Lingnan University 嶺南大學 香港 Hong Kong

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

Social Networks

Where can we get such information?

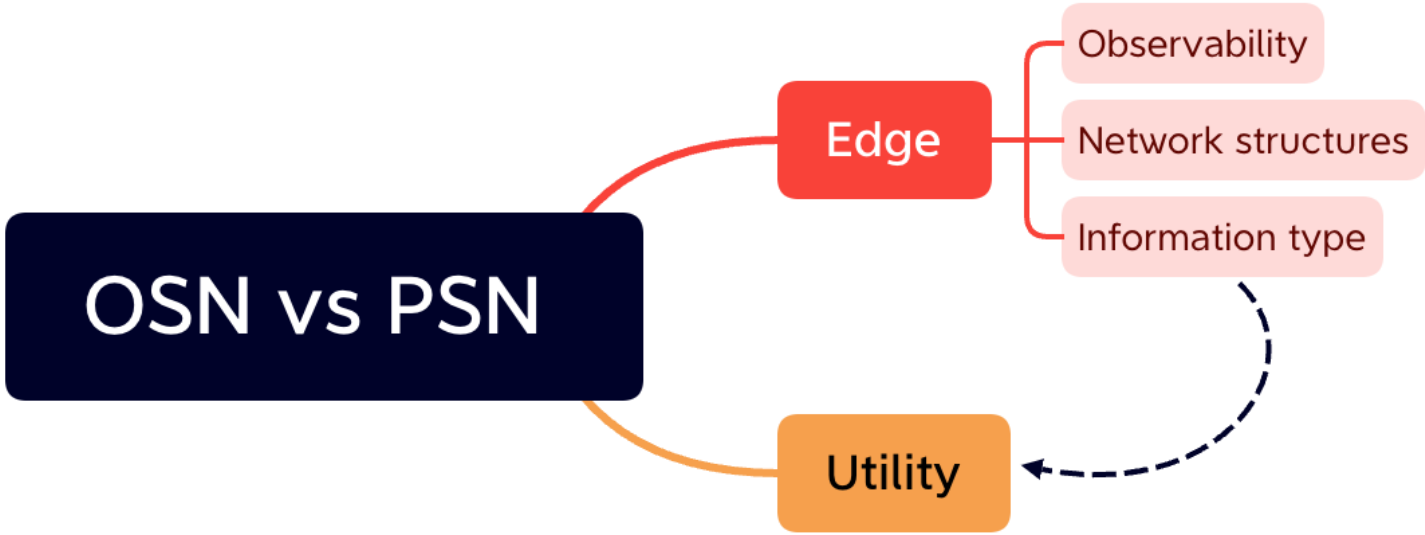
- A social network is a graph with individuals as vertices and their relationships as edges.
- Social networks are important for many disciplines
  - Epidemiology: disease infection
  - Sociology: social segregation
  - Psychology: collective behavior
  - Marketing: recommendation
  - Urban planning: space design

Two Types of Social Networks

- Online social network (OSN)
  - Cyberworld, digital connections
  - Directly observable
  - Examples: Facebook, Twitter, LinkedIn
- Physical social network (PSN)
  - Real-world, in-person connections
  - Cannot be observed at scale
  - Examples: friendships, business networks

From the Internet!

How to get PSN?



Social Network Inference

- Given data D consisting of multiple individuals (vertices), detect their relationships and strengths (edges)

$$\arg\min_{\mathbf{G}} e(\mathbf{T}, \mathbf{T}^*); \mathbf{T} \leftarrow \mathcal{T}(\mathbf{G}, \beta); \mathbf{G} \leftarrow \mathcal{N}(\mathbf{D}, \alpha)$$

