1. (14 points) Submodularity

Q 1.1 (4 points) Prove that the coverage function from our lecture is a monotone submodular function. Recall that the coverage function takes a collection of sets $\{S_i\}_{i=1}^n$ and outputs the size of their union $|S_1 \cup S_2 \cup \cdots \cup S_n|$.

Say that there exists a set $A = \{S_i\}_{i=1}^n$, and also a set $B = \{S_i\}_{i=1}^m$, and $B \subset A$, in other words $\{S_i\}_{i=1}^m \subset \{S_i\}_{i=1}^n$

$f(B) = |S_1 \cup S_2 \cup S_3 \cdots \cup S_m| \leq |S_1 \cup S_2 \cup S_3 \cdots S_n| = f(A)$

for any $B \subset A$, $f(B) \leq f(A)$, so f is monotonic

given $B \subset A$, for any $S_j \notin A$

$f(B \cup S_j) - f(B) = |S_1 \cup S_2 \cdots S_m \cup S_j| - |S_1 \cup S_2 \cdots \cup S_m| = |S_j \setminus (S_1 \cup S_2 \cdots \cup S_m)|$

$f(A \cup S_j) - f(A) = |S_1 \cup S_2 \cdots S_n \cup S_j| - |S_1 \cup S_2 \cdots \cup S_n| = |S_j \setminus (S_1 \cup S_2 \cdots \cup S_n)|$

Since $S_1 \cup S_2 \cdots \cup S_m \subset S_1 \cup S_2 \cdots \cup S_n$ so $S_j \setminus (S_1 \cup S_2 \cdots \cup S_m) \supset S_j \setminus (S_1 \cup S_2 \cdots \cup S_n)$

So $|S_j \setminus (S_1 \cup S_2 \cdots \cup S_m)| \geq |S_j \setminus (S_1 \cup S_2 \cdots \cup S_n)|$

Thus

$f(B \cup S_j) - f(B) \geq f(A \cup S_j) - f(A)$

The coverage function is monotonic submodular

Q 1.2 (5 points) Let f be a monotone submodular function. Show that for any two sets S and T: $f(T) - f(S) \leq \sum_{e \in T} (f(S + e) - f(S))$.

$f(T) \leq f(T \cup S)$ because f is monotone

so $f(T) - f(S) \leq f(T \cup S) - f(S)$

Say $T = \{t_1, t_2, \cdots t_n\}$, according to the definition of f as monotonic submodular

$S \cup \{t_1 \cdots t_i\} \subset S \cup \{t_1 \cdots t_i, t_{i+1}\}$

$f(T) - f(S) \leq f(T \cup S) - f(S) = [f(S \cup \{t_1\}) - f(S)]$
$+ [f(S \cup \{t_1, t_2\}) - f(S \cup \{t_1\})]$
$+ \cdots$
$+ [f(S \cup \{t_1, t_2 \cdots t_n\}) - f(S \cup \{t_1, t_2 \cdots t_{n-1}\})]$

$\leq [f(S \cup \{t_1\}) - f(S)]$
$+ [f(S \cup \{t_2\}) - f(S)]$
$+ \cdots$
$+ [f(S \cup \{t_n\}) - f(S)]$
$= \sum_{i=1}^{n} f(S \cup \{t_i\}) - f(S)$
$= \sum_{e \in T} (f(S + e) - f(S))$

Q 1.3 (5 points) Let f be a monotone submodular function. Show that the following function is also submodular: $h(A) = \min(f(A), f(V/A))$ where V is the universal set of all elements.

Q 1.4 (5 points) **[Bonus]** Let f be a monotone submodular function and let g be a concave function. Show that the following function is also submodular: $h(A) = g(f(A))$.

