A Survey of Social Network Analysis Tools and Techniques enhancing Player Scouting

Vishak Lakshman Sanjeevikani Murugesh,
University of North Carolina at Charlotte,
Charlotte, NC, USA

***Abstract*— Social network analysis (SNA) is the process of plotting and measuring of relationships between people, groups, organizations, and other connected information/knowledge entities even computers. From mapping social dynamics in social networks like Facebook, Twitter to measuring the evolution and changes in organisations, SNA acts as the unifying theme of research between computer science, economics, sociology and biological science. In this study, SNA algorithms are used for the betterment of the scouting system of soccer players. Players attributes and their factor of adaptivity into a new team act as the features to be analysed. Boosted Inductive Matrix Completion (BIMC), a ranking model for blog recommendations, is used as a recommendation tool to find players who have similar attributes but with a meagre transfer market value. The objective of the report is to have a thorough understanding of this algorithm, its performance metrics, applications and shortcomings, so they can be extrapolated to other forms of sports and athletics.**

Keywords—human relationships, player scouting; Boosted Inductive Matrix Completion; ranking model; low-rank structure

I. INTRODUCTION

A social network represents the connections between people, how they are related to one another. It indicates the closeness and how familiar are the people with one another. It can also be used to find people with similar tastes. [1]

Scouting is a player recruiting technique predominantly used in every sport on the planet. There are several factors that decide how a player is scouted or what type of player fits the team. To build a network, that assists in scouting, we can consider the nodes in the network as the players, while team chemistry acts as the link between the players i.e. how well the teammates gel with each other. The main challenge faced by scouts is the money efficiency. Finding quality players for a meagre sum defines the standard of the scout. The goal of the model is to recommend easily affordable players with same quality and traits of an already establish world class superstar.

II. SCOUTING MODELS

Different clubs across the world implement a wide variety of scouting models in developing young athletes. Some clubs sign up players in a relatively young age, then groom them into the type of players they want, while others look out for well established players that can be snatched up from the market for a bargain. Here are few scouting models that are implemented by the different football clubs:

A. Moneyball

Originally implemented in baseball, later it crawled its way into almost every game in the planet. The principle of this model solely depends on goal conversion rate. The various attributes like strength, age and pace are all overlooked if the player has a very high goal conversion rate. So, in this model the goal conversion rate form as the link between the various nodes. It has attained a zeal of success in clubs like Liverpool. The neural network that is used in predicting the recommendations have a very simple structure because of the very few constraints.

B. Adoptive Model

La Masia, the famous academy of Football Club Barcelona is the perfect example for Adoptive model of scouting. This model involves recruiting young talented players in their early teens and then molded into the type of player that best suits the club's requirements and its style of play. In this model, the key factor for the player is his adaptability. The player should be versatile and flexible in playing in different positions.

Even though, at surface the recommendation criteria look straight forward, but to attain such a versatility in playing different positions, the player should be dominating in player features like passing, dribbling, vision and technique. So, these features form the substructure of the SNA while the position forms the primary link between the players.

C. TIPS

TIPS is arguably the best youth development and scouting model in the football world. The birthplace of this model is in Amsterdam and is put into practice efficiently by the juggernaut in Eredivisie, Ajax Amsterdam. TIPS is the acronym for Technique, Insight, Personality and Speed. Only the players who make through these filters have a shot at making into the academy.

The results produced by this model are astounding as Ajax have become one of the biggest feeder club for farming young and bright prospects. Some famous names as result of this model include Johan Cruyff, Marco van Basten, Ronald Koeman, Wesley Sneijder etc This is one of the complex models as there are four primary constraints to be met and established as a link between the players. This leads to the usage of more hidden layers while developing the deep networks.

III. Social Network Properties

Some of the social network properties form a key feature in building the models that can be used in generating recommendation of the similar players while scouting. The integral properties are briefed as follows:

A. Centrality and Power

Sociologists define power as the core building block of social structures. As defined by Mohsen Jamali and Hassan Abolhassani in [7], there are three aspects of power that determine how well structured a social network is built. Table I summarizes what aspects of scouting parameters correspond to the different aspects of power and what influence they have in the predictive model.

