ES404 Exam-1

Yashraj J Deshmukh
21110245
https://github.com/YYashraj/NCS/Exam1

February 27, 2024

1 Question 1:

For Questions 1-6 is performed on the data provided in the trade-2021-flows. and trade-countries.csv files.

Since the data in both sheets exist in concatenated form in one column, we use tab separation to divide the data into appropriate columns. And we would also need to convert the *amount* column datatype to numeric otherwise it would pose problem in graph generation and node strength calculation.

```
nodes_countries = pd.read_csv('trade-countries.csv', sep='\t')

flows = pd.read_csv('trade-2021-flows.csv')
flows[['from','to','amount']] = flows['from\tto\tamount'].str.split("\t",expand=True)
flows.drop(columns=['from\tto\tamount'], inplace=True)

flows['amount'] = pd.to_numeric(flows['amount'])
```

The trade network is visualised using *geopy* and *Basemap* libraries.

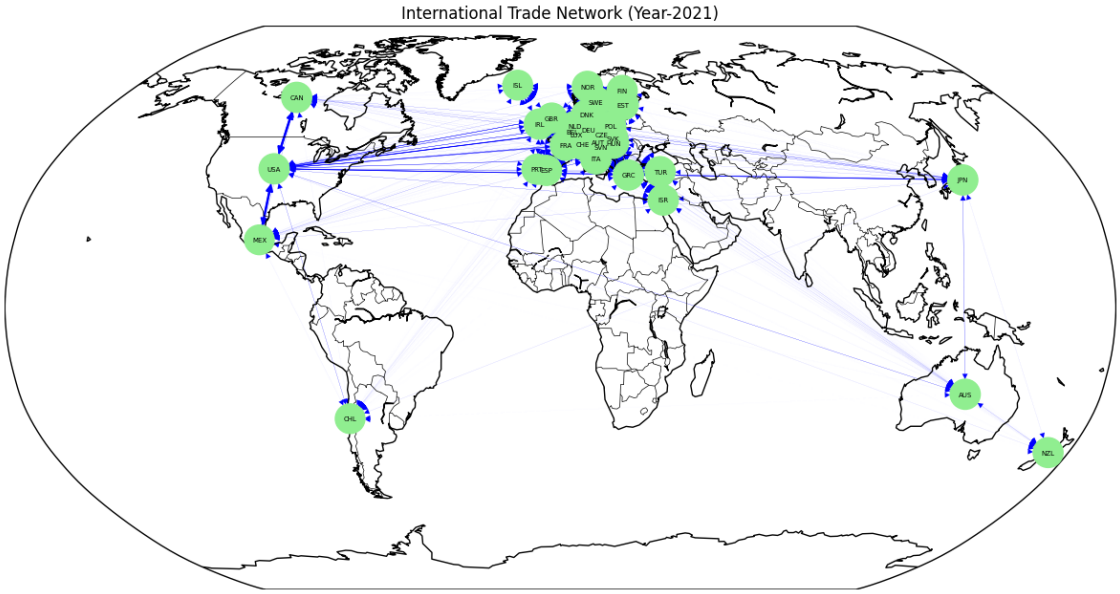

Figure 1: World Plot of Trade

2 Question 2:

Let's take a glance at some basic characteristics of the trade network.

```
DiGraph with 33 nodes and 1049 edges
```

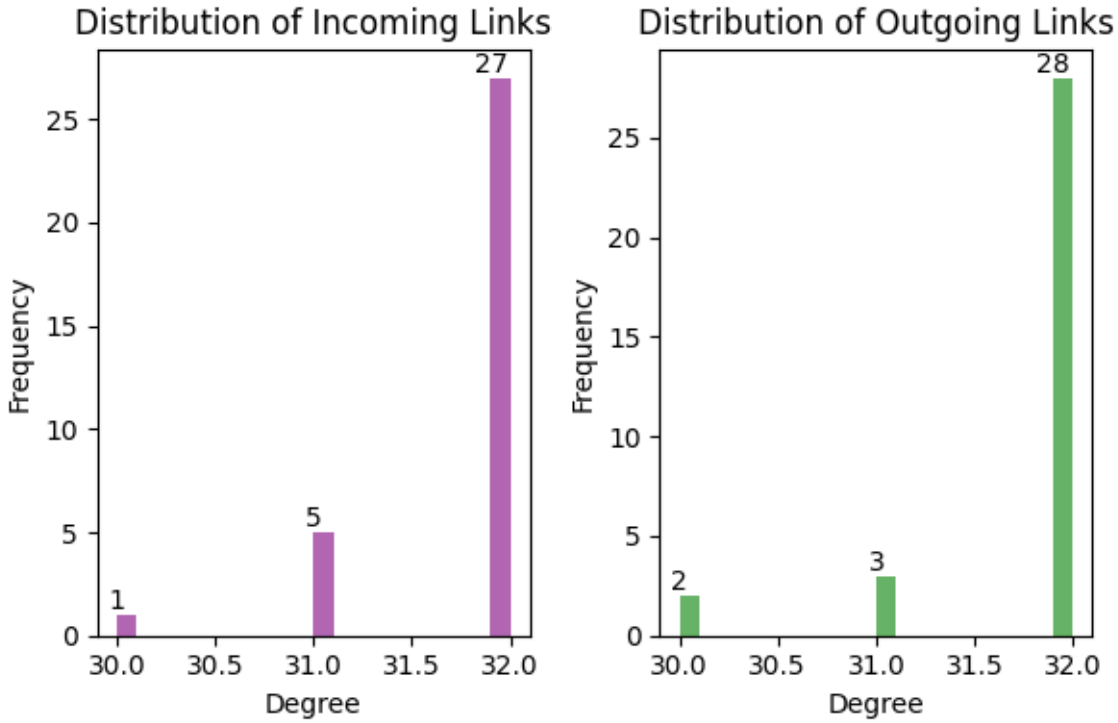

Figure 2: Degree Distribution across Network

Since the network only has 3 distinct degrees, there is no point in plotting the degrees on log scale or performing binning on it.

Next, we perform some hypothesis testing to characterise our distribution. Let our null hypothesis be that the degree distribution follows power-law.

```
In-degree distribution:
Kolmogorov-Smirnov test statistic: 0.8181818007228523
P-value: 1.0910276146255694e-24

Out-degree distribution:
Kolmogorov-Smirnov test statistic: 0.848484834163746
P-value: 2.071638270759745e-27

Reject the null hypothesis for in-degree distribution.
Reject the null hypothesis for out-degree distribution.
```

3 Question 3:

The following are the results of Pearson correlation for the following centrality measures:

```
Pearson correlation between In-Degree and Betweenness Centrality: 0.671515487824284
Pearson correlation between Out-Degree and Betweenness Centrality: 0.6540235839859018

Pearson correlation between In-Degree and Closeness Centrality: 0.9998711235092689
Pearson correlation between Out-Degree and Closeness Centrality: -0.056646214409460154

Pearson correlation between In-Degree and Eigenvector Centrality: 0.9999447512600123
Pearson correlation between Out-Degree and Eigenvector Centrality: -0.05674952711417494
```

The positive correlation coefficient of ¡ 0.65 indicates that nodes with higher degree tend to have higher betweenness centrality, implying that nodes with more connections are likely to lie on more shortest paths between other nodes in the network.

Surprisingly, the negative correlation coefficient of -0.057 indicates a weak negative linear relationship between a node's out-degree and its closeness centrality. This suggests that nodes with higher out-degree may, on average, be slightly further away from other nodes in the network, which contrasts with the expected pattern observed in the in-degree and closeness centrality correlation.

And high eigenvector centrality implies that nodes with higher in-degree tend to have higher eigenvector centrality, indicating their influence within the network.

4 Question 4:

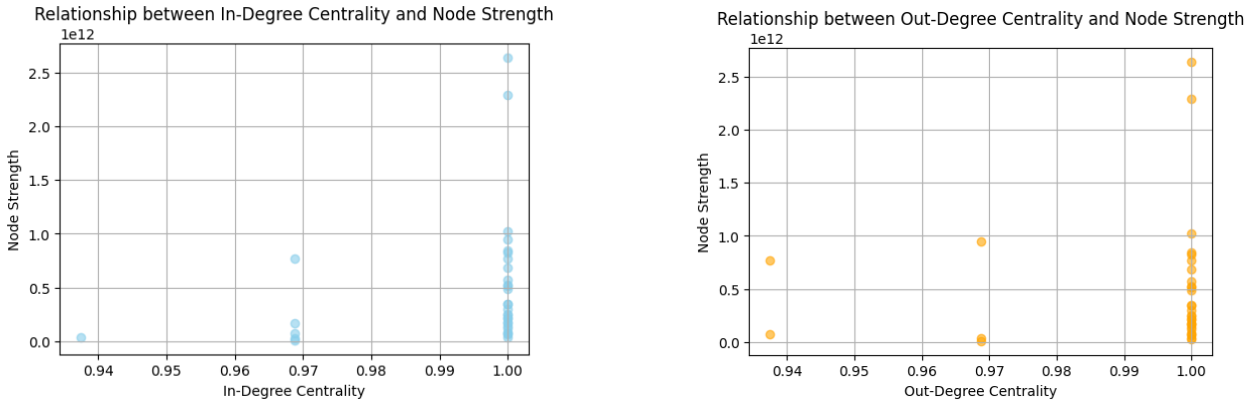

Figure 3: Degree Centrality vs Node Strength

In the network, the nodes which have a higher degree there is often a corresponding increase in node strength. This relationship arises from the fact that nodes with more connections (higher degree) tend to accumulate more interactions or resources, resulting in higher overall strength. As the number of nodes with higher degrees increases, so does the maximum node strength, reflecting the cumulative effect of multiple connections on the strength of individual nodes.

5 Question 5:

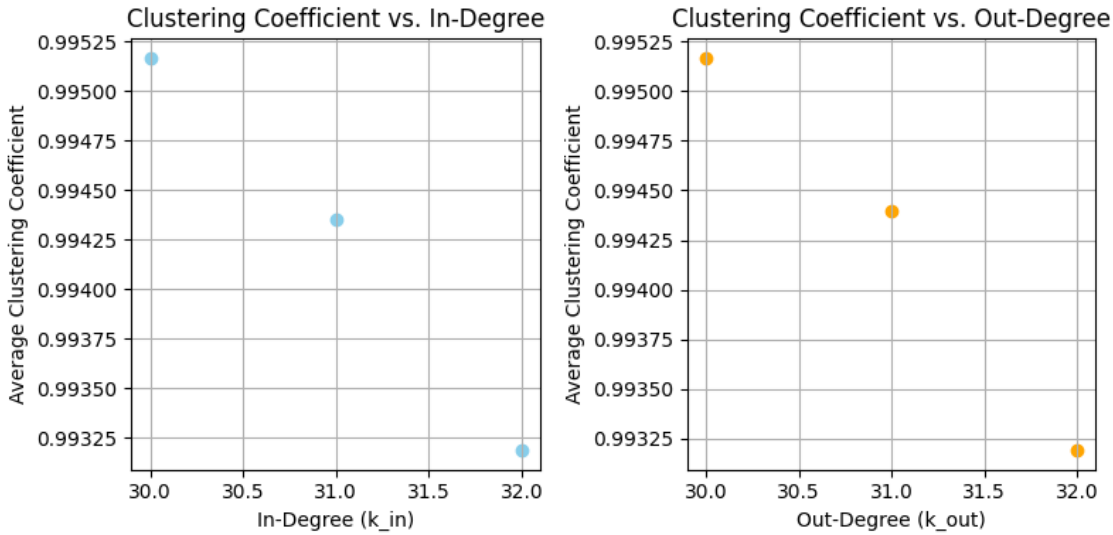

Figure 4: Variation of clustering coefficient with k

This shows that the average clustering coefficient for both in and out degrees is decreasing when increasing degree. But point to note is that there are only 3 distinct continuous degrees so the result may not be what one would typically expect.

6 Question 6:

The average clustering and average shortest path for the networks are as follows:

```
In-degree:
Average Shortest Path Length: 1.0066287878787878
Average Clustering Coefficient: 0.993423556416187

Out-degree:
Average Shortest Path Length: 1.0066287878787878
Average Clustering Coefficient: 0.993423556416187
```

From this one can easily deduce that this is an ultra-small-world network. This comes to no surprise as it was pretty evident from the visualisation that almost all nodes are connected with each other.

7 Question 7:

Models generated by Barabási-Albert (BA) networks and Erdős-Rényi (ER) networks are probabilistic. While BA networks rely on preferential attachment, where new nodes are more likely to connect to well-connected existing nodes, ER networks are characterized by random link formation, with each pair of nodes having an equal probability of forming a connection. These contrasting probabilistic mechanisms play a pivotal role in shaping the structural properties of the resulting networks, including their clustering coefficients.

Mathematically, for a given node i, let k_i be the number of connections (degree) of node i, and let e_i be the number of edges between the neighbours of node i. The local clustering coefficient C_i is defined as:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

In probabilistic models, we can work on expectation to derive a term. Since the link formation is stochastic, we can calculate the number of edges between the neighbours of a node l in terms of the expected number of triangles formed between three nodes in the network, provided that one of the node is l, using the probabilities of each link being realised.

7.1 Clustering Coefficient in BA network

From the above consideration, clustering coefficient for a node in BA network is given as:

$$\begin{aligned} C_l &= \sum_i \sum_j \frac{p(i,l) \cdot p(i,j) \cdot p(j,l)}{k_l(k_l - 1)/2} \\ &= \int_i \int_j \frac{p(i,l) \cdot p(i,j) \cdot p(j,l)}{k_l(k_l - 1)/2} \, \mathrm{d}j \, \mathrm{d}i \end{aligned}$$

Here, $p(a, b)$ is the probability that there is a link between nodes a and b, given that node-b was present when node-a arrived.

Now, $$p(a,b) = \frac{m \cdot k_b}{\sum k_d} = \frac{m \cdot k_b}{2ma - 2m} \approx \frac{m}{2ma} \cdot m \left(\frac{a}{b}\right)^{1/2} = \frac{m}{2(ab)^{1/2}}$$

So, the number of triangles involving l is

$$\begin{aligned} &\int_i \int_j p(i,l) \cdot p(i,j) \cdot p(j,l) \, \mathrm{d}j \, \mathrm{d}i \\ &= \int_i \int_j \frac{m}{2\sqrt{ij}} \cdot \frac{m}{2\sqrt{jl}} \cdot \frac{m}{2\sqrt{li}} \, \mathrm{d}i \, \mathrm{d}j \\ &= \frac{m^3}{8l} \int_1^{N-1} \mathrm{d}i \int_1^{N-1} \frac{1}{ij} \, \mathrm{d}j \\ &\approx \frac{m^3}{8l} \cdot \ln(N) \cdot \ln(N) \end{aligned}$$

Therefore, $$C_l \;=\; \left(\frac{m^3}{8l}\right) \cdot \frac{\ln(N) \cdot \ln(N)}{k_l \cdot (k_l - 1)/2} \;\approx\; \left(\frac{m}{4}\right) \cdot \frac{(\ln(N))^2}{N} \qquad [\text{ because } k_l = m\left(\frac{N}{l}\right)^{1/2}]$$

