

Rapport de Projet : Analyse des transferts des joueurs de Football

Data Mining

par VINCENT Yann et TANG Kévin

fait le 14 Novembre 2023

Université Claude Bernard



Lyon 1

1 Présentation des données

Nature des Données :

Les données que nous avons collectées pour cette analyse nous donnent un aperçu approfondi des transferts de joueurs entre différents clubs de football. Chaque enregistrement comprend des informations essentielles, notamment :

- Informations sur les Joueurs : Noms, positions, nationalités, âge, sa valeur marchande, etc.
- Détails des transferts : Dates des transferts, montants financiers impliqués, transfert simple ou prêt, etc.
- Contexte des Ligues : Informations sur les ligues de départ et d'arrivée, transfert sur le mercato d'été ou d'hiver.

Étendue Temporelle :

Les données que nous avons couvrent la période de 2009 à 2021, offrant ainsi une vision complète des tendances sur plusieurs saisons. Cette plage temporelle étendue nous permettra d'avoir un grand nombre de données à notre disposition, et nous permettra de potentiellement voir des évolutions dans le temps.

Échantillon Représentatif :

Afin de garantir une représentation diversifiée, la base de données répertorie plus de 70 000 transferts de joueurs. Cela couvre non seulement des transferts européens, mais également des joueurs internationaux provenant de ligues au-delà des cinq grands championnats. De plus, ces données rassemblent des mouvements de joueurs, non seulement des ligues professionnelles, mais aussi des ligues amateurs, assurant ainsi une variété de scénarios de transferts pour obtenir une représentation réaliste du marché et repérer des stratégies de recrutement réelles.

Source :

Les données ont été recueillies à partir de Transfermarkt, réputé pour leur fiabilité en matière d'actualités sportives et de transferts de joueurs.

Ces données riches et variées serviront à analyser les transferts dans le football mondial en passant par les coûts, leurs nombres, leurs évolutions et autres.

2 Préparation des données

La préparation des données joue un rôle essentiel dans notre exploration des transferts de joueurs de football. Cette étape est décomposée en 2 phases : un éventuel nettoyage des données et la transformation des données en fonction de nos besoins. Dans notre cas, nous souhaitons faire une analyse graphique. Nous allons donc créer des graphes dirigés à analyser ensuite sur Gephi.

Nettoyage des Données :

L'examen approfondi de nos données a révélé plusieurs anomalies nécessitant une attention particulière. Parmi celles-ci, des montants de transfert aberrants ou incorrects ont été identifiés, représentant seulement une vingtaine de joueurs. Afin de résoudre ces problèmes, nous avons consulté le site de Transfermarkt et conclu qu'il s'agissait d'erreurs de saisie de données. Ces transferts ont été supprimés pour garantir la fiabilité de notre analyse.

Saison	Date	Venant de	Allant à	VM	Montant de transfert	
10/11	30 juin 2011	Catania Calcio	Ascoli	700 K €	550 €	>

Dans cet exemple, le montant du transfert de ce joueur est de 550 euros sur Transfermarkt et le montant est de 550 millions d'euros dans nos données alors que la valeur réelle de ce transfert est de 550 000 euros.

league	season	window	team_id	team_name	team_country	dir	player_id	player_name	player_age	...	counter_team_id	counter_team_name	counter_team_country	transfer_fee_amnt	
21001	IT1	2010	s	1627	Calcio Catania	Italy	left	41312	Simone Pesce	28.0	...	408	Ascoli Calcio 1898	Italy	55000000.0

Notre exploration de ce dataset n'a pas révélé d'autres problèmes.

Création d'un Graphe Dirigé :

Dans le processus de transformation des données en un graphe dirigé via la bibliothèque NetworkX, nous avons rencontré une particularité notable. Chaque joueur était représenté deux fois, une fois pour le club de départ et une fois pour le club d'arrivée lors d'un transfert. Plutôt que de supprimer ces duplicitas, nous avons choisi de les conserver pour refléter fidèlement les transferts bidirectionnels car dans certains cas, comme les départs à la retraite, nous n'avons pas le transfert dans le sens opposé. Lors de la construction des graphes nous avons pris en compte cette particularité et nous avons adapté nos liens pour ne pas avoir de doublons.