Power Aspect Name	Definition	Influences
Degree	Number of ties for a player	Possibilities of more connections and alternatives.
Closeness	Length of paths to other players	Direct exchange with other players.
Betweenness	Lying between each other pairs of players	Determining the chemistry between different players.

TABLE I
COMPARSION OF THREE ASPECTS OF POWER (DEGREE, CLOSENESS, AND BETWEENNESS)

B. Substructures

Though power and centrality and power form the core of SNA, it's the substructures present in the networks that have a significant say in the efficiency and accuracy of the SNA predictive models. Almost every approach to understand a network structure emphasize how the dense connections are compounded and extended to develop larger cliques or sub-groupings.

There are a wide range of algorithms that identify how complex network structures are compounded from the small and less complex ones. Cliques are one such aspect of a network that determine the complexity of the network. For example, consider two networks where one network has two non-overlapping cliques while another network has overlapping cliques. There is a high possibility of conflicts in these networks that have overlapping cliques than that of the one with independent cliques. Also, mobilization and diffusions tend occur more frequently and rapid in this networks with the overlapped cliques [7]. So, these cliques and sub graphs form the synopsis for the inspection of the nature and characteristics of a certain graph;

- Are the sub-graphs independent from each other?
- Do these sub groups share members, or do they or factionalize the network?
- In case the subgraphs are connected with each other, the how complex are these sub-graphs? Are there a few big groups, or a larger number of small groups?

Substructure Name	Description
Clique	Players who have possible links among themselves
N-Clique	Players who are connected to every other player of the team at a maximum distance of N
N-Clans	N-Cliques that all paths among players occur by the way of other players of N-Clique
K-Plex	Clique in which players have links to all but k of players
K-Core	Players are connected to k of members of the group
Cut Points	Nodes which if removed, the structure becomes divided into un-connected systems
Block	The divisions formed by the cut points by the removal of nodes.
Lambda Set	Set of players who if disconnected, would most greatly disrupt the flow among all the players

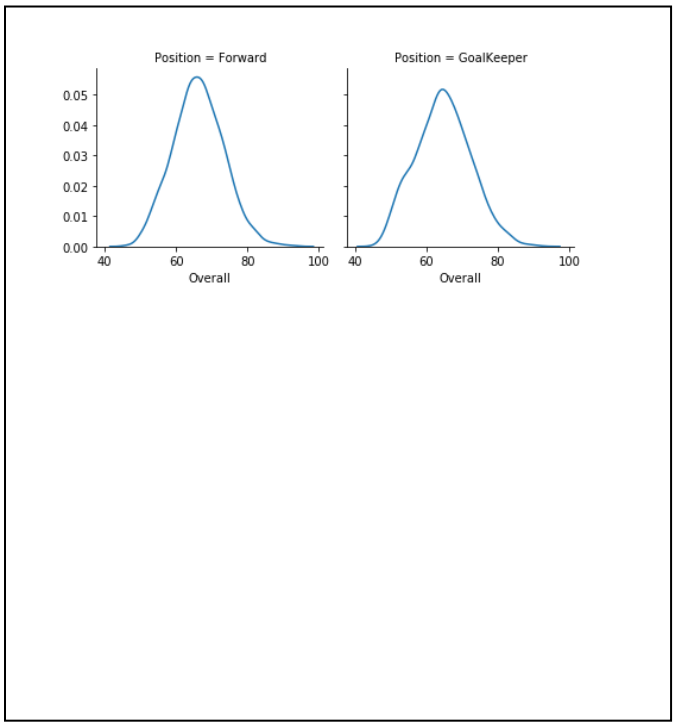
TABLE II
COMPARING DIFFERENT APPROACHES FOR DEFINING SUBSTRUCTURES
AND GROUPS IN SOCIOGRAMS

