香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

DSAA 5002 Final Project

2025.10.27

Feiyu Huang
Han Linghu

General Requirements

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

1. **Topic**: Related to the course material, e.g., data preprocessing, classification, clustering, anomaly detection.
2. **Format**: KDD, double-column, maximum 6 pages including reference and appendix.
   Refer to https://kdd2025.kdd.org/research-track-call-for-papers/
3. **Sections**:
   At least contain: introduction, related work, method, experiments.
   Encouraged: theoretical derivation.
4. **Writing:** Use concise language and avoid complex sentences.
5. **Novelty**: Show how you differentiate from related work.
6. **Code**: Required.

Grading: 50% (Presentation 25% + Report 25%)
Report Due: 2025.12.19 (End of Day)
Presentation Dates: Dec 8th, 15th
Students with **even** ID numbers will present on Dec 8th,
and those with **odd** ID numbers will present on Dec 15th.
The report and presentation file need to be submitted to Canvas

General Requirements

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Individual Project.
- No previously published papers are allowed. Submissions based on published work will receive 0 points.
- Plagiarism in any form is strictly prohibited. Violations will result in 0 points.
- At the beginning of **both** presentation and report, clearly state:
  1. your **project title**; and
  2. the **specific course topic** your project relates to

  Failure to do so will incur a 20% penalty on the total project score.

About Report Writing (25%)

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Clear contribution (differentiation from previous existing works)

Well-structured format

Well-organized narration logic

Make it complete.

About Presentation (25%)

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

General: 5 min presentation for each person.

Format: Due to the large enrollment this semester, project presentations will be conducted in two formats:

In-person: A random half of each class section will present in class during scheduled class time. The presentation schedule and order will be announced in advance.

Online: Within three days of the presentation date, remaining students must upload a 5-minute recorded presentation video to Baidu Cloud (or another file-storage service) and submit only the video link on Canvas (include the extraction/access code if required).
Students with even ID numbers must upload a video by Dec 11th.
Students with odd ID numbers must upload a video by Dec 18th.

All presentations, live or recorded, must adhere to the same time limit and content expectations.

Default Topic - Customer Segmentation

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Topic website: https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering/notebook

If you choose this default topic, you don't need to specify which topic it relates to.

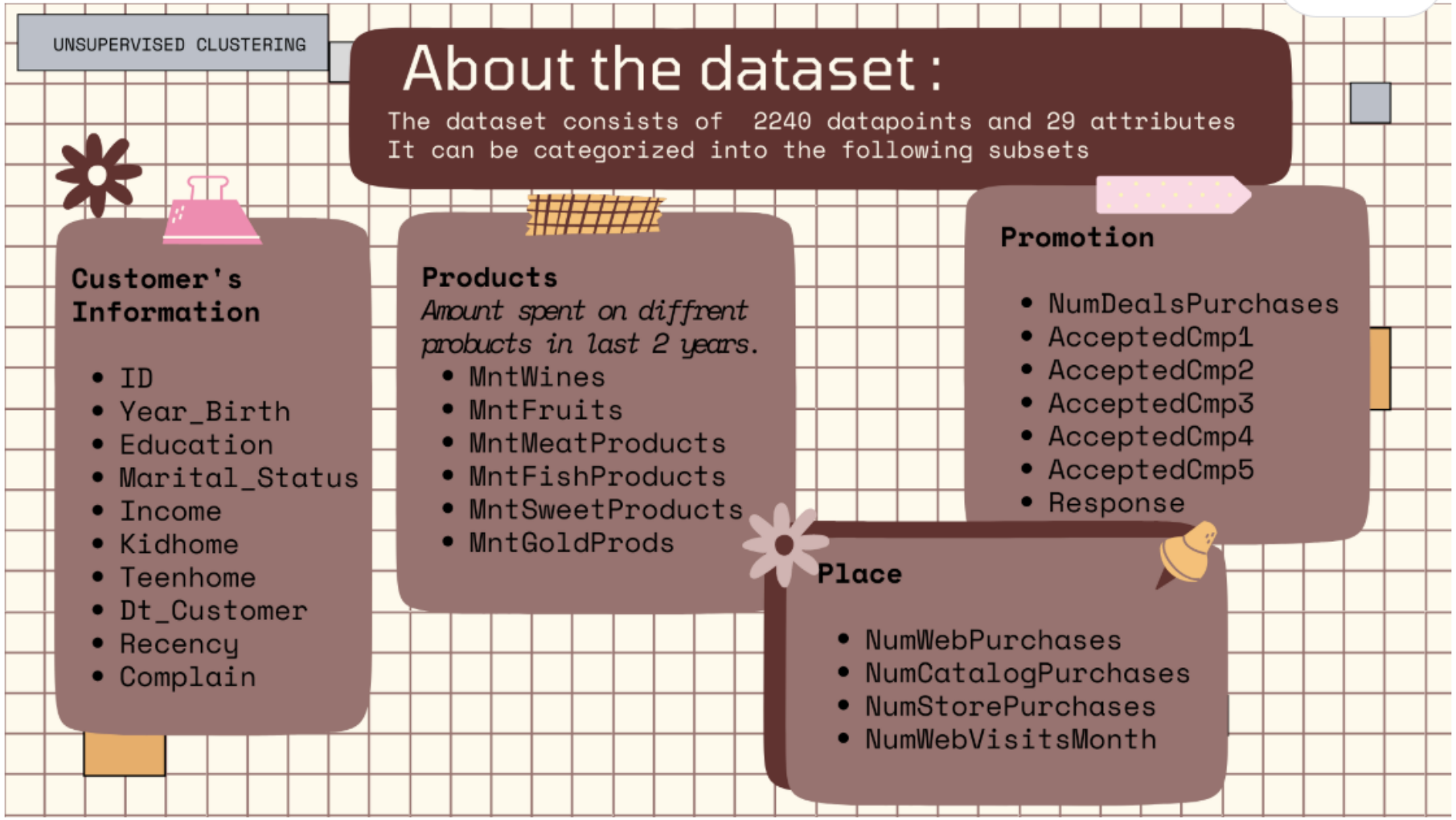
Default Topic

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- This project offers a basic approach to customer segmentation, aiming to understand customer groups better and support targeted marketing efforts.
- It serves as a **60-point baseline**, encouraging students to apply **innovative ideas for improved customer analysis**, such as exploring new clustering methods or models.
- All creative extensions should still be related to the core content of this course.

Default Topic

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)



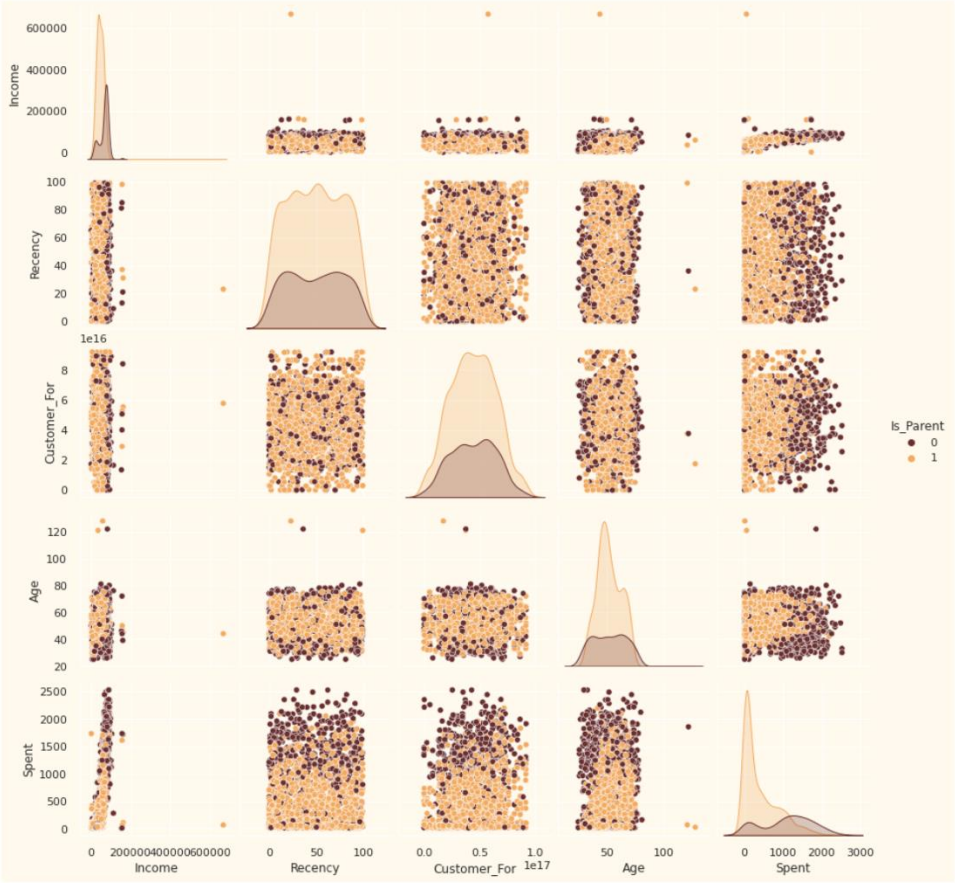
Default Topic

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

1. Data Preparation & Feature Selection

Data Cleaning: Handle missing and abnormal values.

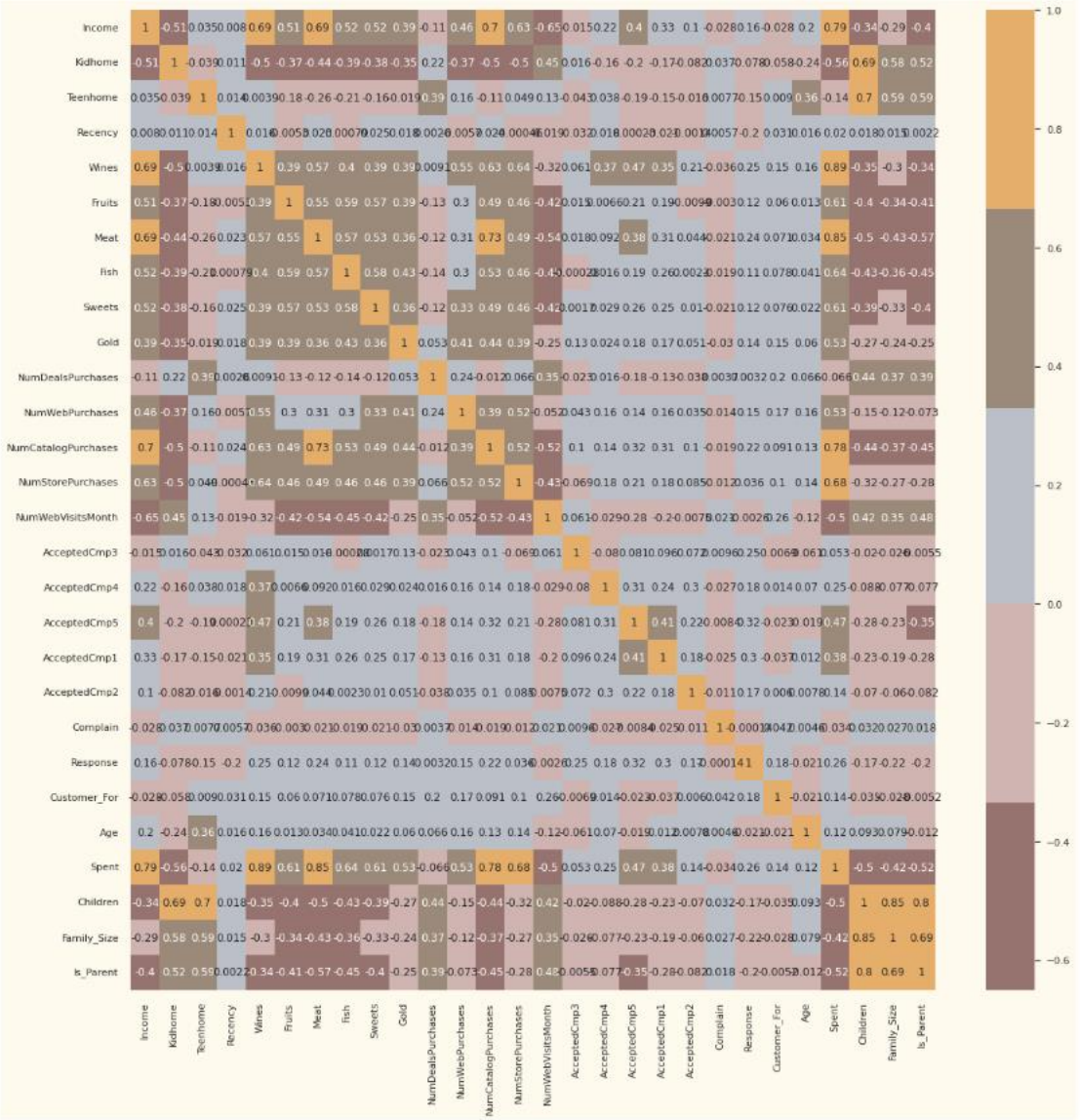
Feature Selection: Use key indicators such as Recency (last purchase time), Frequency, and Monetary value as the basis for clustering.



