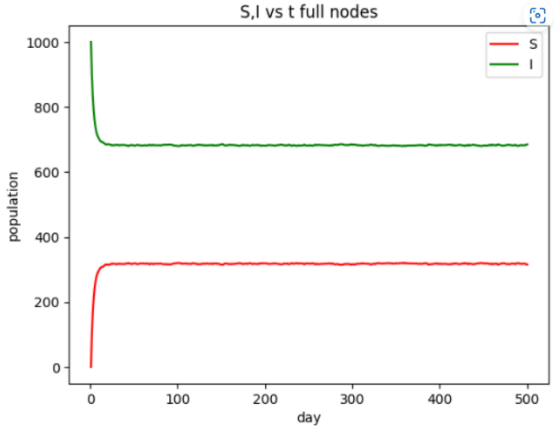
1. (12 points) Immunization in network models

Consider the so-called acquaintance immunization policy (where we pick a uniformly random neighbor of a uniformly random node) and the uniformly random strategy (of picking a node at random). For each policy, you should keep sampling till you have picked k different nodes (where k is the budget). Let us call the former as the FRIENDS policy and the latter as the RANDOM policy. We want to understand the relative performance of these two policies.

Generate a undirected unweighted graphs: G_{ba}. G_{ba} is a Barabasi-Albert preferential model graph with $n = 1000$ nodes (steps) and the number of edges to attach at each step $m = 2$ (You can set up as `G_ba = nx.barabasi_albert_graph(1000,2)`). In the Barabasi-Albert preferential model the probability that a new vertex attaches to a given old vertex is proportional to the (total) vertex degree.

Q 1.1 (5 points) Use can use the implementation from `sis_model.py` provided in canvas. Set $\beta = 0.2, \delta = 0.2$, and `max_time`=500. Initialize the model with all nodes as infected at time-step 0. Run it 200 times on G_{ba}. Report the average number of infected nodes at each step till max_time in the report PDF.

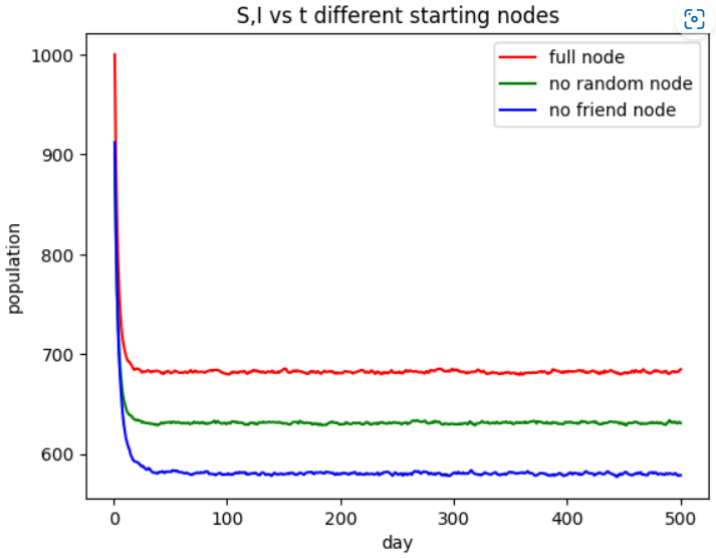
Code in q1/q1.jpynb



Q 1.2 (2 points) Use your implementation of FRIENDS and RANDOM in HW3 or sampling functions provided in `util.py`. Given the budget $k = 100$, report the nodes chosen according to each policy in the report PDF.

```
random nodes chosen:
 [930 521 582 537 590 134   1 445 620 297 305 821  99 740 850 917 633 694
 418 524 426  20 638 608 299 703 628 519 417 236 616 432 230 648 311 874
 194 530 291 880 249 842  84 190 879 832 710 451 293 924 599 677 125 989
 847 968 660 884 466 845 424 705 553 222 198 982 458 317  13 822 912 472
 848  82 706 350 491 378 867 485  74 934 154 739  72 954 640 196 441 580
 556 526  41  32 477 862  61  39 233 803]
friend nodes chosen:
 [56, 477, 50, 977, 252, 185, 402, 936, 146, 665, 171, 4, 136, 21, 470, 288, 198, 464, 635, 10, 128, 354, 876, 542, 95, 440, 996, 810, 15, 243, 20
1, 79, 888, 824, 43, 18, 627, 46, 256, 712, 112, 1, 582, 82, 677, 559, 17, 506, 641, 52, 227, 163, 4, 51, 420, 33, 54, 244, 609, 171, 4, 3, 42, 91
9, 14, 64, 115, 98, 573, 444, 187, 171, 1, 146, 800, 90, 8, 32, 693, 2, 61, 44, 51, 39, 41, 139, 210, 524, 459, 49, 3, 44, 314, 841, 7, 53, 187, 9
87, 15, 10]
```

Q 1.3 (4 points) Run the SIS model with $\beta = \delta = 0.2$ on G_{ba} from Q1.1. Pick $k = 100$ nodes according to both FRIENDS and RANDOM policies (use the nodes from Q1.3). Remove these nodes from the G_{ba}, and re-run the SIS model on the new (smaller) versions of each graph. Generate a plots: plot the average number of infections vs time when (a) no nodes have been removed (b) when nodes have been removed according to FRINEDS and (c) when nodes have been removed according to RANDOM (use different colors for each line (a)-(b)-(c)). Attach the plots in the reported PDF. Note: You should run 50 times and take the average for each line a-b-c)



Q 1.4 (2 points) What do you observe w.r.t. the performance of RANDOM and FRIENDS? Explain your observations in the report PDF.

The performance of infection count without friend is slightly lower than the infection count without random nodes and greatly lower than the infection count with full nodes, so as to say removing the friend nodes is more effective than removing random nodes from the whole graph on slowing down transmission. In another word, the importance of nodes in friend strategy is greater than the ones in random strategy.

2. (40 points) Vaccination interventions in ODE model using Multi-arm bandits

In this question, we look at a simplified toy example to study formulating vaccination policy planning as a multi-arm bandit problem. We will first setup a variant of SIR model that accounts for influence of vaccination rate with disease spread. Next we will setup the problem of choosing an optimal vaccination strategy for our SIR model. Finally, we will use various multi-arm bandit stratgies to solve for optimal vaccination strategy.

We will simulate a variation of SIR model to study the effect of different levels of vaccination intervention. Let $S(t), I(t)$ and $R(t)$ be the fraction of the population in susceptible, infected and recovered state. We will divide the population into vaccinated and unvaccinated.

Let $S_1(t), I_1(t)$ and $R_1(t)$ be fraction of total population that are susceptible, infected and recovered as well as are unvaccinated. Similarly, let $S_2(t), I_2(t)$ and $R_2(t)$ be fraction of population that susceptible, infected and recovered as well as are vaccinated. Therefore, we have $S_1(t) + S_2(t) = S(t)$ and similarly for infected and recovered states.

We define the SIR model via the following ODE equations:

$$\begin{aligned}
\frac{dS_1}{dt} &= -\beta S_1(I_1 + I_2) \\
\frac{dI_1}{dt} &= \beta S_1(I_1 + I_2) - \gamma I_1 \\
\frac{dR_1}{dt} &= \gamma I_1 \\
\frac{dS_2}{dt} &= -\beta(1-\rho) S_2(I_1 + I_2) \\
\frac{dI_2}{dt} &= \beta(1-\rho) S_2(I_1 + I_2) - \gamma(1-\rho) I_2 \\
\frac{dR_2}{dt} &= \gamma(1-\rho) I_2
\end{aligned} \tag{1}$$

where β and γ are the usual SIR model parameters and ρ determines effectiveness of the vaccine in both reducing rate of infection and probability of transitioning to R state among the vaccinated.

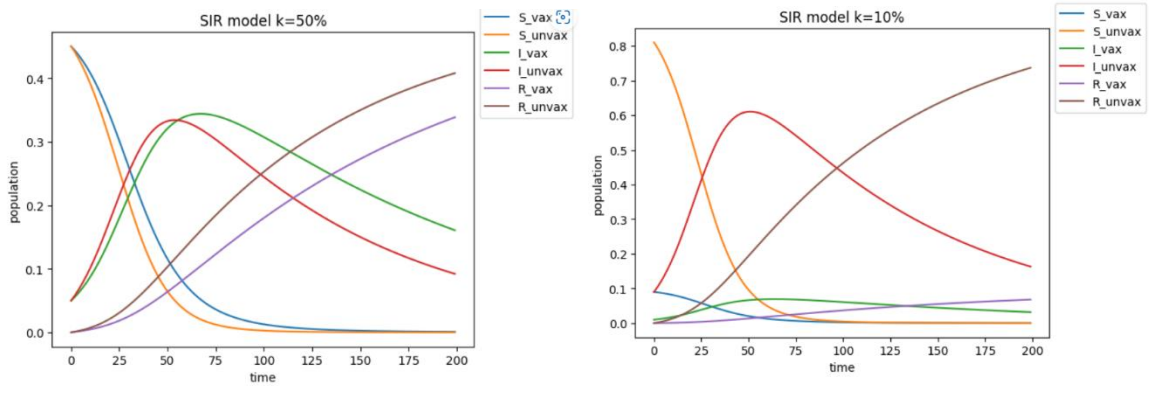
Code in q2/q2.jpynb

We have provided a boilerplate code in `datasets/q2.ipynb` notebook and you only need to fill in the requested portions.

Q 2.1 (6 points) Implement the above defined SIR model and submit the code. Specifically complete the `model_ode` function in the notebook.

Let 95% of the population be susceptible initially and the rest 5% be infected. Set $\beta = 0.1$, $\gamma = 0.01$ and $\rho = 0.3$. Let $k = 50\%$ of both infected and susceptible population be vaccinated (i.e, $S_1(0) = S_2(0) = 0.45$ and $I_1(0) = I_2(0) = 0.05$. Set $T = 200$ and plot the fraction of each compartment for time-steps from 0 to T. Now set $k = 10\%$ and repeat the ODE simulation and plot the fraction of each compartment for time-steps from 0 to T. How does the fraction of the population $R(T) = R_1(T) + R_2(T)$ at the end ($T = 200$) change with k?

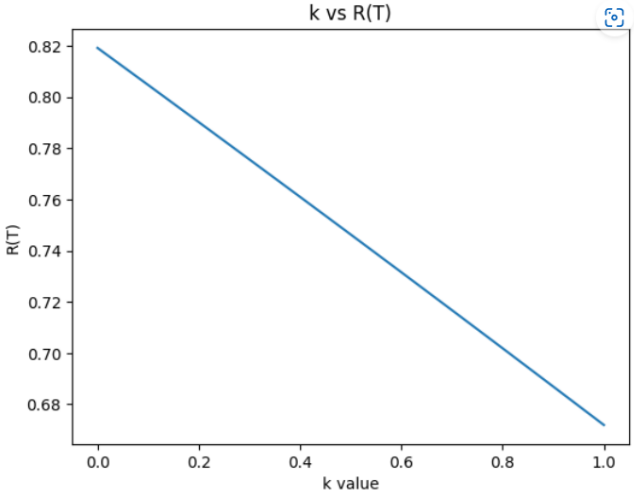
Hint: Refer to Q1 in HW 1.



R(T) changes from 0.746 to 0.805, as k decreases, R(T) increases

Q 2.2 (6 points) In many cases, we are not certain about the parameters β, γ, ρ of the model. We model them as random variables. Assume that $\beta \sim Uniform(0.05, 0.15)$, $\gamma \sim Uniform(0.005, 0.015)$ and $\epsilon \sim Uniform(0.1, 0.3)$. Complete the `stochastic_model_oracle` function.

For each of $k = [0\%, 10\%, 20\%, \ldots, 90\%, 100\%]$ run the SIR model for 1000 runs, sampling the parameters at beginning of each run. Compute the average of $R(T)$ for each of the values of k. Submit a plot with x-axis as k and y-axis as mean $R(T)$ (averaged over 1000 runs for eack values of k).



Q 2.3 (4 points) Using your implementation of the ODE model in Q2.2, write a function that samples the cost given the arm number as input. Specifically complete `cost_function` function. Assume that 90% of the population are susceptible anf rest are infected.

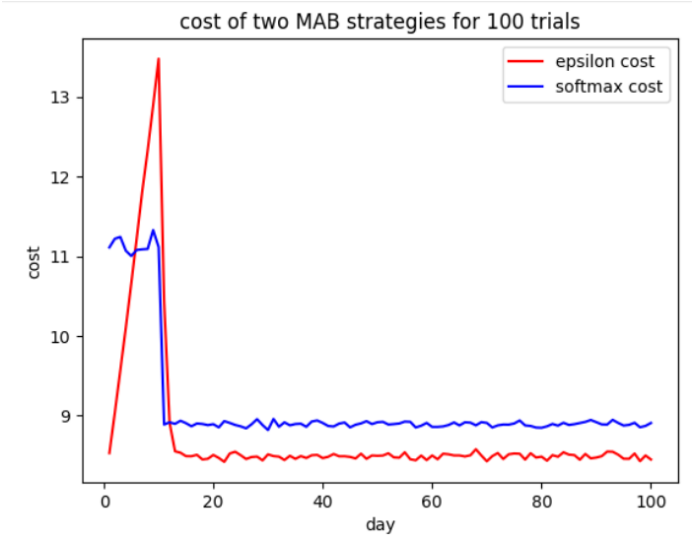
Q 2.4 (20 points) **Multi-arm bandit strategies:** We will use following two MAB strategies

1. ϵ-`GREEDY`: For each trial, we will choose a random arm with probability ϵ. Otherwise we will choose the arm with minimum estimated cost $\arg\min_k V(k)$ where V is estimated from past trials.
2. `SOFTMAX`: For each trial, we choose arm k with probability $\frac{\exp(-V(k)/\tau)}{\sum_{k' \in \mathcal{K}} \exp(-V(k')/\tau)}$ where τ is the temperature.

Implement both ϵ-`GREEDY` and `SOFTMAX` policy. Specifically complete the functions `epsilon_greedy` and `softmax`.

Set $\epsilon = 0.1$ for ϵ-`GREEDY` and $\tau = 1$ for `SOFTMAX` strategies. A single run of the MAB algorithm consists of running `run_bandit` for 100 trials (set `max_time = 100` in `run_bandit`). Perform 1000 independent runs of MAB for each of the two strategies and plot the average cost output by the oracle for each of the 200 trials i.e., run `run_bandit` for 1000 runs and average the output cost over 1000 runs. Submit a plot with x-axis being 1-100 time-steps of running MAB and y-axis being the average cost (averaged over 1000 runs) Which strategy performed better?

Note: This may take over 20 minutes to complete depending on code efficiency and compute resources.



epsilon works better,
it converges to a lower cost

Q 2.5 (4 points) Plot the average values of the cost estimate function V for both strategies. How does it compare to the cost computed from average $R(T)$ calculated in Q2.2?

In 2.2, k is negatively related to R(T). for cost estimate V, k is positively correlated to V.

3. (40 points) Forecasting

Let's try to build a couple of simple ensemble models for forecasting. We have uploaded a csv file to Canvas. It shows the COVID-19 mortality, cases and a couple of auxiliary signals (mobility and testing) for the US national level at a weekly level in 2020. Our goal is to compare two ensemble models on how well they predict mortality for the month of September 2020 (epiweek 202036 to 202039), after training them on data from Mar-Aug 2020 (epiweek 202010 to 202035).

Q 3.1 (15 points) Create an ARIMA (2, 0, 2) model to forecast mortality. You will need to do rolling predictions (i.e., start with the training set (Mar-Aug 2020), create the ARIMA model, use it to predict one week ahead, then add the prediction to your training set, retrain and then predict next week and so on. For example, you first use epiweek 202010 to 202035 as the training data to forecast 202036 mortality, and then use both epiweek 202010 to 202035 and your forecasted 202036 as the new training data to forecast 202037 mortality). Report the average RMSE error between your prediction and ground truth for epiweek 202036 to 202039 in the report PDF.

Note 1: You can use a off the shelf implementation of ARIMA. For python, we recommend using `statsmodel` package (Like: `from statsmodels.tsa.arima_model import ARIMA`[1].

Code in q3/q3.jpynb

rmse of arima: 1812.0965227401564

Q 3.2 (10 points) Create a simple linear regression model, which takes in the number of cases, mobility, testing from a week and predicts mortality for the next week. *Note:* You can the OLS model in statsmodel (`from statsmodels.regression.linear_model import OLS`). Repeat what you did in Q2.1 for rolling predictions, and report the average RMSE error between your prediction and ground truth for epiweek 202036 to 202039 in report PDF.

rmse of ols: 1992.0691496008976

Q 3.3 (10 points) Now create two ensemble models EM1 and EM2. EM1 is just the average of your ARIMA and OLS models i.e.

$$EM1 = \frac{ARIMA + OLS}{2}.$$

EM2 is a weighted average of your ARIMA and OLS models i.e.

$$EM2 = \frac{w_1 \times ARIMA + w_2 \times OLS}{w_1 + w_2}.$$

Make the weight of each model in the ensemble equal to its (1/RMSE) error on the training set (epiweek 202010 to epiweek 202035), i.e., $w_1 = \frac{1}{RMSE_{ARIMA}}$ and $w_2 = \frac{1}{RMSE_{OLS}}$. Measure EM1 and EM2's average RMSE on epiweek 202036 to epiweek 202039 and report the RMSE error for epiweek 202036 to 202039 in the report PDF.

rmse of EM1: 1624.75162518402

rmse of EM2: 1507.8709869242582

Q 3.4 (2 points) What do you observe comparing the test performance of your 4 models: ARIMA, OLS, EM1, EM2? Comment and try to explain the performance you observe in 1-2 lines in the report PDF.

