Who is the Most Important Playermaker on a Team? An Application of PageRank to Soccer

Rohan Sanda
Stanford University - Math 104
rsanda@stanford.edu

ABSTRACT

Spanish soccer is known for its emphasis on passing. Rather than play direct-soccer – as is more customary in the English Premier League – La Liga teams often rely on maintaining possession through indirect play to tire an opponent and create goal-scoring opportunities. In this paper, I seek to determine who the most important passers (ie. playmakers) are on selection of La Liga teams using the PageRank Algorithm. The Github code can be found by clicking here.

1 INTRODUCTION

Context and Motivation. Passing is one of the most important aspects of soccer. A team's choice of how to move the ball around the field can be consequential in creating game-winning opportunities. The question we seek to answer is: who are the most crucial passers for a given team?

Proposed Approach. We focus our analysis on the first-division league in Spain, known as La Liga. Given Spanish soccer's emphasis on passing, we will seek to understand who are the most important passers on a given La Liga team. To do so, we will construct a directed graph network of passes between players on a team where an edge from player 1 to 2 represents the number of passes player 1 made to player 2 over a given period of time. Each player is represented by a node. The novel aspect of the paper is to then apply the well-known graph ranking algorithm – PageRank – to this passing network to rank players. By doing so, we can identify the most crucial nodes – or players –in the passing network.

In addition to the standard PageRank algorithm, we implement our own version of PageRank to weight different passes based on their quality - which is a qualitative metric included in our dataset. Doing so allows us to factor in other information to our passing network that may better inform "player importance" - or "playmaking ability." Finally, we compare these rankings to professional player rankings to determine whether the most strategic passers are actually considered the best players.

2 DATASET

We use the open-source dataset released by Pappalardo et al. in their 2019 *Nature* containing a collection of data from the 2017/2018 season of five national soccer competitions in Europe. Specifically, we use data related to the Spanish first division league - called La Liga. This dataset contains 380 matches that have been spatio-temporally tagged by event. In other words, every event in the game – such as passes, shots, penalties, and goals – are logged and stored in a large file. In total, the Spanish Events dataset contains 628,659 events and tracks 619 players. In our analysis, we analyze the following teams: Barcelona and Athletic Club (aka Athletic Bilbao). Barcelona won La Liga that year and Bilbao finished 16th out of 20 teams.

Math 104: Applied Matrix Theory,
2022.

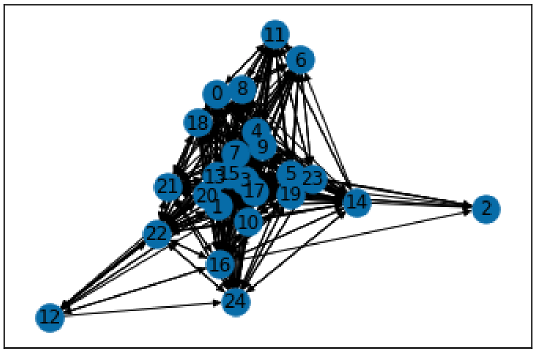

Figure 1: Example of a passing network for Barcelona. Edges represent passes between players (nodes).

3 MODEL

Theory. The PageRank algorithm was invented by Sergey Brin and Larry Page. It can be defined in the context of perturbing an adjacency matrix as follows:

$$W' = \alpha W + (1 - \alpha)\frac{J_n}{n}$$

where W is a sparse Markov matrix, α is the damping factor, J_n is an nxn all-one matrix. By construction, the columns of the adjacency matrix W have been normalized (sum to 1). After perturbing W according to the above formula, the resulting W' is still Markov.

The unique ranking vector $\mathbf{r}$ (which is also a probability vector) is the eigenvector of the 1-eignspace of W'. Since we have perturbed W, the entries of W' are all positive. Therefore, by the Perron-Frobenius Theorem, we know W' has 1) 1 is an eigenvalue and 2) 1 is the largest eigenvalue and all other eigenvalues have absolute value smaller than 1, and so the following holds:

$$\lim_{k\to\infty} W'^k = \begin{pmatrix} \mathbf{r} & \dots & \mathbf{r} \end{pmatrix} \text{ and } W'\mathbf{r} = \mathbf{r}$$

Using this information, we can compute $\mathbf{r}$ iteratively by multiplying in the following chain:

$$\mathbf{r_0} \to \mathbf{r_1} = W'\mathbf{r_0} \to \mathbf{r_2} = W'\mathbf{r_1} \to \dots \to \mathbf{r}$$

where $\mathbf{r_i}$ is a probability vector for the i-th iteration. By the theorem above, this process will eventually converge to the equilibrium vector $\mathbf{r}$ for large i.