Les nœuds du graphe symbolisent les clubs, tandis que les arêtes représentent les transferts de joueurs entre ces clubs. Cette représentation graphique nous permet de visualiser précisément les différents liens entre les clubs.

Pour essayer de maximiser les résultats nous avons choisi de construire trois graphes qui se différencient par le poids des liens.

1. Poids basé sur le montant total des transferts. Par exemple, si 3 transferts ont lieu entre 2009 et 2021 de Lyon vers Paris, le poids du lien correspondra à la somme des 3 transferts. Cette approche permettrait de repérer les acteurs centraux sur le marché des transferts en termes de puissance financière.
2. Poids basé sur le nombre de transferts entre 2 clubs. Voici un exemple où on voit que le Genoa a transféré 16 joueurs à l'AC Milan. Le but étant de voir les clubs qui transfèrent beaucoup de joueurs entre eux, être regroupés.

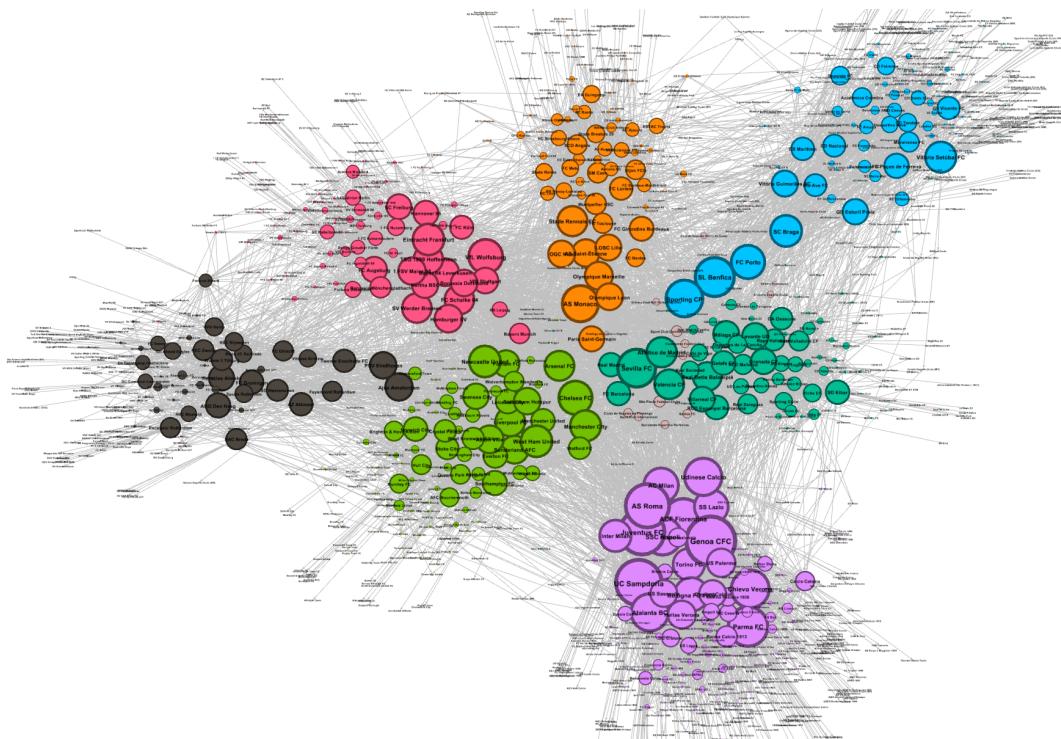
index	team_name	counter_team_name	dir	count ▼
597	AC Milan	Genoa CFC	in	16

- Poids basé sur les 2 approches précédentes. Nous avons comme poids le prix moyen des transferts entre 2 clubs (somme total / nombre de transfert)

3 Clustering

Pour détecre les clusters nous avons utilisé le logiciel Gephi. En utilisant les spatialisations Force Atlas 2 et Fruchterman Reingold ainsi que divers filtres.

Nous avons utilisé les graphes dirigés, préalablement construits pour trouver des structures et des regroupements au sein des transferts de joueurs. Nous avons utilisé le graph basé sur le nombre de transferts (celui qui ne prend pas en compte les montants de transferts).



