# Data Challenge 1
# Group-fair Influence Maximization

Columbia University, Fall 2019
Xi Chen, Amalia Mayorga, Zhuangyu Ren (TAs)
A. Chaintreau (instructor)

## 1 Motivation

Social influence maximization is to strategically choose an initial set of individuals that maximizes the diffusion of information. For example, an outreach program tries to influence as many people as possible with useful information (e.g. available health services) while the total resource is limited. The outreach program must choose a set of $k$ individuals such that they can maximize the spread of information. To achieve this goal, it is natural for the outreach program to strategically choose a set of people from large and densely connected groups. As a result, small and marginalized groups could receive less assistance and attention. This data challenge invites you explore how to spread influence while making sure each group receives a fair allocation of the communitys resources.

## 2 Task Description

You works for a non-profit organization and your task is to provide people in the Antelope Valley community in California with information on HIV prevention. The dataset comes from Wilder et al. (2018). Due to your budget limit, you can only choose 25 people to visit in person and then rely on them to spread the information across the social network. Antelope Valley community is divided into different regions such as Ancaster, Leona Valley, Littlerock, and etc. We model influence using the independent cascade model and take geography into account since information is more likely to spread among people from the same regions. We set the influence probability for connected individuals from the same regions to be 0.2 and otherwise 0.1.

### Task 1

Your task is to spread influence to as many people as possible irrespective of the regions they come from. How many people can you influence with 25 seeds?

1. Construct an independent cascade model
2. Implement algorithm(s) that maximizes influence with 25 seeds. For evaluation, run the influence propagation 500 times and compute the average number of people influenced. Submit the 25 people you choose and the average number of people influenced.
3. For your best performing seed set, compute the proportion of people receiving the information for each region averaged across 500 simulations. What do you see? Report the result.

### Task 2

You wish to maximize influence such that people in each region receive a fair allocation of resources. In particular, you wish to maximize the minimum influence received by any of the region, as proportional to their population. This leads to maximizing the following utility function defined in Tsang et al. (2019):

$$U^{Maximin}(A) = \min_i \frac{\mathbf{I}_{G,C_i}(A)}{|C_i|}$$

where $A$ is the set of chosen nodes, $\mathbf{I}_{G,C_i}(A)$ is the expected number of influenced nodes in group $C_i$, and $|C_i|$ is the size of region $C_i$.

1. Implement and describe your algorithm(s) that maximizes the above utility function. For evaluation, run the influence propagation 500 times and compute the average minimal fraction of people influenced. Submit the 25 people you choose and the average minimal fraction of people influenced.
2. Compute the average number of people influenced and compare it with your previous best result. How much price do you need to pay for fairness? Report the result.

## 3 Dataset

The graph is stored in a pickle file. Below is a sample code to load the graph and access the node attributes:

```
import pickle
import networkx as nx
G = pickle.load(open('graph.pickle', 'rb'))
G.node[0]['region']
```

```
'northwest_antelope_valley'
```

## References

Tsang, A., Wilder, B., Rice, E., Tambe, M., and Zick, Y. 2019. Group-fairness in influence maximization. *arXiv preprint arXiv:1903.00967*.

Wilder, B., Ou, H. C., de la Haye, K., and Tambe, M. 2018. Optimizing network structure for preventative health. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 841–849. International Foundation for Autonomous Agents and Multiagent Systems.