IV. Dataset Exploration and Preprocessing.

The dataset used for analysis is obtained from Kaggle and it has basic information about the players such as Name, Age, Current Club, and Country. The players are classified into four major categories according to the positions they are well suited for. They are as follows: Forwards, Midfielders, Defenders and Goalkeepers. Each of the player's attributes act as 32 features such as acceleration, stamina, strength, finishing, curve, passing etc. and certain features like Goalkeeper Diving, reflexes, Position etc are confined only to Goalkeepers. The attributes are rated from 0-100 and the aggregate of all these attributes are calculated and tabulated as "Overall'. Scouts who have watched the players play also add another column called "Potential" which is a remark of the capability of the player. For example, it is better to recruit a player of age 23, overall 78 and potential 91 rather than recruiting a player aged 32 and overall 92, given the factor that the younger player would come cheap too.

The next set of features include the events related to a player like shots taken, dribbles completed, pass completed, goals scored, assists and just as attributes some special features pertaining only to goalkeepers like shots saved, goal conceded, and penalties saved. These add a further layer for our analysis. Finally, the final ingredient is the value of the player in the transfer market. For a scouting network to work efficiently, it should scrap for players with maximum potential but with very little capitation involved.

Once these data are cleaned and preprocessed, initial plots are visualized for understanding the dataset. All data cleaning, preprocessing and visualizations are done in Python.



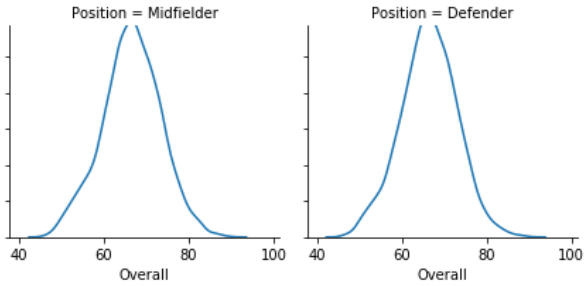


Fig 4.1. PLAYER OVERALL ACROSS DIFFERENT PLAYER CATEGFORIES

This plot represents the spread of various players' Overall stat across the different categories of forward, goalkeeper, midfielder and defender. It gives an idea of the lowest rated player and highest rated player across the different categories.

Next, to determine the features that best influence in categorizing of groups, we can split the four categories of players into two for analysis. Forwards, midfielders and defenders are grouped together as Outfield players while the Goalkeepers are kept separately for analysis.

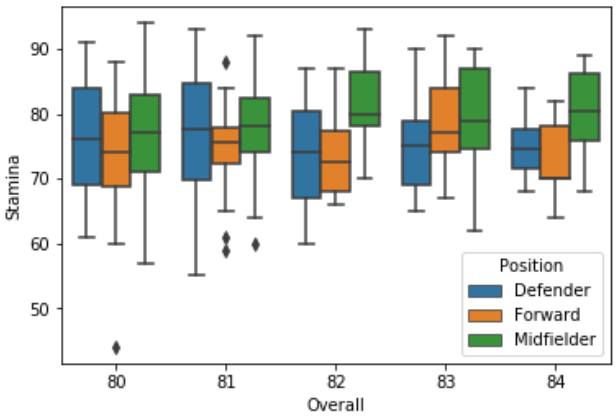


Fig 4.2. OUTFIELD PLAYER ATTRIBUTES ACROSS PLAYER CATEGFORIES

The boxplot shows which position requires the maximum usage of stamina. So, this can be used as a major factor that influences the categorization. It is clear the plot that, the midfielders are require the most stamina then will be defenders and finally the forwards.

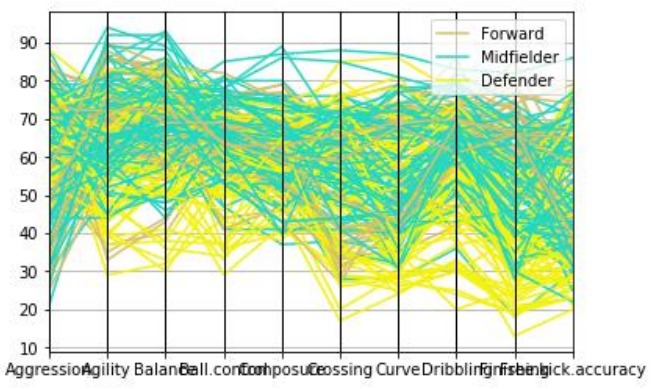


Fig 4.3. ATTACKING PLAYER ATTRIBUTES ACROSS PLAYER CATEGFORIES

This multivariate plot shows the impact of other features than may have role in defining the classification criteria. We can infer the defenders have considerably low values for features like Agility, Composure, Curve, Dribbling and Finishing. These factors are very high for Forwards and just as expected moderate for midfielders.

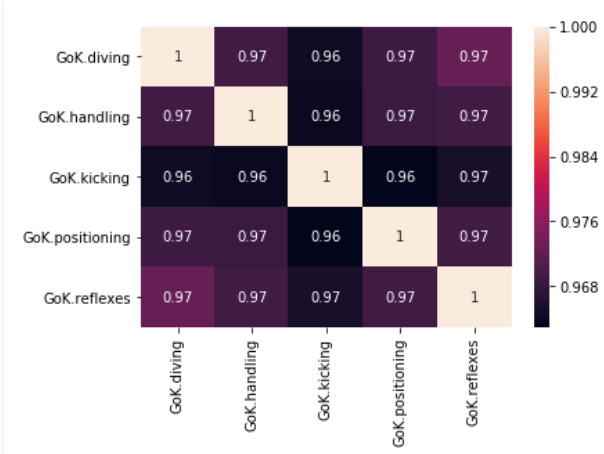


Fig 4.4. GOALKEEPING ATTRIBUTES OF DIFFERENT PLAYERS

Now that there is a clear picture about the outfield players, this heatmap can be used for evaluating the goalkeeper features of Diving, Handling, kicking, positioning and reflexes. A spread of the rating across all the goalkeeper samples are shown here. Now that the dataset is completely cleaned and preprocessed, the algorithm to predict similar players can be implemented.

V. BOOSTED INDUCTIVE MATRIX COMPLETION

The Boosted Inductive Matrix Completion can be implemented in three levels. First level involves the building of an inductive matrix completion method for only the attributes of the players. This gives a recommended set of players with similar traits and attributes. Then, the Boosted Inductive Matrix Completion method is used on this resultant to find players who meet secondary criteria like team playing style and team chemistry.

Notations:

Directed Graph, G	$G = (V1, V2, \varepsilon)$
Set of players and their attributes, V1, V2	V1 (m = |V1|) V2 (n = |V2|)
Edges that form between the players in the graph, ε	$\varepsilon = \{e_{ij} \mid i \in V1, j \in V2\}$

Let $A \in R^{m \times n}$ be the adjacency matrix of G, where each row corresponds to a player and each column adheres to various attributes of the players.

Let $X \in R^{m \times f(u)}$ and $Y \in R^{n \times f(b)}$ denote the player and attributes feature matrices, respectively. [4]

5.1 Matrix Completion

A low rank Matrix Completion method is the most common solution used for recommendation system in most of the applications we use in our daily day life. In a general matrix completion method, a matrix with all the users and features is generated and holes in linking the different players, which are normally represented by the null values in the matrix, are bridged by computation. Similarly, the goal at this stage of the algorithm is to identify the low rank matrix formulated as follows:

$$\min_{U,V} \sum_{(i,j)\in\Omega} (A_{ij} - (UV^T)_{ij})^2 + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2),$$

where,

$U \in R^{m \times r}$ and $V \in R^{n \times r;}$

r is the dimension of the latent feature space;

Ω is the set of observed entries and $\Omega \in$ m x n;

λ is a regularization parameter.