Graph avec poids sur le nombre de transferts

Nous avons attribué des couleurs distinctes aux nœuds en fonction de leur pays d'origine, tandis que la taille de chaque nœud représente son degré. Une observation intéressante est la formation de clusters allongés qui convergent vers le centre du graphique.

Deux éléments clés peuvent être interprétés à partir de cette visualisation. Premièrement, les clubs de même pays ont tendance à s'échanger les joueurs entre eux. Deuxièmement, plus on se rapproche du centre, plus les transferts prennent une dimension internationale. On remarque par exemple que le Paris Saint Germain, Arsenal, le Real Madrid, le FC Barcelone, Chelsea ou encore Manchester City qui sont connues pour être des clubs qui récupèrent des joueurs de toutes les nations sont au centre du graphe.

De manière spécifique, le cluster des clubs italiens apparaît très compact, suggérant que les transferts de joueurs en Italie se réalisent principalement à l'intérieur du pays plutôt qu'avec des clubs étrangers.

Certains clubs des Pays-Bas et du Portugal tels que l'Ajax, PSV, Sporting ou Benfica, sont beaucoup plus proches du centre que les autres clubs de leurs pays. Ces clubs présentent une dimension internationale marquée.

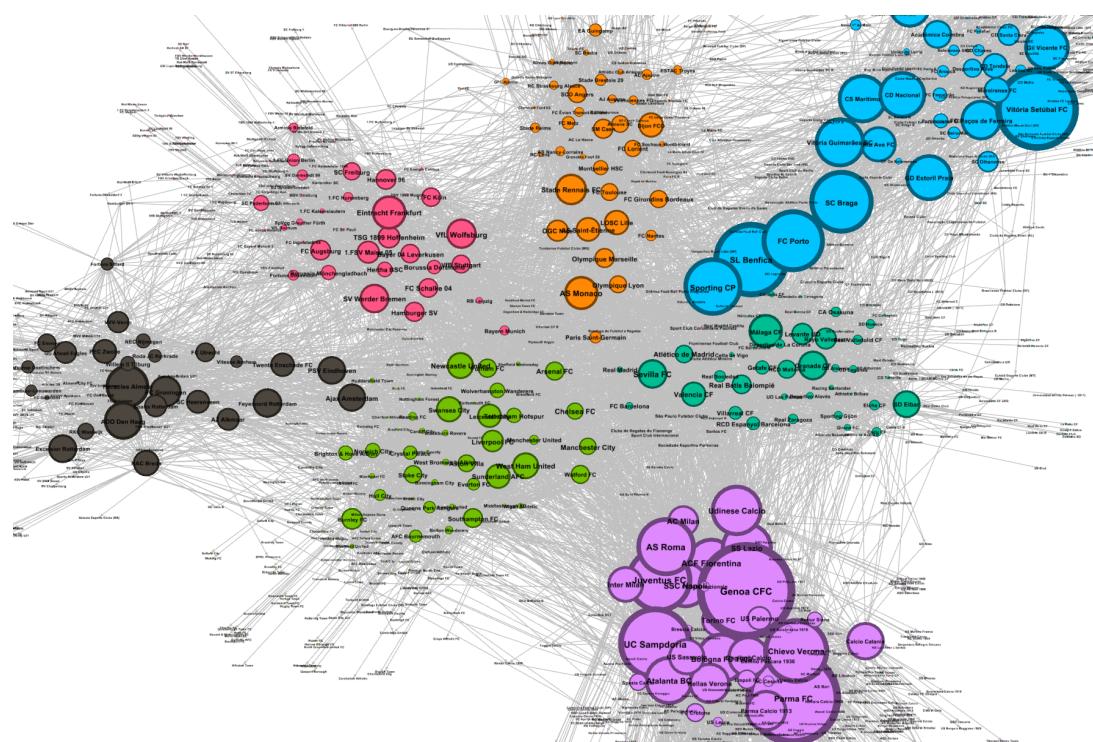
On observe également que les clubs brésiliens, représentés en beige, affichent une proximité notable avec l'Espagne et le Portugal, indiquant ainsi des portes d'entrées vers l'Europe pour ces joueurs.