Clearly, there are a few outliers in the Income and Age features, and deleting the outliers in the data.

Default Topic

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)



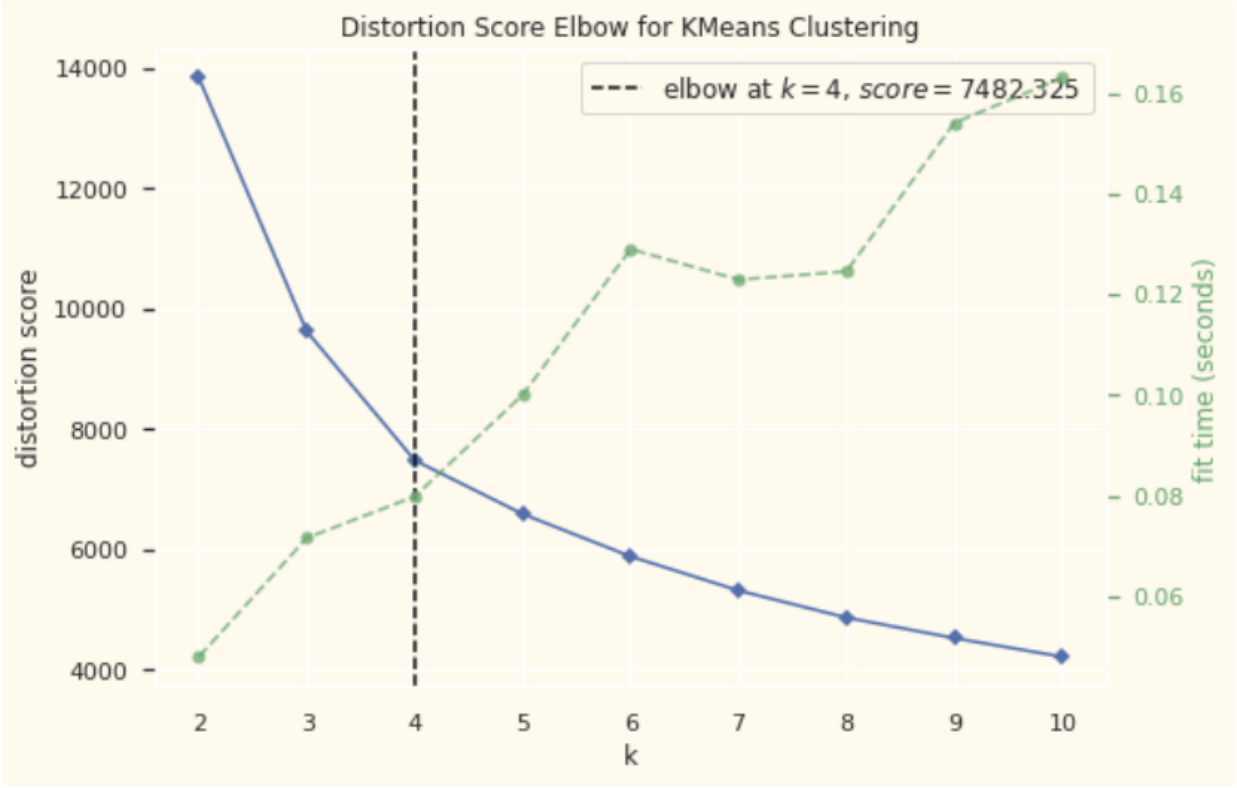
The data is quite clean and the new features have been included. We will proceed to the next step.

Default Topic

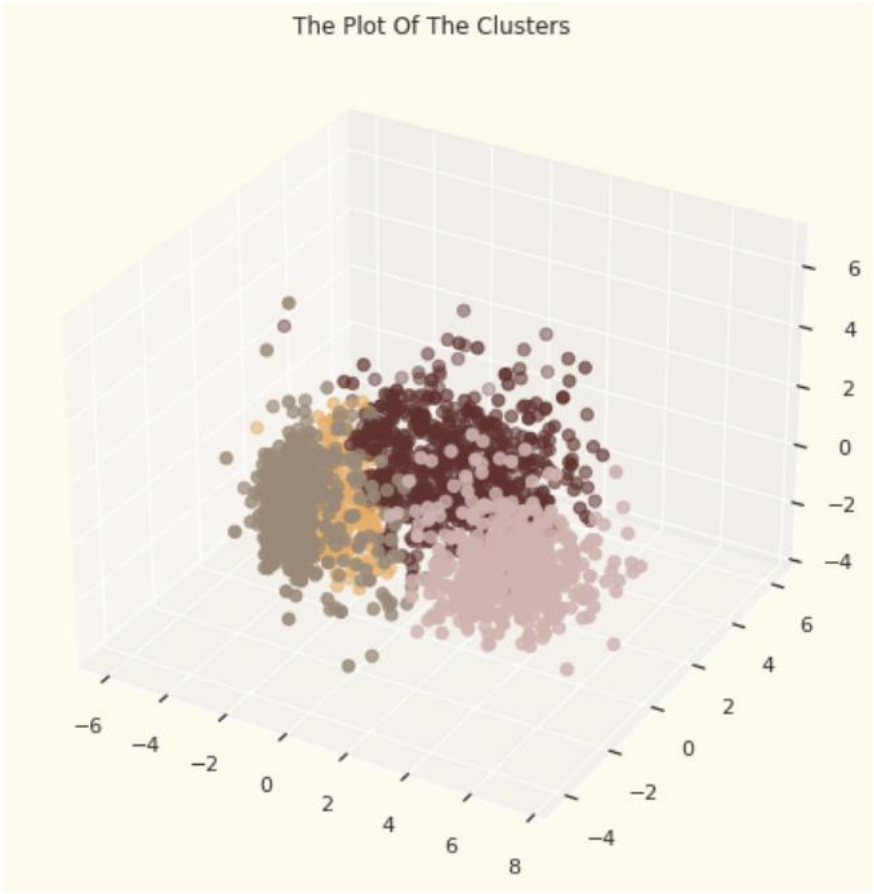
香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

2. Clustering Method

K-Means Clustering: The most common partition-based clustering method used in this



Elbow Method to determine the number of clusters to be formed



To examine the clusters formed let's have a look at the 3-D distribution of the clusters.

3. Evaluating Models

Since this is an unsupervised clustering. We do not have a tagged feature to evaluate or score our model. The purpose of this section is to study the patterns in the clusters formed and determine the nature of the clusters‘ patterns. For that, we will be having a look at the data in light of clusters via exploratory data analysis and drawing conclusions. **Firstly, let us have a look at the group distribution of clustering**

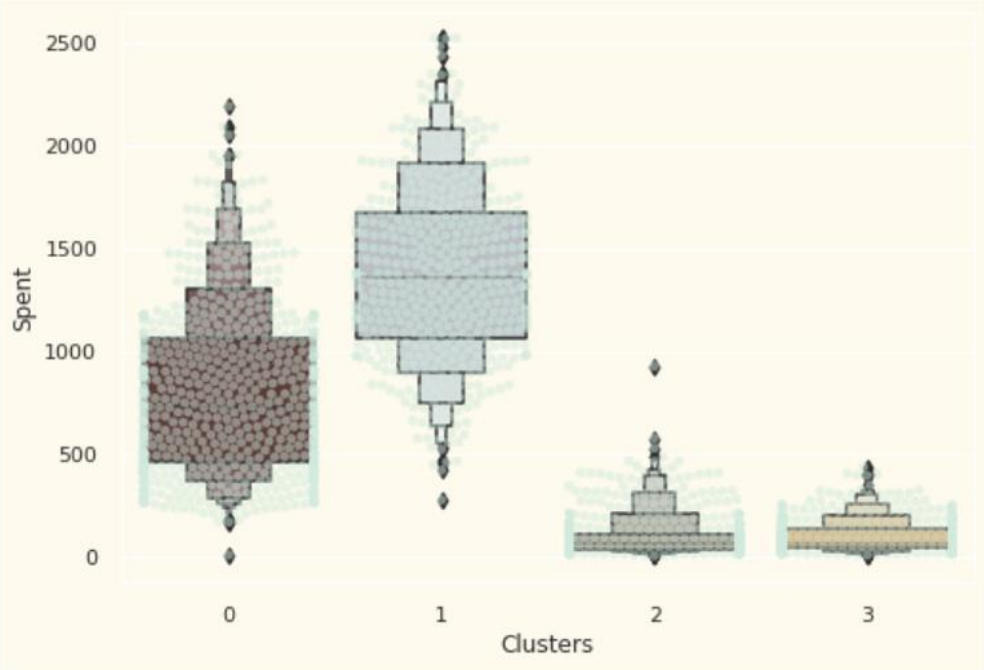
The clusters seem to be fairly distributed.

