Structural properties and a simple dynamical evolution of the football players transfer market network*

Author: Ivan Casanovas Rodríguez[1, *]. Professors: M. Ángeles Serrano,[1, 2] Marián Boguñá [1, 2]
[1] *MSc in Physics of Complex Systems and Biophysics, Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.*
[2] *Departament de Física de la Matèria Condensada, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.*

Abstract: Football player transfers between clubs are analyzed as an economic network using a straightforward approach. The fundamental structural properties reveal the connectivity of this network and its similarity to other international trade ones. Additionally, the SIS model is introduced as a network dynamics tool that can rudimentary illustrate the financial health of clubs in the market.

I. INTRODUCTION

Football moves the world, or at least that's what people say. It is also reported to never talk about football at a family gathering and even less about politics. In this paper, we will delve into the intriguing intersection of these two topics by studying the network of football player transfers between clubs.

Historically, political and economic relationships between communities of all scales (continents, countries, or cities) have been characterized from a geo-commercial perspective. To explore how football can fit into this equation, we propose establishing trade relationships between clubs based on player transfers that have taken place. In such that way, we build an economic network that will be the subject of study from a network science perspective. This approach is powerful and useful to comprehend economic systems in which emergent properties of complex systems, such as heterogeneity or strong interactions, are present.

To construct the network, we begin by defining the nodes and links. Football clubs are considered agents, and a link is established when a player is transferred from one club to another. The most suitable network for this scenario is a directed, weighted network. Each connection indicates the direction of the transfer, depending on whether the player is bought or sold. Furthermore, the weight of this link can be represented by the transfer fee for that player.

Since a single club can buy multiple players from another club, multiple links with the same direction but different weights might exist. To consolidate these links into a single one, it is useful to consider a multiplex network or a combined weight that accounts for both variables: the number of players transferred between the two clubs and the total transfer fee (in one direction).

Although there may be many more variables to consider when characterizing the network, we will adopt the simplest possible approach: an undirected and unweighted network. This means that it doesn't matter whether a player is bought or sold (undirected), and the fee paid or the number of players transferred between clubs are irrelevant variables (unweighted). Therefore, we establish a single connection between clubs whenever at least one player is transferred. A reduction of complexity implies the loss of information, but the study will provide insight into the system's connectivity as well as a measure of the most important clubs within the transfer market.

The assignment is divided into two parts: the first part involves studying the structural properties of the network, while the second part focuses on implementing the Susceptible-Infected-Susceptible (SIS) dynamics to observe how the network evolves. On the one hand, properties such as degree distributions and higher-order correlations are computed to understand how clubs are interconnected. On the other hand, although the SIS model is commonly used to study the transmission of infectious diseases, we apply it to the transfer market to understand the financial health of the clubs. [4]

II. DATA AND METHODS

A. Data description

The original dataset is sourced from the public GitHub repository at `https://github.com/d2ski/football-transfers-data`. This dataset contains essential information for scraping data from the Transfer-Markt website and preprocessing it. It covers transfers from the six major European football leagues from 2009 to 2021. The dataset includes detailed transfer information such as transfer fees, type of transfer (loan, free transfer, deal), season, or transfer window (summer, winter). Additionally, it provides player features like age, position, and nationality. However, we focus exclusively on the origin and destination clubs. It's important to note that our interest lies in the market relationships between clubs. Thus, we do not consider players who are free agents (free transfers).

*Electronic address: icasanro40@alumnes.ub.edu

B. Structural properties

1. Degree distributions

A network is composed of nodes connected by links. The edge list is a list containing The degree, k_i, of a node i is a local variable that quantifies the number of connections that the node has, indicating the count of its nearest neighbors. Therefore, the degree distribution, $P(k)$, refers to the probability of a node having k neighbors. For computational purposes, it is useful to define the complementary cumulative degree distribution, $P_c(k)$. This distribution expresses the probability of finding a node with a degree greater than K, denoted as $P(k \geq K)$.

In real networks, we usually find scale-free power law distributions, $P(k) \propto k^{-\gamma}$, such that the typical exponent lies between 2 and 3. These distributions are characterized by the presence of hubs, which are nodes with a significantly larger number of connections compared to other nodes in the network.

