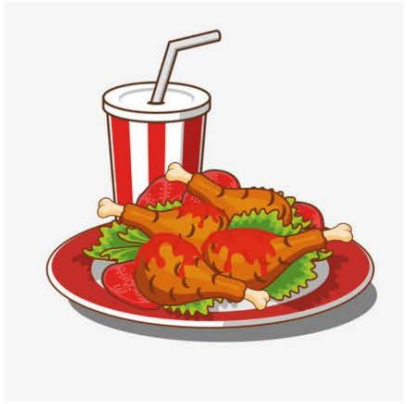
Learning Stable Graphs from Multiple Environments with Selection Bias

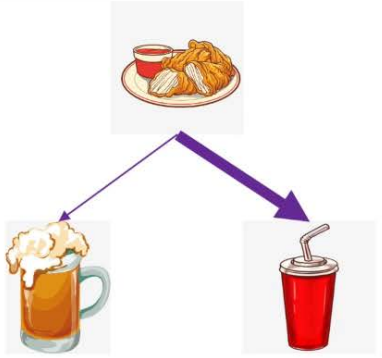
Yue He, Peng Cui, Jianxin Ma, Hao Zou, Xiaowei Wang, Hongxia Yang, and Philip S. Yu. 2020. Learning Stable Graphs from Multiple Environments with Selection Bias. In Proceedings ofthe 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)

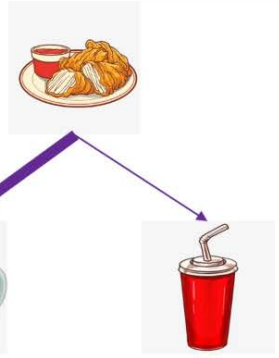
Zekun Cai 20200821

Introduction

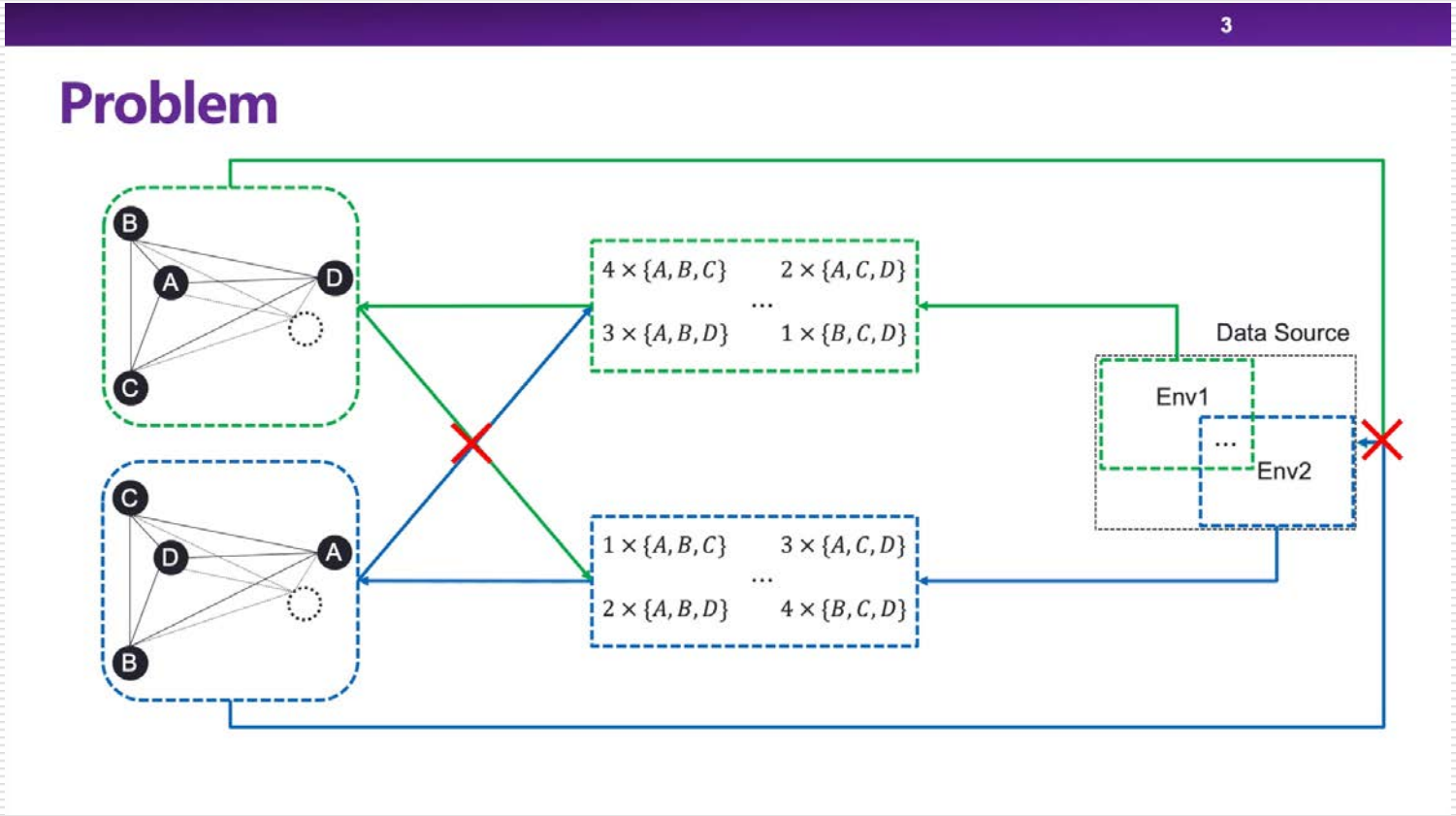
Background







Problem



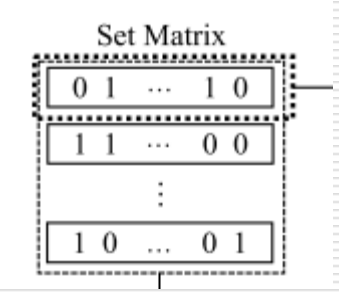
The data collection process is usually full of known or unknown sample selection biases, leading to spurious correlations among entities. (product purchasing graph: female and male; different geographical regions; long time-span)

→We need stable graph across different environments with selection bias.

The learned graph can be applied into multiple, even unknown environments and help to produce stable performances with subsequent tasks.

Concept

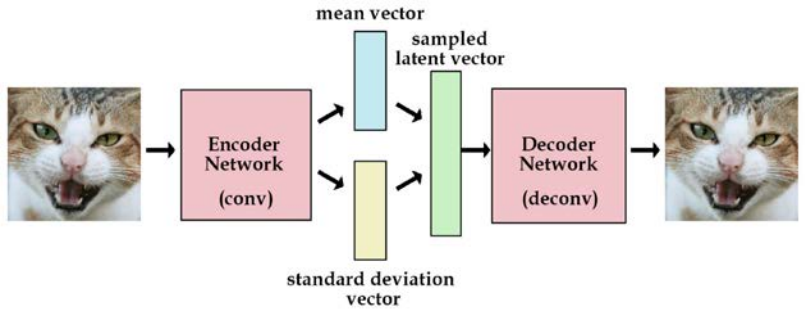
- Set data
  - a data sample is represented by a set (e.g. a shopping basket in recommendation systems)



- Elements
  - Graph nodes(e.g. products in recommendation systems)

- Environments
  - different environments are with different selection bias

- VAE（变分自编码器）
  - Deep Generative Model



$$l_i(\theta,\phi) = -E_{z\sim q\theta(z|x_i)}[log(p_\phi(x_i|z))] + KL(q_\theta(z|x_i)||p(z))$$

重建 loss+KL散度

Notations and Problem

- 给出M个不同环境的graph和set data $\{(G^{(m)}, S^{(m)})\}_{m=1}^{M}$ 我们的任务是在其之上学习出一个稳定的图结构（如图的邻接矩阵），可以在不同的环境中都有良好的效果。

- Method
  - 由于环境的不同，不同环境的图结构以及节点的条件概率分布也不同

$$P^{(m)}(I_k|\mathbf{s}) = h(G^{(m)}, I_k, \mathbf{s}).$$

  - 没有数据偏见的图结构和概率分布，应该是所有环境的随机选择。

$$P_S(I_k|\mathbf{s}) = \sum_{m=1}^{M} \frac{P^{(m)}(I_k|\mathbf{s})}{M}.$$

  - 因此，若一个图结构满足稳定图结构，应该满足上面的非偏见条件概率分布
    →我们需要把图结构原始的参数空间(邻接矩阵)变为概率空间，然后在概率空间优化。

METHOD

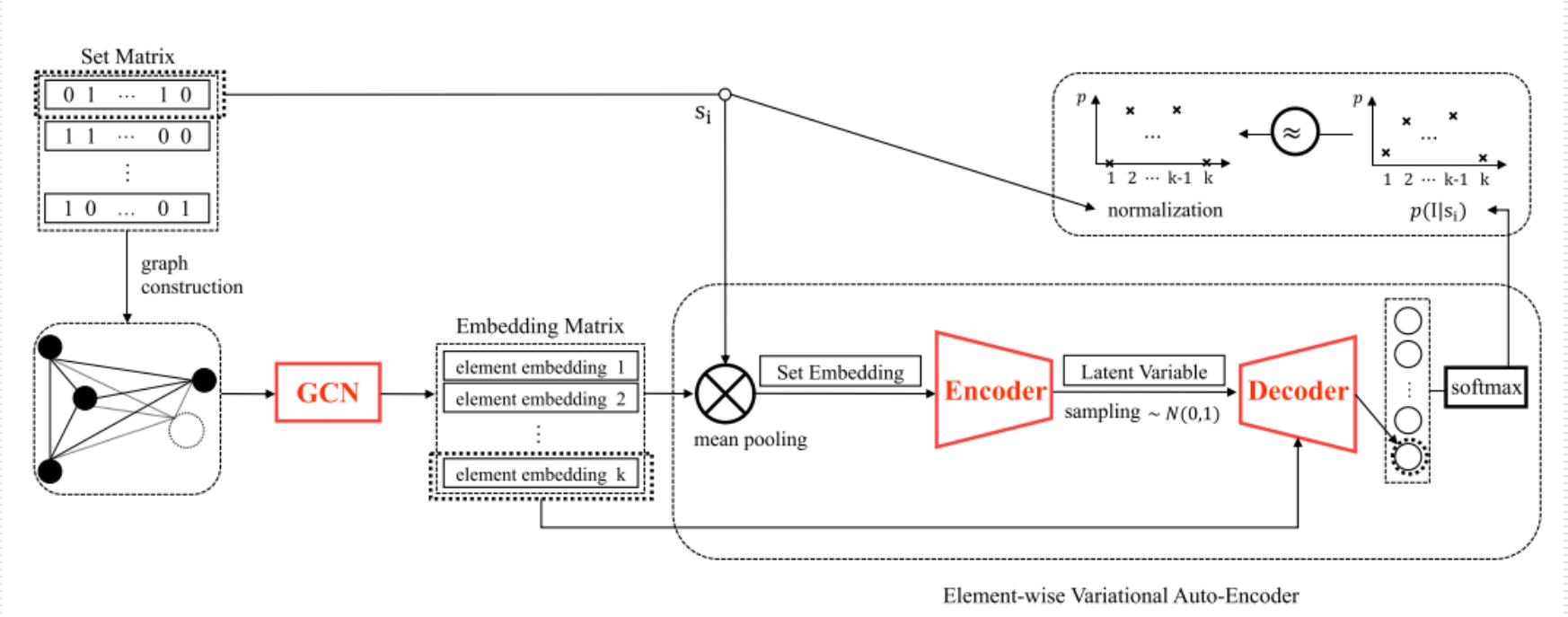
1. Graph Based Set Generation

- Model the function h()
  - GCN学习node的表示得到node的embedding matrix
  - 使用embedding matrix对set编码得到set embedding
  - 将set embedding输入到VAE，得到条件概率分布

$$\mathbf{x}_k^{(l+1)} = \sigma\Big(\sum_{j \in N_k} \hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{x}_j^{(l)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}\Big),$$

$$\frac{\mathbf{s}_i^{(m)} \cdot \mathbf{X}^{(m)}}{\mathbf{s}_i^{(m)} \cdot \mathbf{1}},$$