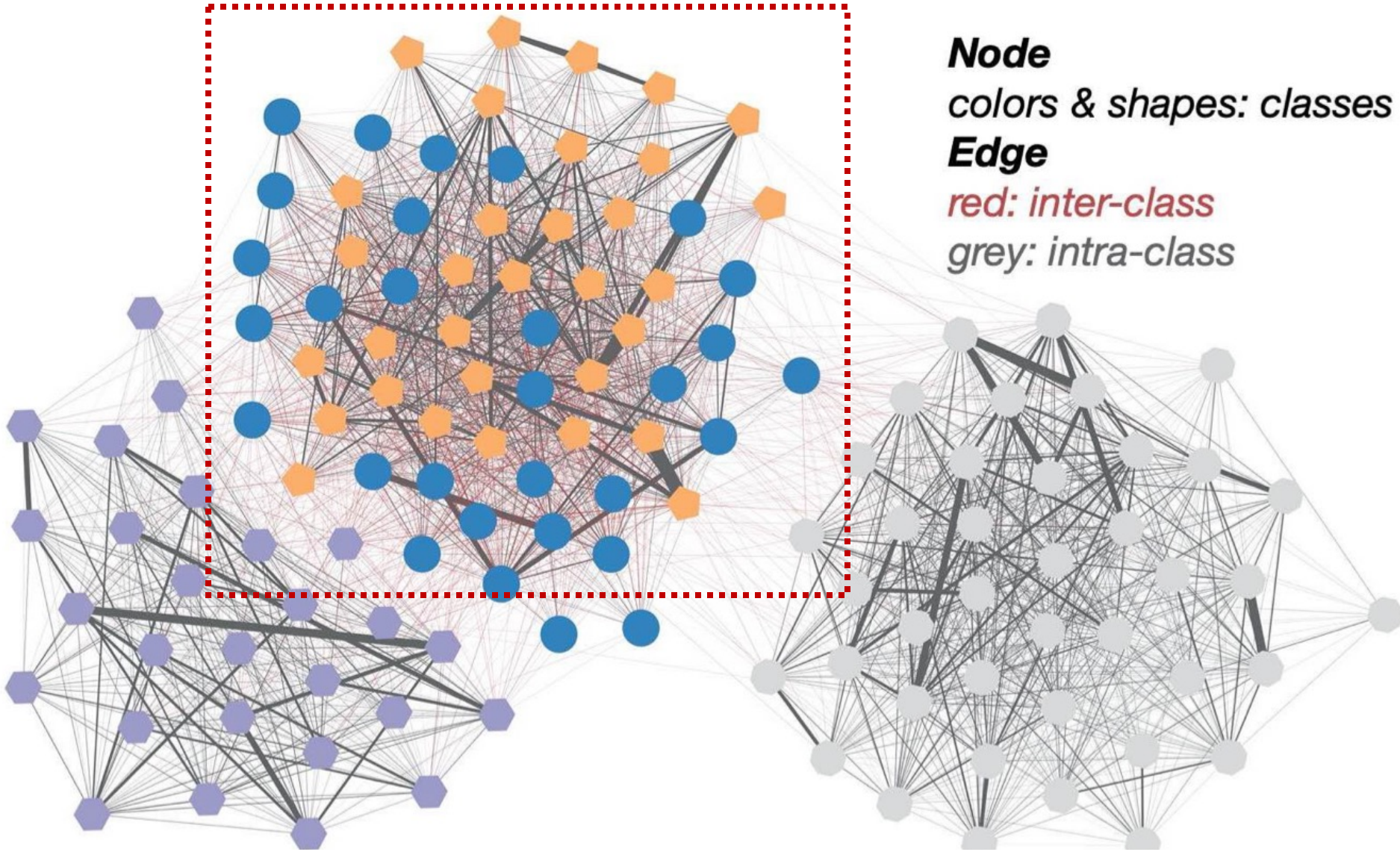
- G is the result of network inference model on input data D and parameters α.
- T is generated by the task with constructed network G and parameter set β.
- e(·) measures the difference between predicted task outputs T and true task outputs T∗

Related Works

- Data sources
  - Check-in data, geo-tagged photos, proximity information of smartphones
- The common assumption
  - Individuals appearing in the same place at the same time simultaneously (i.e., cooccurrence) may have a latent social relationship.
- PSN could be approximated by co-occurrence network
  - Links are constructed by measuring the significance of their co-occurrences
  - Important cooccurrences -> larger weights
  - Strangers who occasionally encountered are filtered out by a predefined threshold

Limitations of Existing Methods

- The approximation using cooccurrence networks can hardly address **individual differences** and **familiar strangers**.
- Individual differences in social strategy
  - A large network but weak links
  - A small network but strong links
- Family strangers
  - Strangers who regularly co-appeared
  - Due to similar daily routines
  - Examples: students living in the same dormitory and employees working in the same office building


Node
colors & shapes: classes
Edge
red: inter-class
grey: intra-class

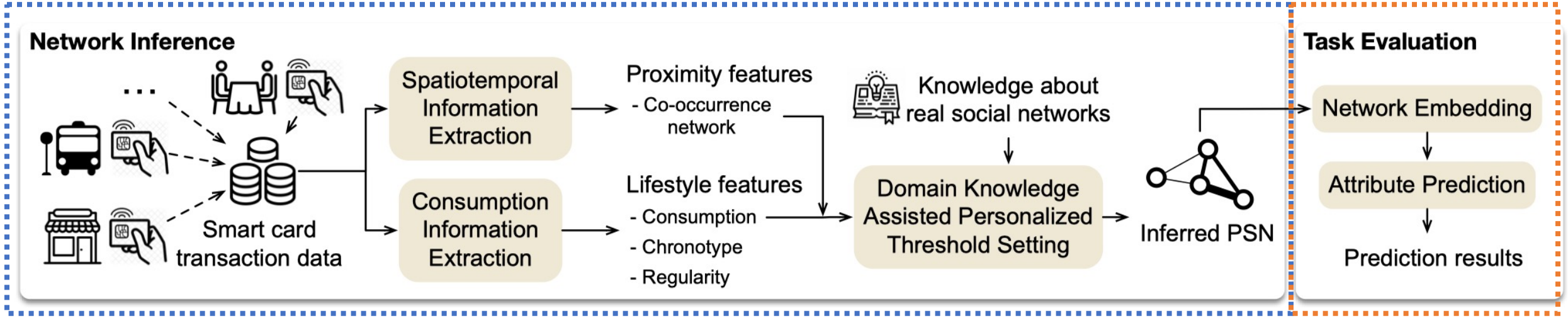
Observation of Heuristics

- With the proliferation of e-payment, a huge amount of smart card transaction data (SD) has been accumulated which brings a great opportunity for accurate PSN inference.
- People with different chronotypes have different social preferences [1].
- Real friends not only co-appear frequently but also have similar lifestyles [2, 3, 4]

[1] T. Aledavood, et al. “Social network differences of chronotypes identified from mobile phone data,” EPJ Data Science, 2018.
[2] N. Eagle and A. S. Pentland, “Eigenbehaviors: Identifying structure in routine,” Behavioral Ecology and Sociobiology, 2009.
[3] R. Di Clemente, et al “Sequences of purchases in credit card data reveal lifestyles in urban populations,” Nature communications, 2018.
[4] T. Fuchikawa, et al “Potent social synchronization can override photic entrainment of circadian rhythms,” Nature communications, 2016.