2. Average nearest neighbors degree

The next natural step in the characterization of the network is to measure two-point degree correlations. Calculating the probability of a link connecting two nodes with degrees k and k' would be very costly. Instead, we can reduce dimensions by working with the conditional probability $P(k'|k)$. In this context, the average nearest neighbors degree, $\overline{k}_{nn}(k)$, measures the probability that a node of degree k is connected to a node of degree k'. By grouping the nodes according to their degree, we can define this metric as follows:

$$\overline{k}_{nn}(k) = \sum_{k'} k' P(k|k') = \frac{1}{N_k} \sum_{i \in \Upsilon(k)} \frac{1}{k_i} \sum_j a_{ij} k_j \quad (1)$$

where N_k is the number of nodes with degree k, $\Upsilon(k)$ stands for the set of that nodes, and a_{ij} is the element of the adjacency matrix connecting node i with j. Notice that we are essentially assuming that all nodes with the same degree are statistically equivalent.

Taking into account that the detailed balance condition must be fulfilled [1], we re-scale the average nearest neighbors degree using the factor $\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle}$. This allows us to measure the probability of finding a link connected to at least one node of degree k and makes the values of $\overline{k}_{nn}(k)$ fluctuate around 1.

3. Clustering

Similarly to the previous case, we can also provide the three-point degree correlations. Clustering measures the probability that a node of degree k is simultaneously connected to two nodes of degree k' and k''. To do this, we count the number of triangles T_i that can be formed with each node i and then normalize it by the maximum number of triangles that can be formed $\left(\frac{k_i(k_i-1)}{2}\right)$ [2]. Again, if grouping the nodes according to their degree:

$$\overline{c}(k) = \frac{1}{N_k} \sum_{i \in \Upsilon(k)} c_i = \frac{2}{k(k-1)N_k} \sum_{i \in \Upsilon(k)} T_i \quad (2)$$

4. Equilibrium models

Equilibrium models play a fundamental role in assessing the significant quantities of computed properties, as they can be used as base-line models. The concept involves randomizing the original network to disrupt correlations while preserving certain properties. For instance, in neuroscience, surrogate models hold particular significance. Two primary methods are used to uncouple the original network: random rewiring (RW) and the configuration model (CM).

In the RW process, two distinct links are selected randomly, and their connections are swapped if possible, repeating this procedure a number of times equal to the number of edges (E) in the network or more. This method conserves the degree distribution and just randomize connections. We anticipate that $\overline{k}_{nn}(k)$ and $\overline{c}(k)$ will decrease as the number of rewirings increases, eventually stabilizing at lower values.

The CM generates networks that are maximally random from a predetermined sequence of degrees $P(k)$. First, a stubs (half-links) list is created with a length of $2E$, where each node appears as many times as its degree. In this list, two stubs are randomly selected, and if possible, they are connected. The process continues until all connections are established.

When implementing both algorithms, it is crucial to avoid self-loops (a node connecting to itself) and duplicate connections (setting a connection that already exists).

C. SIS dynamics

1. The model

The Susceptible-Infected-Susceptible (SIS) model is a representation of the population dynamics used in epidemiology to understand the spread of infectious diseases. In this model, individuals can either adopt two states: susceptible, for the ones that can contract the disease; or infected, for the ones that have contracted the disease and can transmit it to susceptibles. Therefore, the dichotomous variable describing the state of each node i at time t is given by:

$$n_i(t) = \begin{cases} 1\ , & \text{if node } i \text{ is infected at time } t \\ 0\ , & \text{if node } i \text{ is not infected at time } t \end{cases} \quad (3)$$

It is assumed that the population size N (number of nodes) is fixed, so that there are not births and deaths.

The dynamics of the SIS model can be described using a stochastic differential equation:

$$n_i(t+dt) = n_i(t)\xi(dt) + (1-n_i(t))\eta_i(dt) \tag{4}$$

where $\xi(dt)$ and $\eta_i(dt)$ are stochastic variables that determinate the state of the node i after dt. Supposing that the infection and recovery process are both Poisson processes, we can define an infection rate λ and a recovery rate δ. In this case,

$$\xi(dt) = \begin{cases} 1\ , \text{ with probability } 1-\delta dt & \text{(no recovery)} \\ 0\ , \text{ with probability } \delta dt & \text{(recovery)} \end{cases} \tag{5}$$

$$\eta_i(dt) = \begin{cases} 1\ , \text{ with probability } 1-\lambda dt v_i(t) & \text{(no infection)} \\ 0\ , \text{ with probability } \delta dt v_i(t) & \text{(infection)} \end{cases} \tag{6}$$

where $v_i(t) \equiv \sum_j a_i j n_j(t)$ is the number of infected neighbors of node i.