2. (30 points) Anomaly Detection

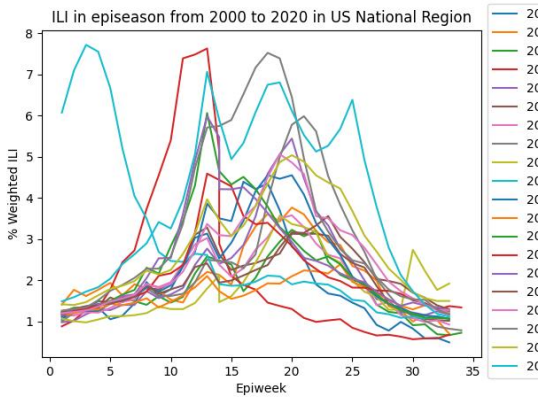
We will try to implement a simple anomaly detection algorithm for detecting outbreaks using ideas similar to the subset scan methods we saw in class. For this question we will use the data from the CDC about the ILI burden across the 10 HHS regions and US national level for the past 21 years (2000-2020).

Typically, epidemiologists focus on specific weeks of a year where the flu is prevalent called a *flu season*. A flu season starts at week 40 of a given year and ends at week 20 of the next year. We enumerate the weeks of a Flu season as *Epiweeks*.

For example, the 2019-20 season starts at week 40 of 2019 (Epiweek 1) and ends at week 20 of year 2020 (Epiweek 33). Since an year can have 52 or 53 weeks, a flu season can have 33 or 34 Epiweeks. The data for ILI is uploaded on canvas as `ILINet.csv`. In this question, we refer to `% WEIGHTED ILI` of the csv file as ILI values.

Q 2.1 (3 points) Plot all the seasons in the dataset for the US National region starting from 2000-2001 season to 2019-2020 season. Y axis is ILI and x axis is Epiweek number of the season. (Plot all the weeks in one graph)

Question 2 code in q2/



Q 2.2 (10 points) We will use the 2019-2020 season as the test season for our outbreak detector. ILI usually varies from 0 to 8. Let's focus on only data from the US National region for this part. For each Epiweek, fit a Gaussian distribution over all values of ILI for that Epiweek from all the training seasons (2000-2001 season to 2018-2019 season), Output the mean and std-deviation of each Epiweek you get in a table. Also submit the code.

Result in 2.2.csv in q2/

Q 2.3 (5 points) In statistics, you may have heard about the '3 sigma' rule for anomalies i.e., anything beyond 3 standard deviations from the mean is unlikely and hence an anomaly [1]. Hence, use the mean and std-deviation you get in Q2.2 for each Epiweek (again focusing on US national) and apply this rule to compute which weeks in 2019-20 season will be marked anomalous.

Week: [9, 23, 24, 25, 26, 27, 28, 29]

Q 2.4 (2 points) Do these weeks 'make sense'? Anything surprising? Any speculation about what happened during these weeks which made them anomalous? Why were the other weeks possibly not anomalous?

These week very much make sense, because we can see from 2.1 graph. 2019-2020 is very different from all other curves. It is most likely due to COVID-19 outbreak. Covid19 has flu symptoms and add anomality to the data. These weeks correspond to the 2.1 graph very well, indicating the weeks where covid 19 play a very important part.

Q 2.5 (10 points) Repeat the same steps in Q2.2, and 2.3 for regions 2, 4, 7, 9, 10 and finally write down the weeks you found anomalous for each of these regions.

```
output_anomaly('Region 2')
```

```
these weeks are anomaly:  [9, 24, 25, 26, 27, 28, 29, 30]
```

```
output_anomaly('Region 4')
```

```
these weeks are anomaly:  [6, 7, 8, 9, 24, 25, 26, 27, 28, 29]
```

```
output_anomaly('Region 7')
```

```
these weeks are anomaly:  [24, 25, 26, 27, 31]
```

```
output_anomaly('Region 9')
```

```
these weeks are anomaly:  [25]
```

```
output_anomaly('Region 10')
```

```
these weeks are anomaly:  [12, 13, 25, 28]
```

