Math 2301: A Model for Ranking UEFA Clubs

Connor Fitch

Collaborator: Alex Burns
Advisor: Professor Thomas Pietraho

May 2020

"[Favorite football club] is the best! We just got unlucky this year!"

— *Every football fan at some point*

1 Introduction

1.1 Purpose and Data

Athletics, especially at the professional level, occupy a prominent role in the lives of many today. Football (read: soccer) is by most accounts the most popular sport on the planet, and countless hours a year are spent watching matches, keeping up with trade rumors, and unsuccessfully emulating a professional player's signature move when in the park with friends. Perhaps chief among these activities is arguing that your team is better than the rest. Because of course they are! The aim of this paper is to use methods from linear algebra and graph theory to produce a model for both ranking football clubs and for predicting the outcomes of matches. There will also be considerable references to probability and real analysis.

One of the largest annual football competitions is the men's UEFA Champions League which pits the top clubs from every European nation's highest league of professional football against each other in the hopes of claiming continental supremacy. For reasons to be explored later, the inter-league nature of this competition is crucial for creating a ranking system that spans the much of Europe. Intuitively, this should make sense; imagine there are two different leagues with teams that only play against teams in their own league. Without any games between the two leagues, for all we know, the very best team from one league may still be worse than the worst of the other! What exactly is meant by "best" and "worst" shall also remain vague for the time being, but if all is right in the world, one expects the "better" team to win more often than not.

Since the current season of the UEFA Champions League is incomplete, we shall use the data from the 2018-2019 season instead, taking the matches from the group stage and beyond. In addition to these UEFA matches, the data set also includes the matches from the regular seasons of each of the national leagues with clubs that made it to the group stage. In total, there are 254 clubs from 15 countries represented in the 4,381 games that form the data set. The data were gathered from `fbref.com` and `soccerway.com`.

2 Graphs and Markov Chains

2.1 A Brief Introduction to Graph Theory

The general idea for this ranking comes from graph theory, so we must familiarize ourselves with the basics thereof.

Definition: A **graph** is a collection of vertices and edges which connect these vertices. A **directed graph** is one for which the edges have a sense of direction; that is, an edge connecting the vertices v_1 and v_2 will either be from v_1 to v_2 or vice-versa, but not both.

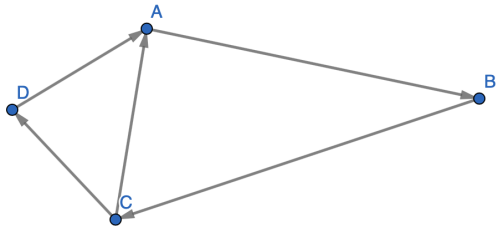


Figure 1: A simple directed graph where arrows indication directionality

A common interpretation of graphs is that the edges form the "roads" that connect the vertices together. Under this interpretation, the edges on directed graphs are like one-way streets. Note that even in the case of a directed graph, there is still a way to go back and forth between two vertices as long as there are two edges of opposite direction that connect the vertices. But what happens if there are no edges for a vertex? Can you ensure that you can go from a given vertex to another?

Definition: A graph is said to be **irreducible** if for any two vertices, there exists a path from one to the other and vice-versa.

It should not be too hard to see that the graph in Figure 1 is irreducible. To further illustrate this property, some examples of graphs that are not irreducible are provided below. The second one, in particular, is useful to examine.

Figure 2: Graphs that are not irreducible

Returning to our interpretation of graphs as road networks, we can introduce another property of a graph using intuition from roads. Not all roads are created equal; an interstate highway is going to have far more people traveling along it than a country back road. Likewise, there is a similar property for edges of graphs.

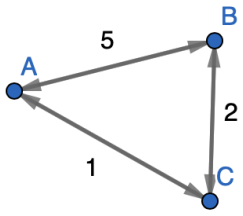
Definition: The **weight** of an edge is a non-negative value associated with a particular edge that, generally speaking, indicates some quantity or probability of going from one vertex to another.

Although far from comprehensive, this should give enough of a general background in graph theory that the rest of the paper shall be approachable.

2.2 Representing Graphs with Matrices

Since our methods for producing a ranking will rely on linear algebra too, it would be useful if there were a way to bring these concepts from graph theory into a linear algebra setting. Luckily there is!