The Proposed Solution

$$\arg\min_{\mathbf{G}} e(\mathbf{T}, \mathbf{T}^*); \mathbf{T} \leftarrow \mathcal{T}(\mathbf{G}, \beta); \mathbf{G} \leftarrow \mathcal{N}(\mathbf{D}, \alpha)$$



Proximity Feature-Based Solution

- Co-occurrence network (CN)
- The more individuals 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.

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.
    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*

Lifestyle Features

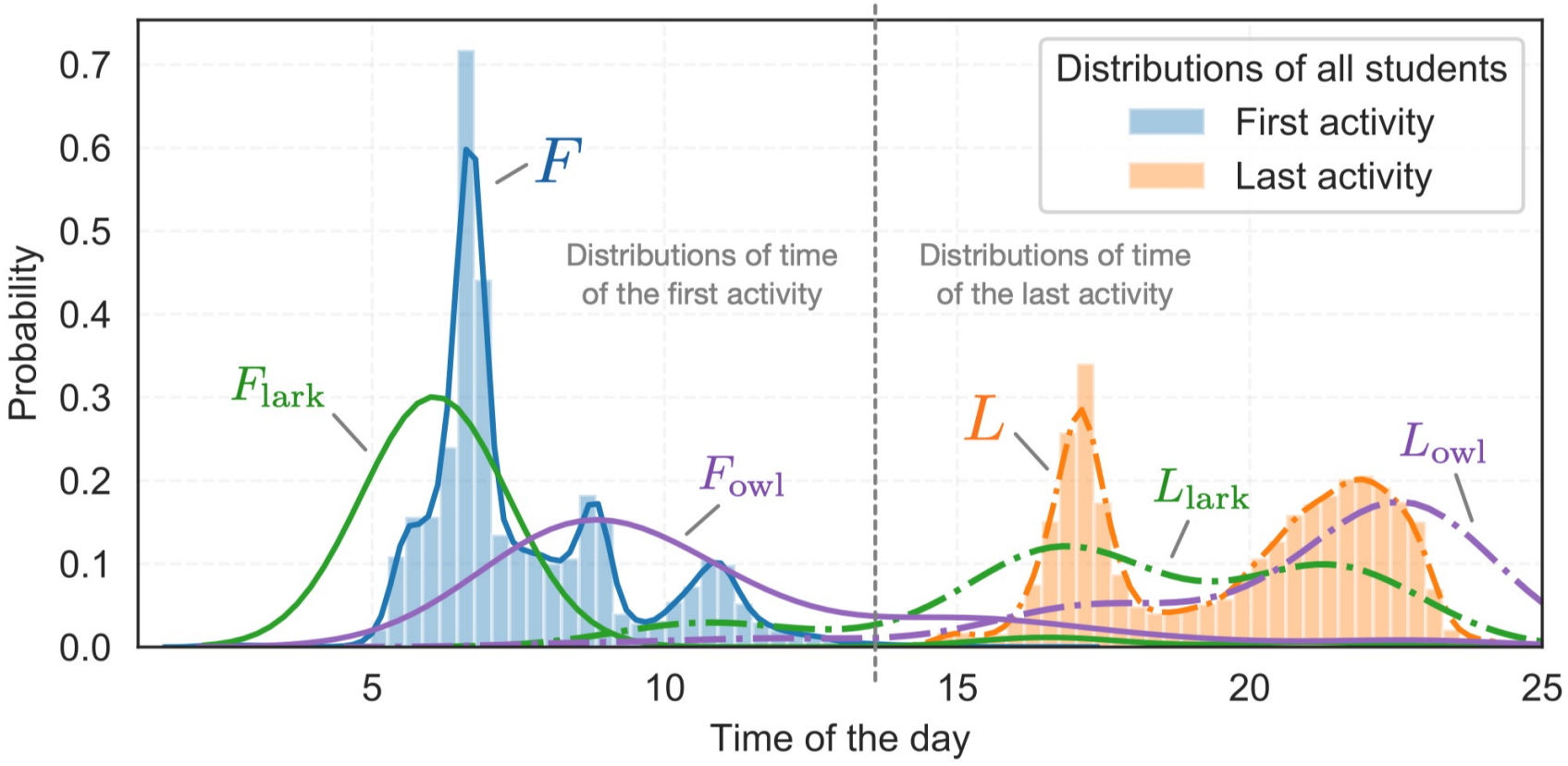
- Consumption
  - Features related to consumption habits, food preferences, brand preferences.
- Chronotype (categorical)
  - An individual's natural inclination regarding the times of day when they prefer to sleep or when they are most alert or energetic.
- Regularity (continuous)
  - Regularity refers to the predictability of biological and behavioral patterns. It is an important aspect of the internal circadian clock.

