifestyle features, our proposed framework outperforms baseline approaches by a significant margin.

References

[1] E. Lee, F. Karimi, C. Wagner, H.-H. Jo, M. Strohmaier, and M. Galesic, "Homophily and minority-group size explain perception biases in social networks," *Nature human behaviour*, 2019.

[2] P. Block, M. Hoffman, I. J. Raabe, J. B. Dowd, C. Rahal, R. Kashyap, and M. C. Mills, "Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world," *Nature Human Behaviour*, 2020.

[3] I. Brugere, B. Gallagher, and T. Y. Berger-Wolf, "Network structure inference, a survey: Motivations, methods, and applications," *ACM CSUR*, 2018.

[4] K. Subrahmanyam, S. M. Reich, N. Waechter, and G. Espinoza, "Online and offline social networks: Use of social networking sites by emerging adults," *Journal of applied developmental psychology*, 2008.

[5] A. Godoy-Lorite and N. S. Jones, "Inference and influence of network structure using snapshot social behavior without network data," *Science Advances*, 2021.

[6] N. Eagle and A. S. Pentland, "Eigenbehaviors: Identifying structure in routine," *Behavioral Ecology and Sociobiology*, 2009.

[7] A. Pentland, N. Eagle, and D. Lazer, "Inferring social network structure using mobile phone data," *PNAS*, 2009.

[8] M. Ye, X. Liu, and W.-C. Lee, "Exploring social influence for recommendation: a generative model approach," in *SIGIR*. ACM, 2012.

[9] W. Dong, B. Lepri, and A. Pentland, "Automatic prediction of small group performance in information sharing tasks," *arXiv preprint arXiv:1204.3698*, 2012.

[10] B. Hashemian, E. Massaro, I. Bojic, J. M. Arias, S. Sobolevsky, and C. Ratti, "Socioeconomic characterization of regions through the lens of individual financial transactions," *PloS one*, 2017.

[11] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, "Superspreading and the effect of individual variation on disease emergence," *Nature*, 2005.

[12] H. Pham, C. Shahabi, and Y. Liu, "Ebm: an entropy-based model to infer social strength from spatiotemporal data," in *SIGMOD*. ACM, 2013.

[13] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *PNAS*, 2010.

[14] T. Aledavood, S. Lehmann, and J. Saramäki, "Social network differences of chronotypes identified from mobile phone data," *EPJ Data Science*, 2018.

[15] S. Milgram, "The familiar stranger: An aspect of urban anonymity," *The individual in a social world*, 1977.

[16] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang, "Understanding metropolitan patterns of daily encounters," *PNAS*, 2013.

[17] R. Di Clemente, M. Luengo-Oroz, M. Travizano, S. Xu, B. Vaitla, and M. C. González, "Sequences of purchases in credit card data reveal lifestyles in urban populations," *Nature communications*, 2018.

[18] T. Fuchikawa, A. Eban-Rothschild, M. Nagari, Y. Shemesh, and G. Bloch, "Potent social synchronization can override photic entrainment of circadian rhythms," *Nature communications*, 2016.

[19] V. Sekara and S. Lehmann, "The strength of friendship ties in proximity sensor data," *PloS one*, 2014.

[20] I. Psorakis, S. J. Roberts, I. Rezek, and B. C. Sheldon, "Inferring social network structure in ecological systems from spatio-temporal data streams," *Journal of the Royal Society Interface*, 2012.

[21] H. Yao, M. Nie, H. Su, H. Xia, and D. Lian, "Predicting academic performance via semi-supervised learning with constructed campus social network," in *International Conference on Database Systems for Advanced Applications*. Springer, 2017.

[22] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, 2005.

[23] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, "Explaining the power-law distribution of human mobility through transportation modality decomposition," *Scientific reports*, 2015.

[24] R. Levandovski, E. Sasso, and M. P. Hidalgo, "Chronotype: a review of the advances, limits and applicability of the main instruments used in the literature to assess human phenotype," *Trends in psychiatry and psychotherapy*, 2013.