3. Evaluating Models

Income vs spending plot shows the clusters pattern

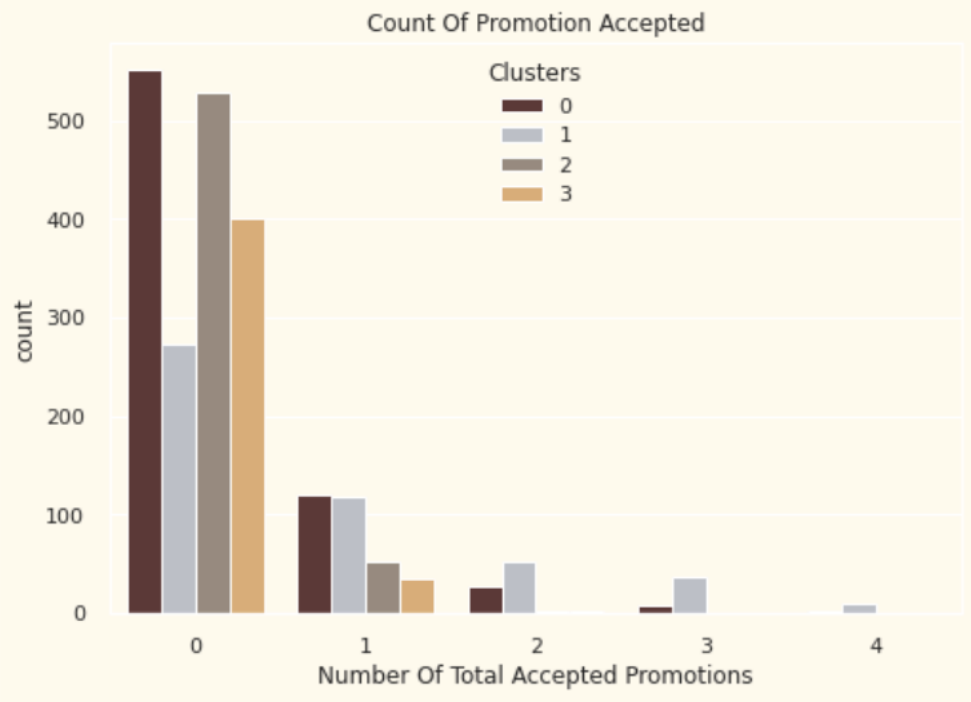
- group 0: high spending & average income
- group 1: high spending & high income
- group 2: low spending & low income
- group 3: high spending & low income

3. Evaluating Models



From the above plot, it can be clearly seen that cluster 1 is our biggest set of customers closely followed by cluster 0. We can explore what each cluster is spending on for the targeted marketing strategies.

3. Evaluating Models

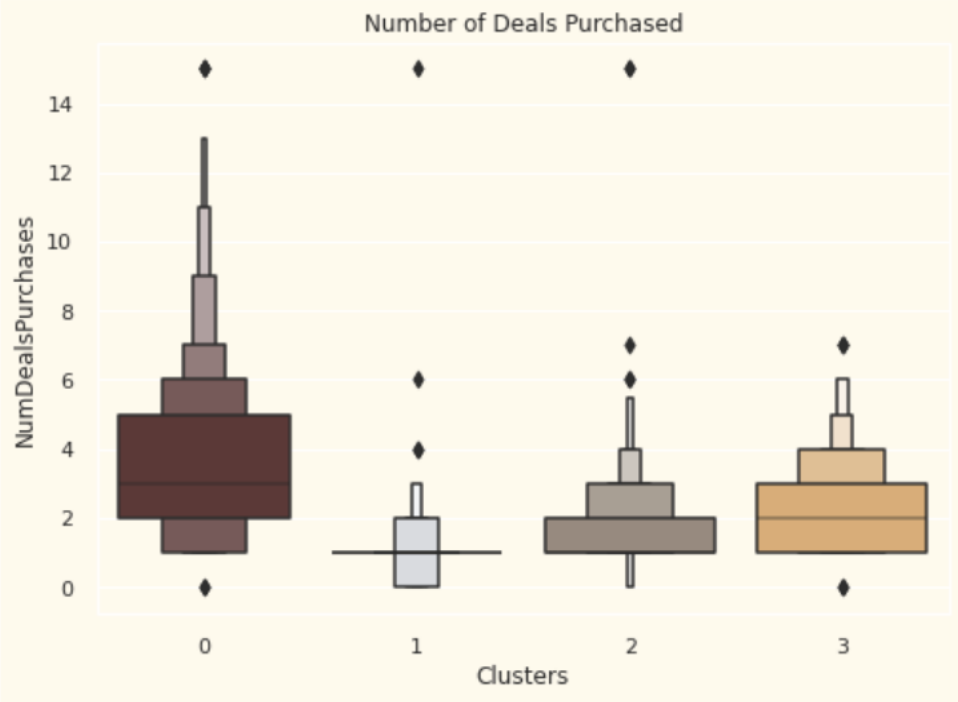


There has not been an overwhelming response to the campaigns so far. Very few participants overall. Moreover, no one part take in all 5 of them. Perhaps better-targeted and well-planned campaigns are required to boost sales.

Default Topic

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

3. Evaluating Models



Unlike campaigns, the deals offered did well. It has best outcome with cluster 0 and cluster 3. However, our star customers cluster 1 are not much into the deals. Nothing seems to attract cluster 2 overwhelmingly

Default Topic

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

4. Profiling

Now that we have formed the clusters and looked at their purchasing habits. Let us see who all are there in these clusters. For that, we will be profiling the clusters formed and come to a conclusion about who is our star customer and who needs more attention from the retail store's marketing team.

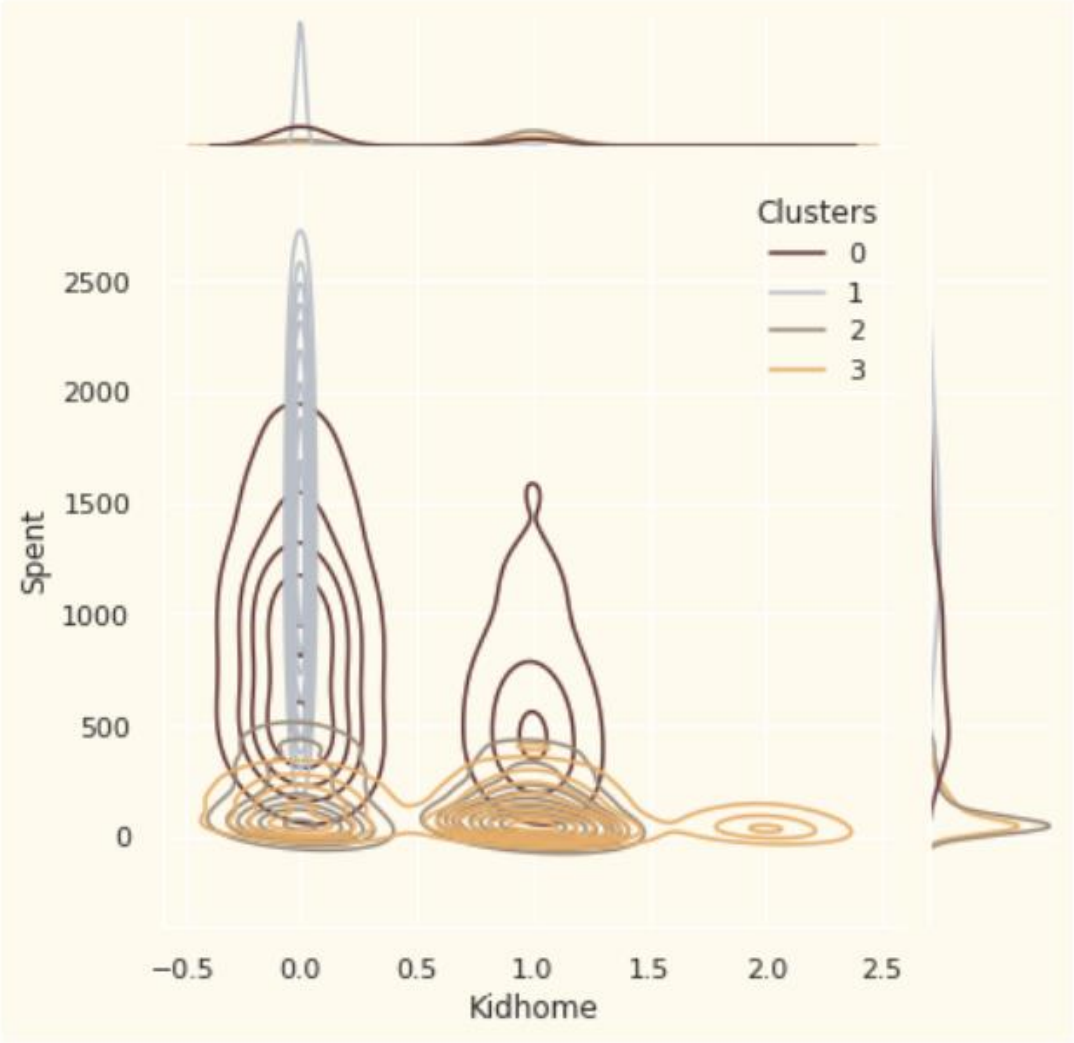
To decide that I will be plotting some of the features that are indicative of the customer's personal traits in light of the cluster they are in.

Default Topic

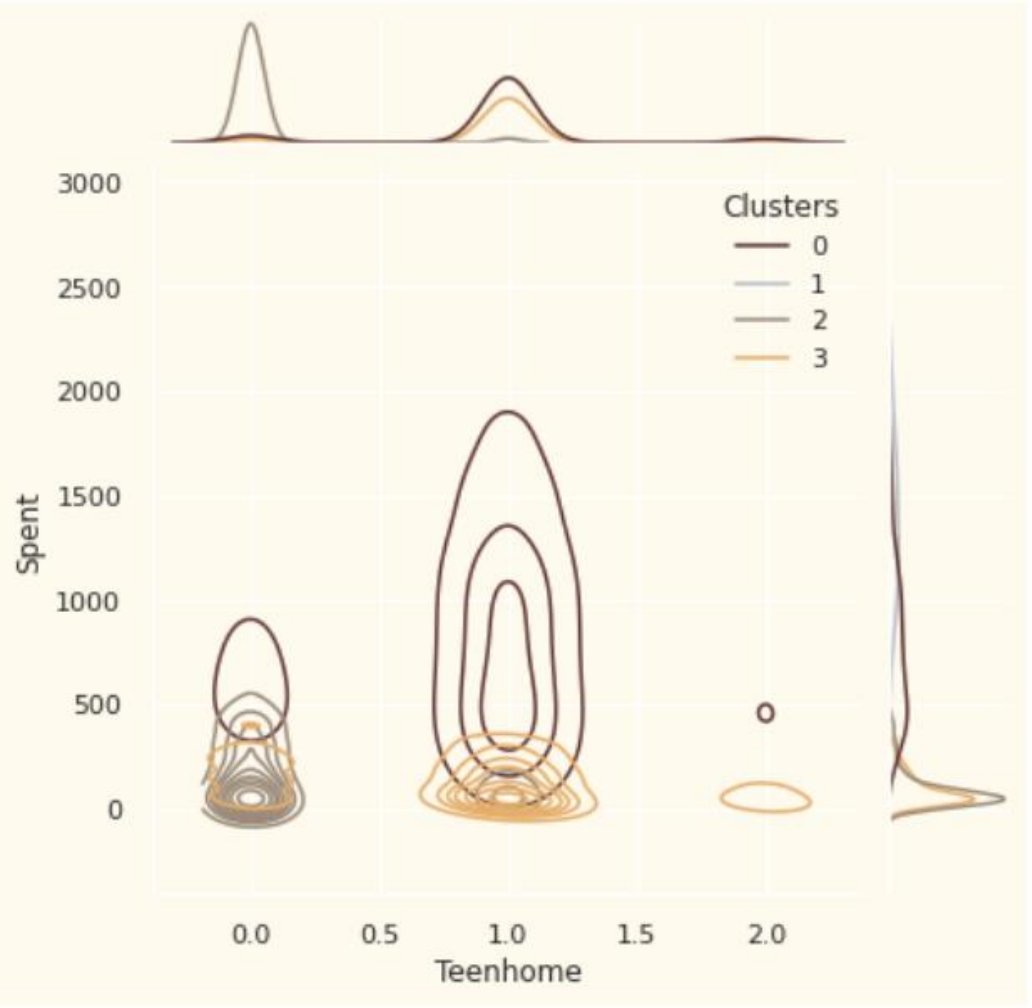
香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