Definition: An **adjacency matrix** is a matrix representation of a graph where the entry in row i and column j is 1 if there exists an edge from the jth vertex to the ith vertex and a 0 otherwise. An adjacency matrix may be **weighted**, meaning that instead of a 1 as the entry, it is now the weight of the associated edge.



$$\begin{bmatrix} 0 & 5 & 1 \\ 5 & 0 & 2 \\ 1 & 2 & 0 \end{bmatrix}$$

Figure 3: A graph with its associated weighted adjacency matrix

Note that in the figure above, the weights for the directed edges are symmetric, but they need not necessarily be that way! For instance, one could give the edge from A to B a weight of 4 but the edge from B to A a weight of 5. Another quick observation is that the adjacency matrix for an undirected graph is always symmetric since the edges are like two-way streets. Further, for a directed graph, if there are multiple edges of the same direction connecting two vertices, then those edges can be represented as a single edge with weight equal to the sum of the individual weights. Lastly, observe that an edge with a weight of zero is identical to not having an edge there altogether!

Furthermore, we can understand other properties of graphs by looking at their associated matrix. For instance, we can determine whether a graph is irreducible if for all i, j there exists an $n \in \mathbb{N}$ such that $(A^n)_{ij} > 0$, where $(A^n)_{ij}$ is the i, jth entry in the adjacency matrix raised to the nth power. Note that if all entries in an adjacency matrix are positive, then it is trivially irreducible.

2.3 Introduction to Markov Chains and Markov Matrices

A common and useful mathematical subject related to the study of linear algebra and probability is that of Markov chains.

Definition: A **Markov chain** is a sequence of random variables $\{X_t\}$ that satisfies the following condition:

$$P(X_t = x_t \,|\, X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots, X_1 = x_1, X_0 = x_0) = P(X_t = x_t \,|\, X_{t-1} = x_{t-1})$$

where P represents a probability measure and $P(A|B)$ means the probability of A given B.

The necessary condition for a Markov chain can be seen as the chain having only "short term memory." That is, the probability of getting a certain value in the next iteration only depends on the value of the current iteration. All other prior values are irrelevant. When studying Markov chains, a common question concerns the probability of having a particular value at some time.

Definition: The set of all potential values — often called states — that an iteration of a Markov chain may take is called the **state space** of a Markov chain.

For Markov chains with finite or countably infinite state spaces, the probability distribution over the state space is often stored in a column vector with the value in the ith entry corresponding to the probability of being in the ith state at the current iteration. This vector is referred to as a probability vector and has some notable properties.

Definition: A **probability vector** is a vector with non-negative entries where the sum of the entries is 1.

This requirement comes from the Law of Total Probability which states that for a given probability space Ω, we get $\sum_{\omega \in \Omega} P(\omega) = 1$ (or, for uncountably infinite state spaces, $\int_\Omega P(\omega)\, d\omega$). Even if this is a first introduction to probability, this should make some intuitive sense; after all, if there are k things that can happen, and we are 100% sure that *something* has to happen, then if we add up the percent chances of each of the k things happening, it must come out to 100%.

With some intuition for the probability of being at different states of a Markov chain in place, we can now discuss the probability of going from one state to another in the next iteration.

Definition: A **Markov matrix** is a matrix of non-negative values for which the entries of each column sum to 1. Equivalently, we say the matrix A is a Markov matrix if

$$\sum_i a_{ij} = 1 \quad \forall j \text{ where } a_{ij} \text{ is the } ij\text{th entry of } A$$

Having established what a Markov matrix is, we should also address how to interpret the values within one. The ijth entry of A, written as a_{ij}, represents the probability of going from state j to state i. Also as a note, Markov matrices are sometimes referred to as stochastic or left-stochastic.

Thus, if we start with some initial distribution over the state space written as a probability vector v_0, we can get the probability vector associated with the nth iteration of the Markov chain using the following formula:

$$v_n = A^n v_0$$

2.4 Graphs for Representing Markov Chains

Often graphs are used to represent Markov chains. Say we have a Markov chain with k different states in its state space. Then we can represent this Markov chain using a graph with k vertices with each corresponding to an element of the state space.

Furthermore, the weights of the directed edges represent the probabilities of transitioning from one state to the other. Thus, the weighted adjacency matrix for the graph is also the Markov matrix for our chain, since a_ij is both the weight of the edge from vertex j to vertex i and the probability of transitioning from state j to state i in the next iteration of the Markov chain.

This lets us introduce the concept of irreducibility from graph theory to our notion of Markov chains. A Markov matrix A is said to be irreducible if for all i, j there exists an $n \in \mathbb{N}$ such that $(A^n)_{ij} > 0$. In other words, if the chain starts at any given state, it can eventually reach any other state.

2.5 Properties of Markov Matrices and Probability Vectors

Markov matrices, together with probability vectors, have some useful properties that shall now be proven.

Property 1
Claim: If A is a Markov matrix and v is a probability vector, then $w = Av$ is also a probability vector.

