**CS-E5740 Complex Networks**
Aalto University
Autumn 2023
Luca Santagata

# Final Course Project

## Preface

*The topic of this project work that we have been running for a long time has, unfortunately, become all too real in the recent years. This shows that pandemics are not entirely unexpected, and the way how they spread through the globe has also been foreseen for a long time. In fact, if one would use global air transport data, a more complex disease model (SEIR), and the real geographic origin of spread, one would get a timeline that eerily matches early 2020.*

## Introduction

In the project work, your task is to implement a Susceptible-Infected (SI) disease spreading model and run it on top of a temporal network from air transport data, containing information on the departure and arrival times of flights. You will study the dynamics of spreading and how it depends on where the process starts as well as the infectivity of the disease, and use static network centrality measures to understand the roles that specific nodes play.

## Model specifications

In the SI model, each node is either Susceptible (S) or Infected (I). When an Infected node is in contact with a Susceptible node, the Susceptible node may become infected with some probability $p \in [0; 1]$, reflecting the infectivity of the disease. Infected nodes remain Infected forever.
In our model that mimics the spreading of disease through the air transport network, nodes are airports and time-stamped connections are flights between them. Initially, only one airport node (called the seed node) is set to the Infected state, while all other airports are Susceptible. Now, following the SI process, a flight from an Infected source airport infects its Susceptible destination airport with probability $p \in [0; 1]$. Note that a flight can carry the infection only if its source airport is infected at the time of the flight's departure! Infected airports remain infected for the rest of the simulation. +

# 1 Data description

The flight data are located in the file `events_US_air_traffic_GMT.txt`, where each row contains the following fields:

- `1st column -> Source [0-278]`
- `2nd column -> Destination [0-278]`
- `3rd column -> Start Time (GMT) [seconds after Unix epoch time]`
- `4th column -> End Time (GMT)`
- `5th column -> Duration [Same as (EndTime-StartTime)]`

You can find the information about the airports in file `US_airport_id_info.csv`. `US_air_bg.png` is an image of the USA map used as a background image in some visualizations.

# Task 1: Basic implementation

Implement the SI model using the temporal air traffic data in `events_US_air_traffic_GMT.txt`. Use the provided visualization module to check that your implementation works reasonably. Assume first that $p = 1$, *i.e.*, the disease is always transmitted.

a) **If Salt Lake City (SLC, node-id=27) is infected at the beginning of the data set, at which time does Anchorage (ANC, node-id=41) become infected?**

Anchorage become infected at time: 1229283600.0

# Task 2: Effect of infection probability $p$ on spreading speed

Run the SI model 30 times with each of the infection probabilities $[0.01, 0.05, 0.1, 0.5, 1.0]$. Again, let Salt Lake City (node-id 27) be the initially infected node. Record all infection times of the nodes, and answer the following questions:

a) **Plot the averaged prevalence $\rho(t)$ of the disease (fraction of infected nodes) as a function of time for each of the infection probabilities. Plot the 5 curves in one graph. You should be able to spot stepwise, nearly periodic plateaus in the curves.**