4 Analyses statistiques

Betweenness

La betweenness centrality mesure le nombre de fois qu'un nœud se trouve sur le chemin le plus court entre deux autres nœuds dans un graphe, indiquant ainsi son importance dans la communication entre différentes parties du réseau.

Pour analyser cette donnée nous l'avons calculée sur notre Graph et nous avons changé la taille des nœuds en fonction de la valeur de betweenness.



Graph avec poids sur la sommes des transferts (taille de noeud sur betweenness)

On remarque que les clubs italiens et portugais affichent des valeurs de betweenness particulièrement élevées par rapport à ceux d'autres nations. Cette observation suggère que ces clubs agissent comme des points de passage importants pour les joueurs. On peut aussi interpréter ceci dû à la forte connexion historique entre le Portugal et le Brésil, ainsi que l'Italie et l'Argentine, que ce sont des pays qui accueillent les joueurs d'Amérique Latine.

Cependant, en ce qui concerne l'Italie, nous devons exercer une prudence dans l'interprétation des valeurs de betweenness. Nous avons identifié que les échanges entre

clubs italiens sont fréquents, et le volume de transferts impliquant des clubs italiens est le plus élevé dans notre ensemble de données. Cette surabondance peut conduire à une surestimation de la betweenness de ces clubs, influençant ainsi les conclusions. Par conséquent, une approche plus nuancée est nécessaire pour évaluer la centralité des clubs italiens dans le contexte des transferts de joueurs.

On peut aussi se poser les mêmes questions sur les clubs portugais, une betweenness élevée pourrait être simplement dû au fait que les joueurs au Portugal naviguent beaucoup entre les clubs portugais dans leur carrière ce qui augmente sensiblement la betweenness.

Nous avons analysé aussi la closeness et le PageRank mais aucune interprétation pertinente n'a pu en être tirée.

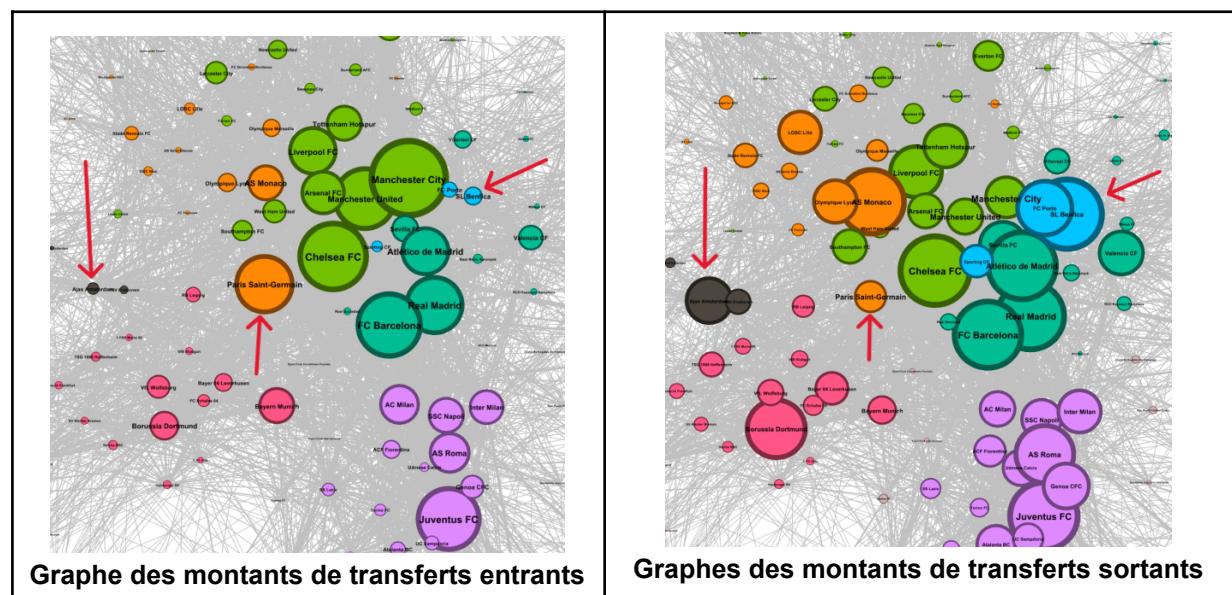
Degrés pondérés

Pour cette analyse nous avons pris le graphe qui prend en compte les sommes totales de transfert et non leurs nombres.

