Facebook based complex network structure, simulation of a cyber-attack and the correlation between fame and ratings

Author: Javier Castillo Uviña
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.[*]

Teacher: Maria Ángeles Serrano & Marián Boguñá

Abstract: This project will talk about a complex network created based on TV shows official accounts on *Facebook* linked by mutual *likes*. We will study and characterize the network behavior. In addition, we simulate a cyber-attack governed by an epidemic model. IN the las section we will try to see some correlations between being active on social media and having good rates on *Google*. The network behavior will be described based on its degree distribution, the average nearest neighbors degree and clustering coefficient. The epidemic model used is the SIS model with a numerical and a Gillespie algorithm interpretations.
It is obtained that this network has a social network-like properties. We also obtained some positive correlations between the number of edges and the TV show rates. The bottom line is that the numerical approach has the same behavior as Gillespie's algorithm, but is not super accurate.

I. INTRODUCTION

For the audiovisual production world is very important to have popularity, the bigger the audience, the greater the profit. Having a high presence on social networks is important due the fact that is another way to get publicity. In addition, this exposure to socials let the costume feel involved to the television show, the audience become to have an active role. This activity ranges from a like in a post to a hole event organization. The public accessibility through social media has a positive impact to the television shows [1].

As we can see in [2], according to the Nielsen's report "The Relationship between social Media buzz and TV ratings" an average increase of 9%-14% on the online buzz implies an increasing of the 1% on the TV show ratings. Moreover, *likes*, comments and shares on *Facebook* are indicators of how well-rated is going to be a TV show. It is also demonstrated the direct relation between posts of the official accounts and the increasing on the rates. In conclusion having an active official account generates social media buzz around the show.

We have a network based on the *Facebook* TV shows accounts interaction, it has been extracted from a database *The Network Data* [3], the network is unweighted and undirected. Data were collected in November 2017. Each node is an official, *blue verified*, TV show account of different categories. Edges represent the mutual *likes* between these accounts. Apart from the edge list, we also have a list with the shows names and its corresponding node.

The main objective of the project is to characterize the structure of the network. In addition, we will make a different approach to the relationship between the ratings of TV shows and the fact of being active with other shows. We will try to find a correlation between popularity in this network and how well rated they are. In this case we will understand by "activity" the importance they have in this network of mutual *likes* in *Facebook*.

On the last section we will simulate a virus propagation through the Susceptible-Infectious-Susceptible (SIS) epidemic model. The objective of this simulation is trying to estimate the damage of a cyber-attack in this CN. In addition, this will give us a picture of how correlated the nodes are. Looking at the infected node behavior we will extract some information about the CN structure.

II. NETWORK STRUCTURE

A. Degree distribution

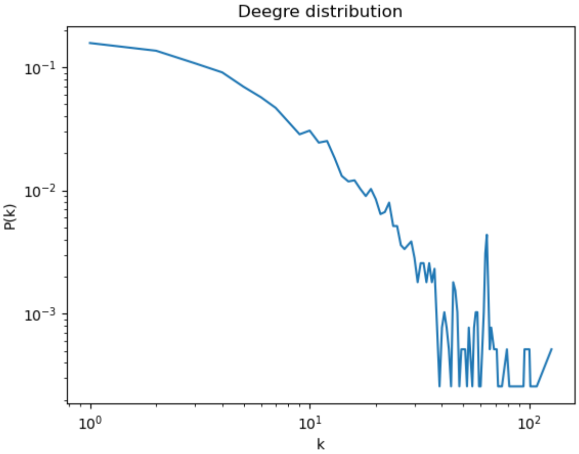

FIG. 1: Normalized degree distribution for the nodes degree in the network based on the TV shows related through *Facebook likes*.

[*]Electronic address: jcastiuv7@alumnes.ub.edu

The first thing to work with is the degree distribution, as we can see in FIG. 1 we get a law $P(k) \propto e^{-\frac{k}{\langle k \rangle}}$, which perfectly describes the behavior for low k, but there is noise for high k. This exponential behavior is characteristic of growing random graphs. This type of network usually has few nodes with high degree. In addition, due to the randomness we expect to have a small world network, it makes sense because we are studying data from social media.

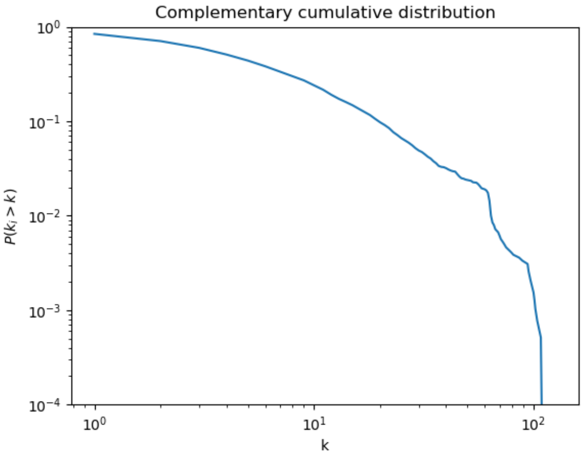

FIG. 2: Normalized complementary cumulative degree distribution for the nodes degree in the network based on the TV shows related through *Facebook likes.*

In FIG. 2 the Y range has been cut off because of it tends to 0 in a log-scale axis. We also can observe an anomaly after $k = 50$.

B. Average nearest neighbor degree

The average nearest neighbor degree (ANND) is a mechanic used in order to know correlations between different degrees in a network.

$$\bar{k}_{nn}(k) = \sum_{k'} k' P\left(k' \mid k\right) = \frac{1}{N_k} \sum_{i \in \mathcal{V}(k)} \frac{1}{k_i} \sum_j a_{ij} k_j \quad (1)$$

k is the node degree, k' is the neighbor node degree, $P(k'|k)$ is the probability of having a neighbor of degree k' being at a node of degree k, N_k are all the nodes with degree k and a_{ij} is the adjacency matrix.

The FIG. 3, despite some exceptions, shows an assortative mixing behavior. So the bigger the degree the bigger of its neighbors degree. This is the expected for the social media based networks, we can find this kind of ANND in biological and technological networks as well. These ”exceptions” and the big value for the degree variance $Var(k) = 157.7$ we can conclude it is a heterogeneous network. In social networks there is a tendency of generating this heterogeneities due to the fact that a nodes ”fame” does not have to affect to its neighbors. Despite this erratic behavior referring to some specific k, the tendency is to have neighbors with high degree if we are in a node with a big degree. Moreover, we expect to have few nodes with high due to the exponential degree distribution. Joining this two statements we can suppose that in this network occurs the rich club phenomenon. It is a common property for the social networks [4].

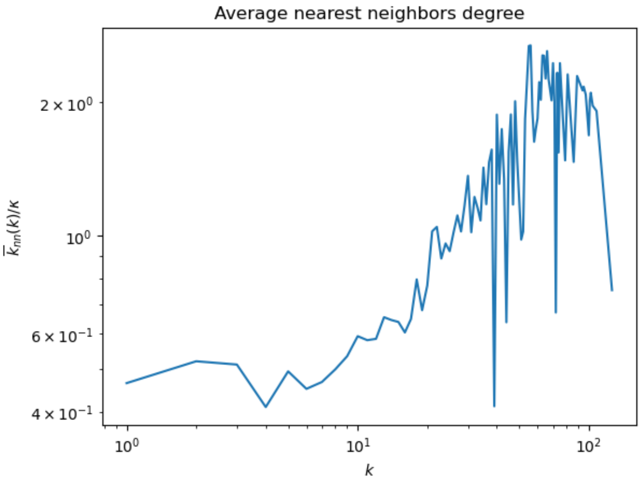

FIG. 3: Average nearest neighbor degree of the network. We plotted it with the $\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle}$ factor dividing, in order to standardize it to be able to compare it with other networks.

C. Clustering

The local clustering is the ratio of triangles versus the number of triplets we can find in average choosing a node of degree k. It helps us to know the nodes of degree k tendency to form clusters or groups. The clustering coefficient computed over the node i is:

$$c_i = \frac{2T_i}{k_i\left(k_i - 1\right)} \quad (2)$$

T_i is the number of triangles it forms and k_i its degree. The global clustering coefficient is the local clustering average and has a value of $C = 0.5479$. A high value for the global coefficient gives us information about the strength of the CN structure. A low value means low clustering o high randomness on its links. In this specific case a high connectivity (high clustering) mean the TV shows will have a lot of publicity through posts sharing. We will see below if we have a high or a low clusterization.

As we can see in FIG. 4 the CN tends to grow slowly with k despite it has specific k with low clusterization.

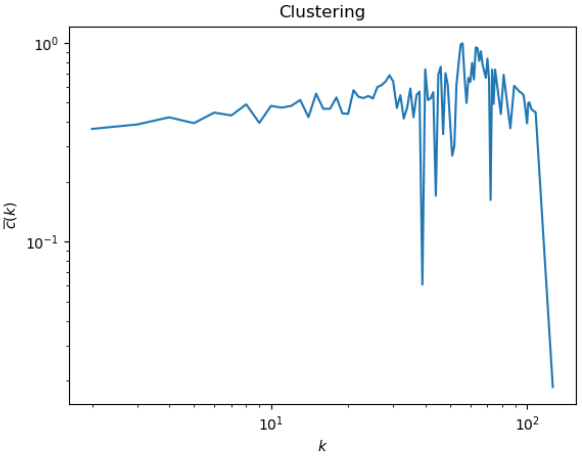

FIG. 4: Averaged clustering coefficient computed for each degree.

In this case this means that popularity calls popularity, we notice a tendency of creating clusters of "famous" TV shows.

III. COMPARISON TO THE CONFIGURATIONAL MODEL (CM)

The Configurational Model is a way to have a "ground state" for a network. We randomize the complex network in order to know if it has a certain structure. We will average it a 100 times simulation in order to minimize the noise associated of the stochastic CM confection. Due to the fact that the *structural cut off* is:

$$k_s(N) \sim (\langle k \rangle N)^{1/2} = 186 \tag{3}$$

being N the number of nodes. The highest degree is $k_{max} = 126$ so the CM nodes degree are conserved from the original CN. It means the degree distribution will be the same as the previous section, we just rewired the edges. If the k_{max} had been higher, we would have had to estimate again the degree of the nodes with a degree higher than the *structural cut off*, since otherwise the CN would be forming a very solid structure that would be difficult to break. Thank to the random version of our CN we now can compare and notice if the magnitudes are high or low.

A. Average nearest neighbor degree

As we can see in FIG. 5 the CM has a very constant and low value compared to the real CN. So we assume the construction of our network it is not random and some TV Shows tend to be correlated by degree. In this case, the trend is as follows: the degree of the node and its neighbors are simmilar.

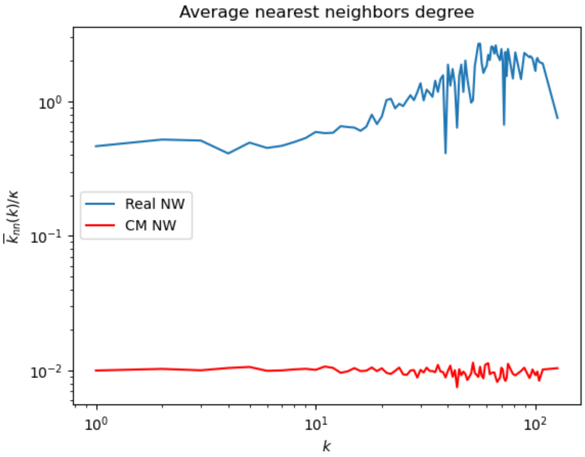

FIG. 5: Average nearest neighbor degree of the network. We plotted it with the $\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle}$ factor dividing, in order to standardize it to be able to compare it with other networks. Here appears another ANND which corresponds to the CM in red. Real NW means the real network constructed with the data and CM NW means the configurational model noetwork created with the real nodes degrees.

B. Clustering

The global clustering coefficient for the CM is $C = 0.0002$ and it is almost constant. Compared to $C = 0.5479$ for the real network we can assume that this node connection it is not casual. They are forming communities, it can be related to the actors they share, the genre, the stream platforms, the producer...

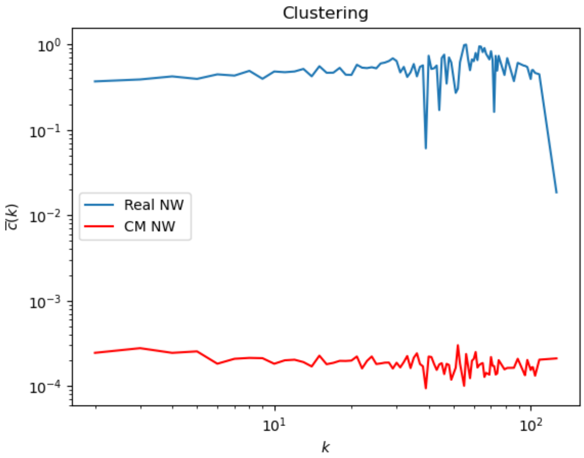

FIG. 6: Averaged clustering coefficient computed for each different degree. In blue for the real network (Real NW) and in red for the configurational model network (CM NW).

IV. SUSCEPTIBLE-INFECTIOUS-SUSCEPTIBLE (SIS) EPIDEMIC MODEL

We will simulate a computer attack in which they will manage to hack 15 accounts at initial time. The attack will be propagated via private message with the accounts that have mutual *likes*. From time to time *Facebook* will return the account to the original owner and that will simulate the recoveries. The messages are sent in time intervals distributed with a Poisson probability density with λ as a expected time (measured in recovery rates δ).
From the literature we know the system will evolve following this expression:

$$\frac{d\rho_i}{dt} = -\delta\rho_i + \lambda \sum_j a_{ij}(\rho_j - \rho_{ij}) \tag{4}$$

$\rho_i \equiv \langle n_i(t)\rangle$, δ is the recovery rate and we will consider it 1, λ is the infection rate and a_{ij} is the adjacency matrix, of course $\rho_{ij} \equiv \langle n_i(t)n_j(t)\rangle$. In our simulation we will use a mean field approximation which neglects the correlations between nodes but describes precisely the behavior, $\rho_{ij} \equiv \langle n_i(t)n_j(t)\rangle \simeq \langle n_i(t)\rangle\langle n_j(t)\rangle$

We will solve the resultant equation below with a Runge-Kutta 4 numerical method, and we will call it the *Mean-Field* solution:

$$\frac{d\rho_i}{dt} = -\delta\rho_i + \lambda \sum_j a_{ij}\rho_j(1 - \rho_i) \tag{5}$$

We will also use a Gillespie algorithm in order to simulate the stochastic process (we will summarize the algorithm in section VII).

A. Constant infection rate λ

We choose a representative λ to see how the epidemic evolves with time. As we can see in FIG. 7, in the long time regime the numerical solution is always higher than the stochastic simulation. It is noticeable that the bigger the λ the accuracy of the analytical approach. We can also see the rate of infected nodes have a stationary value and it depends on the infection rate. So we plotted this stationary ρ value versus λ we obtain the FIG. 8. In fact it depends on the quotient λ/δ.

We can see how the numerical solution does not fit exactly the Gillespie algorithm due to the fact that the algorithm represents the stochastic behavior. This means it adds the effect of noise to the system. But we can see a threshold in both situations and a monotonous growth until reaching $\rho^{st} = 1$. This threshold represents the health regime where no matter the value of λ, the stationary value of infected nodes holds on 0. The other regime, the endemic, comes after the critical value of $\lambda_c \simeq 0.03$, seen in the graphic. Analytically we can estimate it [5] $\left(\frac{\lambda}{\delta}\right)_c = \frac{\langle k\rangle}{\langle k^2\rangle} \simeq 0.0375$. Both numerical and Gillespie simulation also coincide in this critical behavior. Due to the fluctuations, the noise, the Gillespie simulation can still remain at $\rho^{stat} = 0$ after the λ_c, in fact we are seeing this behavior here it stills at 0 until $lambda = 0.04$

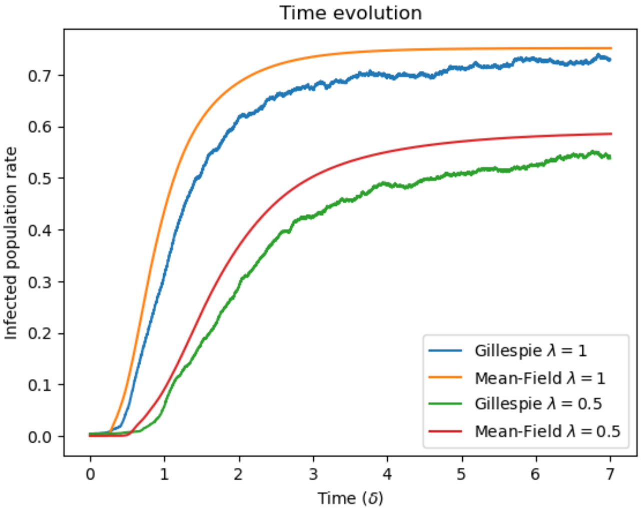

FIG. 7: Time evolution for the numerical integration and the Gillespie algorithm. We plotted them for two different values of lambda