Implementation: Standard PageRank. After performing data cleaning and processing, we construct the adjacency matrix W as follows. First, we filter through all events for a given team and isolate successful pass-events. Since events do not record the player to whom a pass was passed, we define a successful-pass-event as follows: the event that immediately succeeds the pass-event in question corresponds to 1) the same team and 2) a different player who has control of the ball. We then construct an n x n matrix (where n is the number of players on the team - including substitutes), where the w_{ij} entry corresponds to the total number of passes player-j passed to player-i throughout the 2017/18 season. We then normalize each column and apply the PageRank algorithm to W for 100 iterations with a damping factor of 0.85 – a standard factor.

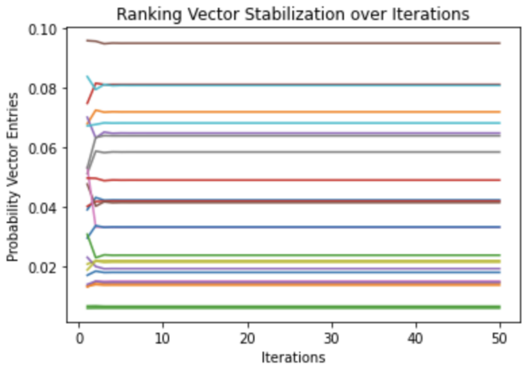

Figure 2: A graph showing the rapid stabilization of the pagerank vector r when running the standard Page Rank algorithm on the Barcelona data.

Implementation: Strategic Passing PageRank. After constructing the adjacency matrix in the same way as above, we modify the PageRank algorithm as follows to incorporate more information about passes and improve ranking recommendations. Specifically, while the standard PageRank method weights all passes between players the same, this modified Strategic Passing PageRank (SSP) weights passes differently based on the type of pass and its significance:

$$W' = \alpha(W + \Delta) + (1 - \alpha)\frac{J_n}{n}$$

$$\text{where } \delta_{ij} = \epsilon + \gamma$$

$$\text{and } \epsilon = \begin{cases} 5 & \text{if "Cross"} \\ 15 & \text{if "Smart Pass"} \end{cases}$$

$$\text{and } \gamma = \begin{cases} 15 & \text{if "Key Pass"} \\ 10 & \text{if "Opportunity"} \\ 5 & \text{if "Through"} \\ 30 & \text{if "Assist"} \end{cases}$$

Here, we define an additional n x n matrix Δ that has additional weights δ_{ij} that are added to the w_{ij}'s based on whether the pass-event is tagged as one of the descriptions listed above. These tags are a feature of the dataset. The standard iterative stationary-vector solving process is applied as before to yield updated recommendations. We believe that by weighting with these tags, we can get a better idea of how significant and strategic a player's passing is.

4 RESULTS AND DISCUSSION

Here, we present the results from our PageRank implementations for the two La Liga teams analyzed: Barcelona and Athletic Club. Each table contains an entry for each player (and their position) with their corresponding PageRank score, Stragetic Passing PageRank (hereby abbreviated as SSP) score, and player rating relative to their team according to the official La Liga WhoScored rankings (abbreviated by WS). The WhoScored rankings are a well-cited ranking system in the media. Each player's WhoScored ranking relative to the whole La Liga is shown in parentheses, next to their team-relative rating. While the exact algorithm used to compute a score considers over 200 statistics and is not publicly available, we will use these metrics as a rough point of comparison for our model results. WhoScored rankings are only computed for players who played more than the average number of minutes of all players in La Liga - in other words, only starting players are ranked. Note that the tables have entries that correspond to the ranking of the player relative to their the team for the given algorithm. We have ordered the table rows by the player's SSP score.