En utilisant les degrés pondérés entrants et sortants des nœuds, nous sommes en mesure de distinguer les clubs qui se sont avérés les plus efficaces dans leur stratégie de recrutement. Les illustrations ci-dessous mettent en évidence des exemples tels que le Paris Saint-Germain, qui investit massivement pour attirer des joueurs, mais ne parvient pas à réaliser une plus-value significative. À l'inverse, des clubs comme le SL Benfica ou l'Ajax Amsterdam parviennent à obtenir des bénéfices intéressants, malgré des dépenses moins importantes. Ces observations offrent des indices importants sur la stratégie de recrutement et sur l'équilibre économique de chaque club.

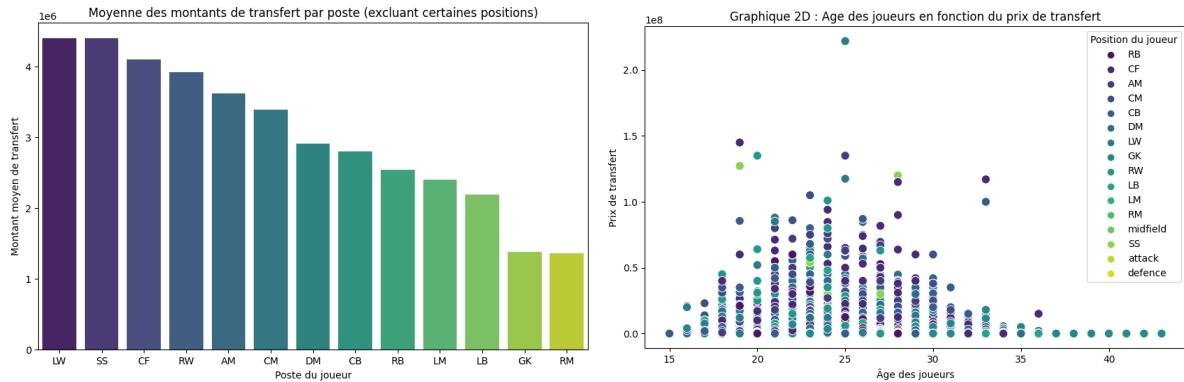
On peut en déduire que le but de certains clubs est de faire du profit sur le marché des transferts (exemple de l'Olympique Lyonnais qui aurait pu être cité et qui est visible graphiquement) en achetant des joueurs pour des petites sommes pour les revendre plus cher ensuite. Cela peut aussi être dû à un centre de formation performant qui vend ses jeunes joueurs à un prix élevé.

Nous ne pouvons pas donner d'interprétations sur l'influence de ces stratégies de recrutements sur les résultats sportifs car nous n'avons pas de données sur ce point.



5 Autres analyses

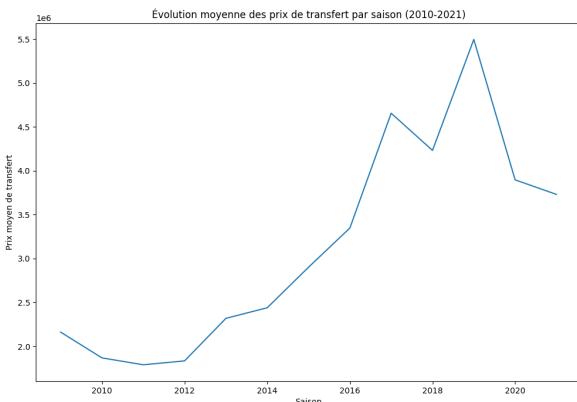
Les données que nous avons nous permettent d'autres analyses sur les joueurs. Nous avons voulu regarder les prix des joueurs en fonction de leurs âges. Ainsi que les prix en fonction de leurs postes.



On remarque que le prix moyen des joueurs est corrélé à leur position sur le terrain : plus un joueur joue haut sur le terrain (vers l'attaque), plus son prix est élevé.

On voit aussi que le prix en fonction de l'âge suit une distribution normale : le prix le plus élevé des joueurs se situe entre 21 et 26 ans, avec un pic à 24 ans et redescend ensuite.