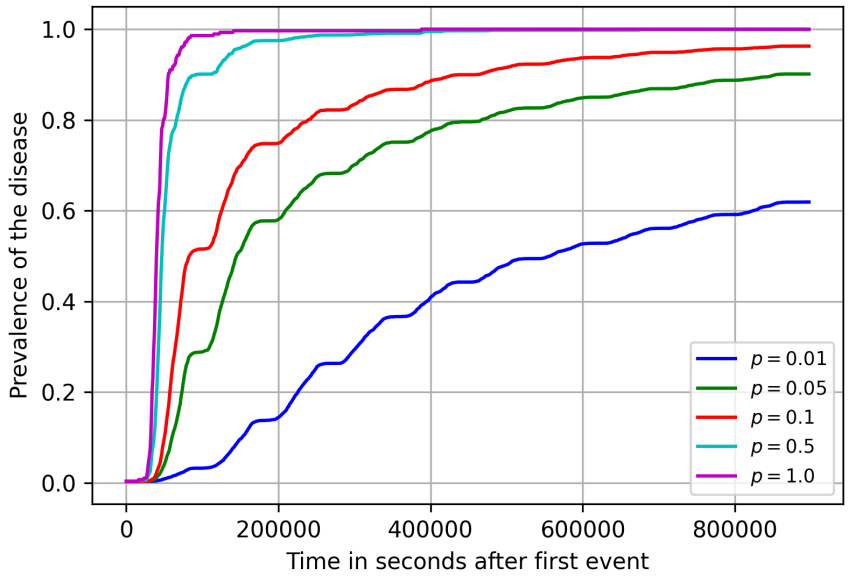



Figure 1: Effect of infection probability on spreading speed

b) **For which infection probabilities does the whole network become fully infected? How can we explain the periodic "steps" in the curves?**

The entire network becomes completely infected (*i.e.* the prevalence of the disease = 1) in both cases of $p = 0.5$ and $p = 1$. The periodic "steps" observed in the epidemic spread curves can be attributed to the discrete nature of air travel events, specifically the landing of airplanes. This discrete phenomenon leads to plateaus in the curves when most passengers are in flight, causing a slowdown in the epidemic spread. Subsequent increases in the curves occur when passengers land, enabling the epidemic to spread rapidly. Analyzing Figure 1 indicates that the step behavior is nearly simultaneous, suggesting concentrated flight landings in different cities during those periods, accelerating the disease spread beyond typical rates.

# Task 3: Spread time with different seed infected nodes

Next, we will investigate how the selection of the initially infected seed node affects the spreading speed.

a) **Use nodes with node-ids [5, 38, 118, 134, 139] (DTW, SMF, CRW, DHN, GGG) as seeds and $p = 0.1$, and run the simulation 30 times for each seed node. Then, plot the average prevalence of the disease separately for each seed node as a function of time.**

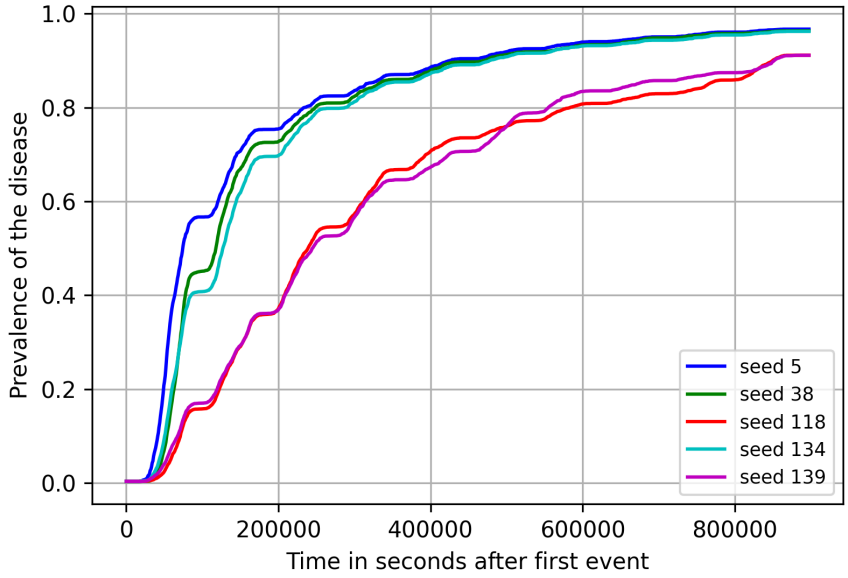



Figure 2: Effect of infection probability on spreading speed

b) **The differences in spreading speed between seeds should be mostly visible in the beginning of the epidemic. Explain why.**

This is because the only infected node is the seed node at the beginning. So the spread is mostly based on the seed node and its connections. As time passes, more and more nodes get infected and infect other nodes. From the figure 2 it is possible to observe how the spreading speed of nodes 5, 38, and 134 is much higher than that of nodes 118 and 139 due to the difference in their degrees. From the table 1, it can be noticed that the first three nodes have a higher degree (especially node 5 in blue with the highest degree) compared to the last two, and this, of course, influences the speed of epidemic propagation, in fact a node with more connections has a higher probability of infection.

| Node | Degree |
|---|---|
| 5 | 110 |
| 38 | 24 |
| 118 | 2 |
| 134 | 13 |
| 139 | 1 |

Table 1: Degrees of the seed nodes

c) **Why is it important to average the results over different seeds? What kind of problems could follow from using only a single seed, for example in the next task where we'll inspect the vulnerability of a node for becoming infected with respect to various network centrality measures?**

As explained in the previous section, the choice of the seed node significantly influences the initial spread of the infection. Consequently, averaging across various seed nodes would diminish the sensitivity of the results to the initial seed node selection, thereby enhancing the robustness of the outcomes.