Proof. Before beginning this proof, some notation shall be introduced. Let $a_{i\cdot}$ be the ith row of A expressed as a row vector. Although it will not be used in this proof, it is worth noting that $a_{\cdot j}$ is the jth column of A expressed as a column vector.

Let us now begin the proof itself.

$$w = Av = \begin{bmatrix} a_{1\cdot}\, v \\ a_{2\cdot}\, v \\ \vdots \\ a_{n\cdot}\, v \end{bmatrix}$$

Let us now consider the sum of the entries of w.

$$\sum_i w_i = \sum_i a_{i\cdot}\, v$$

Notice that each term in the sum can be seen as the inner product — sometimes called dot product — of the two column vectors $(a_{i\cdot})^T$ and v. Thus

$$\begin{aligned} \sum_i w_i &= \sum_i \langle (a_{i\cdot})^T, v \rangle \\ &= \sum_i \sum_j a_{ij}\, v_j \\ &= \sum_j v_j \sum_i a_{ij} \\ &= \sum_j v_j (1) && \text{by } A \text{ a Markov matrix} \\ \boxed{\sum_i w_i = 1} && \text{by } v \text{ a probability vector} \end{aligned}$$

Thus, the sum of the entries of w is 1. Furthermore, since w is the result of the product of a non-negative matrix and a non-negative vector, it is also true that w is non-negative. Thus, w is a probability vector. □

Property 2
Claim: If A is a Markov matrix, then one of its eigenvalues is 1.

Proof. First we shall show that a matrix has the same eigenvalues as its transpose. Recall that the eigenvalues for a matrix are given by the values λ that solve

$$\det(A - \lambda I) = 0$$

where I is the identity matrix of the correct size. Observe that A and A^T share the same entries along their diagonal. Thus since $A - \lambda I$ and $A^T - \lambda I$ are only removing λ from the entries along the diagonal, we can say that $(A - \lambda I)^T = A^T - \lambda I$. This is useful because we know that the determinant of a matrix equals that of it's transpose. A quick verification of this can be done by picking the first column for the matrix in the determinant expansion and picking the first row for the transpose's determinant expansion. Clearly, these are the same, and if you continue to compute the determinant expansions in this manner, you will get the same ultimate result for both of them.

Thus, since $(A - \lambda I)^T = A^T - \lambda I$, we know that $\det(A - \lambda I) = \det(A^T - \lambda I)$, meaning that they will have the same solutions for λ when set equal to zero. This ultimately means that A and A^T will have the same eigenvalues.

Let's now consider the following product where a^T_{ij} represents the ijth entry of A^T:

$$A^T \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \sum_j a^T_{1j} \\ \sum_j a^T_{2j} \\ \vdots \\ \sum_j a^T_{nj} \end{bmatrix} = \begin{bmatrix} \sum_i a_{i1} \\ \sum_i a_{i2} \\ \vdots \\ \sum_i a_{in} \end{bmatrix}$$

However, since A is a Markov matrix, we know

$$\sum_i a_{ij} = 1 \quad \forall j$$

Thus,

$$A^T \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Therefore, the column vector consisting of all ones is an eigenvector for A^T with eigenvalue $\lambda = 1$. Since we demonstrated that A and A^T have the same eigenvalues, this means that A must also have an eigenvalue of $\lambda = 1$. Note that this does not necessarily imply that the column vector consisting of all ones is an eigenvector for A. □

Property 3
Claim: If A is a Markov matrix, then all other eigenvalues satisfy $|\lambda| < 1$.

Proof. Since A and A^T have the same eigenvalues, we will look at A^T to explore the eigenvalues. Considering "all other eigenvalues" means that we are excluding the eigenvalue associated with the eigenvector of all ones used in the proof of Property 2. Therefore, consider some eigenvalue λ such that

$$A^T v = \lambda v \quad \text{where } v \neq \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Let's explore the product $A^T v$ more closely.

$$\begin{aligned} A^T v &= \lambda v \\ \begin{bmatrix} a_{1\cdot}^T v \\ a_{2\cdot}^T v \\ \vdots \\ a_{n\cdot}^T v \end{bmatrix} &= \lambda v \\ \begin{bmatrix} \langle a_{\cdot 1}, v \rangle \\ \langle a_{\cdot 2}, v \rangle \\ \vdots \\ \langle a_{\cdot n}, v \rangle \end{bmatrix} &= \lambda v \\ \begin{bmatrix} \sum_i a_{i1} v_i \\ \sum_i a_{i2} v_i \\ \vdots \\ \sum_i a_{in} v_i \end{bmatrix} &= \lambda v \end{aligned}$$

