Exercise 1: Analysis the copying model  

Through the analysis of the Yule process, we have seen in class the consequence of reinforcement. Reinforcement here denotes the fact that a difference between two entities (e.g. the size of two genus, the number of links received by two webpages) is itself biasing the dynamics so that the difference continues to increase. As a consequence, even starting from a small initial set of equivalent entities, minor difference created by randomness could further lead to major differences. In the case of the Yule process, it provided a simple model explaining the imbalance of species among genus which is characterized by a power law.

In this exercise, we conduct a very similar analysis to model edges created in a graph. The main result is to show that a very simple copying strategy leads to big imbalance, characterized by a power law degree distribution of nodes’ in-degree.

The copying model  

We analyze the following dynamics. We start from a directed graph containing $N_0$ nodes such that each of these nodes has exactly one outgoing edge. We introduce at each time step, denoted by $t = 1, 2, \ldots$, a new node $v(t)$ with a single outgoing edge $e(t)$ that is initially connected to another node chosen uniformly at random (that we denote by $u(t)$). We then assume the following evolution:

- with probability $p$, the process stops there, and the new edge connects $v(t)$ to $u(t)$.
- otherwise (hence, with probability $1-p$), $v(t)$ examines the edge that is starting at $u(t)$ and decides to copy this edge. This means that the edge from $v(t)$ to $u(t)$ is changed to one that goes from $v(t)$ to the destination of the edge starting in $u(t)$.
Evolution of node’s degree  Since the graph is directed all nodes both have an out-degree and an in-degree. The out-degree of all nodes in the graph remains constant equal to 1. The interesting problem is to analyze the evolution of the in-degree of nodes in the graph as \( t \) becomes large.

Let us denote for any \( i \geq 0 \) by \( X_i(t) \) the number of nodes in the graph with an in-degree equal to \( i \).

1 (\( \rightarrow \)) How many nodes (denoted by \( N(t) \)) and edges (denoted by \( E(t) \)) are there in the graph as a function of \( t \)?

2 (\( \sim \)) Assuming that \( X_0(t) \) (i.e., the number of nodes with no incoming edge) is known, what are the possible values of \( X_0(t+1) \) and what are the probability that these values occur?

3 (\( \sim \)) Derive from the previous question the evolution equation giving the expectation \( E[X_0(t+1)] \) as a function of \( E[X_0(t)] \).

4 (\( \sim \)) Let us introduce, for a given constant \( c_0 \), the sequence \( \Delta_0(t) = E[X_0(t)] - c_0t \). Show that there exists a value of the constant \( c_0 \) such that:

\[
\exists A > 0 \text{ such that for any } t > 0 |\Delta_0(t)| \leq A \ln(t) .
\]

What is the value of \( c_0 \)? (i) \( c_0 = \frac{1}{1-p} \) (ii) \( c_0 = \frac{1}{2-p} \) (iii) \( c_0 = \frac{1}{1+p} \)

(N.B.: Note that the following fact is useful: If \( (x_n)_{n \geq 0} \) is a sequence of real number such that \((\forall n \geq 0, x_{n+1} = x_n r_n + s_n, \text{ and } |r_n| \leq 1 )\), then we have \( |x_n| \leq |x_0| + \sum_{i=1}^{n} |s_i| \).)

5 (\( \rightarrow \)) Deduce that the following hypothesis is true for \( i = 0 \):

\[
\forall \varepsilon > 0, \exists A > 0 \text{ such that } |\Delta_i(t)| \leq A t^\varepsilon .
\]

6 (\( \sim \)) For a sequence of constant \( c_0, c_1, \ldots \), let us define \( \Delta_i(t) = E[X_i(t)] - c_i t \). Show that for any \( i > 0 \), if the sequence satisfies \( c_i = c_{i-1} \left( 1 - \frac{2-p}{(1+p)+i(1-p)} \right) \) then we have:

\[
\Delta_i(t+1) = \Delta_i(t) \left( 1 - \frac{p+i(1-p)}{N(t)} \right) + \Delta_{i-1}(t) \frac{p+(i-1)(1-p)}{N(t)} + \frac{A}{N(t)}
\]

where \( A \) is a constant.

7 (\( \sim \)) Deduce by recurrence that the hypothesis of question 5 is true for all \( i \geq 0 \).

8 (\( \sim \)) We admit the following concentration result on the random sequence \( (X_i(t))_{t>0} \) for any \( i \geq 0 \)

Lemma 1. For any \( M > 0 \), we have \( \mathbb{P} \left[ |X_i(t) - E[X_i(t)]| > M \right] \leq 2 \exp \left( -\frac{M^2}{8t} \right) \)

Prove that almost surely there exists a value \( T \) such that for \( t > T \) we have

\[
|X_i(t) - E[X_i(t)]| \leq 4 \sqrt{t \ln(t)} .
\]

Conclude that for any \( i \geq 0 \), almost surely, the fraction of nodes with in-degree equal to \( i \) converges to \( c_i \) as \( t \) grows large.
9 (→) Assuming that $p < 1$, show that for $i > 0$ we have:

$$c_i = c_{i-1}(1 - \frac{\beta}{i} + \varepsilon(i)) \text{ where } |\varepsilon(i)| \leq \frac{A}{i^2} \text{ and } \beta = \frac{2 - p}{1 - p}.$$  

As a consequence, as shown in the lecture, if we neglect the error term $\varepsilon(i)$ we have that $c_i$ is approximately following a power-law with coefficient $\beta$.

For which values of $p$ does the power law becomes the most imbalanced? Does this correspond to your intuition about the dynamics of copying.

10 (←) Assuming now $p = 1$, how could you characterize the decrease of $c_i$ as a function of $i$? Relate this behavior to the dynamics of the copying model.