# Task 4: Where to hide?

Now, consider that you'd like to be as safe from the epidemic as possible. How should you select your refuge? To answer this question, run your SI model 100 times with $p = 0.5$ using different random nodes as seeds and record the median infection times for each node.

a) **Run 100 simulations, and create scatter plots showing the median infection time of each node as a function of the following nodal network measures:**

i) *Unweighted* clustering coefficient $c$

ii) Degree $k$

iii) Strength $s$

iv) *Unweighted* betweenness centrality

**To compute the above structural measures of the network, you first need a static network. Construct an aggregated, undirected, weighted network where the weight of each link corresponds to the number of flights between the nodes (in both directions). Then, compute the centrality measures for the aggregated network using ready-made NetworkX functions.**

The scatter plots showing the median infection time of each node as a function of the following nodal network are showed in figure 3

b) **Use the Spearman rank-correlation coefficient for finding out which of the measures is the best predictor for the infection times.**

| **Network Measure** | **Spearman r** |
|---|---|
| Median Infection Time vs. Clustering | $-0.113$ |
| Median Infection Time vs. Degree | $-0.812$ |
| Median Infection Time vs. Strength | $-0.883$ |
| Median Infection Time vs. Betweenness | $-0.634$ |

Table 2: Median Infection time w.r.t. different network measures

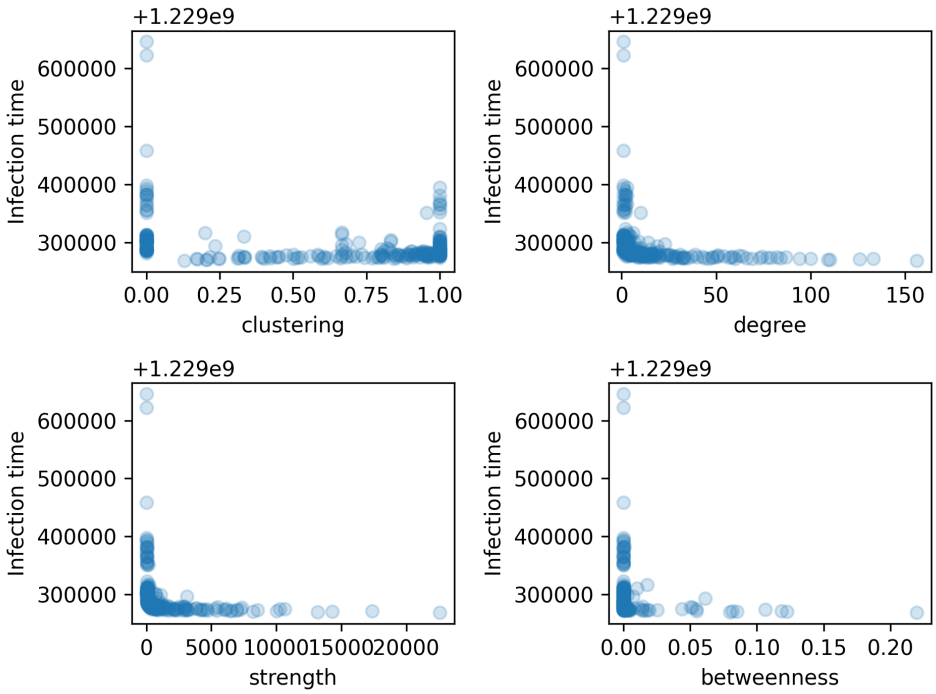



Figure 3: Median Infection time w.r.t. different network measure

c) **Based on your results, answer the following questions:**

- **Which measure(s) would you use to pick the place to hide, i.e., which measure best predicts node infection time? Why are these measures good in predicting the infection time?**