4. Profiling

Kidhome vs Spent



Teenhome vs Spent

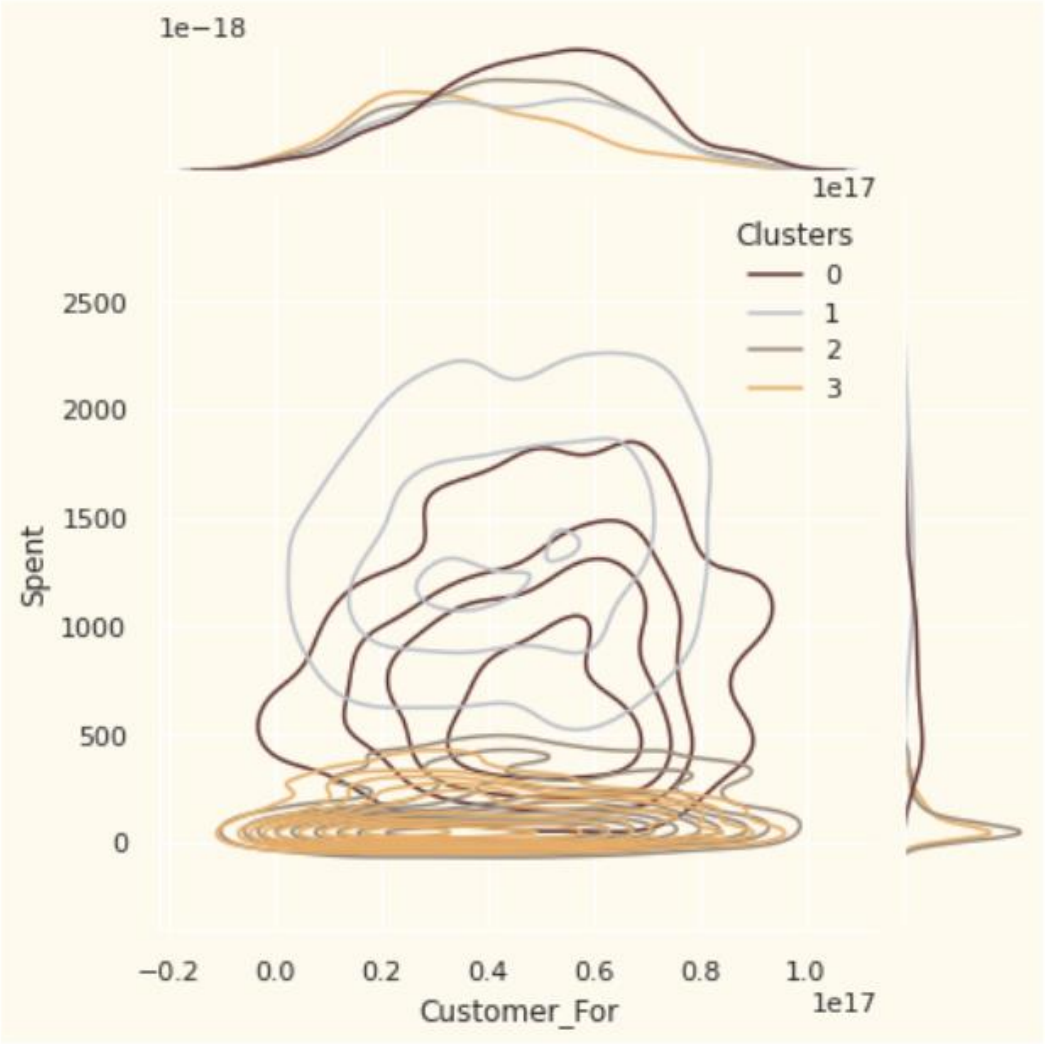


Default Topic

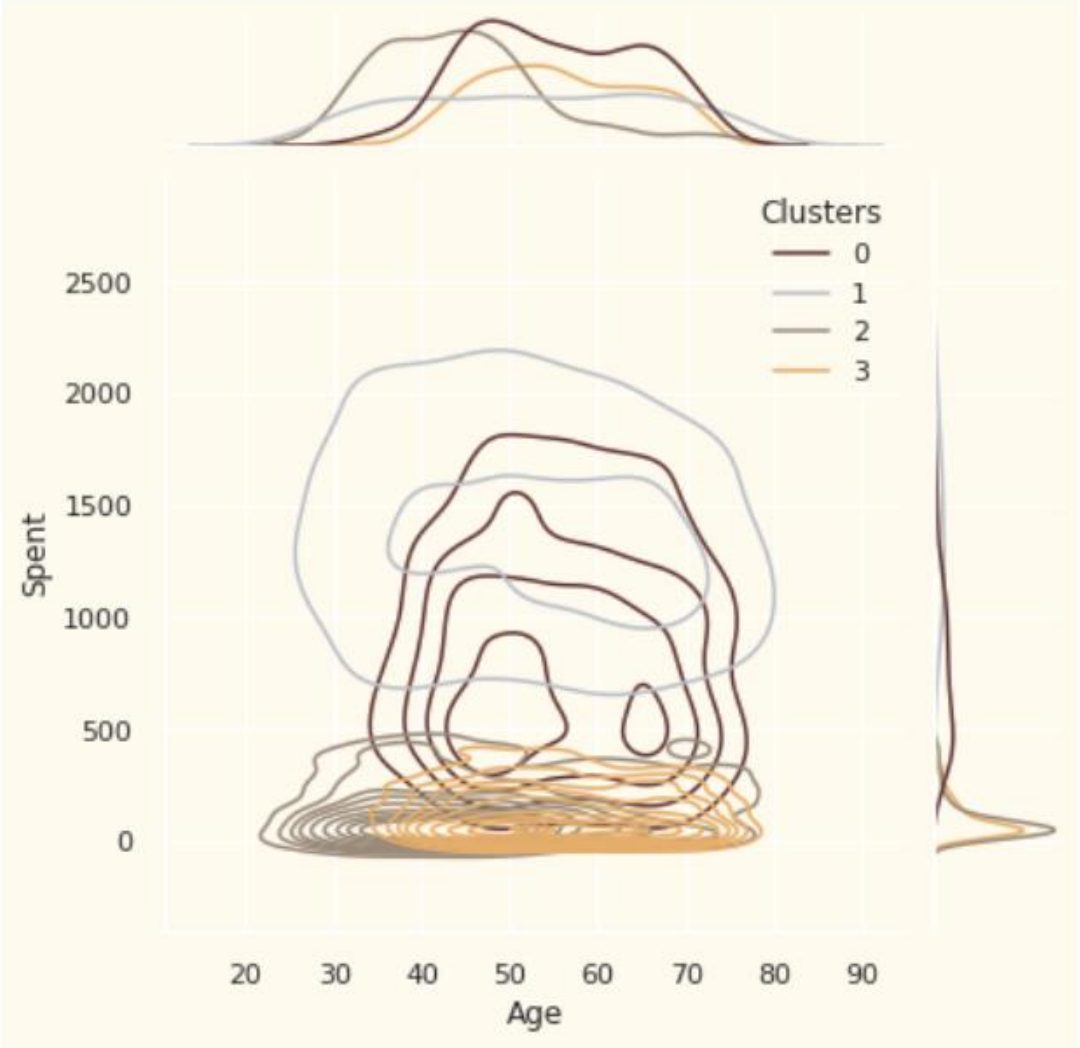
香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

