**DS 320: Homework 4**

Due: Monday, November 29th 2021, 11:59pm (EST)

Name: Josiah Kim

| Question | Points | Score |
|---|---|---|
| 1 | 10 | |
| 2 | 10 | |
| 3 | 10 | |
| 4 | 10 | |
| 5 | 10 | |
| Total | 50 | |

1. (10 points) Write an HLRT Wrapper to extract the information (Speaker, Date) from the following HTML page.

## Data Science Seminar Speakers (Fall 2018)

- **Sharon Huang** - *October 2*
- **Naomi Altman** - *November 3*
- **Zihan Zhou** - *December 4*

Head

Body

Tail

```
<html><body>
<h1>Data Science Seminar Speakers (Fall 2018) </h1>
<ul>
<li><b>Sharon Huang</b> - <i>October 2</i> <br>
<li><b>Naomi Altman</b> - <i>November 3</i> <br>
<li><b>Zihan Zhou</b> - <i>December 4</i> <br>
</ul>
</body></html>
```

```
procedure getSpeakerDate
    skip h1
    while <li> before </ul> do
        extract Speaker between <b> and </b>
        extract Date between <i> and </i>
    end while
```

2) (10 points) Compute universal and core solution instances for the following Data exchange setting and source instance:

D

Source S has one relation
$DeptEmp(dept\_id, mgr\_name, eid)$ — no constraints

The target schema T has two relations
$Dept(dpt\_id, mgr\_id, mgr\_name)$
$Emp(eid, dpt\_id)$

The mapping from S to T is:
$DeptEmp(d, n, e) \rightarrow \exists_M Dept(d, M, n), Emp(e, d)$

I

I(S) is $\{DeptEmp(IST, Liza, E005), DeptEmp(CSE, John, E010)\}$.

## Universal Solution:

$\{ Dept(IST, M_1, Liza), Emp(E005, IST), Dept(CSE, M_2, John), Emp(E010, CSE) \}$

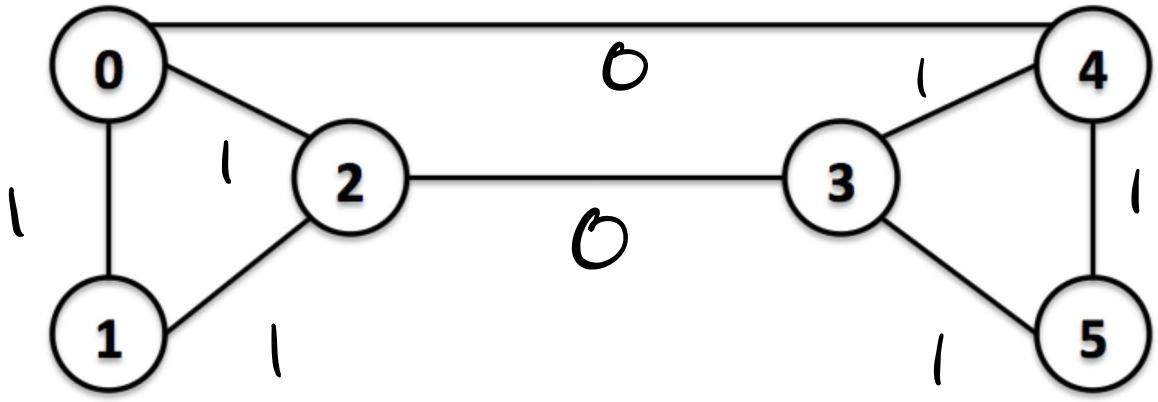
## Core Solution:

$\{ Dept(IST, M_1, Liza), Emp(E005, IST) \}$

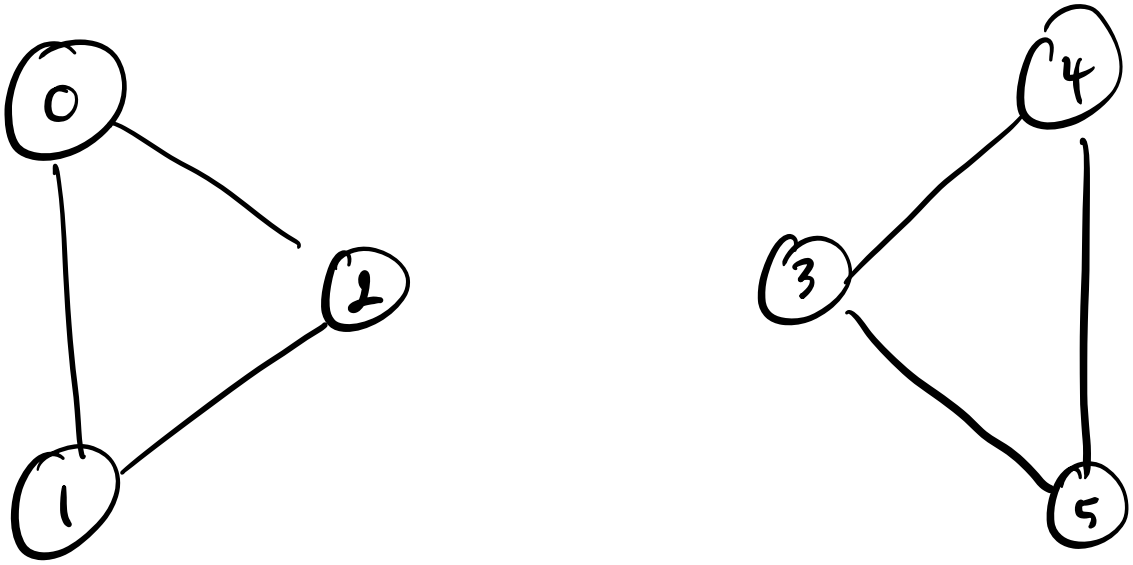
3) (10 points) what is the deep Web? What are the advantages and challenges of surfacing the deep Web?

The deep web is a collection of databases that are accessed by users entering values into HTML forms. This content is not accessible via search engines. Some advantages to surfacing are that there are no mediated schema and mappings, and reduced load on target websites. Disadvatages to surfacing include predicting the correct input combinations and predicting appropriate text inputs to retrieve data.

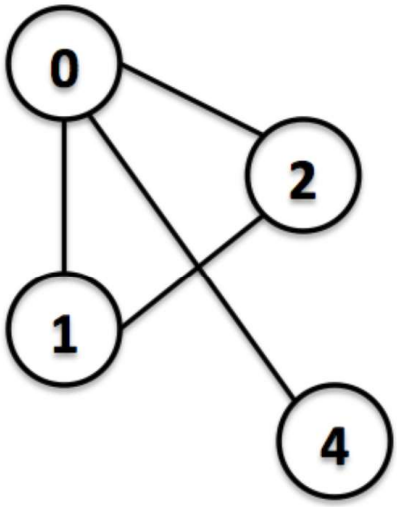
4) (10 points) Cluster the following graph using SNN and a threshold equals one.

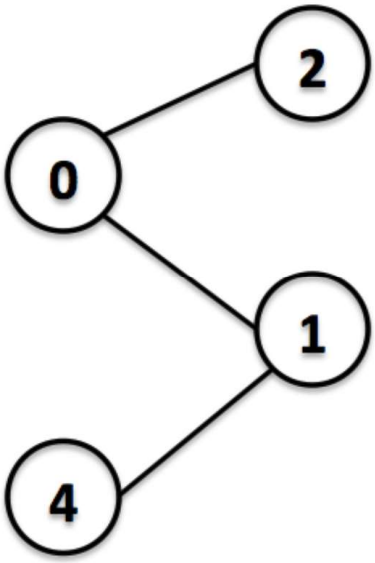



$\tau = 1$:



5) (10 points) Fuse the following two graphs using two different methods.



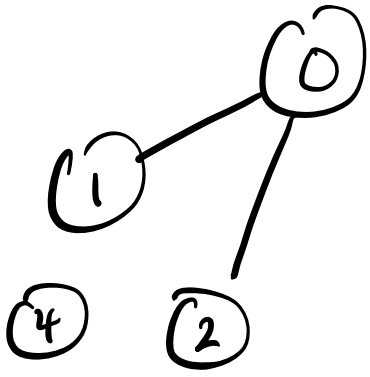

## FUSING ADJACENCY MATRICES:

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 | 0 |

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 |

AND

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow$$



# Fusing Adjacency Matrices: