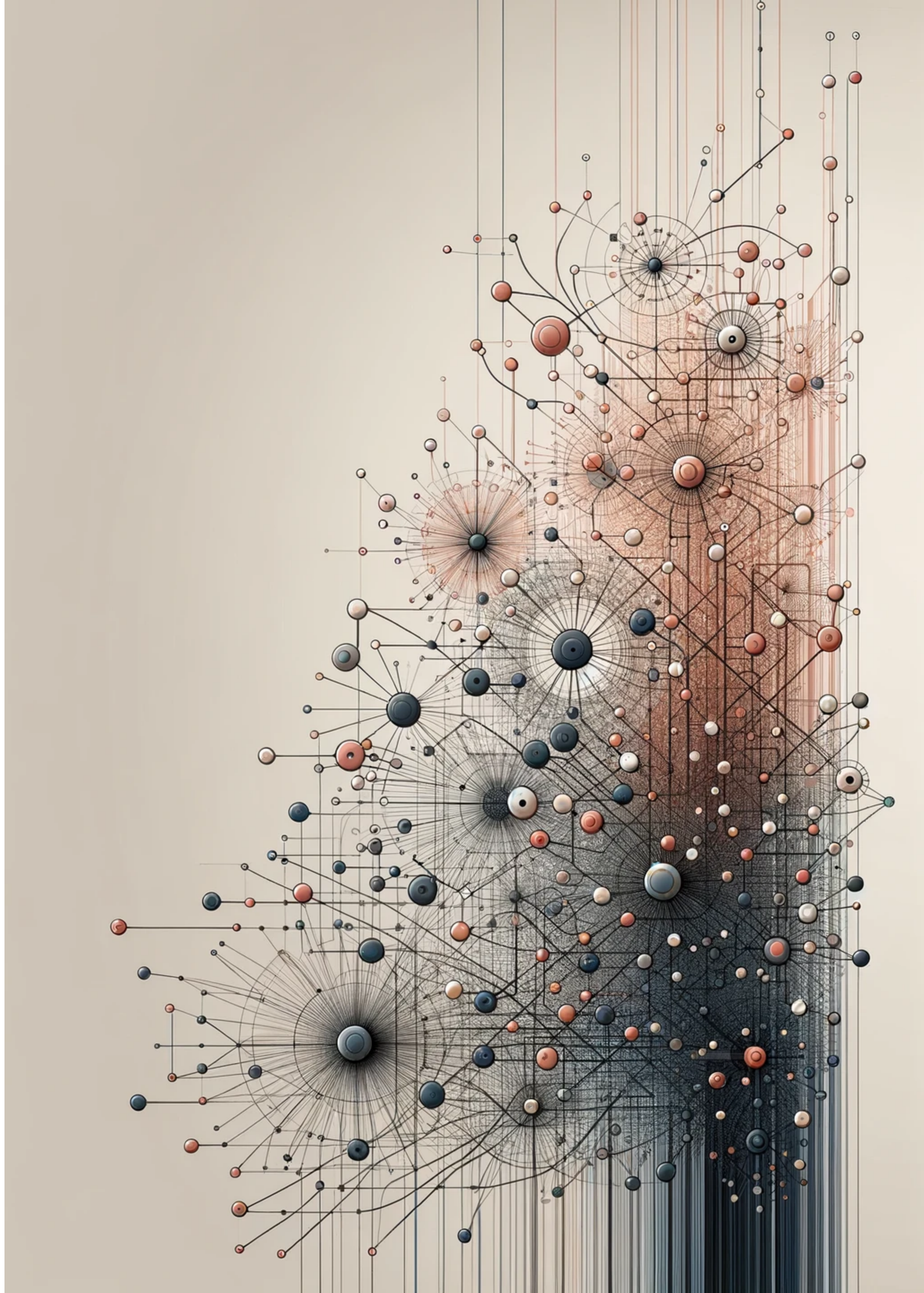
UNIVERSITÀ
DI TORINO

Analisi e Visualizzazione delle Reti Complesse

NS10 - Analysis of Rich-Get-Richer Processes

Prof. Rossano Schifanella



Analysis of Rich-get-Richer processes

Objectives

- $f(k)$ fraction of nodes with degree k
- Goal: $f(k) \propto k^{-c}$
- Why does this happen with the Rich-Get-Richer model?
- What is the role of c?