In numerical simulations,a relevant variable to monitor as a function of time is the empirical prevalence, ρ_{emp}, defined as:

$$\rho_{emp}(t) = \frac{1}{N}\sum_i n_i(t) = \frac{N_I(t)}{N} \tag{7}$$

where $N_I(t)$ is the number of infected nodes at time t. If considering the mean-field approximation and the fully-mixed hypothesis [3], the dynamics of $\rho_{emp}(t)$ leads to a Langevin equation:

$$\frac{d\rho_{emp}(t)}{dt} = -\delta\rho_{emp}(t) + \lambda\langle k\rangle\rho_{emp}(t)(1-\rho_{emp}(t)) +$$

$$+\sqrt{\frac{D(\rho_{emp}(t))}{N}}\xi(t) \tag{8}$$

where $\xi(t)$ is a stochastic variable.

By allowing the system to evolve for sufficient time, it reaches a state of equilibrium, and we can define the stationary prevalence as ρ_{st}. Since it is a purely stochastic process, it is important to perform multiple realizations and take measurements of the averages:

$$\langle\rho_{st}\rangle = \frac{1}{M}\sum_{j=1}^{M}\rho_{st,j} \tag{9}$$

where M is the number of realizations. This magnitude will be analyzed for different values of the ratio $\frac{\lambda}{\delta}$, with the expectation of identifying a phase transition that distinguishes between the epidemic and non-epidemic phases at the critical point $\left(\frac{\lambda}{\delta}\right)_c$. Since the process depends only on the ratio $\frac{\lambda}{\delta}$, we will consistently set $\delta = 1$ in the simulations.

2. *Gillespie algorithm*

The Gillespie algorithm is a widely used method for generating stochastic sequences in systems where multiple Poisson processes (non-Markovian) occur simultaneously. It is efficiently applied in the SIS model, which involves two types of events that can alter the network: infection (at rate λ) and recovery (at rate δ). At each time step, an infection or recovery event is chosen based on their respective probabilities:

$$\begin{cases} P(n=0,t+dt|n=1,t) = \frac{N_I(t)\delta}{N_I(t)\delta+E_A(t)\lambda} \\ P(n=1,t+dt|n=0,t) = \frac{E_A(t)\lambda}{N_I(t)\delta+E_A(t)\lambda} \end{cases} \tag{10}$$

letting $E_A(t)$ represent the number of active links at time t. These links are the edges that connect an infected node with a susceptible node.

If an infection event is chosen, an active link is randomly selected, and the susceptible node connected by this link is infected. Otherwise, a random infected node recovers. It is important to note that the list of infected nodes and active links changes dynamically at each time step. Therefore, both lists must be efficiently managed to minimize computational costs.

3. *Model application*

Under the assumption that the mere existence of the market is a sufficient condition for its development, we propose a SIS model application to the football clubs by having the transfer market connections.

Some clubs might be unable to make their historically expected moves due to various circumstances, such as the economic impacts of a pandemic or failing to achieve financially rewarding football objectives. A clear example is FC Barcelona, which currently cannot participate in the transfer market as it once did. Such clubs, with restricted market actions, can be considered "infected."

Due to the self-regulating nature of the market, new clubs may also become infected if they rely on transactions with already infected clubs to stay afloat. Conversely, if a financially stable club enters the market to buy players, it injects money into the market, potentially reaching some of the infected clubs. This financial influx can help these clubs recover.

III. RESULTS AND DISCUSSION

A. Structural properties

1. *Network description*

The economic network consists of $N = 2866$ clubs that have transferred at least $E = 17329$ players among themselves over more than a decade. Although a few clubs

belong to the six major European leagues (about 20 per league), the high number of nodes involved suggests two things:

- It is possible that the more powerful clubs have good scouting systems and occasionally sign players from small, unknown clubs. Consequently, these small clubs will appear infrequently in the transfer market equations and will have a low degree unless they are a consistent source of professional players. To determine this accurately, we would need to account for the directionality in the network.
- There may be significant heterogeneity in the promotions and relegations within the major leagues, which would considerably increase the number of nodes.

Considering these two variables, we can ensure that the average number of players transferred between clubs is $\langle k \rangle = \frac{2E}{N} = 12$. This means that, on average, each team has conducted one buying or selling operation per year. However, this macroscopic measure does not capture the reality with sufficient precision. Therefore, we will next study the results of the degree distribution.