Chronotype

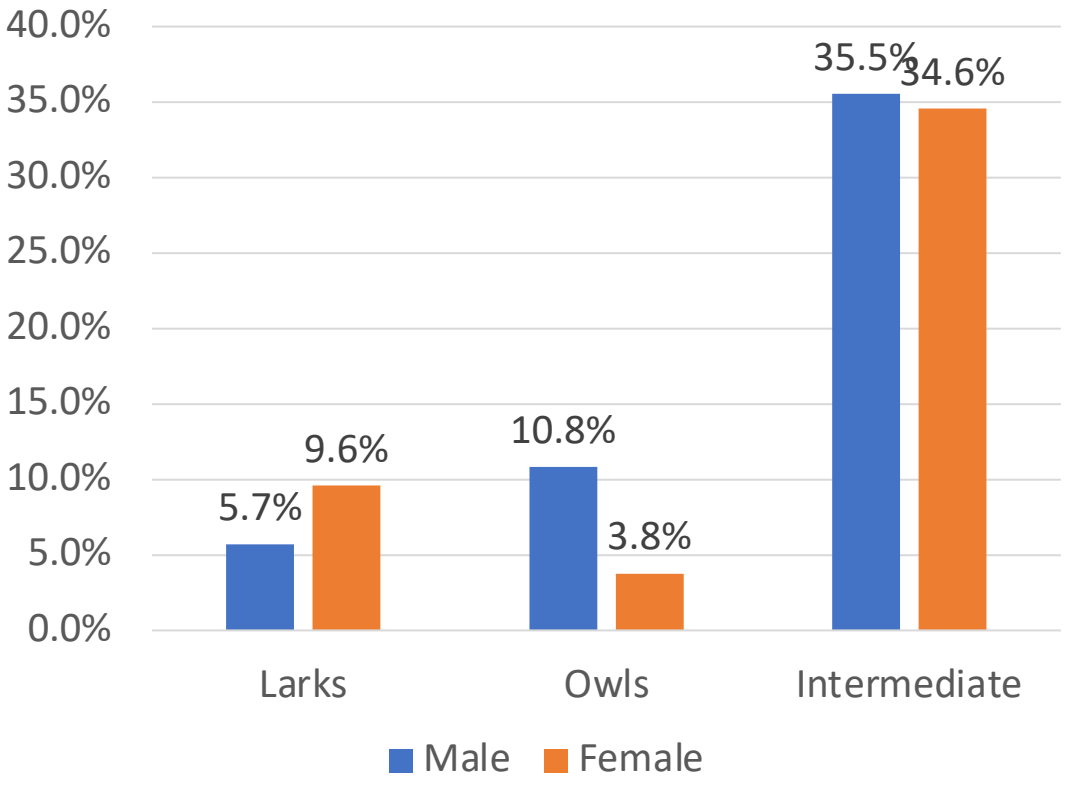
- **Morning types**, or "larks", are most alert in the morning and prefer to go to bed early.
- **Evening types**, or "owls", feel most alert in the evening, and prefer to go to bed late.
- **Intermediate types** fall somewhere in between.

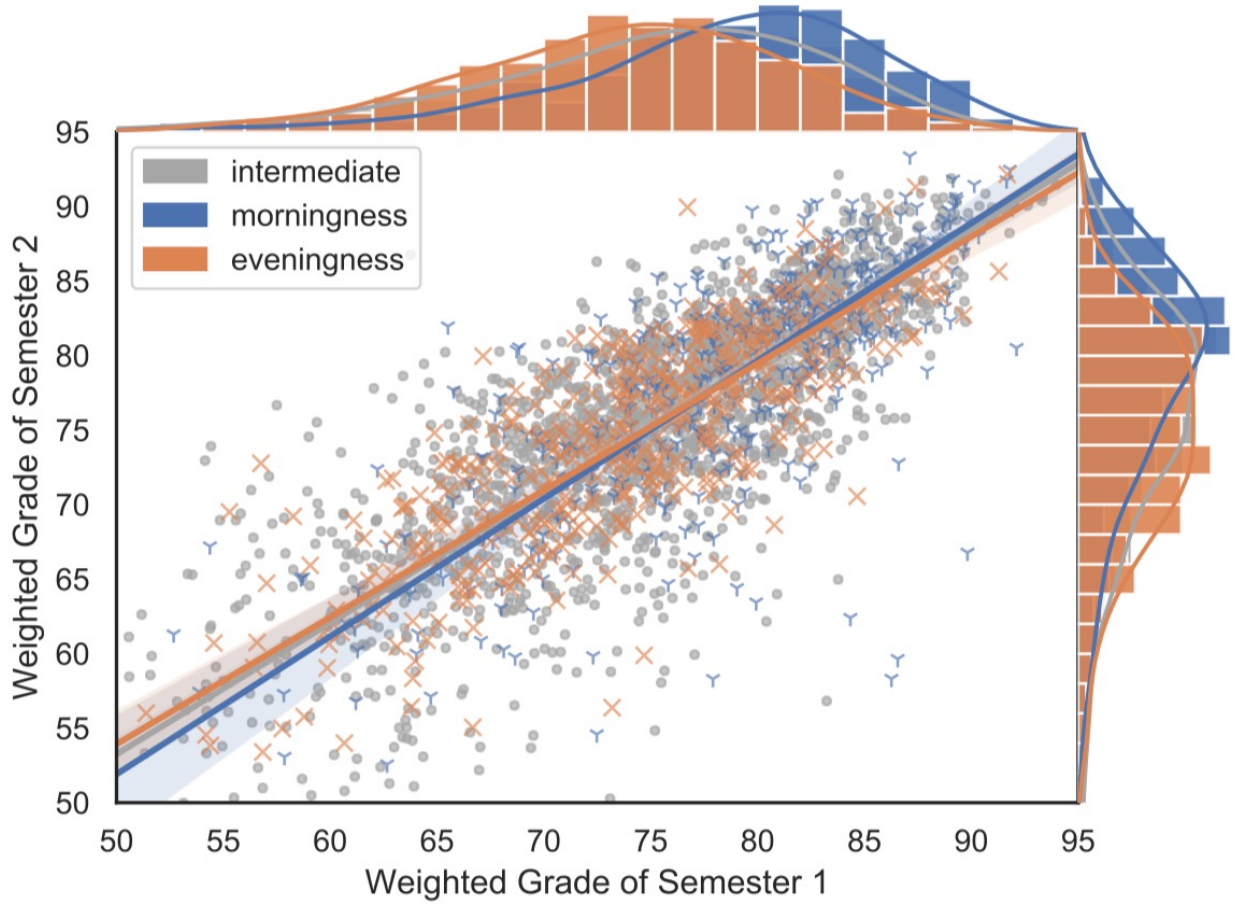
$$\begin{cases} \text{Morningness (Lark)} : (F_i < F) \wedge (L_i < L) \\ \text{Eveningness (Owl)} : (F_i > F) \wedge (L_i > L) \\ \text{Intermediate} : \text{other situations} \end{cases}$$