7.2 Clustering Coefficient in ER network

The calculation for average clustering coefficient for ER network is fairly similar to that for BA network. Here too we will calculate the estimated number of triangles through a node l and normalise it by all possible pairs between its neighbours. The only difference is that the numerator term will be divided by a factor of

two so that we don't end up counting the triangle ijl and jil twice. So,

$$
\begin{aligned}
C_l &= \frac{1}{2} \cdot \sum_i \sum_j \frac{p(i,l) \cdot p(i,j) \cdot p(j,l)}{k_l(k_l-1)/2} \\
&= \frac{1}{2} \sum_{j=1}^{N-1} \sum_{i=1}^{N-1} \frac{p \cdot p \cdot p}{k_l(k_l-1)/2} \quad \text{[as the probability of any edge existing b/w two nodes is the same for all pairs of nodes]} \\
&\approx \frac{1}{2} \cdot \frac{(N-1)^2 \cdot p^3}{((N-1)p)^2/2} \quad \text{[because the expected degree of a node in a random network is } p(N-1)\text{]} \\
&= p
\end{aligned}
$$

8 Question 8:

Two necessary ingredients for emergence of scale-free networks in the BA model are:

8.1 Preferential Attachment

The preferential attachment mechanism dictates that new nodes entering the network are more likely to connect to well-connected nodes. This "rich-get-richer" phenomenon amplifies the probability of high-degree nodes gaining even more links. Thus, this facilitates growth of hubs in the network and leads to a power-law degree distribution.

Mathematically, the probability $\Pi(k)$ that a new node attaches to a node of degree k is proportional to k. This is often expressed as:

$$\Pi(k) = \frac{k}{\sum_j k_j}$$

where k_j represents the degree of node j, and $\sum_j k_j$ sums over all existing nodes in the network.

The absence of preferential attachment will lead to new node attaching to existing nodes randomly, without considering their degrees or connectivity. This means that each existing node has an equal probability of receiving a connection from the new node. As a result, the network generated would be an ER network, not a scale-free one. Let's see how:

Since there is no preferential attachment in the case in consideration, the new incoming node is equally likely to form an edge with any of the existing nodes. Hence, the chance that degree of a node increases is the uniform probability times the number of stubs of new node.

$$So, \quad \frac{dk_i}{dt} = m \cdot \Pi(k_i) = m \cdot \frac{1}{\text{number of nodes at time } t} = m \cdot \frac{1}{t} = \frac{m}{t}$$

$$
\begin{aligned}
\text{Integrating (for when } t \gg 0), \quad \int_m^{k_i(t)} dk_i &= \int_{t_i}^{t} \frac{m}{t} dt \\
=> \quad k_i(t) - m &= m \ln\left(\frac{t}{t_i}\right) \\
=> \quad k_i(t) &= m + m \ln\left(\frac{t}{t_i}\right)
\end{aligned}
$$

This result is different from the usual relation we observe in case of BA model. On further solving this for $p(k)$ we will observe that it is an exponential function. From this we can conclude that without preferential attachment, the BA model will be similar to a random network instead of being scale-free.

8.2 Network should grow

Network can't be stagnant; new nodes must come (or atleast new edges must emerge between already present nodes). The network must expand over time with the continual addition of new nodes and their formation of links with other nodes in network (with preferential attachment). This growth is essential for the dynamic range of node degrees necessary for a scale-free structure.

If the network is not growing or stops to grow after a short while then all the assumptions that we made to prove that BA model generates a scale-free network won't hold true anymore.

Thus, the combination of preferential attachment and network growth ensures that a power-law degree distribution is followed, the emergence of scale-free properties, and the formation of hubs in the network.

9 Question 9:

Let $p_k^{\text{in}}(t)$ and $p_k^{\text{out}}(t)$ denote the probability of a node having in-degree k and out-degree k, respectively, at time t. Let the number of stubs be m.

9.1 Out-degree Distribution

The out-degree distribution is actually pretty straight-forward. Since all nodes come with m number of stubs that is their out-degree, and this value won't change as any new link will count towards their in-degree since they can't produce any new outwards edge. So, $p^{\text{out}}(t) = m$ for all $t > m$.

Note that the above relation is only true for $t > m$. This is because before that critical point in time there won't be sufficient nodes in the network to completely utilise the stubs of a node. In this time period, the out-degree = number of nodes - 1, or, $t - 1$.