$$P^{(m)}(I_k|\mathbf{s}_i^{(m)}),$$

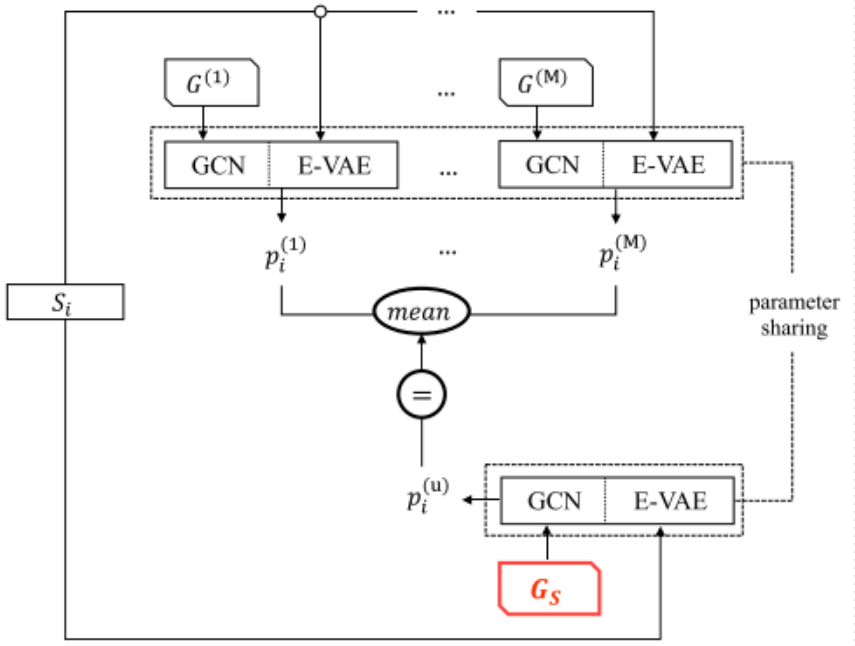


METHOD

2. Stable Graph Learning

- 使用训练好的GCN和VAE得到set在每个环境的条件概率分布 $\{P^{(m)}(I_k|\mathbf{s})\}_{m=1}^{M}$
- 初始化稳定图结构G_S，将set输入得到稳定图中的条件概率分布 $P_S(\mathbf{I}|\mathbf{s}_i^{(m)})$
- 优化G_S使得稳定图条件概率等于每个环境的概率分布的均值

$$\mathcal{L}_{rf} = ||P_S(\mathbf{I}|\mathbf{s}_i^{(m)}) - \sum_{j=1}^{M} \frac{P^{(j)}(\mathbf{I}|\mathbf{s}_i^{(m)})}{M}||_2$$



EXPERIMENT

- Baselines and Data
  - 由每个环境生成自己的图结构 $\{G^{(m)}\}_{m=1}^{M}$
  - 将上面所有的图求平均 $G_A = \sum_{m=1}^{M} \frac{G^{(m)}}{M}$
  - 将所有环境的数据拼接到一起生成图结构 G_C
  - 本研究提出的稳定图结构 G_S

  - Cloud Theme Click Dataset
    - 淘宝APP中云主题场景的用户点击日志
    - more than 4 million purchase histories of users for one month

EXPERIMENT

Real Data Experiment

- 根据purchasing behavior prediction判断不同图结构的好坏
- 依据是否经常购买流行商品/用户性别换把整个数据集划分成两个不同的环境

	Mean ACC	STD	Env1:Env2=0:10	1:9	2:8	3:7	4:6	5:5	6:4	7:3	8:2	9:1	10:0
$G^{(1)}$	12.93%	**0.0050**	13.54%	13.42%	13.62%	12.62%	13.49%	12.87%	12.92%	12.81%	12.16%	12.17%	12.62%
$G^{(2)}$	15.96%	0.0485	23.12%	22.21%	20.64%	18.83%	17.91%	16.11%	14.04%	13.67%	11.62%	9.46%	7.56%
G_A	18.09%	0.0347	23.21%	22.74%	21.76%	19.68%	19.46%	17.71%	16.68%	16.87%	14.98%	13.34%	12.57%
G_C	16.15%	0.0310	20.50%	20.57%	19.23%	17.49%	17.47%	16.08%	14.79%	15.24%	13.26%	11.78%	11.23%
G_S	**18.64%**	**0.0288**	22.90%	22.44%	21.81%	19.71%	19.98%	18.53%	17.31%	17.57%	15.91%	14.68%	14.22%

Table 4: Purchasing behavior prediction with exposure bias using item embeddings learnt from commodity network. The environment 1 consists of shopping logs mainly with unpopular items and env 2 consists of logs mainly with popular items.

	Mean ACC	STD	Env1:Env2=0:10	1:9	2:8	3:7	4:6	5:5	6:4	7:3	8:2	9:1	10:0
$G^{(1)}$	17.69%	0.0148	15.85%	16.03%	16.44%	16.64%	16.94%	16.67%	18.22%	18.87%	18.87%	19.99%	20.03%
$G^{(2)}$	17.46%	**0.0063**	16.86%	16.97%	16.56%	16.87%	17.17%	18.16%	16.99%	18.07%	18.10%	18.09%	18.27%
G_A	18.51%	0.0132	16.79%	16.94%	17.24%	17.60%	17.94%	18.50%	18.24%	19.78%	19.43%	20.46%	20.70%
G_C	18.56%	0.0127	16.84%	16.97%	17.34%	17.63%	17.81%	18.68%	18.94%	19.50%	19.53%	20.33%	20.62%
G_S	**20.17%**	**0.0092**	19.09%	19.01%	19.14%	19.51%	19.77%	20.02%	20.29%	20.84%	21.02%	21.62%	21.53%

Table 5: Purchasing behavior prediction in different gender groups using item embeddings learnt from commodity network. The environment 1 consists of shopping logs of females and env 2 consists of logs of males.