The performance of arima is better than ols, but their performance is complementary, having bias towards opposite way. So using ensemble learning decrease the overall rmse. Especially for em2 that takes into account weight, the rmse would be the smallest.

4. (8 points) Ethical and Societal Issues

Choose any one of the many facets of data science in epidemiology discussed in class (e.g. forecasting, surveillance, modeling, interventions, data collection etc etc) and discuss various societal challenges associated with it such as ethics, privacy, anonymity, consent, equity, etc. Submit a short 500 word essay. This is an open-ended question, therefore feel free to read various resources and formulate you own points. Make sure to cite relevant works when you are making any factual claims.

Covid-19, the one-word that occupied people's life in the recent few years, is making a huge change to everyday life around the whole world. Regarding the pandemic, different countries come up with various strategies to arrange the medical resources, control the transmission, and develop vaccines. To realize these steps, surveillance is crucial, since it is important to know the real time data to administrate appropriate policies and distribute resources. To make the problem simpler, strategies can be concluded into two ways: live with the virus or clear out the virus. One can imagine that these two ways of strategies corresponds to very different level of surveillance—there will be way heavier surveillance if the country decides to clear out the virus. Living in the US and having families in China, I will use these two countries as examples for the two distinct strategies to address the challenges for surveillance during and after the pandemic.

Mentioned above that surveillance is crucial to effective policy administration. Moreover, the accuracy and thoroughness of the surveillance are the key points to the quality of surveillance. Data from a city will obviously not be represented enough for the whole country. Thus, countries spent huge amount of money to apply the surveillance to as many people and as detailed data as need. Back in March 2021, Biden administration announced to invest more than 12 billion dollars to expand covid-19 testing to aid school reopening and for underserved population [1]. However, people still question the accuracy of the surveillance data. Alwan claimed that the surveillance is underestimating the burden of covid-19 [2]. People may feel unwell but not get tested, or people can get positive tests but do not report. Due to the limitation of hospital resources, the actual cases are underestimated. On the contrary, in China, people are required to do covid testing once every two days in major cities. Given the huge population, the spending can be tremendous for a more real time and accurate surveillance data.

The huge spent on surveillance leads to complains and worries. On one hand, some people worry that unthorough and inaccurate surveillance is threating people's health. On the other hand, other people complain about the heavy surveillance, the inconvenience, and the privacy violation.

To build accurate models for forecasting and to accurately distribute resources, countries have applied the up-to-date technology for efficient surveillance. With the help of internet, it is very easy to collect

people's traces and spendings. Locating through WIFI connection, video cameras, and GPS location, government can easily get these data without individual permission or even one step closer: to force people to fall into certain traces. Before universities reopen, students need to get tested to keep the code green to enter universities. In China, for people tested positive, their traces are public for the good of everyone else. Moreover, each person has a 'health code'. One need to upload negative scanning result to keep the code green. Only through scanning the code one can enter each public facility, and if the code turns red, the facility will refuse the entrance of that person. In this way, the government can not only monitor, but also control the trace of each individual. To comply to the policies, people in different countries are handing in different levels of privacy and right. But are those privacy and right going to get back after the pandemic? As Barriga et al. argued in their paper "as one door opened, it hardly closes again" [3].

We can expect that after the pandemic, people will worry more about how to retrieve the privacy and right back. But due to the coverage of internet and its significance in life, one cannot escape from internet, a place where people can be vague about the privacy borderline. Whichever strategies countries apply, governments do not need to give up the convenience of internet. In this way, although with different policies, such pandemic events would inevitably sacrifice people's privacy irreversibly in the name of accurate surveillance for public health.

Reference:

1. N. Division, "Biden administration to invest more than $12 billion to expand COVID-19 testing," HHS.gov, 18-Apr-2022. [Online]. Available: https://www.hhs.gov/about/news/2021/03/17/biden-administration-invest-more-than-12-billion-expand-covid-19-testing.html. [Accessed: 15-Nov-2022].

2. N. A. Alwan, "Surveillance is underestimating the burden of the covid-19 pandemic," The Lancet, vol. 396, no. 10252, Aug. 2020.

3. A. do Barriga, A. F. Martins, M. J. Simões, and D. Faustino, "The COVID-19 pandemic: Yet another catalyst for governmental mass surveillance?," Social Sciences & Humanities Open, vol. 2, no. 1, p. 100096, 2020.