On peut aussi remarquer que les gardiens de but ont une plus grande longévité dans leur carrière. En effet, le deuxième graphique montre que les gardiens de but sont les plus représentés au-delà de l'âge de 37 ans.



Enfin cette courbe nous montre que le prix moyen des joueurs a grandement augmenté au fur et à mesure des années mais une chute nette a eu lieu à partir de 2020. On peut associer cela à un événement mondial, la pandémie de Covid-19 qui a bouleversé les transferts dans le football mondial.

6 Conclusion

Dans ce projet, nous avons pu interpréter de manière différente plusieurs résultats. Grâce à la représentation graphique sur Gephi, nous avons détecté des clusters en fonction des pays ainsi que des clubs internationaux. Les analyses statistiques de betweenness et de degré pondéré nous ont permis de mettre en valeur certaines stratégies de recrutement des clubs ainsi que les pays passerelles avec l'Amérique latine. Enfin, à l'aide d'autres analyses plus simplistes, nous avons trouvé des corrélations entre l'âge et les prix des joueurs ainsi que leurs positions sur le terrain. Le dernier graphique nous a permis de voir que l'actualité (le Covid) a aussi son impact sur le monde du football.

Concernant la répartition du travail, nous n'avons pas travaillé séparément. Chacuns des graph et interprétations ont été effectuées ensemble.

7 Annexes

Dataset : <https://github.com/d2ski/football-transfers-data/blob/main/dataset/transfers.csv>

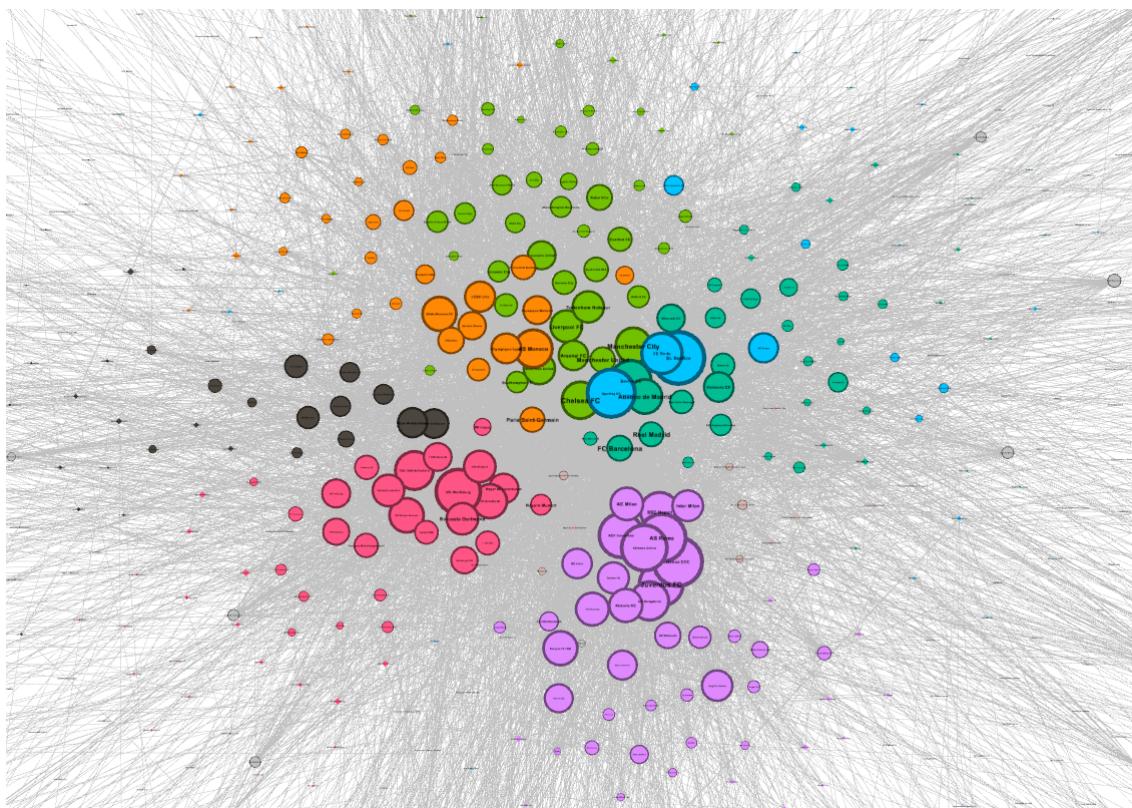
Représentation des différents postes des joueurs sur un terrains.