Distribution of time of the first activity for individual i



15.3% larks and 14.6% owls among all students, which is close to the reported percentages (20%, 20%) in the literature

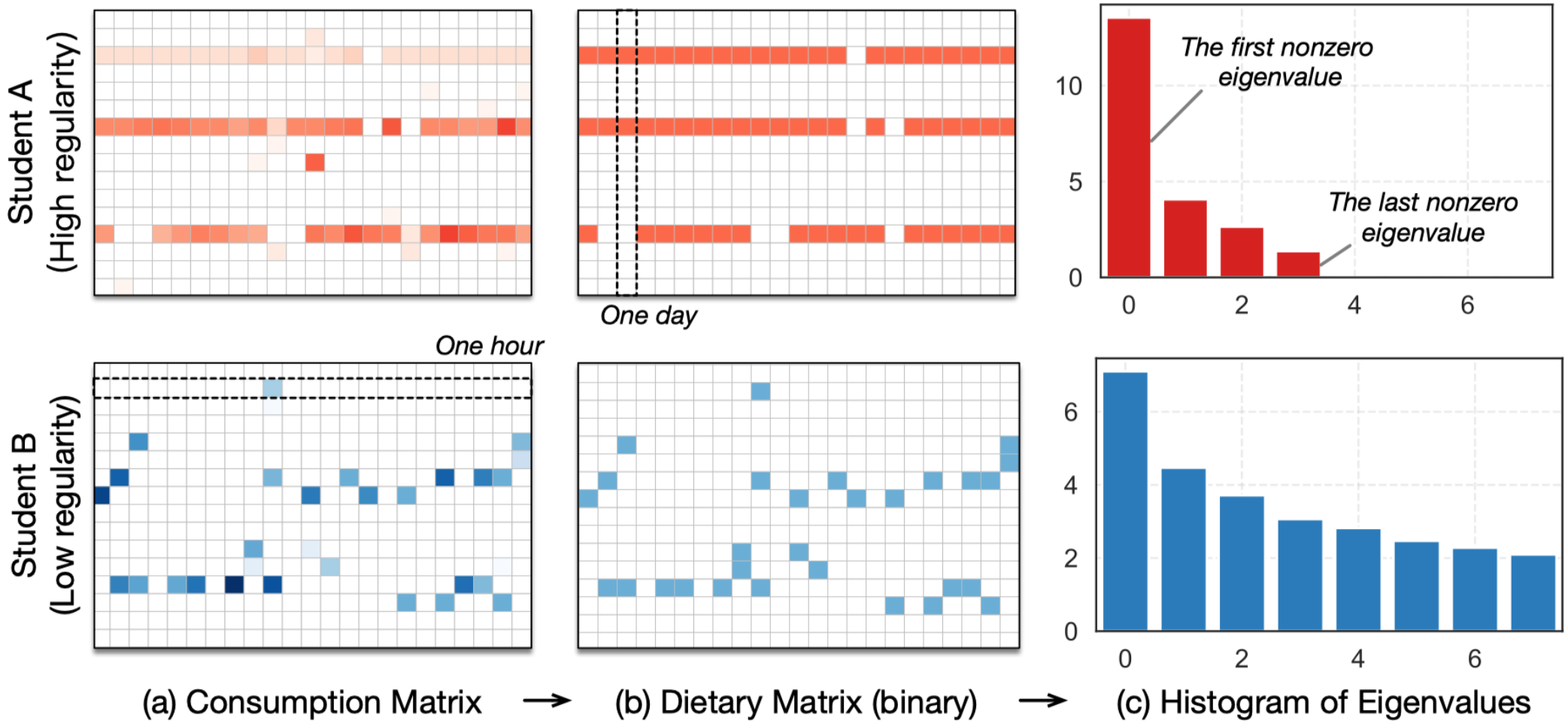




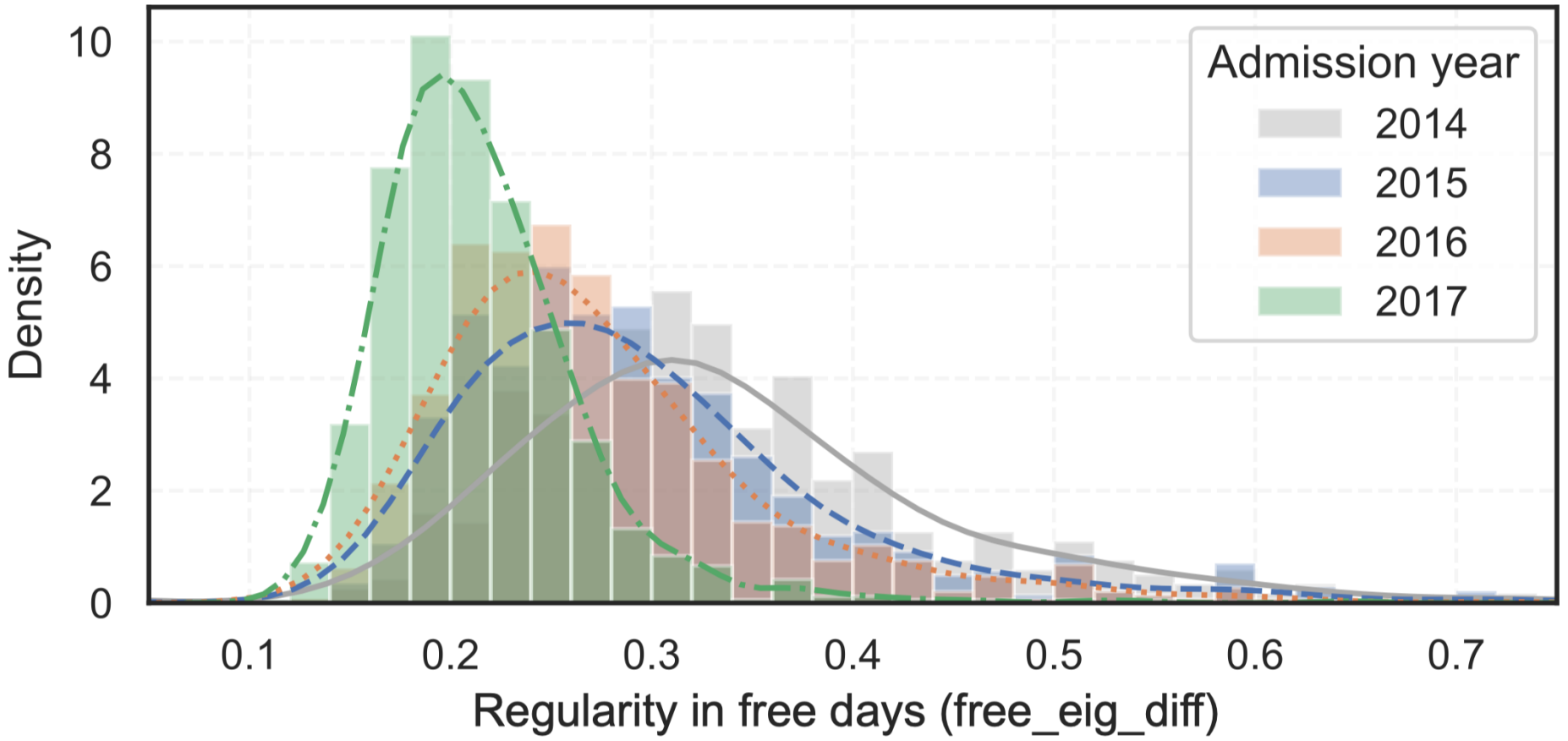
Academic performance follows the order:
Morningness > Intermediate > Eveningness

Regularity

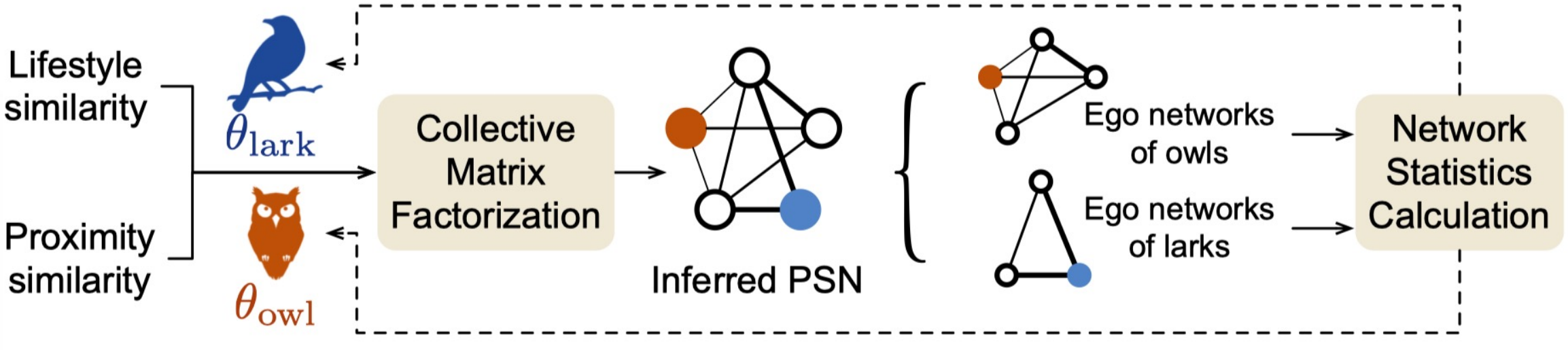
The higher regularity an individual's dietary behaviors, the fewer nonzero eigenvalues



Seniors (admission year: 2014) > juniors > sophomores > freshmen indicating students' dietary routines become more stable over the years on the campus in free days.



Fusing Proximity & Lifestyles



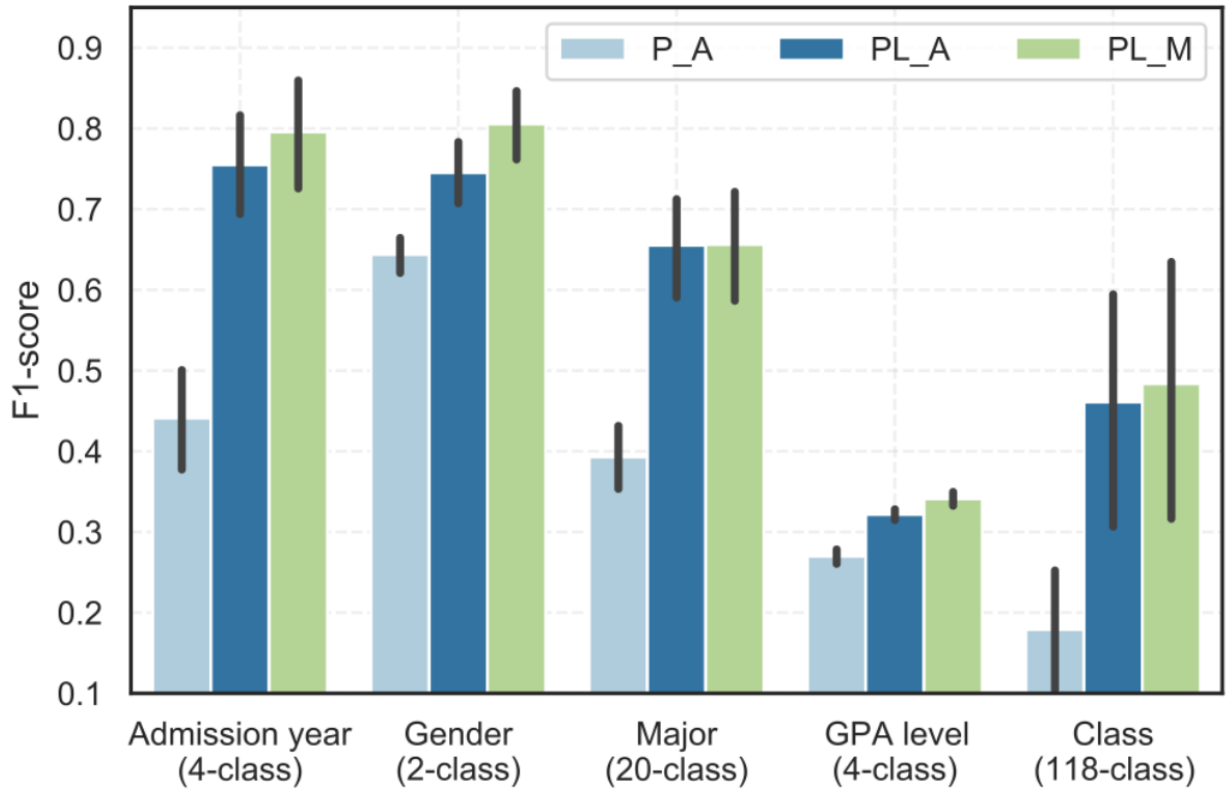
Evaluation

- 633,180 transaction records
- 2,274 students from 6 faculties ranging from freshmen to seniors.
- 3 months, starting from 1 Oct 2017 to 31 Dec 2017

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

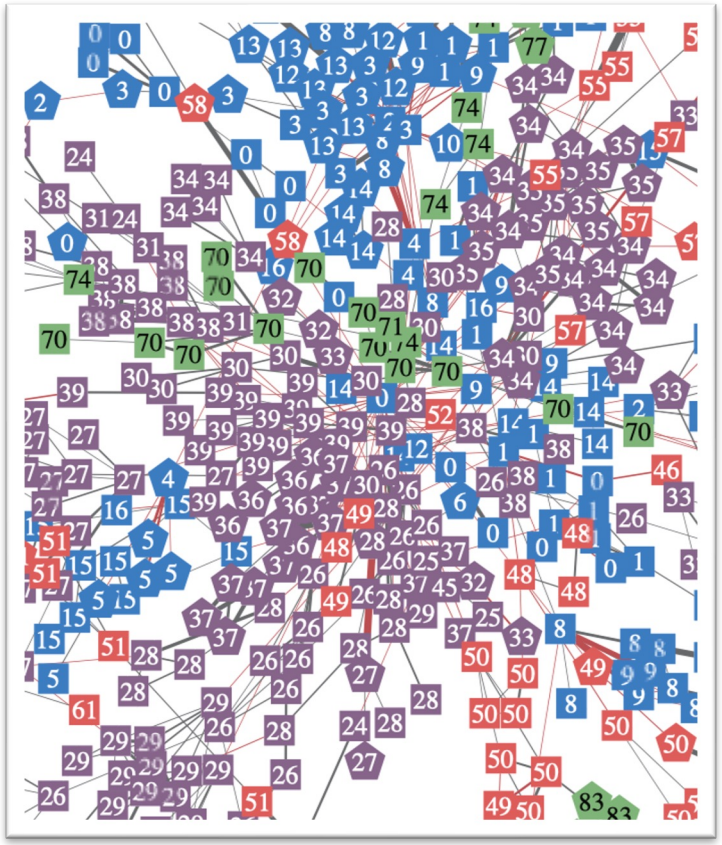
§ : Evaluate lifestyle features

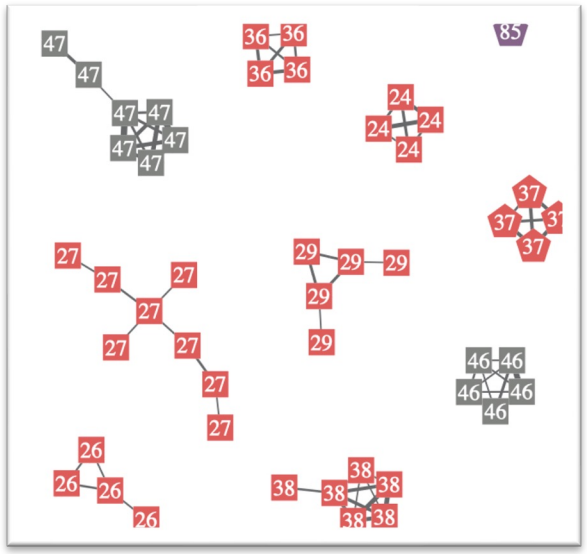
† : Evaluate parameter setting

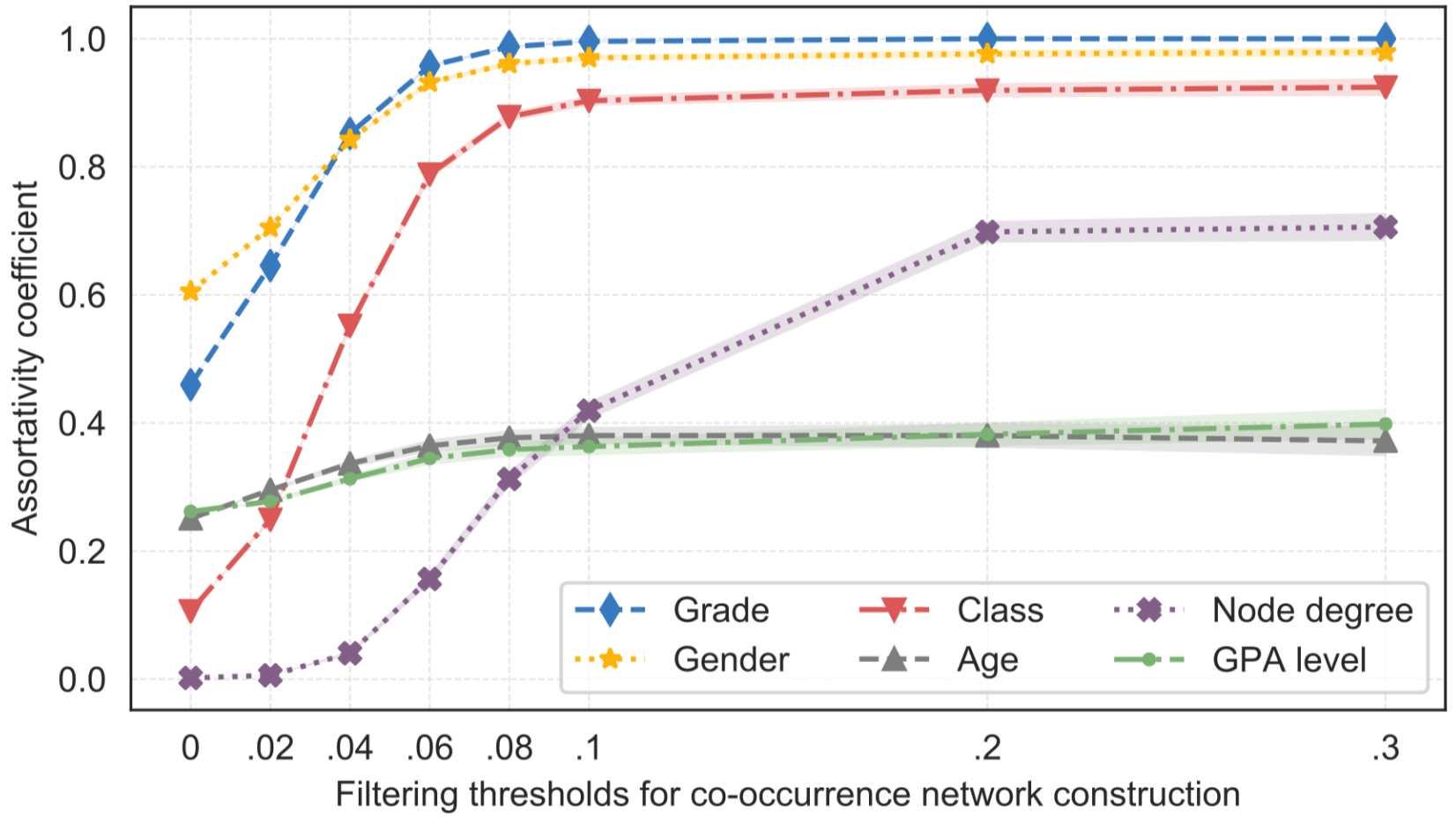


Method\Task	Admission Year			Gender			Major			GPA Level			Class		
	P_A	PL_A	PL_M	P_A	PL_A	PL_M	P_A	PL_A	PL_M	P_A	PL_A	PL_M	P_A	PL_A	PL_M
AdaBoost	0.403	0.69	**0.723**	0.651	0.745	**0.773**	0.37	**0.526**	0.515	0.268	**0.331**	0.328	0.019	0.064	**0.069**
Decision Tree	0.319	0.587	**0.642**	0.587	0.661	**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**

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.







Homophily of different attributes in CN under different thresholds.


thank
you


Q
A