  Given the pronounced anti-correlation observed between the strength and degree of a node and its infection time, it suggests that both strength and degree serve as effective predictors of node infections compared to other measures. Anti-correlation, in this context, signifies that higher strength and degree of a node are associated with lower infection times, indicating a faster spread of the infection. A high strength for a node implies a substantial number of flights connecting that node with others in the network. This heightened connectivity increases the likelihood of rapid infection, as the node can potentially get infected through any of its numerous connections. Similarly, a high degree for a node indicates a large number of direct connections. Nodes with high degrees are more central in the network, facilitating the swift transmission of the infection. The more direct connections a node has, the quicker it can become infected, as it serves as a hub for potential transmission pathways. So, both strength and degree measures prove to be valuable predictors of node infection times. Their anti-correlation with infection time underscores their effectiveness, as higher values in these measures correlate with a faster spread of the infection. The interconnected nature of nodes with high strength or degree positions them as key players in the epidemic dynamics, making them significant factors to consider when selecting a place to hide to minimize infection risk.

- **Why does betweenness centrality behave differently than degree and strength?**

  Betweenness Centrality (BC) serves as a valuable metric for identifying hubs within a network—nodes that act as bridges between different communities. These hubs play a crucial role in connecting disparate parts of the network. Despite having the lowest relative shortest path to other nodes, their significance in infection transmission lies in the fact that infections primarily propagate through direct links to other infected nodes. In the context of epidemic spread, a notable observation is that, despite their pivotal bridging role, hubs with high BC might not necessarily be among the first nodes to get infected. This phenomenon can be attributed to the nature of infection transmission, where the infection predominantly spreads through direct links. As a result, a hub, even with a low degree to other nodes, may experience delayed infection compared to other nodes with higher degrees.

- **Why is the clustering coefficient a poor predictor?**

  The clustering coefficient is a metric that provides insights into the connectedness of a node's neighbors, reflecting the extent to which neighboring nodes are interconnected. In the context of epidemic spread, where infections primarily transmit through direct links, the significance of the clustering coefficient lies in understanding how tightly knit the local neighborhood of a node is.

  Unlike other centrality measures, such as degree or strength, which focus on the individual node's connectivity, the clustering coefficient delves into the local structure surrounding a node. Specifically, it measures the proportion of connections among a node's neighbors relative to the total possible connections. However, for the transmission of infection, the emphasis is not on how well-connected the neighbors of a node are to each other but rather on how well-connected the node is to other nodes in the network.

# Task 5: Shutting down airports

Now take the role of a government official considering shutting down airports to prevent the disease from spreading to the whole country. In our simulations, shutting down airports corresponds to immunization: an airport that has been shut down cannot become infected at any point of the simulation and therefore cannot transmit the disease.

One immunization strategy suggested for use in social networks is to pick a random node and immunize a random neighbour of this node. Your task is now to compare this strategy against five other immunization strategies: the immunization of random nodes and the immunization of nodes with the largest values of the four node properties calculated in task 4. In this exercise, use $p = 0.5$ and average your results over 30 runs of the model for each immunization strategy (180 simulations in total).

To reduce the variance due to the selection of seed nodes, use the same seed nodes for investigating all immunization strategies. To this end, first select your immunized nodes, then select 30 random seed nodes such that none of them belongs to the group of immunized nodes in any of the 6 different strategies.

a) **Adapt your code to enable immunization of nodes, and plot the prevalence of the disease as a function of time for the 6 different immunization strategies (social net., random node, and 4 nodal network measures), always immunizing 10 nodes.**

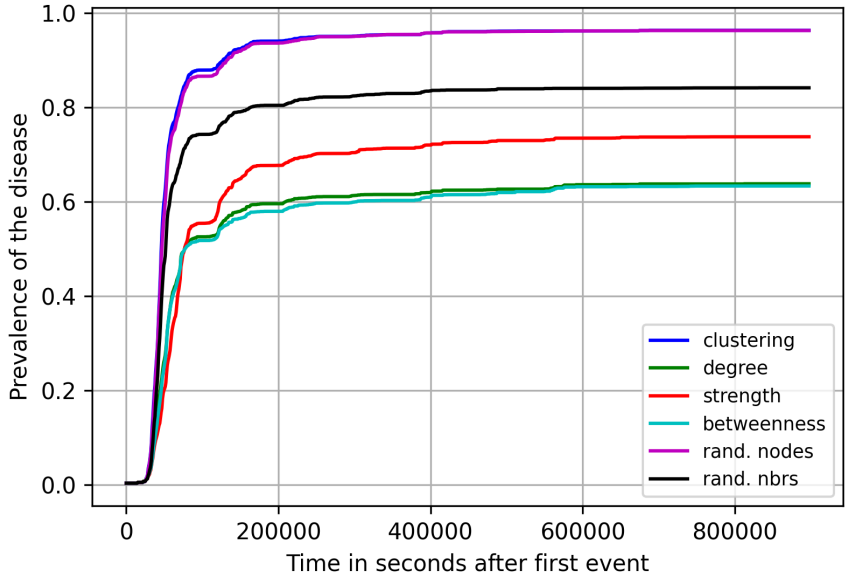



Figure 4: Different Immunization Strategies

b) **Based on your results, answer the following questions:**

- **Which immunization strategy performs the best, and why?**
  The adoption of a Betweenness Centrality (BC) strategy emerges as an optimal approach for immunizing nodes within a network. This strategy proves particularly effective due to the unique characteristics of nodes with high BC, which often function as hubs connecting diverse communities within the network. Nodes with high BC play a pivotal role as bridges, linking disparate parts of the network and immunizing these key hubs strategically disrupts the pathways through which the disease could propagate from one community to another. By rendering these central nodes immune, the potential for disease transmission across different communities is significantly curtailed. The strength of the BC strategy lies in its ability to target nodes with maximal influence over the network's connectivity. Immuning these influential hubs can act as a barrier, impeding the inter-community transmission of the disease. Consequently, this strategic immunization approach not only protects the highly connected nodes themselves but also serves as a proactive measure to contain the spread of the disease at the network level.
- **Why does betweenness centrality perform better as an immunization strategy than as a predictor for a safe hiding place?**
  BC proves to be a superior strategy for immunization rather than a strategy to find a safe place within the network. This effectiveness arises from the fact that nodes with high BC values typically represent hubs, acting as crucial connectors between different communities. Disrupting these hubs through immunization is highly impactful as it hinders the potential spread of the disease from one community to another. However, it's important to note that while BC is effective for preventing inter-community transmission, it doesn't guarantee immunity for the node itself. Nodes with high BC could still be susceptible to infection, considering that the transmission occurs through direct links, and BC doesn't inherently indicate the presence of these direct links. BC quantifies the relative shortest paths to other nodes rather than the specifics of direct connections. As a result, selecting a node with high BC as a place to hide may not be an optimal strategy. Such a node, while influential in connecting communities, could still be affected by the disease transmitted through direct links.

c) **The random neighbour immunization strategy probably worked better than the random node immunization. Let us try to understand why.**

- **First, if the degree distribution of the network is $P(k)$, what is the probability of picking a random node of degree $k$?**
  The likelihood of selecting a node with degree $k$ is contingent upon the probability of choosing one of its connected edges, given by $\frac{k}{2e}$, where $e$ represents the total number of edges in the network. Considering there are $n$ nodes with degree $k$, the cumulative probability of selecting a node with degree $k$ becomes $\frac{nk}{2e}$. Notably, this probability is equivalent to $\frac{n}{N}$, signifying that among the $N$ nodes, there are $n$ nodes with degree $k$.

- **What is the expected outcome if you then pick a random neighbour of the random node?**
  The expected outcome of following a random link of a random node is $\frac{\langle k^2 \rangle}{\langle k \rangle}$
- **Consequently, which of the strategies is expected to be more effective and why?**
  As the expected degree of choosing a random node in the network is $\langle k \rangle$, the expected degree of following a random link is bigger. Therefore it is a better immunized strategy because it selects nodes with higher expected degrees and it would do a better job in preventing the disease from getting spread. This is also called the friendship paradox which means in a network, a neighbour of a node on average has more neighbours.