4. Profiling

Customer_For vs Spent



Age vs Spent

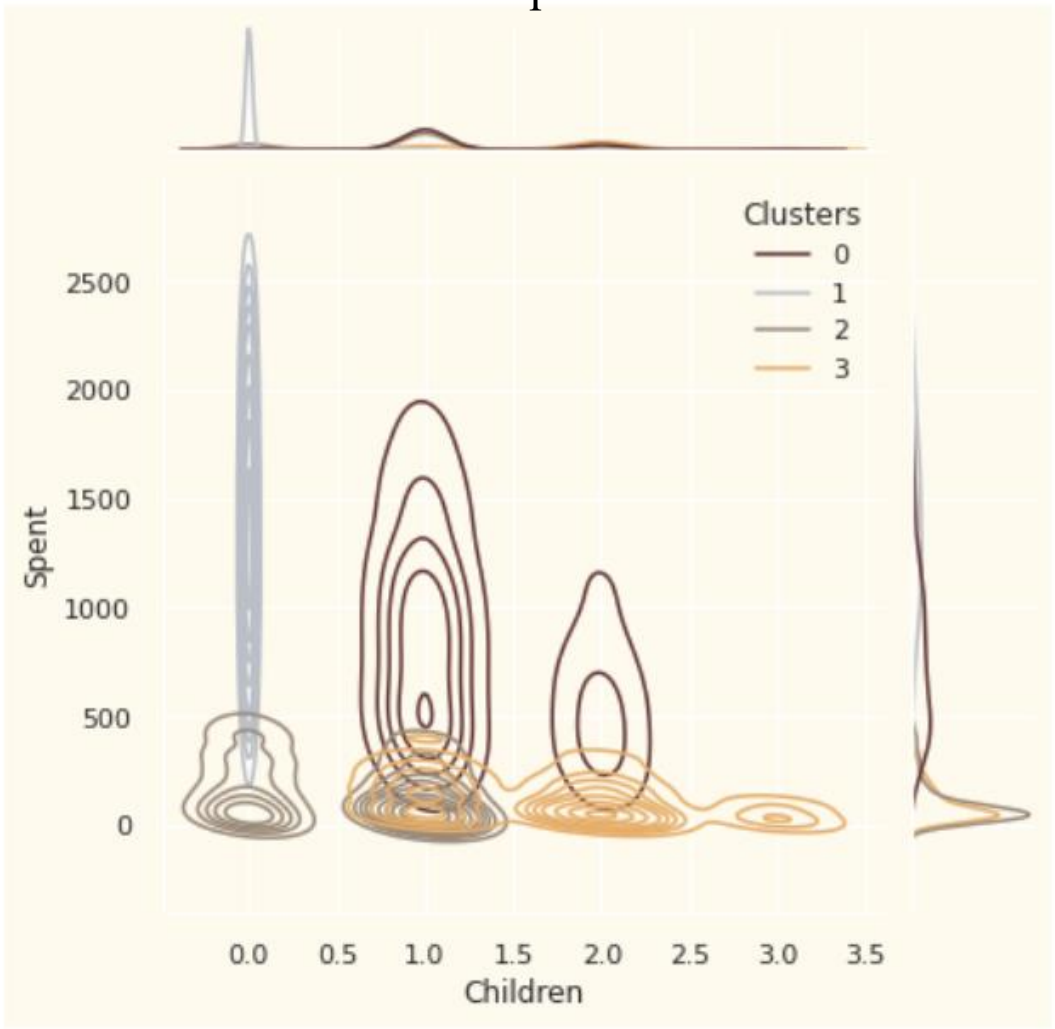


Default Topic

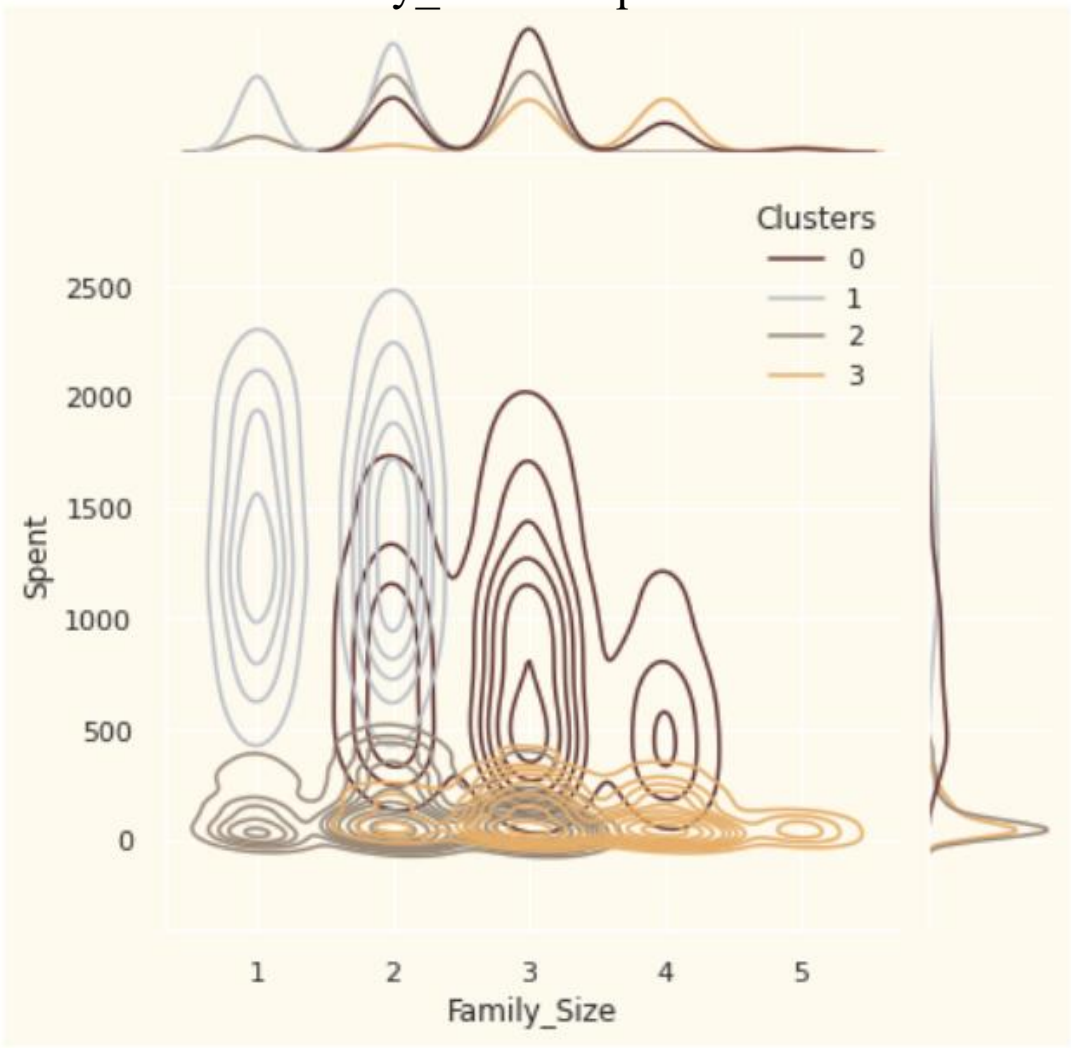
香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

4. Profiling

Children vs Spent



Family_Size vs Spent

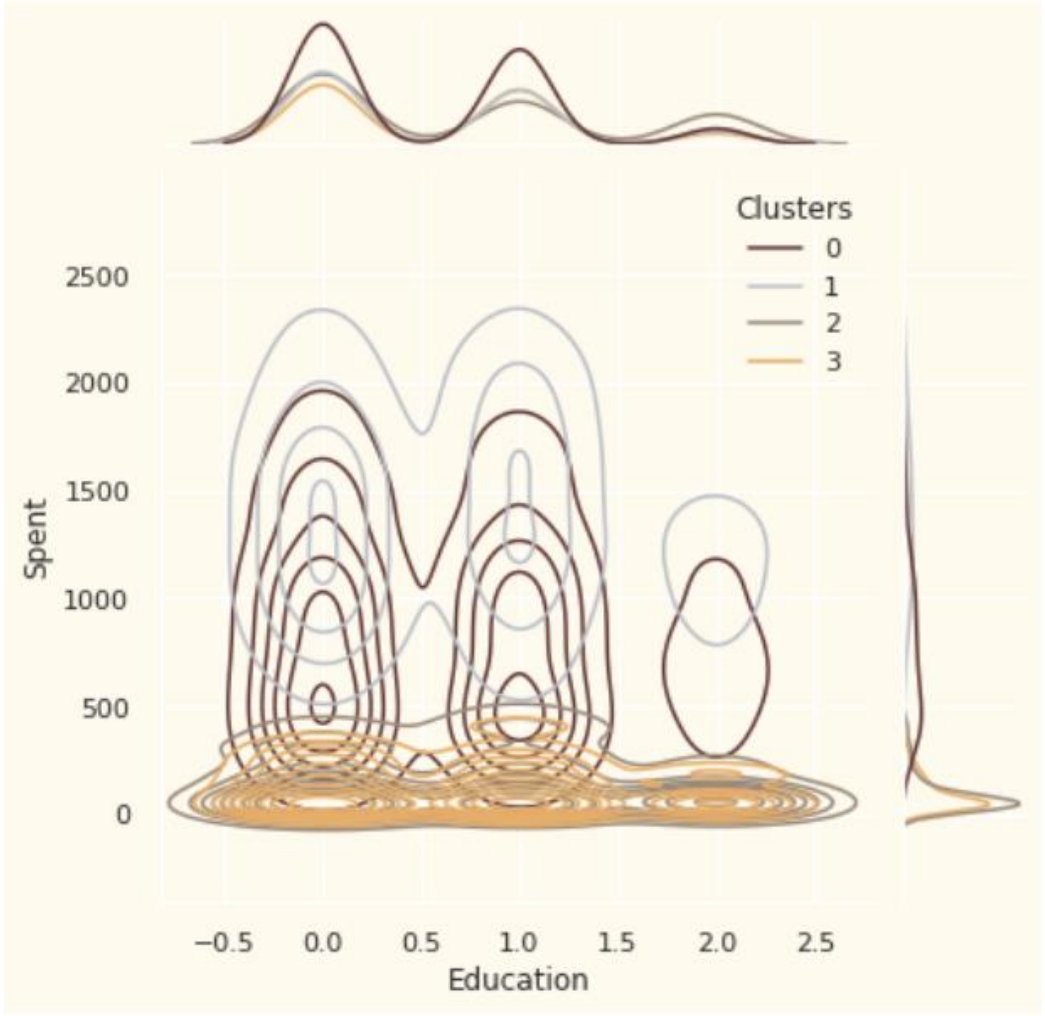


Default Topic

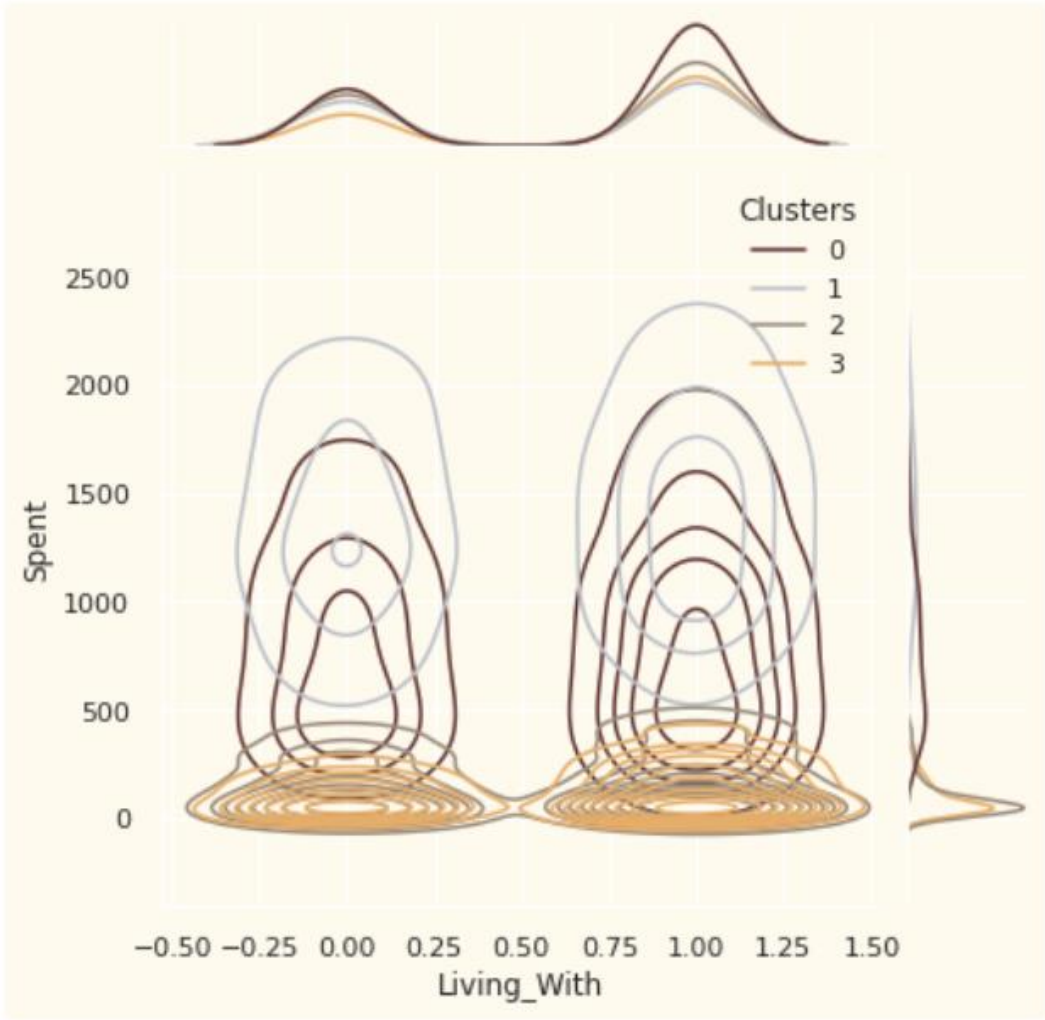
香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