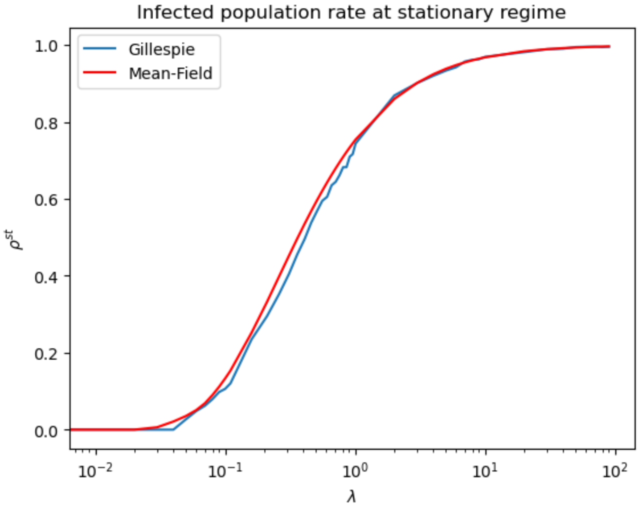

FIG. 8: Stationary value of infected nodes rate versus lambda for the numerical solution and the Gillespie algorithm.

B. Two different infection rates λ_1 and λ_2

In this case we are simulating a case where the λ decreases from 40 to 0.5. The whole dynamic is described as follows: the hackers become to steal accounts at a normal $\lambda = 40$, then someone develop a software capable of detecting this phishing. Now $\lambda = 0.5$. We see it in green at FIG. 9. First of all the green follows the blue. At a time = 2 δ The antivirus is released so the infected rate decreases exponentially until reaching the λ_2 stationary value.

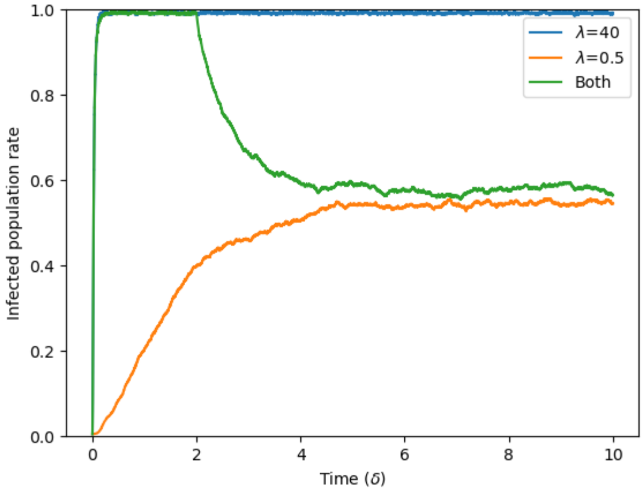

FIG. 9: Time evolution for the numerical integration and the Gillespie algorithm for a value of $\lambda = 40$, for the $\lambda = 0.5$ and a combination of both λ from 40 to 0.5 at a time = 2 δ.

In FIG. 10, despite the value of ρ^{st} is an asymptote and we expect different values depending on if we come from high or low values of ρ in time. We notice there is no an important differences for the one and two λ regimes in the Gillespie algorithm. This behavior can be explained thank to the noise that let the simulation to cross the asymptotic ρ^{st} value.

V. CORRELATIONS RESULTS

On the table I we have put the four nodes with higher degree and the four nodes with lower degree. In order to compare the degree with the rate, it seems to be a little correlation between this two variables. The bigger the degree the higher the Google rate.

VI. CONCLUSIONS

Studying the clusterization we see a social network-like structure, high clustering. This characteristic usually means Small-world property, which is a well knonw property of social networks.
With this tiny sample of comparisons in the table above we can see a low variability and high rates for higher degrees. For the lower degrees we obtain a very varied rates but the tendency is to be lower than the high degree nodes.