2. *Degree distributions*

The direct degree distribution shown in FIG.1a is a scale-free power law distribution with exponent $\gamma = 1.53$. This means that it is not as heterogeneous as the typical real networks which exponents are between 2 and 3. Similarly, the complementary cumulative degree distribution in FIG.1b indicates the presence of hubs, which are those minority clubs involved in a significant number of transfers, generally more than $k_{hubs} = 100$. Therefore, the clubs with the highest degree, which may be the so-called "big teams" (except PSG), predominantly dominate the transfer market network and can afford to negotiate for almost any player.

3. *Correlations*

The two-point and three-point degree correlations are represented in FIG.2 with the average nearest neighbors degree and the clustering coefficient. In both cases, nodes tend to connect to other nodes with different degrees: high-degree nodes are more likely to connect to low-degree nodes and vice versa. Therefore, it is a disassortative network.

This finding suggests that "big clubs" (nodes with higher degrees) do not interconnect among themselves, indicating that they do not form a rich club. It is reasonable when considering that transfers between giant clubs are usually few but of very high value. On the other hand, these clubs are precisely the ones that can significantly influence the market with smaller teams, as they continuously sell or loan out their secondary players to second-tier teams. At the same time, they can afford to buy a standout player since the cost is very low for them: they just can take the risk.

As seen in FIG.2, both the RW and CM models, averaged over 100 different realizations, yield identical results because they both begin with the original degree distribution of the network. Once again, in these cases, the curves $\overline{k}_{nn}(k)$ and $\overline{c}(k)$ flatten out, indicating that in a randomized model, nodes would be more uniformly connected, and the previously described disassortative property would not be valid.

Furthermore, FIG.2b also indicates that the degree distribution results in strong clustering, as all values exceed $\frac{1}{k-1}$, including the ones of the RW and CM. Consequently, we observe that triangles present in the network consolidate, leading to a more pronounced hierarchy of k-cores.

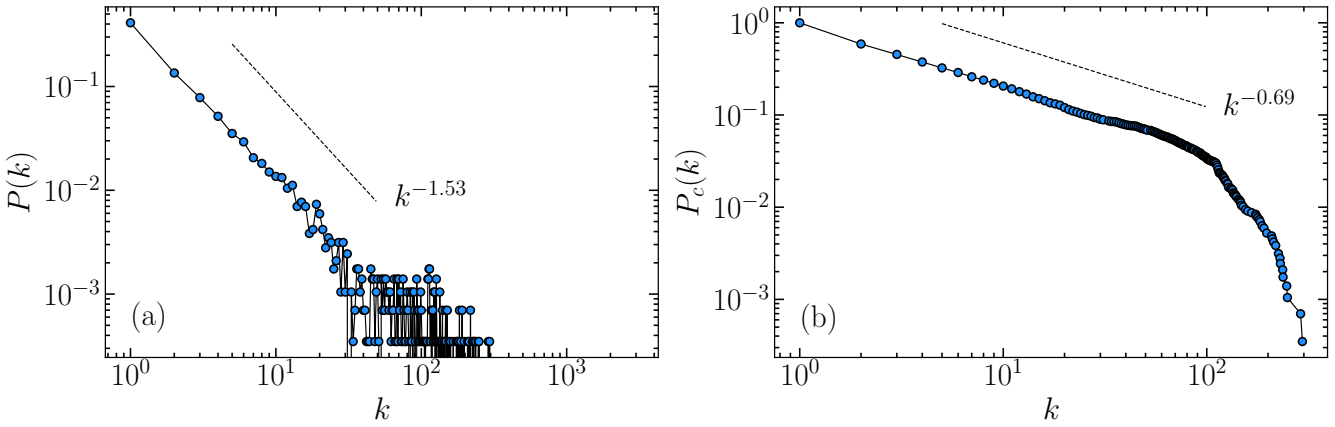

FIG. 1: Direct (a) and complementary cumulative (b) degree distributions, $P(k)$ & $P_c(k)$, of the football transfer market network. Power law fittings with the corresponding exponents are included in each distribution.