d) **Although the random neighbour immunization strategy outperforms the random immunization, it may be less effective than some other immunization strategies. Nevertheless, explain shortly, why it still makes sense to use this strategy in the context of social networks.**

Given the challenges in employing alternative immunization strategies within social networks, the pick-a-neighbor strategy stands out as a pragmatic choice. Other strategies often necessitate access to the full network topology for calculating various measures, which isn't always possible. Moreover, these alternative measures tend to be computationally demanding, posing practical obstacles. In contrast, the pick-a-neighbor strategy shines due to its simplicity in implementation and execution. It doesn't require access to the complete network topology, making it more feasible for real-world applications where such access might be restricted. The strategy's computational efficiency further enhances its appeal, as it can be easily implemented and run without the resource-intensive nature of other approaches. Additionally, the strategy aligns with the observed friendship paradox in social networks, where a node's random neighbor tends to have a higher degree than a randomly chosen node. This inherent property bolsters the strategy's effectiveness compared to random node selection, emphasizing its practical utility in navigating the constraints and complexities of social network dynamics.

# Task 6: Disease transmitting links

So far we have only analyzed the importance of network nodes—next, we will discuss the role of links. We will do this by recording the number of times that each link transmits the disease from one node to another. To this end, adapt your code for recording the static links that are used to transmit the disease. You can do this e.g. by storing for each node where the infection came from. For example, if node $i$ gets infected by a flight arriving from node $j$, you should store the information of $i$ being infected by $j$. This can be done by using a list `trans_nodes`, where `trans_nodes[i] = j`.

Run 50 simulations using random nodes as seeds and $p = 0.5$. For each simulation, record which links are used to infect yet uninfected airports (either by first infection-carrying flights arriving at susceptible airports or by infecting flights arriving before the already recorded infection time).

a) **Run the simulations, and compute the fraction of simulations where each link transmitted the disease ($f_{ij}$). For example, if the disease spread from node $a$ to node $b$ in 10 simulations out of 50, $f_{ab} = 0.2$. Then use the provided function `plot_network_USA` to visualize the network on top of the US map. Adjust the width and opacity of the links according to the fractions $f_{ij}$ to better see the overall structure. Compare your visualization with the maximal spanning tree of the network.**

The visualization refers to figure 5

b) **Explain why your visualization is similar to the maximal spanning tree.**

The similarities observed in these two graphs can be attributed to the algorithm employed in the computation of Minimum Spanning Trees (MST). Specifically, MST generates a tree incorporating edges with the highest weights in the graph, where a high weight signifies an elevated frequency of travels between two airports. In the second graph, the analysis focuses on the fraction of times a specific link interacts with others. Notably, links with higher weights exhibit a pronounced correlation with the potential to infect others. This signifies that the strength of the link corresponds to its influence in spreading infections.

c) **Create scatter plots showing $f_{ij}$ as a function of the following link properties:**

i) link weight $w_{ij}$

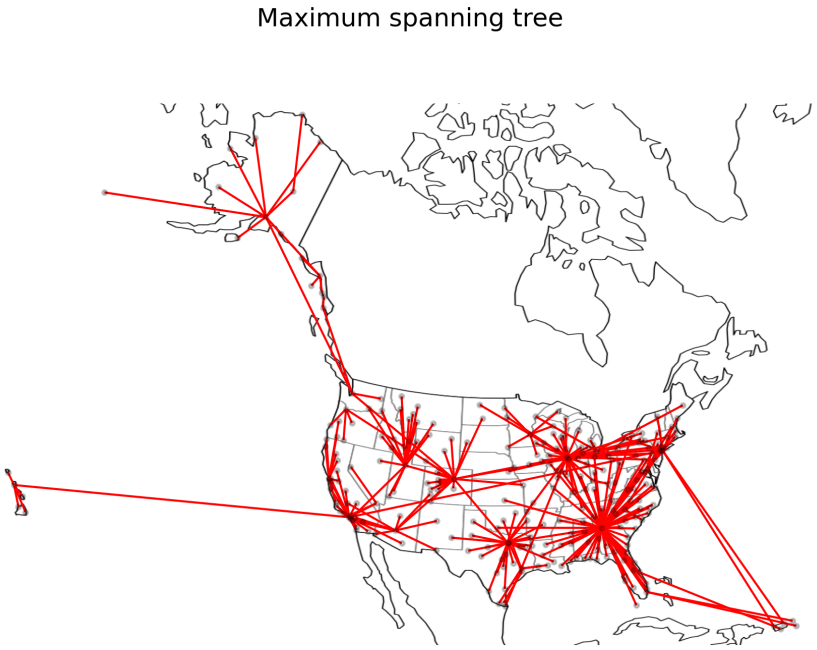
ii) *unweighted* link betweenness centrality $eb_{ij}$ (`edge_betweenness_centrality` in NetworkX)