Consider the entry of v with the largest absolute value. Say it is the kth entry. We can then write

$$\sum_i a_{ik} v_i = \lambda v_k \quad \text{or equivalently put} \quad \left| \sum_i a_{ik} v_i \right| = |\lambda v_k| = |\lambda| \, |v_k|$$

Since $v_k \geq v_i$ for all i, we can create an upper bound for the term on the left hand side. We are allowed to say that this upper bound is strictly greater because the only way that it could be equal would be if $v_k = v_i$ for all i, but at that point we would be dealing with a scaled version of the eigenvector consisting of all ones

used in the proof of Property 2. As outlined at the start of the proof, that is strictly verboten since we are considering all *other* eigenvalues. Thus, we can write

$$\begin{aligned}\left|\sum_i a_{ik}v_i\right| &< \left|\sum_i a_{ik}v_k\right| \\ &< \left|v_k \sum_i a_{ik}\right| \\ &< |v_k|\left|\sum_i a_{ik}\right| \\ \left|\sum_i a_{ik}v_i\right| &< |v_k|\end{aligned}$$

We are able to write the above since A is a Markov matrix, so its columns sum to 1, meaning that the sum goes away in the product on the right hand side. Let us inspect the implications of this result.

$$\begin{aligned}|\lambda|\,|v_k| &= \left|\sum_i a_{ik}v_i\right| \\ |\lambda|\,|v_k| &< |v_k| \\ |\lambda| &< 1\end{aligned}$$

We have thus shown the claim. □

Property 4
Claim: If A is a diagonalizable Markov matrix, then the system approaches a steady state; that is, $\lim_{n\to\infty} A^n v_0 = \lim_{n\to\infty} v_n$ exists. Call this limit v_∞.

Proof. Notice that A can be seen as a function from V to V where V is the vector space associated with the column vectors of the appropriate length. Further, recall that A is diagonalizable if and only if V has a basis $\mathcal{B} = \{w_i\}$ with $Aw_i = \lambda_i w_i$. In other words, A is diagonalizable if and only if the eigenvectors of A form a basis for V. A proof of this is included in the Appendix since it veers off a bit from our current trajectory, but the big takeaway for the purpose of this proof is that the eigenvectors of A give a basis for V.

Suppose that $\{\lambda_i\}$ are the eigenvalues of A and $\{w_i\}$ is a basis of eigenvectors. Choose an ordering of the eigenvalues where λ_1 is the eigenvalue that is guaranteed to be 1. From before, we know that $\{w_i\}$ is a basis for V, so we can write v_0 as a linear combination of the basis elements. Let

$$v_0 = \sum_i c_i w_i$$

Then,

$$\begin{aligned}v_n &= A^n v_0 \\ &= A^n \sum_i c_i w_i \\ &= \sum_i c_i A^n w_i \\ v_n &= \sum_i c_i A^{n-1}(Aw_i)\end{aligned}$$

Notice that since w_i is an eigenvector for A, we can write $Aw_i = \lambda_i w_i$. Thus,

$$\begin{aligned} v_n &= \sum_i c_i A^{n-1} \lambda_i w_i \\ &= \sum_i c_i \lambda_i A^{n-2} (A w_i) \\ &= \sum_i c_i \lambda_i^2 A^{n-2} w_i \\ &= \quad \vdots \\ v_n &= \sum_i c_i \lambda_i^n w_i \end{aligned}$$

Let us now consider the limit as n approaches infinity.

$$\begin{aligned} \lim_{n\to\infty} v_n &= \lim_{n\to\infty} \sum_i c_i \lambda_i^n w_i \\ \lim_{n\to\infty} v_n &= \lim_{n\to\infty} c_1 \lambda_1^n w_1 + \sum_{i>1} c_i \lambda_i^n w_i \end{aligned}$$

Recall that $\lambda_1 = 1$ and $|\lambda_i| < 1$ for all $i > 1$ due to Property 3. This means that $\lim_{n\to\infty} \lambda_i^n = 0$ for $i > 1$. Thus,

$$\begin{aligned} \lim_{n\to\infty} v_n &= \lim_{n\to\infty} c_1 1^n w_1 + \lim_{n\to\infty} \sum_{i>1} c_i \lambda_i^n w_i \\ &= c_1 w_1 + \sum_{i>1} c_i \lim_{n\to\infty} \lambda_i^n w_i \\ &= c_1 w_1 + \sum_{i>1} c_i 0 w_i \\ \lim_{n\to\infty} v_n &= c_1 w_1 \end{aligned}$$

We have thus not only shown that $\lim_{n\to\infty} v_n$ exists but also that it takes the value $v_\infty = c_1 w_1$. □

3 Perron-Frobenius Theorem

3.1 Motivation

At the end of the previous section, we demonstrated that the following limit exists for a diagonalizable Markov matrix A.

$$\lim_{n\to\infty} A^n v_0 = \lim_{n\to\infty} v_n = v_\infty$$

Furthermore, we also found the value it takes on. Let $\{w_i\}$ be a basis of eigenvectors of A and w_1 is the eigenvector with eigenvalue of 1.

$$v_\infty = c_1 w_1 \quad \text{where } v_0 = \sum_i c_i w_i$$

However this value seems to be intrinsically related to our initial vector v_0. Perron-Frobenius Theorem, among other things, helps settle the question of the dependence on the initial distribution.