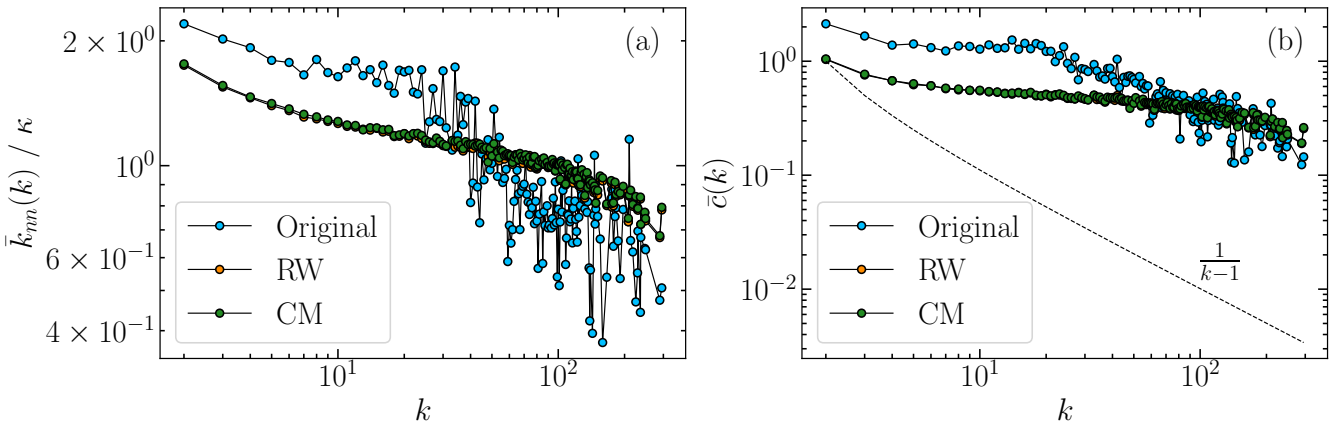

FIG. 2: Average nearest neighbors degree (a) and average clustering coefficient (b), $\overline{k}_{nn}(k)$ & $\overline{c}(k)$, of the football transfer market network. The computations for RW and CM networks . A straight line $\frac{1}{k-1}$ dividing strong and weak clustering

B. SIS dynamics

Given the stochastic nature of the process described, correctly evaluating the measures requires averaging over many realizations. In our case, we have performed at least $M = 100$ distinct realizations for each imposed initial conditions. Furthermore, it is essential to allow the system to achieve equilibrium. We have defined a unit of time as any event occurrence (infection or recovery), and ensured that the total timesteps T are at least as long as the number of nodes in the network ($T \geq N$).

1. *Empirical prevalence*

FIG.3a shows the evolution of the prevalence for single realizations at different ratios λ over δ. The first notable observation is the differentiation between two types of behaviors based on this ratio. On one hand, for low ratios, the prevalence relaxes to the steady state approaching zero, indicating no infected nodes in the system.

2. *Phase transition and critical point: Average stationary prevalence*

Theoretically, the phase transition occurs at the same point regardless of the initial infected nodes, but the curves are not the same for each case shown in Figure 3b. As we increase the initial number of infected nodes, it becomes more likely to reach global infection and more difficult to reach the state where no nodes are infected. This difference arises because we have allowed the same number of events to occur in all simulations with different values of $N_I(0)$. Given enough time, the initial condition would have less influence, and all curves would eventually converge. On the other hand, for high ratios, we observe the prevalence tends to grow as the ratio increases,