The scatter plots are showed in figure 6

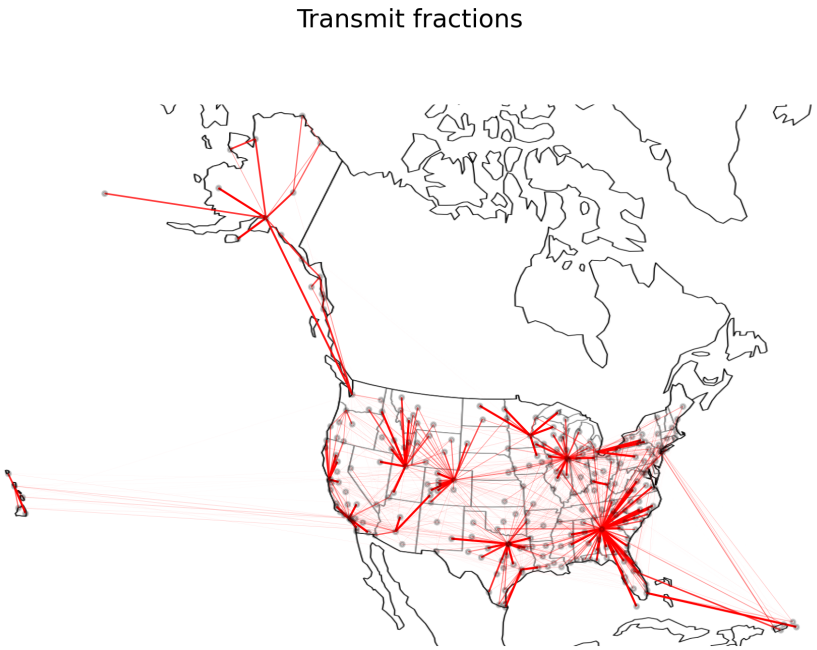
**Compute also the Spearman correlation coefficients between $f_{ij}$ and the two link-wise measures.**

Table 3: Correlation of $f_{ij}$ and network measures

| | **Weight** | **Edge Betweenness** |
|---|---|---|
| **Correlation** | 0.458 | 0.558 |