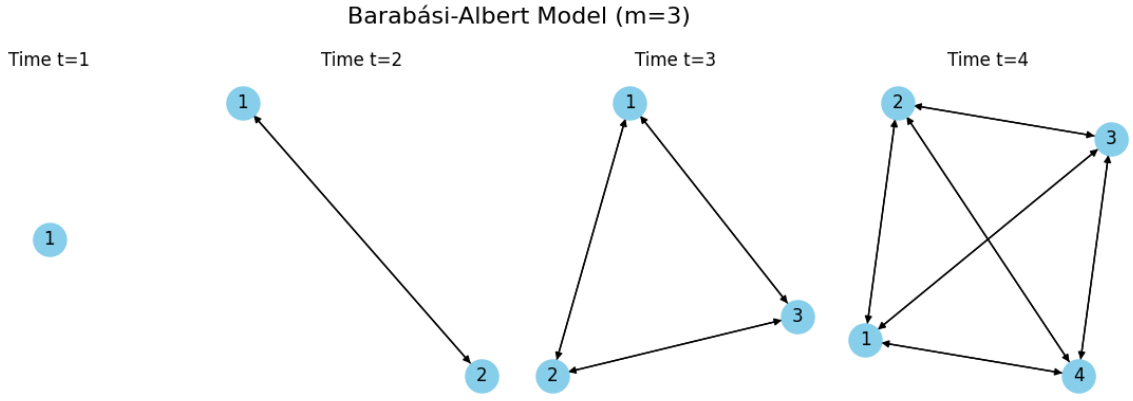

Figure 5: Visual Representation of Network for time $t \leq m$

$$Therefore, \quad p^{\text{out}}(t) = \begin{cases} t-1 & \text{t} \leq \text{m} \\ m & \text{otherwise} \end{cases}$$

9.2 In-degree Distribution

Here, preferential attachment is defined as: $\Pi(k_i^{\text{in}}) = \dfrac{k_i^{\text{in}}}{\sum_j k_j^{\text{in}}} = \dfrac{k_i^{\text{in}}}{2mt}$

Let $N_k(t)$ be the number of nodes with degree k at time t.

Whenever a new node arrives it affects N_k and p_k^{in} .
The in-degree of a k in-degree node can become $k+1$ if the new node links with it.
The in-degree of a $k-1$ in-indegree node can become k if the new node links with it.

At any time, the number of nodes in the network, N, is equal to t.
Now, number of links that are expected to connect to k in-degree nodes is: $m \cdot \dfrac{k}{2mt} \cdot N \cdot p_k^{\text{in}}(t) = \dfrac{k}{2} \cdot p_k^{\text{in}}(t)$

Then at time $t+1$, the number of k in-degree nodes is:

$$(N+1) \cdot p_k^{\text{in}}(t+1) = N \cdot p_k^{\text{in}}(t) + \frac{k-1}{2} \cdot p_{k-1}^{\text{in}}(t) - \frac{k}{2} \cdot p_k^{\text{in}}(t) \quad \forall \;\; k > 0$$

The in-degree for nodes of the network ranges from [0, N). Minimum possible in-degree is 0 as a new node would only have outward directed edges; no edges will be pointed towards. Its in-degree increases only when a new new node attaches to it.

When the size of the network has grown sufficiently large, then the degree distribution barely changes even on arrival of new nodes *i.e.* for $t >> 1$, $p_k^{\text{in}}(t) \approx p_k^{\text{in}}$ [time invariance]