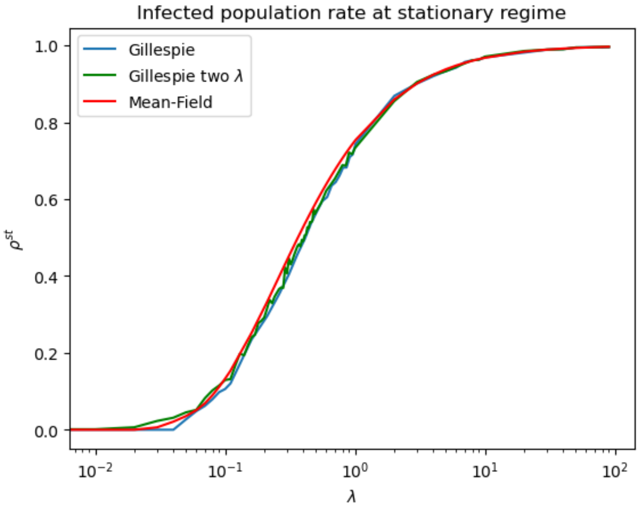

FIG. 10: Stationary value of infected nodes rate versus lambda for the numerical solution and the Gillespie algorithm for the normal time evolution and the Gillespie two λ means the simulation started with $lambda = 40$ then it changed to 0.5 at a time $=2\delta$.

Node	Degree	TV show name	Rate
3254	126	Queen of the South	9.2
2008	126	Home & Family	9.0
3525	108	So You Think You Can Dance	8.5
1177	104	Glee	8.8
2931	1	Adán y Eva	7.7
1794	1	True Detective	9.1
3763	1	Gomorron Sverige	5.8
2727	1	Deutschland-86	9.2

TABLE I: On this table we can se the node identification, the degree it has, the real TV show name and their rate on Google.

Eventually the epidemic simulation shows us that the mean field approximation describes accurately the system for this model, despite there are few differences on the time evolution, the infected population rate versus lambda behavior is well defined by the numerical model. The two λ regimes simulation has the expected behavior, we can conclude for this model that stationary value of ρ only depends on λ not on the path followed or the initial conditions, except, of course if there is a reasonable amount of nodes infected at $t = 0$.

VII. APPENDIX: GILLESPIE ALGORITHM

The Gillespie algorithm is a powerful tool to simulate Poisson processes occurring simultaneously on the same physical system. Events that occur with a certain probability with a time distribution:

$$\Psi(\tau) = \lambda_{event} e^{-\lambda_{event} * \tau} \tag{6}$$

τ is the time between two neighbor events, λ_{event} is the rate of event occurrence. In our case we have a two event process:

- Recovery. With a probability:

$$Prob\{recover\} = \frac{\delta N_I(t)}{N_I(t) + \lambda E_{act}(t)} \tag{7}$$

  The $\lambda_{recover} = \delta N_I(t)$, $N_I(t)$ are the number of infected nodes, $E_{act}(t)$ are the active edges, they both magnitudes change each time step. As we said before δ will be 1 in our simulations.

- Infection. With a probability:

$$Prob\{infection\} = \frac{\lambda E_{act}(t)}{N_I(t) + \lambda E_{act}(t)} \tag{8}$$

  The $\lambda_{infection} = \lambda E_{act}(t)$.

The $\lambda_{event} = \lambda_{recover} + \lambda_{infection}$. Each Gillespie step we will compute this probabilities then we will compare them to a random number generated uniformly. This will choose the process activated. The time simulation will evolve like $t = t + \tau$, τ generated randomly with a Poisson distribution (6).

Acknowledgments

I want to express my gratitude to my professors: Maria Ángeles Serrano and Marián Boguñá for their lectures on Complex Networks. They give us a solid and insightful knowledge on this topic.
I also want to thank my friends: Mireia Olives and Emma Oriol. Thank to discuss the doubts with them I have been able to accomplish this project.

[1] Napoli, P. M. (2014). Measuring media impact. The Norman Lear Center. http://www. learcenter. org/pdf/measuringmedia. pdf.

[2] Cheng, M. H., Wu, Y. C., & Chen, M. C. (2016). Television meets *Facebook*: The correlation between tv ratings and social media. American Journal of Industrial and Business Management, 6(03), 282.

[3] Ryan A. Rossi & Nesreen K. Ahmed. (2015). The Network Data Repository with Interactive Graph Analytics and Visualization. `https://networkrepository.com/fb-pages-tvshow.php`

[4] Zhou, S., and Mondragón, R. J. (2004). The rich-club phenomenon in the Internet topology. IEEE communications letters, 8(3), 180-182.

[5] Boguñá, M. (2023). *Notes on Complex Networks, the SIS epidemic Model.* Física de la matèria condensada, Universitat de Barcelona.