3.2 Statement and Proof of the Theorem

Theorem (Perron-Frobenius) Suppose A is a positive Markov matrix. Then the following are true:

(i) $\lambda = 1$ is an eigenvalue of A.

(ii) If $\lambda \neq 1$ is an eigenvalue of A, then $|\lambda| < 1$.

(iii) A has a unique probability eigenvector with eigenvalue of $\lambda = 1$.

Proof. Depending on one's perspective on things, a sigh of relief may be in order because results (i) and (ii) of the theorem have already been proven. They follow directly from Property 2 and Property 3, respectively, from the previous section. However, for those who were disappointed at that revelation, fret not, for result (iii) still requires our attention!

The proof relies on the Contraction Mapping Theorem from real analysis. A formal statement of the theorem will be provided, but a proof thereof is beyond the scope of this paper. However, for those interested, there are many excellent treatments of the theorem that make for leisurely beachside reading.

Definition: Suppose (X, ρ) is a metric space. A function $f : X \mapsto X$ is a **contraction mapping** if there exists $0 < \alpha < 1$ such that if $x, y \in X$ with $x \neq y$, then

$$\rho(f(x), f(y)) < \alpha\rho(x, y)$$

Theorem (Contraction Mapping) If $f : X \mapsto X$ is a contraction mapping and X is a complete metric space, then there exists a unique $x \in X$ such that $f(x) = x$. We call this point x a **fixed point**.

For those without some exposure to real analysis, the statements above may seem intimidating, but here are the big takeaways: (1) a function is a contraction mapping if applying the function to two distinct elements in the domain makes them "closer" together and (2) if a function is a contraction mapping, we can guarantee that there is a fixed point. For the purposes of Markov matrices and probability vectors, a vector v is a fixed point if it is an eigenvector with eigenvalue of 1 because then $Av = v$.

Let $X = \{x \in \mathbb{R}^n | \sum_i x_i = 1 \text{ and } x_i \geq 0\}$; essentially, X is the set of all probability vectors in $\mathbb{R}^n$. Further, it is worth noting that X is a complete metric space when endowed with the metric $\rho(v, w) = ||v - w||$ where $||\cdot||$ is some vector norm.

We will now show that A is a contraction mapping. As a simplifying assumption, let A be diagonalizable. Consider two vectors $v, w \in X$ with $v \neq w$. We will now show that $||Av - Aw|| \leq \alpha||v - w||$ for some $0 < \alpha < 1$.

$$||Av - Aw|| = ||A(v - w)|| \qquad \text{by } A \text{ a linear transformation}$$
$$||Av - Aw|| = ||Ax|| \qquad \text{letting } x = v - w$$

Since A is diagonalizable, we know that the set of eigenvectors of A, call it $\{u_i\}$, forms a basis of X (and therefore $\mathbb{R}^n$). Order the eigenvectors such that u_1 is the eigenvector with eigenvalue of 1. We are guaranteed its existence by result (i) of this theorem. Further, recall result (ii) of this theorem. We therefore know that all of the eigenvalues λ_i where $\lambda_i > 1$ satisfy $|\lambda_i| < 1$. We can then write:

$$x = \sum_i c_i u_i$$

Which means that we can say

$$\begin{aligned} Ax &= A\sum_i c_i u_i \\ &= \sum_i c_i A u_i \\ &= \sum_i c_i \lambda_i u_i && \text{by } Au_i = \lambda_i u_i \\ Ax &= c_1 u_1 + \sum_{i>1} c_i \lambda_i u_i && \text{by } \lambda_1 = 1 \end{aligned}$$

Consider the set $\{|\lambda_i|\}_{i=2}^n$. Since the set is finite, there exists a maximum entry. Further, since $|\lambda_i| < 1$ for all $i > 1$, we know that this maximum is less than 1. Call this maximum m and let $\alpha = 1 + m/2$. Thus if we take a norm of Ax, we know that

$$||Ax|| < \left|\left|c_1 u_1 + \sum_{i>1} c_i \frac{\alpha\lambda_i}{\lambda_i} u_i\right|\right|$$