UNIVERSITÀ DI TORINO

Recap of the Rich-Get-Richer process

- Lesson from cascades: we assume that people tend to copy the decision of people who acted before them

1. Nodes are created in a sequence: $1, 2, \ldots, N$
2. For each node j that joins the network, repeat:
   i. with probability $p \Rightarrow$ page i is selected uniformly at random, and a link (j, i) is created
   ii. with probability $1 - p \Rightarrow$ page i is selected uniformly at random, l is the page i is connected to, then a link (j, l) is created

- Keep the process simple: only one link is created at every step basic formalization

Basic formalization

- $X_j(t)$ random variable that represents the number of links to j at a time step t
- $X_j(j) = 0$
- $X_j(t+1) = X_j(t) + \frac{p}{t} + \frac{(1-p)X_j(t)}{t}$
  - where the term $\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$ is the expected change in $X_j(t)$

The deterministic argument

Let's suppose that:

- time runs continuously from 0 to N
- $X_j(t) $is a continuous function
- It is like we are ignoring probabilities, and our idealized physical system just starts from a set of initial conditions
- $X_j(t+1) = X_j(t) + \frac{p}{t} + \frac{(1-p)X_j(t)}{t} \Rightarrow \frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}$
- let's set $q = 1 - p$

$$\frac{dx_j}{dt} = \frac{p + qx_j}{t}$$

$$\frac{1}{p + qx_j} \cdot \frac{dx_j}{dt} dt = \frac{1}{t} dt$$

- Integrating both sides:

$$\int \frac{1}{p+qx_j} \cdot dx_j = \int \frac{1}{t} dt$$

$$q\left(\frac{\ln(p+qx_j)}{q} + c_1\right) = q(\ln t + c_2)$$

$$\ln(p+qx_j) = q\ln t + c$$

- Let us set $A = e^c$
- We can exponentiate both sides:

$$p + qx_j = At^q$$

$$x_j(t) = \frac{1}{q}(At^q - p)$$

- Recall initial condition: $X_j(j) = 0$

$$0 = X_j(j) = \frac{1}{q}(Aj^q - p)$$
$$Aj^q - p = 0$$
$$A = \frac{p}{j^q}$$

- We can substitute $A = \frac{p}{j^q}$ with $x_j(t) = \frac{1}{q}(At^q - p)$

$$x_j(t) = \frac{1}{q}\left(\frac{p}{j^q}t^q - p\right) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right]$$

- So we solved the deterministic approximation:
- $x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right]$ is a closed form expression for how each x_j grows over time

Identifying a power law in the deterministic approximation

- For a given value of k and a time t, what fraction of all functions x_j satisfies $x_j \geq k$?

$$x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right] \geq k$$

$$\left[\left(\frac{t}{j}\right)^q - 1\right] \geq k\frac{q}{p}$$

$$\frac{t^q}{j^q} \geq k\frac{q}{p} + 1$$

$$t^q \geq j^q \cdot \left(k\frac{q}{p} + 1\right)$$

$$j^q \leq t^q\left(\frac{q}{p}k + 1\right)$$

$$j \leq t\left(\frac{q}{p}k + 1\right)^{-\frac{1}{q}}$$

- Out of all the functions $x_1, x_2, \ldots, x_t$ at time t, the fraction of values j that satisfies the above inequality is:

$$F(k) = \frac{1}{t} \cdot t \left(\frac{q}{p} k + 1 \right)^{-\frac{1}{q}} = \left(\frac{q}{p} k + 1 \right)^{-\frac{1}{q}}$$

- We have the shape of a power law $F(k) \propto k^{-c}$:
    - $\left(\frac{q}{p} k + 1 \right)$ is proportional to k
    - $-\frac{1}{q}$ is a negative exponent

$F(x)$: fraction of nodes with **at least** in-degree k

but we aim at finding an approximation for

$f(k)$: fraction of nodes with **exactly** in-degree k

that means we can approximate $f(k)$ taking the derivative:

$$
\begin{aligned}
-\frac{dF}{dk} &= -\frac{d\left(\frac{q}{p}k+1\right)^{-\frac{1}{q}}}{dk} \\
&= \frac{1}{q}\cdot\frac{q}{p}\cdot\left(\frac{q}{p}k+1\right)^{-1-\frac{1}{q}} \\
&= \frac{1}{p}\cdot\left(\frac{q}{p}k+1\right)^{-1-\frac{1}{q}} \propto k^{-(1+\frac{1}{q})}
\end{aligned}
$$

Final step

The deterministic approximation of the model predicts that:

$$f(k) \propto k^{-\left(1+\frac{1}{q}\right)}$$

that is a power law with exponent:

$$1 + \frac{1}{q} = 1 + \frac{1}{1-p}$$

Meaning of the exponent

Let's study the behavior of the exponent:

$$\lim_{p\to 1}\left(1+\frac{1}{1-p}\right)=\infty$$

- the exponent is infinity when link formation is mainly governed by uniform random choice ($p \to 1$): very large numbers of in-degree are extremely rare

$$\lim_{p\to 0}\left(1+\frac{1}{1-p}\right)=2$$

- the growth is mainly governed by the preferential attachment process. The power law's exponent decreases toward 2, allowing for nodes with very large in-degree