So, $(N+1) \cdot p_k^{\text{in}} - N \cdot p_k^{\text{in}} = N \cdot p_k^{\text{in}} + p_k^{\text{in}} - N \cdot p_k^{\text{in}} = p_k^{\text{in}}$

$$\begin{aligned}
&\Rightarrow p_k^{\text{in}} = \frac{k-1}{2} \cdot p_{k-1}^{\text{in}} - \frac{k}{2} \cdot p_k^{\text{in}} \\
&\Rightarrow p_k^{\text{in}} \cdot \left(1 + \frac{k}{2}\right) = \frac{k-1}{2} \cdot p_{k-1}^{\text{in}} \\
&\Rightarrow p_k^{\text{in}} \cdot (2+k) = (k-1) \cdot p_{k-1}^{\text{in}} \\
&\Rightarrow p_k^{\text{in}} = \frac{k-1}{k+2} \cdot p_{k-1}^{\text{in}} \qquad\qquad \forall\; k > 0
\end{aligned}$$

For 0-degree node,

$$\begin{aligned}
&(N+1) \cdot p_0^{\text{in}}(t+1) = N \cdot p_0^{\text{in}}(t) + 1 - \frac{(0)}{2} \cdot p_0^{\text{in}}(t) \\
&\Rightarrow (N+1) \cdot p_0^{\text{in}}(t+1) - N \cdot p_0^{\text{in}}(t) = 1
\end{aligned}$$

For time invariance,

$$(N+1) \cdot p_0^{\text{in}} - N \cdot p_0^{\text{in}} = 1 \quad \Rightarrow \quad p_0^{\text{in}} = 1$$

This is a peculiar relation as it would mean that after a long time all nodes have degree 0! But taking a look back we can realise what the cause for this is.

We defined our preferential attachment to be proportional to the in-degree of a node. But the problem here is that a new node would have its in-degree set to 0. The only nodes that would have non-zero in-degree would be the first $m+1$ node (as shown in the example of out-degree distribution). For any time later when a new node arrives, it would have 0 probability of forming links with a node that came after $t = m+1$, so all the m connection would be formed between the $m+1$ nodes only. So, after $t >> 1$ time, the network will be flooded with nodes having zero-indegree and hence why the p_0^{in} tends to 1.

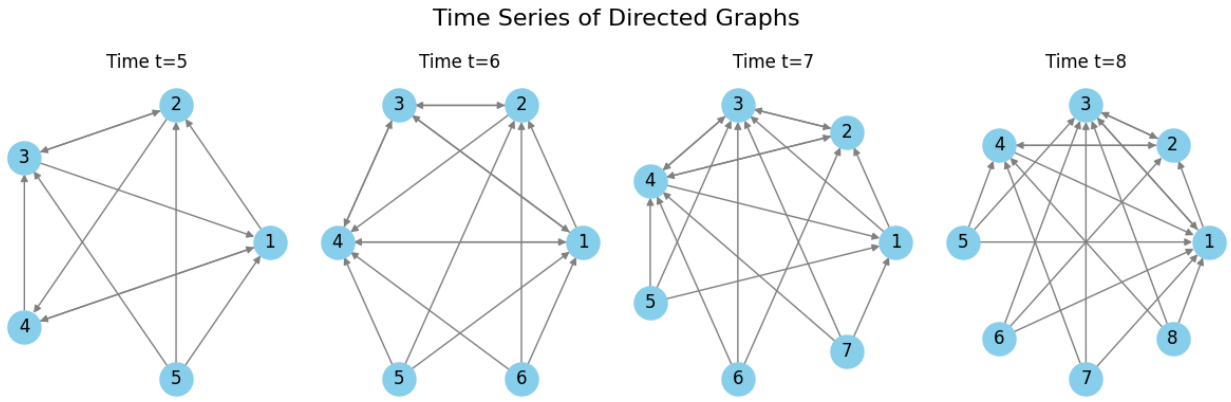

Figure 6: Network after time $t > m$, newer nodes can only have in-degree=0

A good way to avoid this would be to apply Laplace (or simply add-one smoothing). This would give some chance to newer nodes that don't have any in-degree.

Proposed preferential attachment: $$\Pi(k_i^{\text{in}}) = \frac{k_i^{\text{in}+1}}{\sum_j (k_j^{\text{in}} + 1)} = \frac{k_i^{\text{in}} + 1}{2mt + t}$$