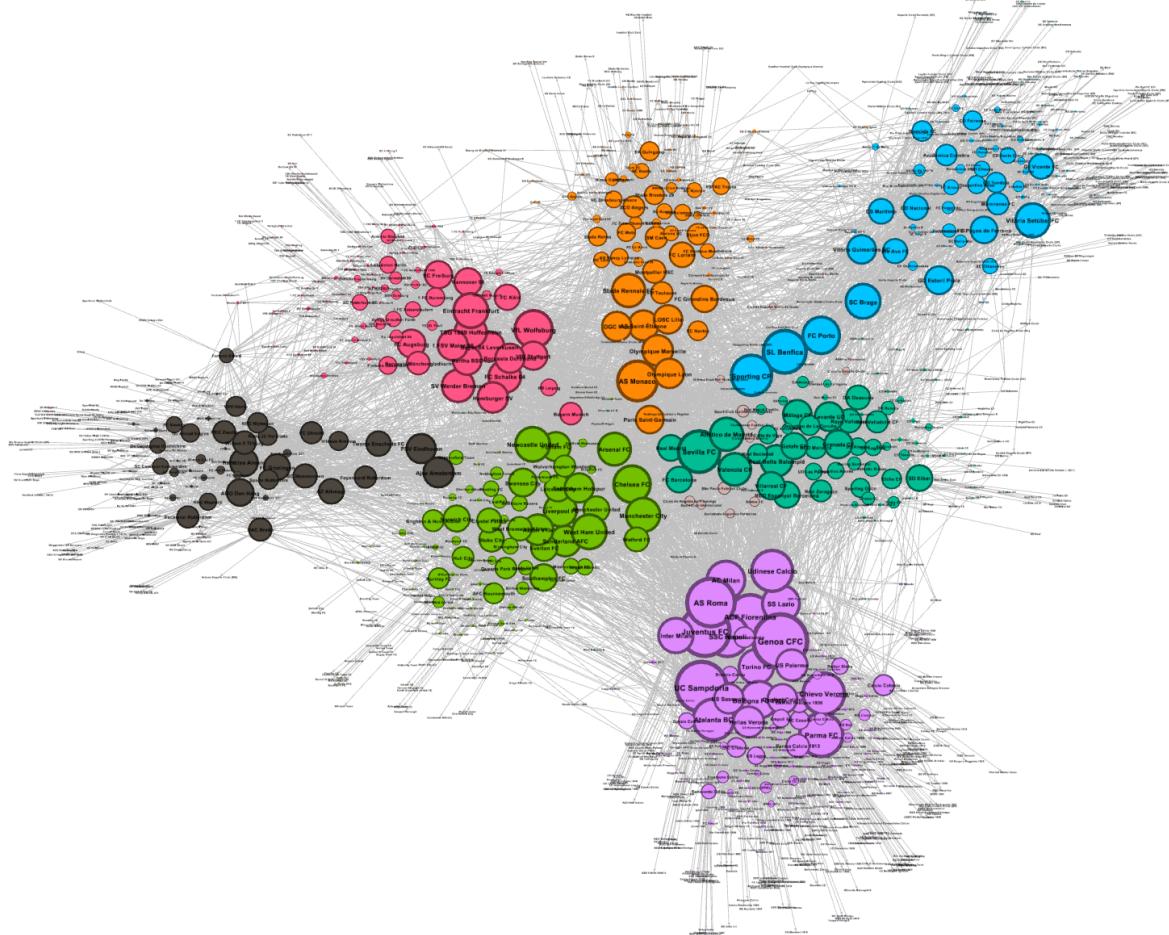
<https://jobsinfofootball.com/blog/soccer-positions/>

Graph :

1. Poids sur la sommes des transferts :



2. Poids sur le nombre de transferts :



3. Poids sur la moyenne des transferts (Somme des transfert / nombre de transferts) :