We include the λ_i/λ_i term to preserve the sign of the eigenvalue. The intuition behind this statement is that we are essentially scaling the components of the vector x not in the direction of u_1 by a factor of $\alpha < 1$. This is a bit hand-wavy, but it should provide some intuition of why $||Av - Aw|| \leq \alpha||v - w||$ for some $0 < \alpha < 1$. From here, we can then say that A is a contraction mapping so there exists a unique $x \in X$ such that $Ax = x$, and it is the eigenvector for which $\lambda = 1$. We call this fixed point v_∞ and successive applications of A to a vector v_0 will approach it regardless of the choice of v_0. □

3.3 Primary Takeaway

The upshot of Perron-Frobenius theorem for the purposes of this paper is that a positive Markov matrix A has a unique stationary probability vector. This means that if you start with any probability vector v_0 and successively apply the matrix, it will approach that stationary vector. In other words, our stationary probability vector is v_∞! This means that a Markov chain governed by A eventually approaches the same stationary probability distribution over the state space, regardless of the initial probability distribution.

4 Application and Results

4.1 Overview of the Approach

The general approach is to treat each team in the data set as a vertex on a graph. The edges between teams will indicate the outcomes of the matches between the teams. For instance, if team x loses twice to team y and once to team z, then, generally speaking, the edge from x to y will have a higher weight than that of the edge from x to z. "Generally speaking" is used because some wrinkles to the weighting function shall be introduced later that, under certain circumstances, may mean that the previous generalization may not hold.

Thus, we can create a weighted adjacency matrix to represent this graph, and if we scale the columns properly, we can also get that the matrix is a Markov matrix. We shall call it M. The idea is to treat the teams as the state space of a Markov chain governed by M. Since the probability of transitioning from one team to another is a function of the outcomes of the matches between the two teams, it follows that, in the long run, the probability of being in a state associated with a team that performed well will be higher than that of a state associated with a team that performed poorly. Colloquially, the Markov chain will have higher probability of being at a team that won a lot of games.

One issue that arises is that we would like to guarantee that a long-run stationary distribution not only exists but is also unique. We shall therefore invoke Perron-Frobenius Theorem. The only complication is that all of the teams do not play each other, meaning that though M is non-negative, it is not necessarily positive since some entries may be 0. To resolve this, we can add a dampening factor $\varepsilon > 0$ to every entry of M before normalization. We are then free to use Perron-Frobenius.

Adding a dampening factor will guarantee that our matrix is positive and therefore irreducible. However, it is desirable to have the matrix as close to irreducible as possible before adding the dampening factor. This is because if there were no paths from one league to another, we would face the issue discussed in the introduction of not being able to distinguish the quality of teams across leagues. Even with the dampening factor added, it would be a glorified and complicated coin flip to see which league's best team can claim the absolute top spot.

Ultimately, we will take our dampened, normalized matrix M and compute the eigenvalues and eigenvectors thereof. By Perron-Frobenius Theorem, we know that the largest eigenvalue will be 1 and the eigenvector associated with it is the steady state vector v_∞ for M. The teams will then be ranked according to their corresponding entry in v_∞.

Clever readers may correctly notice that this approach is highly similar to the PageRank algorithm which was once used by Google for determining the order of search results. This is intentional. While it makes for fascinating reading, a proper discussion of the algorithm and the paper that introduced it would lead us far astray, and thus is better left for another time.

4.2 First Weighting Function

A general description of the weighting function was given above, but it will now be fleshed out in full detail. We define the weight of the edge from team j to team i to be w_{ij}. Before dampening and normalizing, we can write w_{ij} as follows:

$$w_{ij} = \#\text{ times } i \text{ beat } j$$

There is some consideration to be given for our choice of ε. It can be interpreted as some noise that we are adding to the data, which, besides the fact that it lets us apply Perron-Frobenius Theorem, is actually somewhat desirable since it will reduce the chances that our ranking is overfit to the data. However, too much noise is still a problem. Restricting ε to be less than 1 is reasonable because the noise factor should not have the same impact as losing a game. We can then pick a reasonable ε, say $\varepsilon = 0.3$ and produce the

following ranking:

Rank	Team	Rank	Team
1	Tottenham	11	Besiktas
2	Sint Truiden	12	Aves
3	Chaves	13	Feirense
4	Spartak Subotica	14	Manchester United
5	Boavista	15	Mariupol
6	Barcelona	16	Zorya Luhansk
7	Huesca	17	Manchester City
8	Eibar	18	De Graafschap
9	Tondela	19	Fortuna Sittard
10	Sporting CP	20	ADO Den Haag

This ranking is a bit troubling since our UEFA Champions League winners Liverpool are not in the top 20. Further, the second-ranked team Sint Truiden did not even qualify for the Champions League. We shall see if we can refine our methods.