(a) Maximum Spanning Tree




(b) Transmitting Link

Figure 5: Comparing MST and Disease Transmitting Links

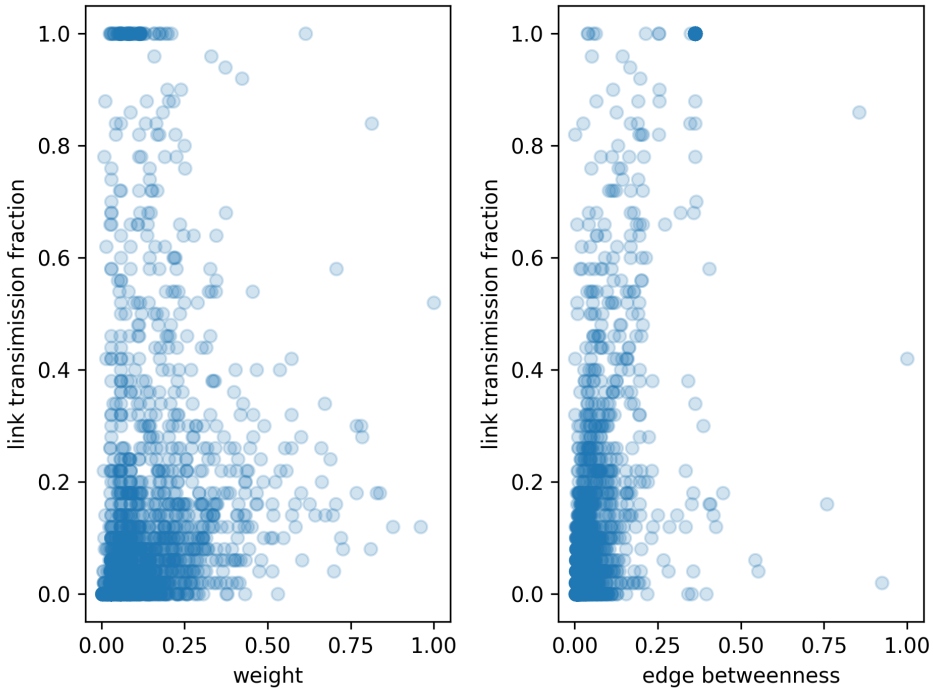



Figure 6: $f_{ij}$ w.r.t Networks measures

d) **How well do the two link properties predict $f_{ij}$ and why? Explain their performance based on how they are defined.**

Referring to the table 3, both the notable correlation values for weight and edge betweenness highlight the predictive power of both link properties in capturing the dynamics of correlations within the network, but the preference for edge betweenness (EB) as a better measure in predicting $f_{ij}$ correlations arises from its heightened correlation compared to edge weight. The elevated correlation underscores EB's effectiveness in capturing the underlying dynamics of network structure. EB proves advantageous due to its ability to identify edges functioning as bridges between distinct communities within the network. These bridges play a crucial role in connecting disparate communities, making them pivotal for the spread of diseases across community boundaries. On the other hand, edge weight solely reflects the numerical count of flights between two nodes, lacking the capacity to discern the structural significance of these connections in the broader network context. It fails to consider the topological position of the connected nodes concerning the rest of the network. Consequently, it becomes less effective in predicting $f_{ij}$ correlations, as it overlooks the crucial aspect of how the interconnected nodes contribute to the overall network structure.

# Task 7: Discussion

Even though extremely simplistic, our SI model can readily give some insights on the spreading of epidemics. Nevertheless, the model is far from an accurate real-world estimate for epidemic spreading.
**Discuss the deficiencies of the current epidemic model by listing at least four (4) ways how it could be improved to be more realistic.**

1) *Dynamic Infection Probability*: rather than assuming a constant infection probability, we can make it dynamic. This means taking into account various factors like the population size of a city, the number of flights, and even region-specific factors like the local behavior of the population.

2) *Incorporating Immunity and Recovery*: instead of sticking with the simple SI model, we could level up to a more realistic SIR (Susceptible-Infectious-Recovered) model. This allows us to consider the recovery and immunity of individuals over time, acknowledging that once someone gets infected, they may develop immunity or recover and return to the susceptible group.

3) *Adaptive Flight Schedules*: instead of assuming a constant flight schedule, we can make it adaptive. If a city gets hit or has a high risk of infection, it makes sense to reduce the number of flights. This dynamic adjustment reflects the real-world scenario where transportation measures are often modified in response to changing epidemiological conditions.

4) *Tailoring Flight Considerations*: our initial model treated all flights equally, but let's get real. Flight duration, the number of passengers, and other factors matter. By factoring in these details, we can better capture how different flights contribute to disease spread. A long-haul international flight might have a different impact compared to a short domestic one.

5) *Incorporate an exposed state*: the current SI model assumes that individuals immediately transition from the susceptible state to the infected state upon contact with an infected individual. This assumption is unrealistic as there is often a period of time between exposure to an infectious agent and the onset of symptoms. To better capture this, the model could be extended to include an exposed state (E). In the SEIR model, individuals first progress from the susceptible state to the exposed state, where they remain asymptomatic but infectious. Once the incubation period ends, they transition to the infected state, where they exhibit symptoms.

6) *Consider heterogeneous transmission rates*: The current SI model assumes a uniform transmission rate, meaning that all infected individuals have the same probability of transmitting the disease to susceptible individuals. This assumption is unrealistic as transmission rates can vary depending on factors such as age, health status, and behavior. For instance, flights to leisure destinations may typically carry younger individuals, which could have a different impact on the spread of the infection.