3. (30 points) Sensors for detection in Network Model

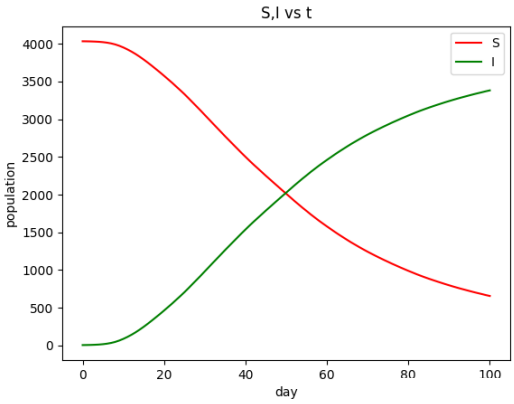
We are going to empirically look at various strategies of selecting social network sensors to detect epidemic outbreaks. We will use the graph in `facebook.txt` [2] which contains a small subset of friendship network of a Facebook group. We will use SI model with $\beta = 0.005$ to simulate the infection.

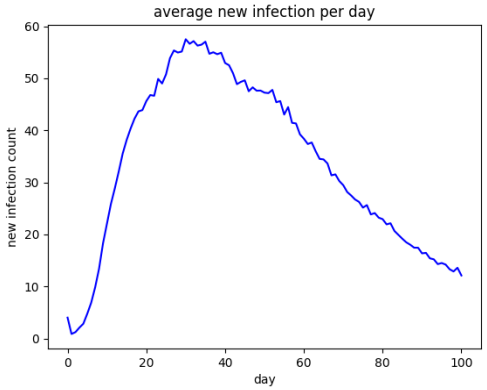
Q 3.1 (4 points) As a warmup, simulate the SI model for upto $T = 100$ time-steps for 100 runs and plot the average fraction of S,I vs t curve. Select 4 nodes at random to be infected at $t = 0$. Also plot the average number of daily infected people for $t = 0, \ldots, 200$. Report the average time t^* at which the maximum daily infections is maximum.

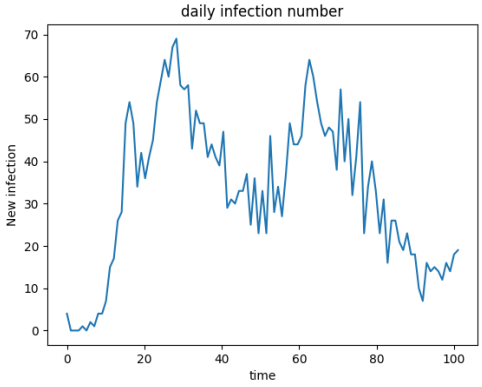
Hint: Use the implementation of SI model in `si_model.py` file. If you use another language you can convert the logic of the code in the file.

`Note:`

all q3 code and output in q3/





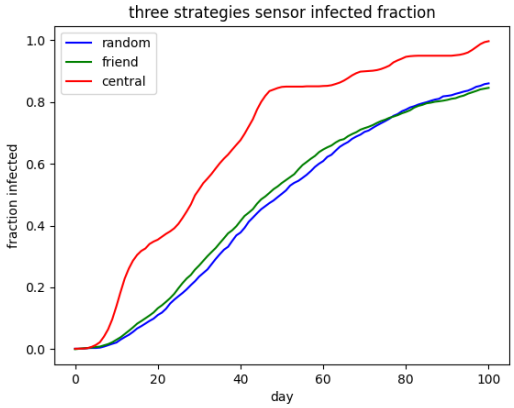


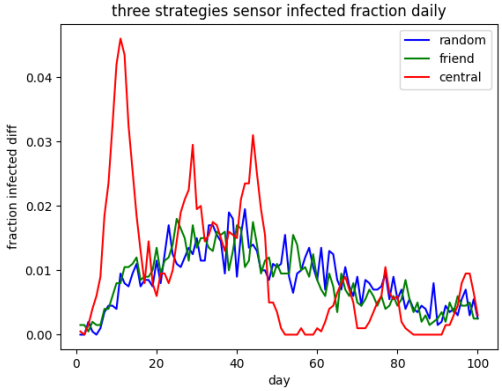
Q 3.2 (10 points) Since the graphs are large, it is not always feasible to keep track of the states of all nodes in practice. Therefore, we will select a smaller subset of nodes to track during the epidemic which we call as sensors. We will implement three strategies for sensor selection:

- `RANDOM`: We choose k nodes uniformly at from the graph
- `FRIENDS`: We choose k nodes uniformly at random and for each of them we select a random friend. We will use these friends as the sensors.
- `CENTRAL`: We select the top k nodes with largest *eigenvector centrality* [3]. You can use functions like `nx.eigenvector_centrality_numpy`.

Set $k = 100$ for all strategies. The file `RandNodes.npy` contains a random list of nodes. Select the first k nodes from the list to implement the `RANDOM` strategy. For `FRIENDS` strategy, use the k selected nodes from `RANDOM` strategy to randomly select their friends. Simulate the SI model for $T = 100$ steps and note the fraction $\tilde{I}(t)$ of the sensors that are infected at each time-step for each strategy. Note that sensors can also be infected at $t = 0$.

Plot average $\tilde{I}(t)$ vs t for all three strategies averaged over 20 runs. Also plot the average number of daily infections $\tilde{I}_d(t) = \tilde{I}(t) - \tilde{I}(t-1)$ over time.





Q 3.3 (6 points) Report the peak time $\tilde{t}^*$ and peak daily infection for all 3 strategies (the time t where $\tilde{I}_d(t)$ is maximum is peak time).

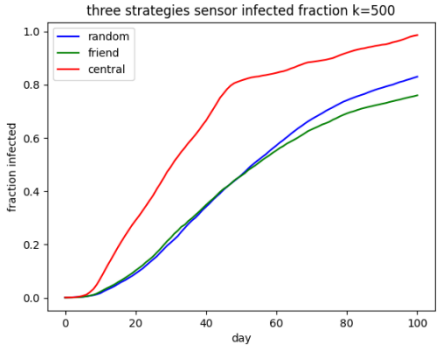
The value $t^* - \tilde{t}^*$ is the *lead time* i.e., the difference in time between detection of epidemic peak among sensors and time when it peaks in entire population. Report the lead-time for all 3 strategies.

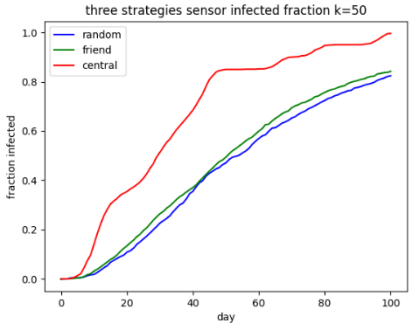
```
random 0.841
friend 0.8279999999999998
central 0.9509999999999998
```

Q 3.4 (4 points) Compare the lead-time of various strategies and explain difference in lead-time of various strategies.

The lead time of random and friend are very similar, but central is way higher. Random and friend strategies basically rely on random choice of nodes. For central strategy, the nodes selected are under higher risk of infection, and thus choosing them as sensor and increase the lead time a lot, indicating the sensitivity of the strategy to discover outbreak.

Q 3.5 (6 points) Now repeat Q 3.2 for $k = 50$ and $k = 500$. For each strategy submit a $\tilde{I}(t)$ vs t plot comparing different values of k. How does the lead time change with value of k for each strategy?





For random and friend strategy, they base on a random mechanism, so when k changes, the lead time can vary much, from 0.74 to 0.84 for friend and 0.80 to 0.84 for random strategy. Theoretically, friend should perform a little better than random because it somehow selects the nodes with one degree of associate with other nodes. The difference is not obvious in different k values of the above graphs, but the friend curve at the most time is a little greater than random. For central, the strategy, due to its nature, is very robust and sensitive. No matter the k value, the lead time is around 0.95.