Player	PR	SSP	WS
Lionel Messi (F)	5	1	1 (1)
Ivan Rakitic (M)	1	2	8 (34)
Jordi Alba (D)	3	3	5 (18)
Andres Iniesta (M)	8	4	10 (87)
Luis Suarez (F)	11	5	2 (3)
Sergio Busquets (M)	2	6	4 (16)
Sergio Roberto (D)	6	7	6 (19)
Gerhard Pique (D)	4	8	7 (25)
Paulinho (M)	9	9	12 (99)
Sam Umtiti (D)	7	10	9 (49)
Philip Coutinho (M)	13	11	3 (4)
Nelson Semedo (D)	12	12	13 (162)
Ousmane Dembele (M)	17	13	-
Marc-Anders ter Stegen (G)	10	14	11 (89)
Lucas Digne (D)	15	15	-
Thomas Vermaelen (D)	14	16	-
Javier Mascherano (D)	16	17	-
Alexi Vidal (D)	22	18	-
Denis Suarez (M)	19	19	14 (189)
Paco Alcacer (F)	23	20	-
Gerard Deulofeu (F)	20	21	-
Andre Gomes (M)	18	22	-
Yerry Mina (D)	21	23	-
Jasper Cillessen (G)	24	24	-
Jose Araiz (M)	25	25	-

Table 1: Barcelona '17-18. A total of 22518 passes were recorded and analyzed.

Player	PR	SSP	WS
Inaki Williams (F)	12	1	-
Raul Garcia (M)	10	2	1 (36)
Inigo Lekue (D)	4	3	8 (192)
Unai Nunez (D)	1	4	2 (76)
Markel Susaeta (F)	6	5	4 (146)
Ander Iturraspe (M)	2	6	9 (197)
Mikel San Jose (M)	5	7	6 (157)
Oscar de Marcos (D)	9	8	5 (154)
Inigo Cordoba (M)	16	9	10 (211)
Aymeric Laporte (D)	3	10	-
Joseba Etxeberria (F)	8	11	11 (232)
Iker Muniain (F)	18	12	-
Artiz Aduriz (F)	19	13	8 (178)
Mikel Balenziaga (D)	13	14	-
Inigo Martinez (D)	15	15	-
Enric Saborit (D)	14	16	7 (163)
Kepa Arrizabalaga (G)	16	17	3 (135)
Mikel Rico (M)	15	18	-
Mikel Vesga (M)	20	19	-
Eneko Boveda (D)	21	21	-
Yeray Alvarez (D)	17	21	-
Iago Herrerin (G)	22	22	-
Xabier Etxeita (D)	23	23	-
Sabin Merino (F)	24	24	-
Ager Barrutia (M)	25	25	-
Andoni Lopez (D)	26	26	-

Table 2: Athletic Club (aka Bilbao) '17-18. A total of 14023 passes were recorded and analyzed.

Barcelona. Our SSP ranking results indicate that Lionel Messi is the most "important" player on Barcelona, when it comes to making strategic passes. In other words, Messi's node in the Barcelona passing network is the most "connected" after weighting passes for their strategic significance. Following Messi are other players known for their play-making ability, such as Ivan Rakitic and Andres Iniesta. In this regard, the SSP model seems to correctly rank more strategic players in the passing network higher up.

How do the PR results compare to the SSP results? From the Barcelona results, we observe that the PR and SSP results are roughly similar. The main difference is that the SSP results give a slight boost to attacking players (midfielders and forwards) who are more likely to make an assist or cross (two of the qualitative metrics of pass quality) which grant larger weights. Additionally, both the PR and SSP algorithms effectively separate players into frequent-starters and frequent-bench players - the latter having fewer passes and thus less-connected nodes in the passing network and consequently showing up in the bottom half of the ranking table (indicated by having a "-" in their WS column since they are not frequent starters). The presence of several defenders being prominently ranked (like Jordi Alba and Sergio Roberto) is likely reflective of the *tiki-taka* style of soccer pioneered by Barcelona. This style of soccer prioritizes small, short passes and often builds up from the back of the field with defenders. For this reason, even the Barcelona starting goalkeeper, ter Stegen is also relatively highly ranked.

When we compare our rankings to the WhoScored results, we immediately observe that there is less similarity. While both models put Messi at the top, other areas of disagreement are likely caused by the WhoScored ratings' prioritization of factors other than passing. For instance, both the PR and SSP models ranked Ivan Rakitic very highly – Rakitic completed the most passes of any player in La Liga in the 2017-18 season. However the WhoScored results place Rakitic at 8th. It is possible that with weight-tuning, the SSP model could produce results that more closely align with the WS results. Therefore, while our model may give an idea of a player's strategic passing ability, it does not match the ranking of a widely-used player rating system. In other words, the who we determine is the best playmaker is not necessarily the best player, according to professional player rankings.

Athletic Club. Our SSP results indicate that Inaki Williams - a striker - is the most "important player" on Athletic Bilbao. The results for Athletic Club follow the same general trends that Barcelona follows. Notably, there is slightly less agreement between the PR and SSP algorithm rankings. For instance, Inaki Williams and Raul Garcia move up over 10 spots in the SSP rankings verus the PageRank rankings - this may indicate that the SSP algorithm weights attacking players too heavily. That being said, the SSP and WS models do perform relatively similar. With the major exception of Inaki Williams, the SSR and SSP rankings do no not differ by more than a few places in the upper half of the table. In the future, it may be interesting to compare the SSP rankings with the player cards from FIFA 2018 - a soccer video game that uses creates player profile ratings based on official FIFA data. These ratings have a passing metric that may align more closely with the SSP rankings.

5 CONCLUSION

In closing, this report has summarized our work on applying the PageRank algorithm to graphs representing passing networks for teams in La Liga. By representing passing networks for La Liga teams – where passing is more heavily emphasized than other leagues – we are able to answer questions about who the most important and strategic players are on a team both in terms of their connectedness to other players in the network, as well as the quality and intelligence of their passes. While both the PR and SSP models returned roughly similar rankings that seemed to align with expected results, these rankings did not match the WhoScored rankings - likely due to WhoScored's prioritization of different factors in player performance other than passing strategy. While in this report we considered the results from two teams - Barcelona and Athletic Club – our code processing and analysis pipeline can be applied to any of the teams in the dataset. Our code can be found here: Click here. Note that this repository does not contain the datafiles, these must be downloaded separately. The repository also contain the code (in the "watermark" folder) for our discussion of image watermarking discussed next.

6 A QUICK DISCUSSION OF IMAGE WATERMARKING

For my project, I focused on applying the PageRank algorithm to the novel domain of soccer passing networks. However, prior to settling on this idea, I also stumbled across the topic of image-watermarking - an image-processing technique whereby an image (known as the watermark image) can be inserted into another image (the host image) to create a new image (the watermarked image) that to the human eye looks identical to the host image but has the watermark image embedded in it. This is a useful form of digital authentication - especially for digital media such as audio, video, or image files. While I did not end up continuing this project, I did play around with image-watermarking and include this work as an extra part of this extra credit report.

In the section, I briefly explain the theory behind an SVD-based watermarking scheme first proposed by Ruizhen Liu and Tieniu Tan in 2002 their IEEE paper titled "An SVD-based watermarking scheme for protecting rightful ownership." I then display some results from my own attempt to recreate the algorithm described in the paper.

Embedding and Extraction Algorithms. The singular-value decomposition of an n x m matrix R is given by $R = USV^T$ where $U \in \mathbb{R}^{nxn}$ and $V^T \in \mathbb{R}^{mxm}$ are orthogonal matrices and $S \in \mathbb{R}^{nxm}$ is a diagonal matrix whose diagonal entries correspond to the singular values of R. To apply watermarking, we use this concept in both algorithms to embed and extract a watermark W image into/from a host image A. While I have modified my code to work with color images, I describe the algorithm in terms of a single-channel grayscale image for simplicity. Additionally, we assume input images are the same size for simplicity. We adopt the notation from the Liu, Tan paper for continuity.

Embedding Algorithm:

- Compute SVD of A (host image): $A \Rightarrow USV^T$
- Dampen W (watermark) with scalar α then add to singular-value matrix of A. We use $\alpha = 0.1$. Then compute the SVD of this new intermediary matrix: $S + aW \Rightarrow U_W S_W V_W^T$.
- To get watermarked image (A_W), compute: $A_W = U S_W V^T$.