4. Profiling

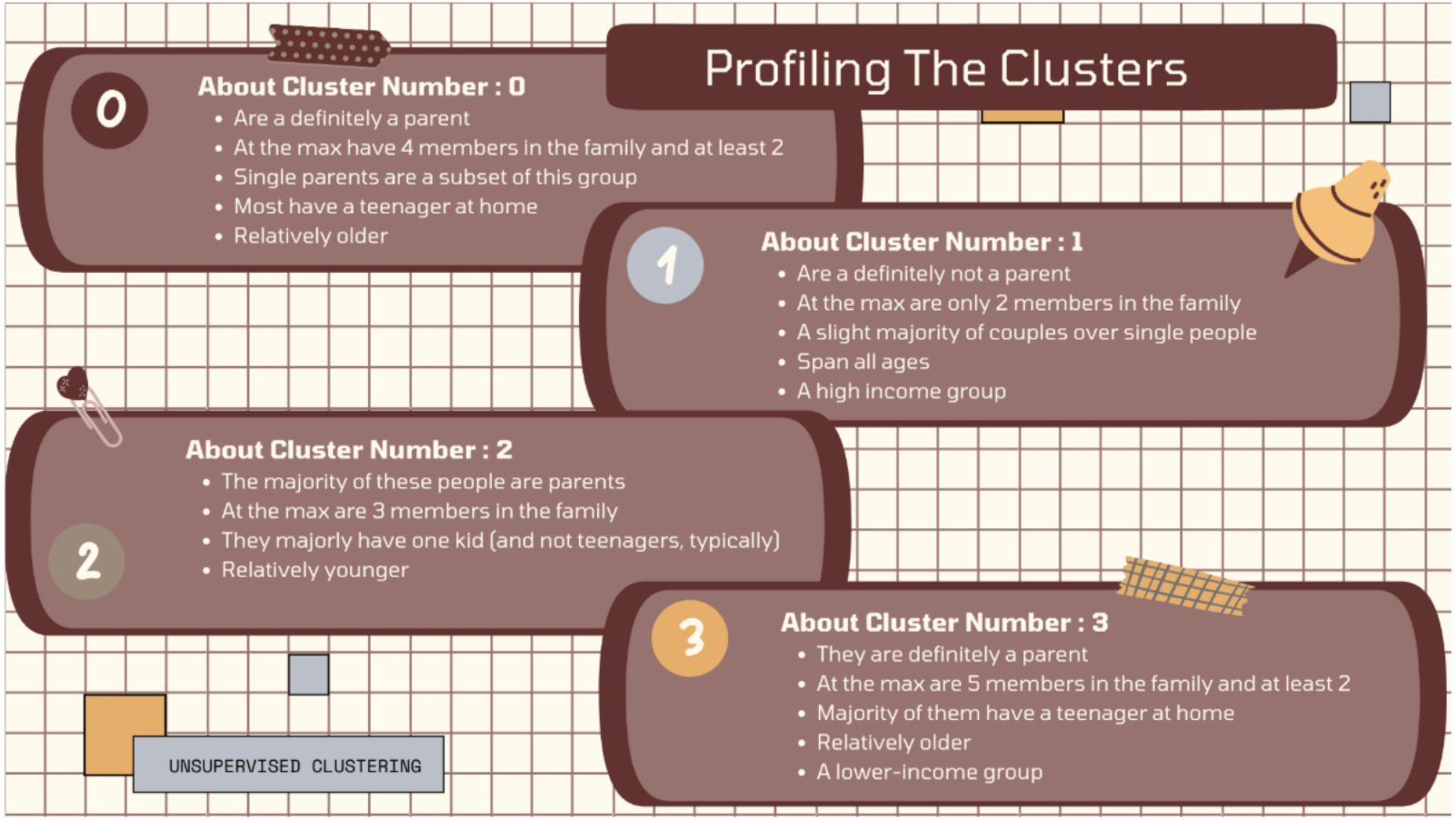
Education vs Spent



Living_With vs Spent



4. Profiling


Profiling The Clusters
0
About Cluster Number : 0
Are a definitely a parent
At the max have 4 members in the family and at least 2
Single parents are a subset of this group
Most have a teenager at home
Relatively older
1
About Cluster Number : 1
Are a definitely not a parent
At the max are only 2 members in the family
A slight majority of couples over single people
Span all ages
A high income group
2
About Cluster Number : 2
The majority of these people are parents
At the max are 3 members in the family
They majorly have one kid (and not teenagers, typically)
Relatively younger
3
About Cluster Number : 3
They are definitely a parent
At the max are 5 members in the family and at least 2
Majority of them have a teenager at home
Relatively older
A lower-income group
UNSUPERVISED CLUSTERING

Reminder

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- This project is an opportunity to think critically and innovate. Please do NOT copy existing solution. Get your hands dirty and explore your own ideas.
- **Plagiarism** in any form is strictly prohibited. Violations will result in 0 points.

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Zebra: When Temporal Graph Neural Networks Meet Temporal Personalized PageRank

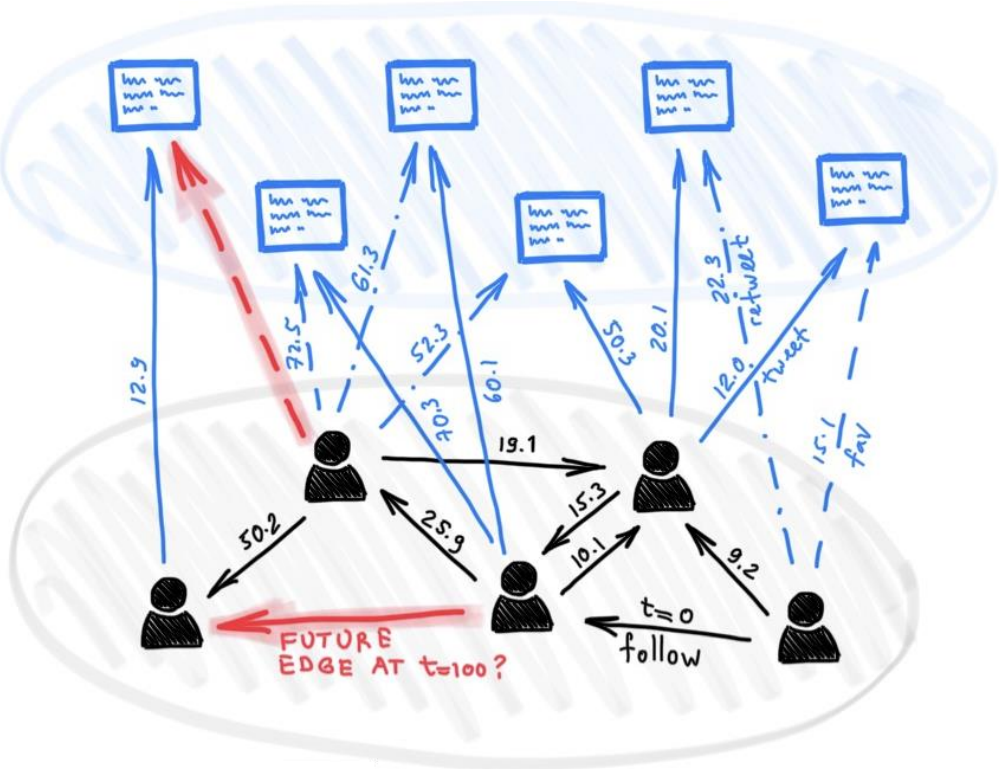
Yiming Li[1], Yanyan Shen[2], Lei Chen[1], Mingxuan Yuan[3]

[1]Hong Kong University of Science and Technology, Hong Kong, China

[2]Shanghai Jiao Tong University, Shanghai, China

[3]Huawei Noah's Ark Lab, Hong Kong, China

Background



A dynamic network of Twitter users interacting with tweets and following each other.

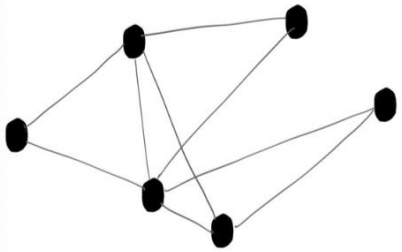
Applications on dynamic graphs

- Temporal link prediction
- Dynamic node classification
- Social network
- Recommendation
- E-commerce
- Fraud detection

Background

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

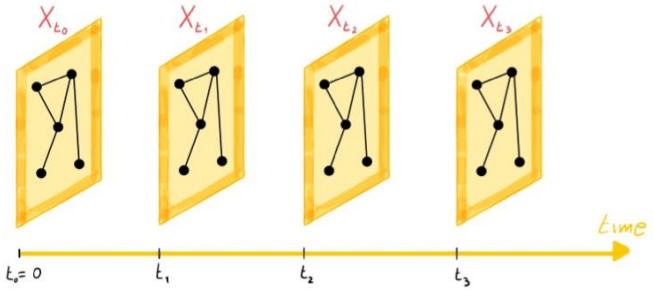
$G = (V, E, X)$



Static Graph

- No notion of time

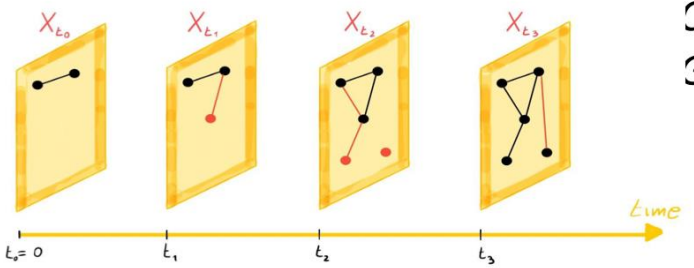
$G_t = (V, E, X_t)$



Spatio-Temporal Graph

- Topology is fixed, but features change over time
- (Usually) observed at regular intervals
- Examples: *traffic forecasting*, *covid-19 forecasting*

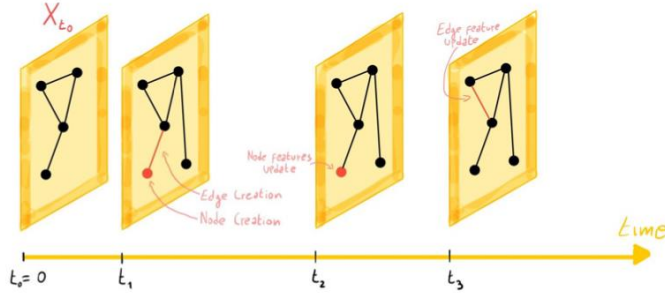
$G_t = (V_t, E_t, X_t)$



Discrete-Time Dynamic Graph (DTDGs)

- Both topology and features change over time
- However, graph is observed at regular intervals (no information about what happens in between)
- Examples: Any system which is observed at regular intervals

$G(t) = \{x_{t_1}, x_{t_2}, \ldots\} \quad 0 \leq t_1 \leq t_2 \leq \cdots \leq t$



Continuous-Time Dynamic Graph (CTDGs)

- Most general formulation
- Each change (‘*event*’) in the graph is observed individually with its timestamp
- Examples: Recommender Systems