4.3 Second Weighting Function

One thing to notice is that not all losses are created equally. If team x loses to team y with a final score of 0-7 and loses to team z with a final score of 0-1, it makes sense that the edge to team y should be given a greater weight. Furthermore, if team x loses to team y with a final score of 2-3, they should be given more "credit" in that loss than if they lost 0-3. This leads us to the conclusion that we should include a term for the goals allowed in the weighting function. For the un-normalized, un-dampened weight of the edge from team j to team i, we write:

$$w_{ij} = \big(\#\text{ times } i \text{ beat } j\big) + c\,\big(\text{avg. goals scored by } i \text{ against } j\big)$$

The value c is some factor that determines how highly we weigh the effect of goals. It makes sense that $c \geq 0$ since teams should not be punished in the rankings for scoring goals. Further, we will restrict $c \leq 1$ since a goal should not count more than a match victory. Choosing $\varepsilon = 0.3$ and $c = 0.65$ seems reasonable enough. That selection produces the following ranking:

Rank	Team	Rank	Team
1	Dynamo Kyiv	11	Manchester United
2	Belenenses	12	Liverpool
3	Leganes	13	Vozdovac
4	de Graafschap	14	Macva Sabac
5	Oostende	15	Karvina
6	Charleroi	16	Pribram
7	Toulouse	17	Hoffenheim
8	Marseille	18	Galatasaray
9	Sporting CP	19	St. Gallen
10	Santa Clara	20	Watford

This seems a bit better since we see more of the clubs that are widely regarded as the best among the top 20, including the overall champions Liverpool. However, we can still do better!

4.4 Metropolis-Hastings Weighting Function

The skeptics may have raised their eyebrows at the seemingly-arbitrary nature of simply "choosing" our values of ε and c. Such skepticism is not necessarily unfounded. Furthermore, a predictive model was promised in the introduction of the paper! We shall run an algorithm to give a distribution of the values for ε and c that work best for creating predictions for the outcomes of games. This algorithm is called the

Metropolis-Hastings algorithm and is a type of Monte-Carlo Markov chain. There is some degree of happy coincidence that we shall use Markov chains to optimize the result of our Markov chain!

The general idea for the predictive model is to give some probability of team i beating team j. For this, we can look to their corresponding entries in v_∞.

$$\text{probability of } i \text{ beating } j = \frac{(v_\infty)_i}{1.2\big((v_\infty)_i + (v_\infty)_j\big)}$$

The factor of 1.2 is included on the bottom to account for the fact that a match can end in a draw. We can then remove some of the data from our data set and try to predict the outcomes based off of v_∞ for our partial-data version of M. The Metropolis-Hastings algorithm will, roughly speaking, find the choices of c and ε that give the v_∞ that best predicts the outcomes of the games subject to some additional prior information. A complete discussion and proof of Metropolis-Hastings is beyond the scope of this paper, but it can be understood as a probabilistic way of exploring the sample space for our values of c and ε that is weighted to favor values of c and ε that give the most accurate predictions. A scatter plot of the distribution of accepted values of ε and c is given below.

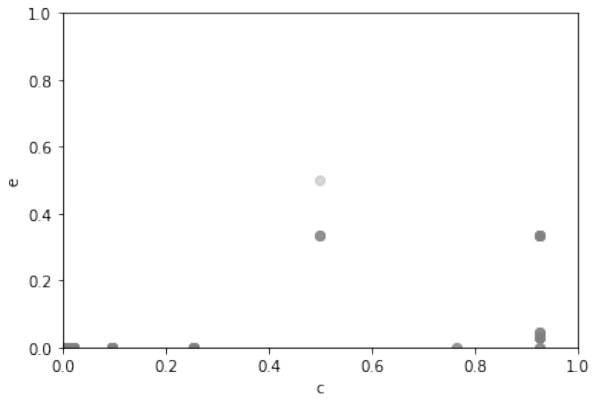


Figure 4: Scatter plot of accepted values of c and ε

From this, we can more thoroughly justify a choice of ε and c. Based off of the distribution, it seems like there are a few clusters to consider, namely around (0.003, 0.001) and (0.92, 0.0001) where the values are a tuple of the form (c, ε). Those values give the following rankings:

Rank	Team	Rank	Team
1	Tottenham	11	Heracles Almelo
2	Liverpool	12	Brighton
3	Amiens	13	Cardiff
4	Real Madrid	14	Montpellier
5	Atalanta	15	Dijon
6	Sassuolo	16	Barcelona
7	Nurnberg	17	Wolfsburg
8	Vorskla Poltava	18	Frankfurt
9	Desna Chernihiv	19	Dusseldorf
10	NAC Breda	20	Bremen