Conclusion

- Rich-Get-Richer processes explain the emergence of power laws and also exponents that in real scenarios are often slightly larger than 2
- Case Study: empirical findings in the Web showed that in-degree distributions can be fitted by a power law with exponent ≈ 2.1

Some practical notes

Plotting empirical distributions

- When we download some data (or a sample), we have a collection of observations
- We should count the observations as a function of a given variable, then we can plot the empirical (probability) distribution
    - For example, we can count how many individuals in our sample have a given height
- If the variable has continuous values, we need to discretize these values into intervals (binning)

UNIVERSITÀ DI TORINO

Example: humans heights

- In this example, we read a dataset stored in a CSV file to create a pandas dataframe
- We can count every occurrence of heights values in different intervals, then divide every sum by the size of the sample (e.g., 10000)
- Python has a lot of pre-boiled methods and function to do that

```
hw_df = pd.read_csv("datasets/weight-height.csv")
hw_df
```

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

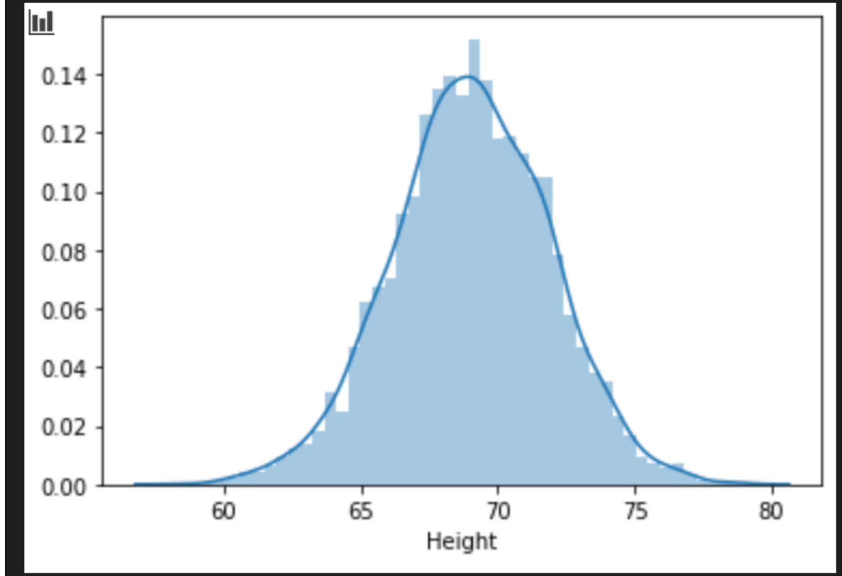
10000 rows × 3 columns

Histograms

- The histogram is a natural choice
- you can also try scatterplots
- Library seaborn has functions that make everything: counting, normalizing, binning, fitting

```
hw_df_male=hw_df[hw_df.Gender=='Male']
sns.distplot(hw_df_male.Height)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcd711ef990>
```



The 3 sigma rule of thumb

- Given an empirical distribution, we can check where the observed data falls
- The three-sigma rule or 68-95-99.7 rule, is a statistical rule which states that for a normal distribution, almost all observed data will fall within three standard deviations (denoted by δ) of the mean or average (denoted by μ)

```
mu = hw_df_male.mean().Height
sigma = hw_df_male.std().Height
sample100 = hw_df_male.sample(100).Height
np.sum((sample100.values >= mu-sigma) & (sample100.values <= mu+sigma))/100
0.62

np.sum((sample100.values >= mu-2*sigma) & (sample100.values <= mu+2*sigma))/100
0.97

np.sum((sample100.values >= mu-3*sigma) & (sample100.values <= mu+3*sigma))/100
0.99
```

Exercises

- Download a sample of the Web graph, for example, from here
- Create a directed graph from the Web sample
- Generate a random graph with an equal number of nodes and edges for comparison
- Calculate degree distributions of both graphs, and plot them
- Estimate heterogeneity of both graphs
- Is the 3 sigma rule useful here?
- Can you say if some degree distribution would be fitted with a power law?

Reading material

[ns2] **Chapter 18 (18.7) Power Laws and Rich-Get-Richer Phenomena**

Please check your general understanding of the topic completing the exercises at the end of the chapter

UNIVERSITÀ DI TORINO

Q&A