Extraction Algorithm: Let A_W be the matrix in outputted from the above process - it could be corrupted so we denote it as A_W^*. We can extract the watermark W as follows (there will be some watermark corruption so we denote the extracted W as W^*.

- Compute SVD of A_W^* (watermarked image): $A_W^* \Rightarrow U^* S_W^* V^{*T}$
- Using the U_W and V_W^T from the intermediary matrix from the embedding process (the true owner of the digital file will have this), compute D^*: $D^* \Leftarrow U_W S_W^* V_W^T$.
- Finally, recover the watermark by subtracting the following. Note that the true owner of the digital file will possess the original S: $W^* \Leftarrow \frac{1}{a}(D^* - S)$.

Results. Here is an example of my code in action. We use a QR code to Gene Kim's website and embed it inside of an image of Stanford. Notice that the peak signal to noise ratio is sort of close to 30 dB, a widely recognized cutoff for an acceptable watermark. Additionally, the 2D correlation is calculated to be quite high: 0.9845! My code also contains the ability to add a Haar wavelet transform to improve results - this addition is recommended by Li, Yap, and Lei in their 2011 IEEE paper titled "A new blind robust image watermarking scheme in SVD-DCT composite domain." The full code is available on my Github link and is all written originally. The QR codes actually work!

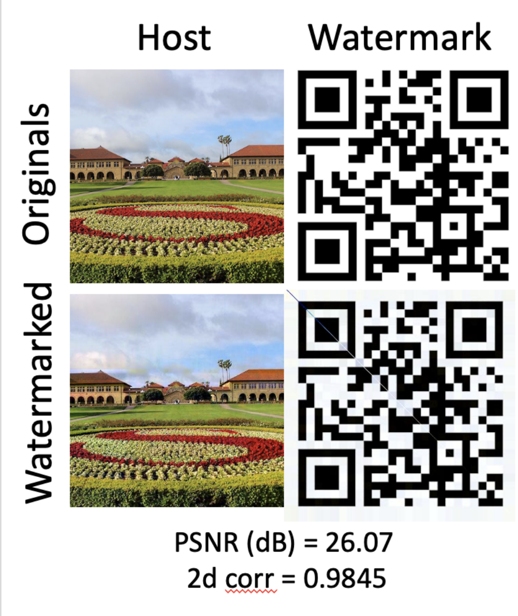

Figure 3: A demonstration of the SVD-based watermarking approach using the Li, Yan algorithm (no discrete wavelet transforms (DWT) are added. Notice the slight image corruption in the recovered watermark. You can view these images in full size if you visit my Github repository.

Multiple algorithms based on the basic SVD-watermarking scheme have since been proposed. I did not have the time to implement some of my own possible modification to this algorithm, but here is one idea: is it possible to improve the watermarked image to original image correlation by first applying principal-componént analysis to the RGB channels of the watermark image, then trim out pixels/colors that do not correlate well with the principal components of the host image, and then embed this modified image into the host?

NOTE: The code for this part of the project is in the same Github repo.

7 ACKNOWLEDGEMENTS

In addition to the papers that are mentioned in the text, I used content from Gene Kim's Math 104 Lecture Notes.

ALL CODE FOR THIS PROJECT CAN BE FOUND HERE.

Pappalardo, L., Cintia, P., Rossi, A. et al. A public data set of spatio-temporal match events in soccer competitions. Sci Data 6, 236 (2019). https://doi.org/10.1038/s41597-019-0247-7

Ruizhen Liu and Tieniu Tan, "An SVD-based watermarking scheme for protecting rightful ownership," in IEEE Transactions on Multimedia, vol. 4, no. 1, pp. 121-128, March 2002, doi: 10.1109/6046.985560.

Z. Li, K. -H. Yap and B. -Y. Lei, "A new blind robust image watermarking scheme in SVD-DCT composite domain," 2011 18th IEEE International Conference on Image Processing, 2011, pp. 2757-2760, doi: 10.1109/ICIP.2011.6116241.