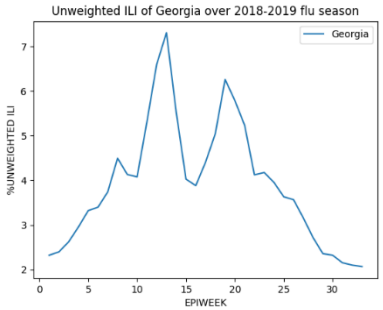
4. (26 points) Flu Surveillance using Google Symptoms Data

We will study the efficacy of Google Symptoms Data [4] as a source of Flu surveillance. We measure the usefulness of a signal using the correlation with ILI signals collected by CDC [5]. Specifically we will look at 2018-19 season (datasets can be downloaded from canvas).

Q 4.1 (3 points) We will start by considering the state of Georgia. Extract the `% Unweighted ILI` from `ILINet_states.csv`. Plot the weekly `Unweighted ILI` of Georgia over 2018-19 flu season. (Refer to Q2 for details on flu seasons.)

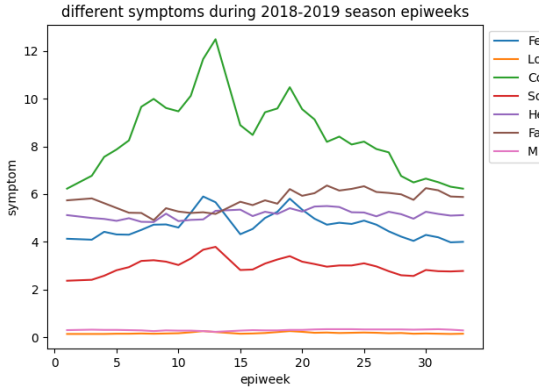
All q4 code and output in

q4/



Q 4.2 (10 points) CDC lists various symptoms for Flu [6]. We will consider the following symptoms: `Fever, Low-grade fever, Cough, Sore throat, Headache, Fatigue, Muscle weakness`.

Extract the Symptoms trends for each of these sympotms from the files `2018_symptoms_dataset.csv` and `2019_symptoms_dataset.csv` over the weeks of 2018-19 seasons. Submit a single plot showing the trends of all the symptoms over 2018-19 seasons with x-axis showing the Epiweeks and y-axis the symptom trend values.



Q 4.3 (3 points) We will use Pearson Correlation Coefficient (PCC) [7] to measure how correlated each of symptom's trend is to ILI. Evaluate PCC for each of the symptoms with ILI for 2018-19 season in Georgia. You may use functions like `scipy.stats.pearsonr` to evaluate PCC. Also, submit the code.

	symptom	PCC
0	Fever	0.949919
1	Low-grade fever	0.828344
2	Cough	0.947621
3	Sore throat	0.889420
4	Headache	0.225429
5	Fatigue	-0.307975
6	Muscle weakness	-0.539034

Q 4.4 (10 points) Repeat Q4.1 and Q4.2 for following states: **California, Texas, New York, Alaska** and **Mississippi**. For each state including Georgia, also report the symptom with the highest PCC along with the value of PCC. Can you think of any reasons for the differences in PCC across these states?

```
generate_pcc('California')
```

```
California Highest PCC symptom: Cough , PCC value:  0.9602297157131077
```

```
generate_pcc('Texas')
```

```
Texas Highest PCC symptom: Low-grade fever , PCC value:  0.9662977722721163
```

```
generate_pcc('New York')
```

```
New York Highest PCC symptom: Low-grade fever , PCC value:  0.9632297029572723
```

```
generate_pcc('Alaska')
```

```
Alaska Highest PCC symptom: Cough , PCC value:  0.9672301858176479
```

```
generate_pcc('Mississippi')
```

```
Mississippi Highest PCC symptom: Fever , PCC value:  0.9329311476905453
```

The highest pcc symptoms are different across states. For these six states, cough, low-grade-fever, and fever each occupies two states. This may actually due to different variants of COVID (or flu variant) that is epidemic in that state, and different variants may lead to different significance of fever of symptoms.