Less General → *More General*

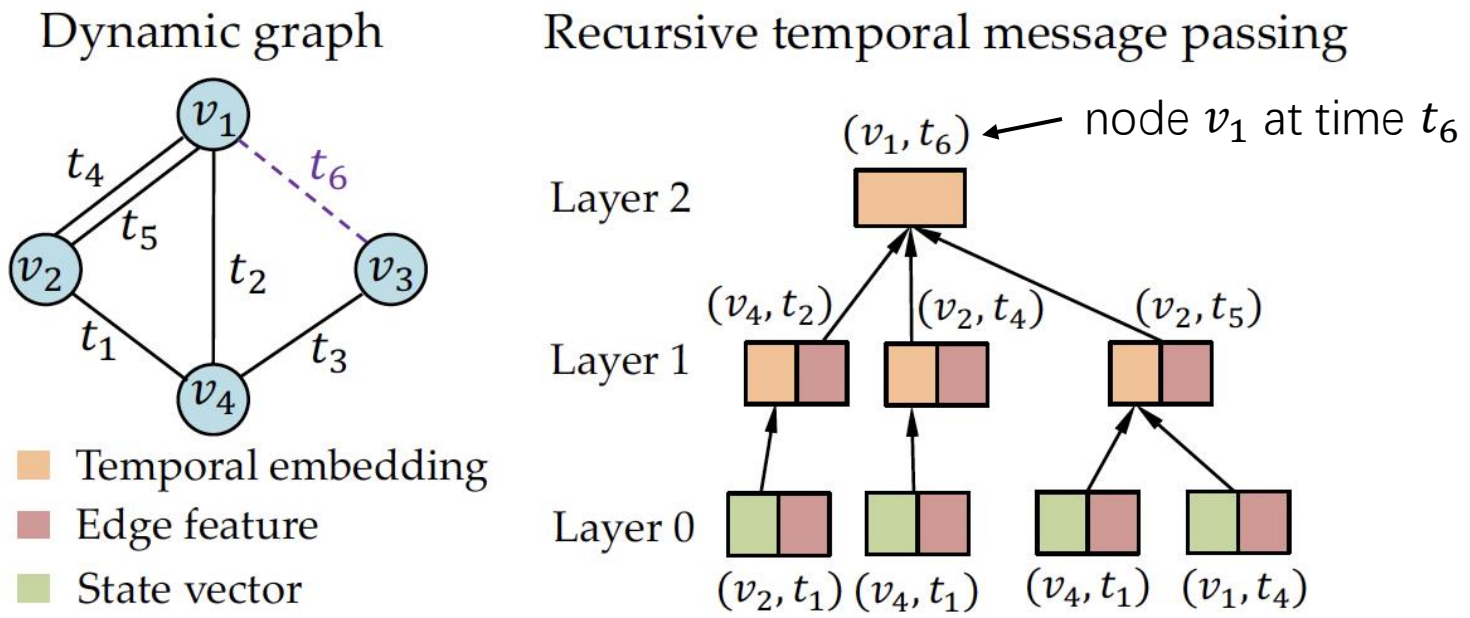
Background

- Using a static GNN for dynamic graphs would mean
  - Inefficiency: computation is repeated each time we want to make prediction
  - Loss of information: not able to take into account how the graph evolved

- T-GNNs learn embeddings on continuous-time dynamic graph
  - Support addition/deletion of nodes, edges, as well as feature changes
  - Make predictions (e.g., classify a node) at any time point

Background

- T-GNN: given a node v at timestamp t, an L-layer T-GNN computes its temporal embedding via recursive temporal message passing



Example of a 2-layer vanilla T-GNN.

Motivation

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- T-GNN identify influential neighbor nodes via **tedious recursive temporal neighborhood aggregation**

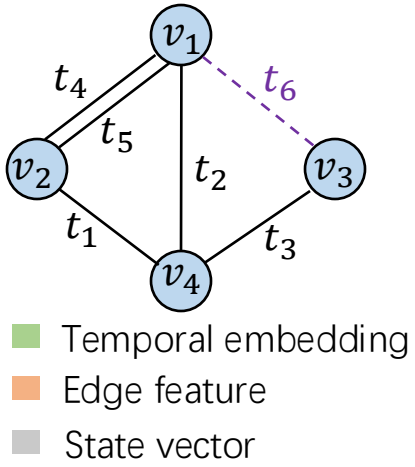
- Do there exist influential nodes that we can utilize to accelerate the computation of T-GNNs without compromising model accuracy?

T-GNN and Temporal Random Walk

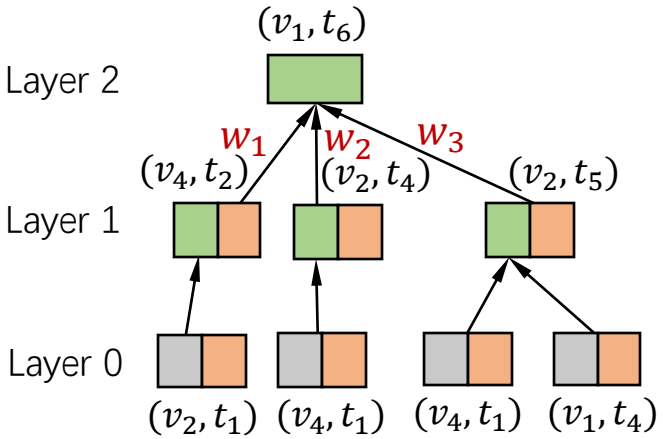
香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- T-GNN mimics the temporal random walk process on dynamic graphs
- Key observation: 1-hop neighbor weight implicitly defines the 1-step temporal random walk probability on dynamic graphs

Dynamic graph



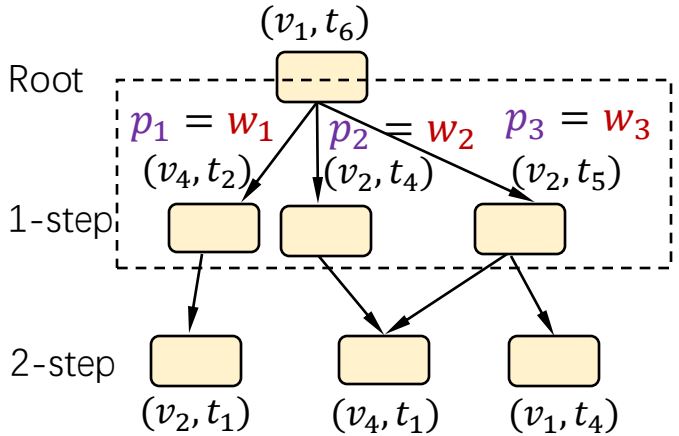
Recursive temporal message passing



1-hop neighbor weighs learned by T-GNN

$w_1 + w_2 + w_3 = 1$

Temporal random walk

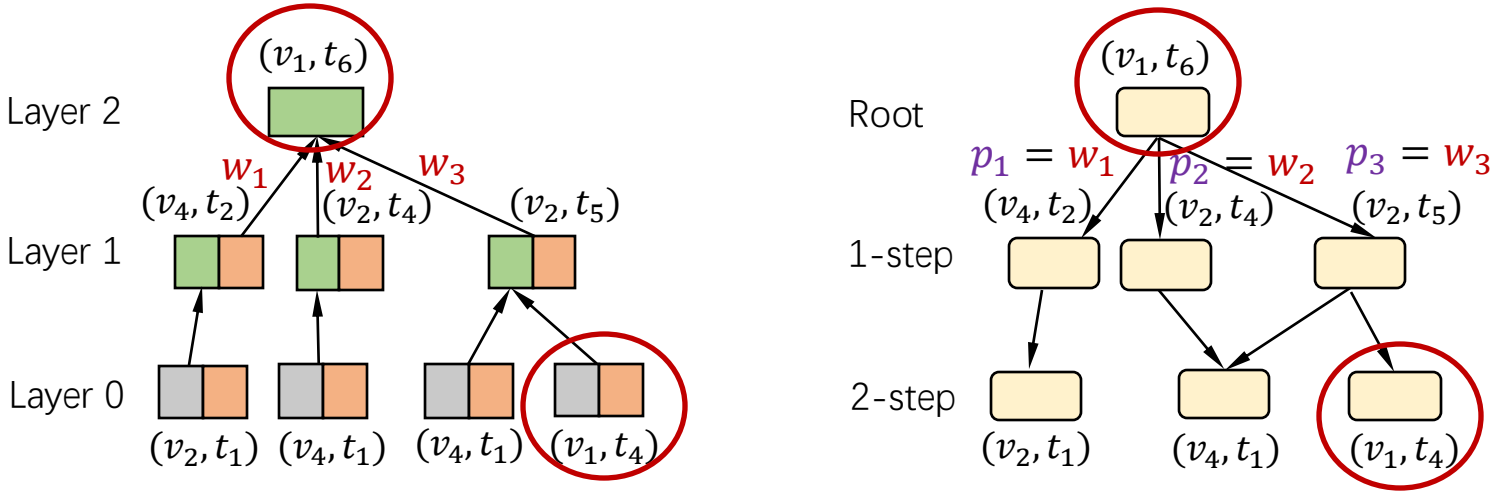


1-step temporal random walk probabilities

$p_1 + p_2 + p_3 = 1$

Insights

- Implication
  - We can identify influential neighbors if there exists a method that is able to sufficiently capture the diffusion process (temporal random walk probabilities) on dynamic graphs
  - In this way, we can prune insignificant neighbors and avoid the laborious process of recursive temporal message passing

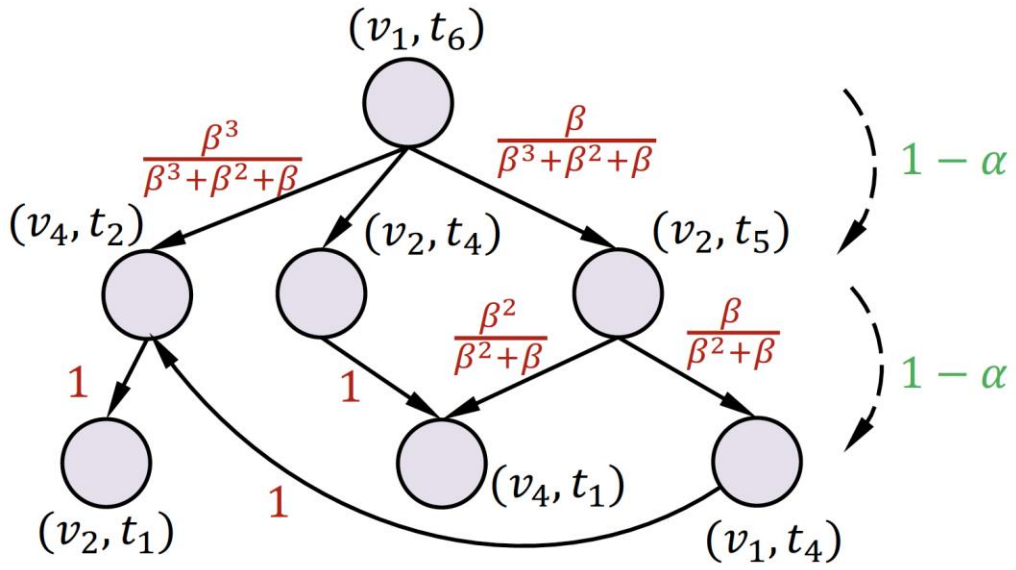


Challenge: As we cannot know the neighbor weights in advance without running T-GNN, how to estimate the corresponding temporal random walk probabilities?

T-PPR

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

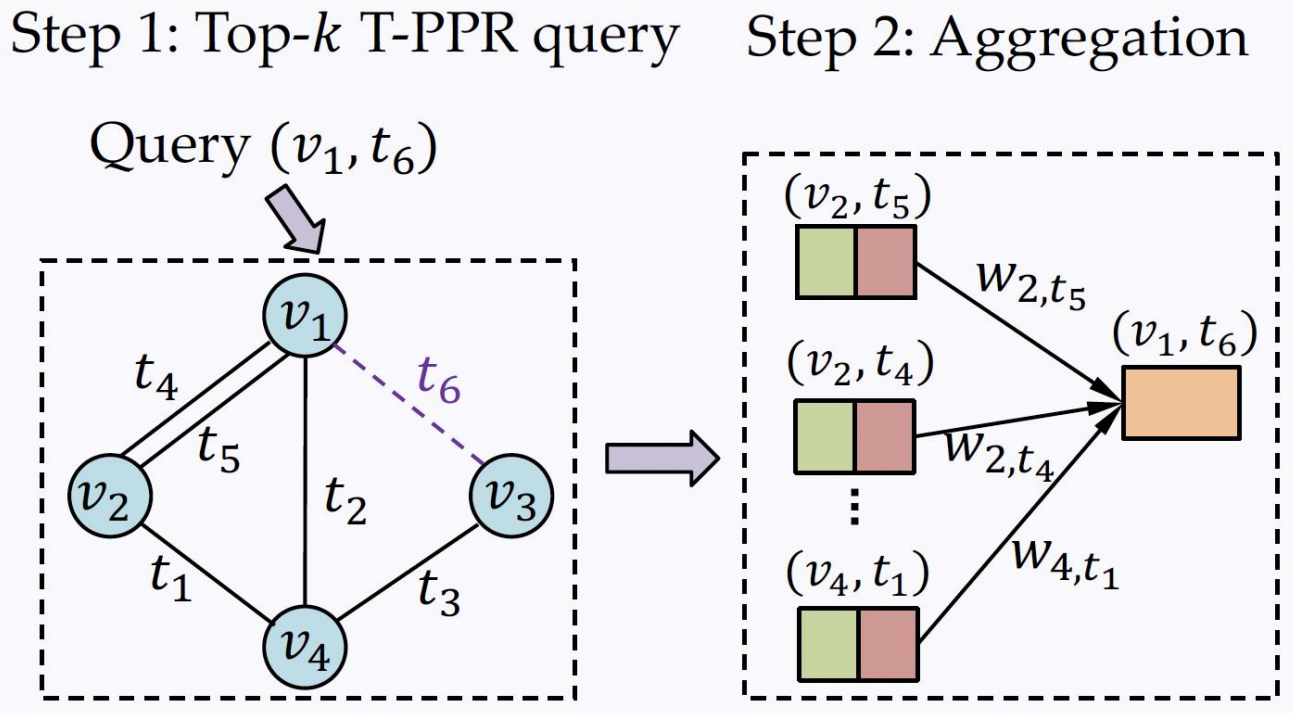
- Temporal Personalized PageRank (T-PPR) is a parameterized metric for estimating the influence score of temporal nodes on dynamic graphs
  - The T-PPR value of (j, τ) with respect to (i, t) is defined as the probability that an α-temporal random walk with β-exponential decay starting from (i, t) terminates at (j, τ)



Methods

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Zebra framework avoids the tedious recursive temporal message passing



Methods

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Existing techniques for PPR cannot be adapted to efficiently solve the top-k T-PPR query due to the highly imbalanced weight distribution

- A straightforward solution——power iteration
  - Recursively traverses the graph and then selects the top-k temporal nodes
  - However, this approach is computationally prohibitive

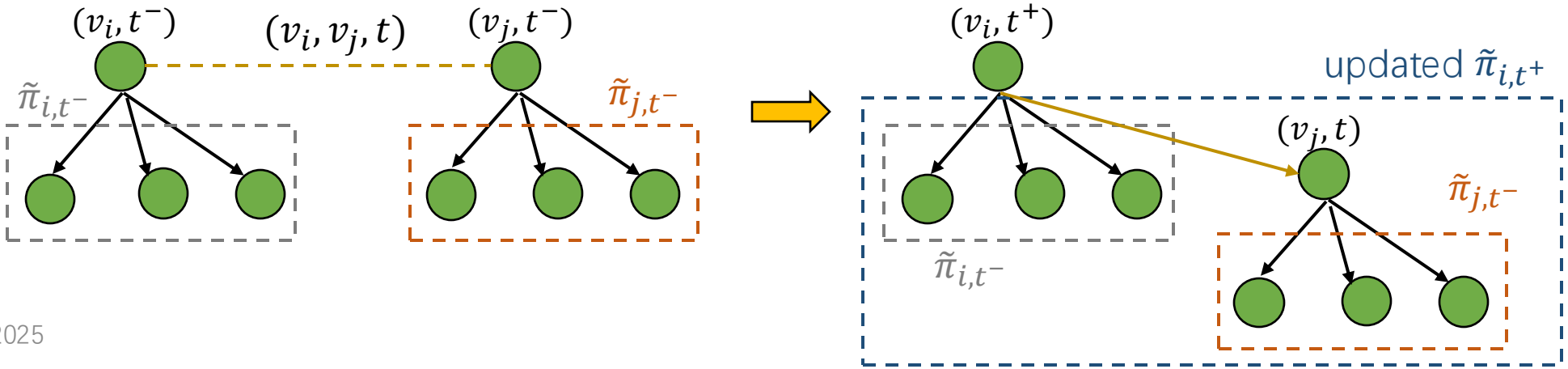
Methods

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Single-scan Top-k T-PPR Algorithm (SANTA)

- Data structure
  - For each node v_i SANTA maintains $\tilde{\pi}_{i,t^-}$, the latest approximated top-k T-PPR dictionary for node v_i before the time current timestamp t

- Single-edge update
  - Given a new interaction (v_i, v_j, t), SANTA updates the top-k T-PPR dictionary for node v_i by merging $\tilde{\pi}_{i,t^-}$ and $\tilde{\pi}_{j,t^-}$ in a streaming fashion



Theoretical Results (Optional)

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Additive error bound: the additive errors are negligible when the temporal random walk probabilities are highly skewed or a large k is adopted

THEOREM 7 (ADDITIVE ERROR BOUND). *Given a new interaction* $\gamma(t) = (v_i, v_j, e_{ij}(t), t)$, *let* $\tilde{\pi}_{i,t^+}$ *be the top-k T-PPR dictionary returned by SANTA. For any temporal node* (z, τ), *we have*

$$\tilde{\pi}_{i,t^+}(z,\tau) \leq \pi_{i,t^+}(z,\tau) \leq \tilde{\pi}_{i,t^+}(z,\tau) + \frac{\theta^{(k)}_{max}}{\alpha(1-\beta)}, \tag{16}$$

where $\theta^{(k)}_{max}$ *is the maximum possible top-k threshold for the T-PPR metric parameterized by* α *and* β.

Datasets & Baselines

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Statistics of dynamic graphs.

Dataset	$\|V\|$	$\|E\|$	d_v	d_e	Timespan
MOOC [22]	7,144	411,749	172	4	30 days
Wikipedia [22]	9,227	157,474	172	172	30 days
Reddit [22]	10,984	672,447	172	172	30 days
AskUbuntu [1]	159,316	964,437	172	0	2613 days
SuperUser [2]	194,085	1,443,339	172	0	2773 days
Wiki-Talk [4]	1,140,149	7,833,140	172	0	2320 days

- **Zebra** identifies essential neighbors via SANTA (Algorithm 1) and computes node embeddings through 1-layer neighborhood aggregation (Eq.11). In addition, Zebra updates the state vectors of nodes upon new interactions as vanilla T-GNN [36].
- JODIE [22] uses RNNs to propagate the information of temporal interactions to update node representations.
- TGAT [54] mimics the message passing scheme of static GNNs and encodes time information through random Fourier features.
- TGN [36] dynamically maintains a state vector for each node and includes previous methods [22, 41, 54] as special cases.
- CAW [50] encodes temporal neighborhood information using anonymized random walk and attention modules.
- APAN [49] accelerates model inference by decoupling graph computation and message propagation.

Experiments

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Zebra can be orders of magnitude faster than the state-of-the-art baselines without sacrificing model performance

Table 4: Comparison of T-GNNs on small dynamic graphs. Zebra is implemented as an ensemble of two top-20 T-PPR metrics. We report model performance in transductive average precision (%), inductive average precision (%), and total training time (s). In addition, we report the number of epochs required for model convergence in parentheses. The best and second-best results in each metric are marked in bold and underlined, respectively. "∗" indicates that the performance improvement of Zebra over the best-performing baseline is statistically significant with the significance level set to be 0.05.

Model	Wikipedia			Reddit			MOOC		
	Trans AP	Induct AP	Training (#)	Trans AP	Induct AP	Training (#)	Trans AP	Induct AP	Training (#)
JODIE	95.16 ± 0.4	93.13 ± 0.5	2356.4 (18)	95.83 ± 0.3	93.20 ± 0.4	9243.64 (14)	83.26 ± 0.5	81.77 ± 0.4	5160.30 (15)
TGAT	94.26 ± 0.1	92.88 ± 0.3	2881.4 (29)	97.80 ± 0.2	96.08 ± 0.3	11933.8 (21)	70.22 ± 0.4	70.83 ± 0.5	7838.5 (25)
TGN	98.58 ± 0.1	98.05 ± 0.1	2182.7 (26)	98.66 ± 0.1	97.55 ± 0.1	11528.9 (26)	88.88 ± 1.7	88.17 ± 2.1	5086.83 (21)
APAN	96.41 ± 0.5	96.06 ± 0.4	1605.0 (21)	98.50 ± 0.2	97.62 ± 0.7	16431.8 (18)	87.02 ± 0.3	86.74 ± 0.5	3374.1 (14)
CAW	98.18 ± 0.1	98.24 ± 0.1	10175.5 (13)	98.54 ± 0.1	97.97 ± 0.1	75585.1 (16)	80.60 ± 0.4	80.18 ± 0.4	34063.9 (14)
Zebra	**98.67 ± 0.1**	**98.59 ± 0.1***	**302.6 (34)**	**98.76 ± 0.1***	**98.28 ± 0.1***	**1342.0 (31)**	**92.45 ± 0.2***	**89.56 ± 0.3**	**619.5 (25)**

Experiments

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Zebra is also memory efficient
  - It reduces the computational cost of T-GNNs and thus save the GPU memory required to perform the corresponding computations

Table 5: Comparison of T-GNNs on large dynamic graphs. Zebra is implemented as an ensemble of two top-20 T-PPR metrics. The evaluation metrics are the same as those in Table 4. Particularly, "OOM" denotes out-of-memory error, and "TLE" denotes time limit exceed such that we cannot finish one epoch of model training in 12 hours. "∗" indicates that the performance improvement of Zebra over the best-performing baseline is statistically significant with the significance level set to be 0.05.

Model	AskUbuntu			SuperUser			Wiki-Talk		
	Trans AP	Induct AP	Training (#)	Trans AP	Induct AP	Training (#)	Trans AP	Induct AP	Training (#)
JODIE	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
TGAT	87.57 ± 0.3	84.21 ± 0.4	11728.5 (21)	86.40 ± 0.5	83.12 ± 0.5	17129.3 (19)	91.72 ± 0.2	85.38 ± 0.3	189420 (34)
TGN	**94.51 ± 0.2**	92.73 ± 0.2	36202.6 (20)	93.18 ± 0.3	91.76 ± 0.2	81747.6 (24)	OOM	OOM	OOM
APAN	89.17 ± 0.1	88.43 ± 0.1	21606.1 (14)	87.07 ± 0.3	85.50 ± 0.5	48724.2 (19)	TLE	TLE	TLE
CAW	90.87 ± 0.2	90.63 ± 0.2	61903.7 (14)	88.92 ± 0.2	88.32 ± 0.1	111744.8 (15)	TLE	TLE	TLE
Zebra	94.47 ± 0.1	**97.91 ± 0.1***	**1362.2 (24)**	**93.21 ± 0.3**	**97.93 ± 0.1***	**2095.1 (23)**	**95.45 ± 0.1***	**97.96 ± 0.1***	**9909.6 (16)**

Report Contents

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Abstract
- Introduction
- Preliminary
- Methodology
- Related Work
- Experiments
- Conclusion

Summary

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Content you need to submit:
  - Source code
  - Report (in required format)
  - Presentation video & PPT

- Pay attention:
  - Clearly state your **project title** and the **specific course topic** your project relates to.
  - Arrive **on time** on your presentation day (even ID/odd ID)
  - Plagiarism in any form is strictly prohibited.

Thanks!

Q & A