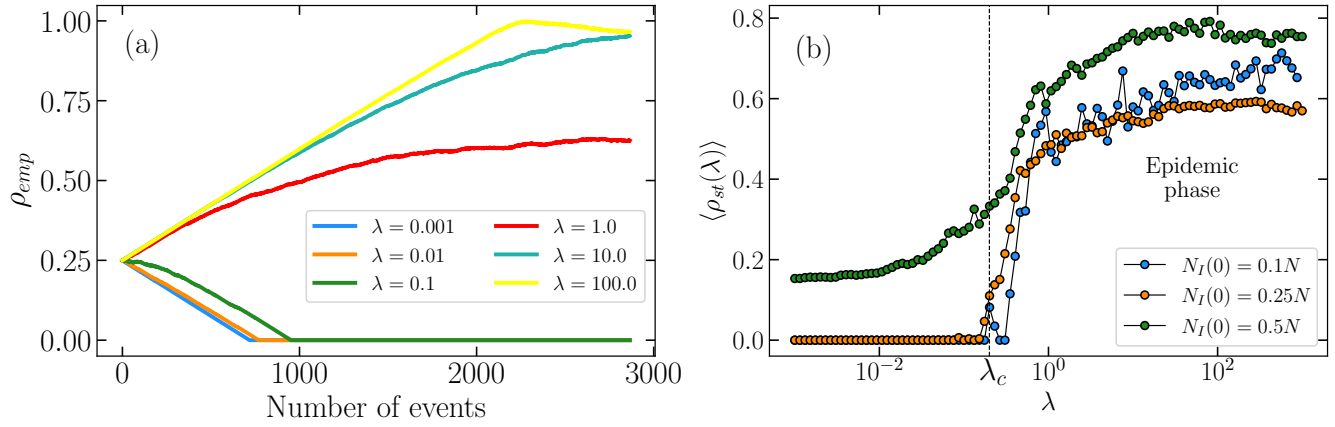

FIG. 3: (a) Empirical prevalence (ρ_{emp}) evolution for single realizations above and below the critical point. (b) Average steady-state prevalences as a function of the ratio $\left(\frac{\lambda}{\delta}\right)_c$ and for different initial number of infected nodes $N_I(0)$. The initial number of infected nodes is $N_I(0) = 0.25N$.

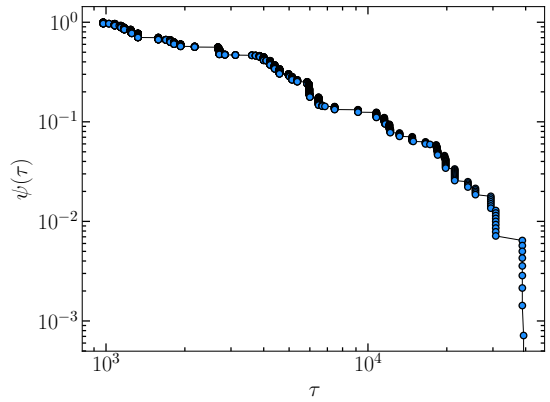

FIG. 4: Life-time distribution $\psi(\tau)$ for dynamics simulated near the critical point, taking values of infected rate between 0.1 and 1.

expecting that, for some values, all nodes will be infected. These two types of behavior are clearly differentiated in FIG.3b, which shows the average steady-state prevalence. In this case, there are distinct epidemiological and non-epidemiological phases separated by a critical point and hence a phase transition.

Based on the system correlations, we could determine the critical point $\left(\frac{\lambda}{\delta}\right)_c$ by diagonalizing the matrix and finding the largest eigenvalue. For simplicity, we observe that this value will certainly lie within the range of 0.1 to 1. What we can firmly assert is that the network exhibits structure and is not uncorrelated, as the critical point value is definitely $\left(\frac{\lambda}{\delta}\right)_c > \frac{\langle k\rangle}{\langle k^2\rangle}$.

3. Life-time distribution

With the initial condition of having just one quarter of the nodes infected ($N_I(0) = 0.25N$), we define the life-time τ as the time it takes for the system to eliminate the infection from the network, thereby restoring a healthy economic system. For different infection rates near the critical threshold (between 0.1 and 1), we simulate various scenarios and analyze the distribution of these times.

The results presented in FIG.4 reveal a power law distribution with a relatively short heavy tail. Although theoretically, at the critical point, one would expect a perfect power law, the observed decay is normal since we are near the critical point but not exactly at it. Furthermore, we are combining times from various λ values.

IV. CONCLUSIONS

In this short study, we have demonstrated how a very simple approach to studying the properties of a network can help us understand what is happening in a football players transfer network through logical reasoning. More complex studies are open for future work, such as incorporating directionality or weights in the network, or even studying the evolution over time. Additionally, we have simulated the dynamics of the Susceptible-Infected-Susceptible (SIS) model using the Gillespie algorithm and verified, through the definition of prevalence, that there is indeed a phase transition in the system. Although this model has its application in the network of player transfer markets in football, it would be interesting to propose other types of dynamics.

[1] M. Serrano, M. Boguñá, R. Pastor-Satorras, and A. Vespignani, *Correlations in Complex Networks* (2007), pp. 35–65, ISBN 978-981-270-664-5.

[2] M. Serrano and M. Boguñá, Physical Review E **74** (2006), ISSN 1550-2376, URL `http://dx.doi.org/10.1103/PhysRevE.74.056114`.

[3] M. Boguñá, C. Castellano, and R. Pastor-Satorras, Physical Review Letters **111** (2013), ISSN 1079-7114, URL `http://dx.doi.org/10.1103/PhysRevLett.111.068701`.

[4] Codes and results are available in an structured way at the GitHub public repository `https://github.com/ivancasanovaas/football-transfer-market-network.git`. Additional information about these codes and how to replicate results is given in the `README` file of the repository.