Table 1: Rankings with (0.003, 0.001) as parameters

Rank	Team	Rank	Team
1	Schalke	11	Genk
2	Club Brugge	12	Fiorentina
3	Barcelona	13	Sevilla
4	Manchester United	14	Rayo Vallecano
5	Manchester City	15	Galatasaray
6	Liverpool	16	CSKA Moskva
7	Huddersfield	17	Real Madrid
8	Chievo	18	Huesca
9	PSV	19	Eibar
10	Antwerp	20	Atletico

Table 2: Rankings with (0.92, 0.001) as parameters

These rankings, especially the first, seem pretty accurate with many of the top clubs and other perhaps-underrated clubs represented.

4.5 Discussion of Results

Although we seem to have ultimately produced a fairly reasonable ranking of the teams across much of Europe, there is still a lot of room for further exploration. First, we can include all of the games and nations represented in the UEFA Champions League which will give us more inter-league matches and better coverage of the continent. Additionally, we could include more years of matches and/or the matches from UEFA Europa League which features the top teams from each national league that barely missed the cut for the Champions League. We can also explore adding other variables to the weighting function. In short, this is a decent start on what may prove to be an interesting project.

5 References

- `fbref.com` for data
- `soccerway.com` for data
- `geogebra.com` for creating figures for graphs
- Math 2301: Intermediate Linear Algebra notes
- Math 2603: Introduction to Analysis notes
- Math 3606: Advanced Bayesian Statistics notes

6 Appendix

6.1 Proof of Diagonalization Fact

In Property 4 from Section 2.5, the following fact about diagonalizable linear transformations was stated:

Claim: A linear transformation $T : V \mapsto V$ is diagonalizable if and only if V has a basis $\mathcal{B} = \{v_i\}$ where each v_i is an eigenvector.

Proof.

($\Longrightarrow$) Suppose that T is diagonalizable and is given by the matrix A; that is suppose $A = T_{\mathcal{E},\mathcal{E}}$ where $\mathcal{E}$ is the standard basis. Then we can express A as the product of a matrix, a diagonal matrix, and the inverse of the first matrix.

$$A = P\,D\,P^{-1}$$

Recall that a matrix P is invertible if and only if its columns are linearly independent. Now consider the product $A\,P_{\cdot j}$.

$$A\,P_{\cdot j} = P\,D\,P^{-1}\,P_{\cdot j}$$

Since $P^{-1}\,P$ produces the identity matrix I, we know that $P^{-1}\,P_{\cdot j}$ is a column vector of 0s except for a 1 in the jth entry. In other words this is the jth column of the identity matrix. all this $I_{\cdot j}$. Thus,

$$A\,P_{\cdot j} = P\,D\,I_{\cdot j}$$

Now notice that the product $D\,I_{\cdot j}$ is really just the jth column of D which is really just $I_{\cdot j}$ scaled by d_j, the jth entry along the diagonal of D. Thus,

$$A\,P_{\cdot j} = P\,d_j\,I_{\cdot j} = d_j\,P\,I_{\cdot j} = d_j\,P_{\cdot j}$$

From this it is clear that the columns of P are eigenvectors of A with eigenvalues corresponding to the entries along the diagonal, and further, since they are n linearly-independent vectors and $\dim V = n$, we can say that they form a basis for V.

($\Longleftarrow$) Suppose that V has a basis $\mathcal{B} = \{v_i\}$ where each v_i is an eigenvector. Since v_i is an eigenvector of T we know that $Tv_i = \lambda_i v_i$. Thus if we express T as a matrix with respect to the basis $\mathcal{B}$, we get

$$T_{\mathcal{B},\mathcal{B}} = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

Suppose T is given by the matrix A; that is suppose $A = T_{\mathcal{E},\mathcal{E}}$ where $\mathcal{E}$ is the standard basis. Then, if we let $I_{\mathcal{B},\mathcal{B}'}$ represent the change of basis matrix from $\mathcal{B}$ to $\mathcal{B}'$, we can write

$$A = T_{\mathcal{E},\mathcal{E}} = I_{\mathcal{B},\mathcal{E}}\,T_{\mathcal{B},\mathcal{B}}\,I_{\mathcal{E},\mathcal{B}}$$

Notice that $I_{\mathcal{B},\mathcal{E}} = (I_{\mathcal{E},\mathcal{B}})^{-1}$ and $T_{\mathcal{B},\mathcal{B}}$, so we have written A in terms of the product of a matrix, a diagonal matrix, and the inverse of the first matrix. Thus A is diagonalizable.

Having proven both directions of the if and only if statement, the claim is proven. □