[25] F. Preckel, A. A. Lipnevich, S. Schneider, and R. D. Roberts, "Chronotype, cognitive abilities, and academic achievement: A meta-analytic investigation," *Learning and Individual Differences*, 2011.

[26] H. Yao, D. Lian, Y. Cao, Y. Wu, and T. Zhou, "Predicting academic performance for college students: A campus behavior perspective," *ACM TIST*, 2019.

[27] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM journal on matrix analysis and applications*, 2008.

[28] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," Georgia Institute of Technology, Tech. Rep., 2008.

[29] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *KDD*. ACM, 2016.

[30] I. Smirnov and S. Thurner, "Formation of homophily in academic performance: Students change their friends rather than performance," *PloS one*, 2017.

[31] M. E. Newman, "Mixing patterns in networks," *Physical Review E*, 2003.

[32] T. Britton and P. D. O'NEILL, "Bayesian inference for stochastic epidemics in populations with random social structure," *Scandinavian Journal of Statistics*, 2002.

[33] N. Du, L. Song, M. Yuan, and A. J. Smola, "Learning networks of heterogeneous influence," in *Advances in Neural Information Processing Systems*, 2012.

[34] E. E. Papalexakis, A. Fyshe, N. D. Sidiropoulos, P. P. Talukdar, T. M. Mitchell, and C. Faloutsos, "Good-enough brain model: Challenges, algorithms, and discoveries in multisubject experiments," *Big data*, 2014.

[35] M. G. Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf, "Uncovering the structure and temporal dynamics of information propagation," *Network Science*, 2014.

[36] M. De Choudhury, W. A. Mason, J. M. Hofman, and D. J. Watts, "Inferring relevant social networks from interpersonal communication," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[37] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys," *PloS one*, 2015.

[38] S.-P. Hong, Y.-H. Min, M.-J. Park, K. M. Kim, and S. M. Oh, "Precise estimation of connections of metro passengers from smart card data," *Transportation*, 2016.

[39] L. Wu, B. N. Waber, S. Aral, E. Brynjolfsson, and A. Pentland, "Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task," *Available at SSRN 1130251*, 2008.

[40] N. J. Yuan, Y. Wang, F. Zhang, X. Xie, and G. Sun, "Reconstructing individual mobility from smart card transactions: A space alignment approach," in *ICDM*. IEEE, 2013.

[41] Y.-A. De Montjoye, L. Radaelli, V. K. Singh *et al.*, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, 2015.

[42] Y.-A. de Montjoye *et al.*, "Response to comment on "unique in the shopping mall: On the reidentifiability of credit card metadata"," *Science*, 2016.

[43] A. K. Poernomo and V. Gopalkrishnan, "Mining statistical information of frequent fault-tolerant patterns in transactional databases," in *ICDM*. IEEE, 2007.

[44] R. Guidotti, A. Monreale, M. Nanni, F. Giannotti, and D. Pedreschi, "Clustering individual transactional data for masses of users," in *KDD*. ACM, 2017.

[45] S. Wang, L. Hu, L. Cao, X. Huang, D. Lian, and W. Liu, "Attention-based transactional context embedding for next-item recommendation." AAAI, 2018.

[46] M. Wan, D. Wang, M. Goldman, M. Taddy, J. Rao, J. Liu, D. Lymberopoulos, and J. McAuley, "Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs," in *WWW*. ACM, 2017.

[47] K. K. A. Chu, "Two-year worth of smart card transaction data–extracting longitudinal observations for the understanding of travel behaviour," *Transportation Research Procedia*, 2015.

[48] V. L. Miguéis, A. S. Camanho, and J. F. e Cunha, "Customer data mining for lifestyle segmentation," *Expert Systems with Applications*, 2012.

[49] Z. Kun, Z. Guangyi, Z. Deshun, Y. Jun, and S. Jinrong, "Link formation in undergraduate students friendship network," in *BESC2014*. IEEE, 2014.

[50] M. Backes, M. Humbert, J. Pang, and Y. Zhang, "walk2friends: Inferring social links from mobility profiles," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017.

[51] C. Krumme, A. Llorente, M. Cebrian, E. Moro *et al.*, "The predictability of consumer visitation patterns," *Scientific reports*, 2013.