5.2 Inductive Matrix Completion

To make more accurate recommendations, an additional layer of information is added to this low rank matrix. As

discussed earlier, the players with similar attributes should have higher chemistry with the team of the player he intends to replace. This chemistry depends on two features: the country he represents and the country the club he currently plays for represent.

To achieve this, an inductive matrix completion (IMC) approach is proposed and the main idea is to model Aij using the player's feature vector xi, his attribute's feature vector y_j and the low-rank matrix $Z \in R^{f(u) \times f(b)}$ as

$$A_{ij} = \mathbf{x}_i^T Z \mathbf{y}_j.$$

where

$x_i \in R^{f(u)}$

$y_j \in R^{f(b)}$

$Z \in R^{f(u) \times f(b)}$

Hence the final step is to model Aij which has the power of MC to reduce the noise level in the input data and the advantage of IMC to incorporate side information of players and their attributes. A step by step depiction of the algorithm can be described as follows:

1. Learn the latent factor matrices U and V of the MC model.

$$A_{ij} = (UV^T)_{ij} + \alpha \mathbf{x}_i^T Z \mathbf{y}_j$$

2. The residual matrix $R = A - UV^T$ illustrates the links that MC failed to capture in the follower graph. Now, R_{ij} is modeled with IMC as

$$R_{ij} = A_{ij} - (UV^T)_{ij} = \mathbf{x}_i^T Z \mathbf{y}_j$$

3. Finally the required resultant matrix that provides the recommendations for alternate player can be obtained by the combinations of both MC and IMC as:

$$\min_{W,H} \sum_{(i,j)\in\Omega} \ell(A_{ij} - (UV^T)_{ij}, \mathbf{x}_i^T W H^T \mathbf{y}_j) + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2).$$

VI. Experiments and Analysis

Before starting with the implementation of the BIMC algorithm, first the choice of classifier is chosen. Random Forest and KNN Classifiers are tested for accuracy from the results below KNN has the better accuracy.

Random Forest Classifier

0.770092226614

	precision	recall	f1-score	support
Defender	0.76	0.79	0.78	1698
Forward	0.78	0.77	0.78	959
GoalKeeper	0.85	0.87	0.86	620
Midfielder	0.69	0.67	0.68	2118
avg / total	0.77	0.77	0.77	5395

KNN Classifier

0.8328081557

	precision	recall	f1-score	support
Defender	0.87	0.86	0.87	1698
Forward	0.79	0.72	0.75	959
GoalKeeper	0.92	0.96	0.95	620
Midfielder	0.78	0.81	0.79	2118
avg / total	0.83	0.83	0.83	5395

Now factorization is performed is done using this matrix A_{ij} and the resultant matrix is formulated by:

$$\min_{W,H} \sum_{(i,j)\in\Omega} \ell(A_{ij}, \mathbf{x}_i^T W H^T \mathbf{y}_j) + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2),$$

where $Z = WH^T$ and $W \in R^{f(u) \times r}$, $H \in R^{f(b) \times r}$

5.3 Boosted Inductive Matrix Completion

IMC has the drawback of being too rigid and as a result there are performance issues with alarming levels of noise in the resultant matrix. To overcome this shortcoming, a combination of both standard matrix completion and inductive matrix completion is implemented and thereby utilizing the power of both the methodologies.

Now, we can go on with the implementation of BIMC. The algorithm is implemented is Python. The input is the player that we want to find a replacement for. Upon running the code, the possible alternate players with similar players are displayed. The players are sorted based on the decreasing order of the chemistry with the team ie. the player who can attain maximum chemistry with the team is shown at the top. Here is sample for one of the players.

```
Player to replace
-----------------
Player    Age Overall

G. Higuain   29   90

Recommended Alertnate Players
------------------------------

D. Sturridge   27   82

Morata   24   84

Leo Baptistao   24   80

K. Benzema   29   86

C. Thereau   34   77
```

Fig 6.1. RECOMMENDED ALTERNATE PLAYERS

We can further make use of the results to find the players who are undervalued or overvalued for their overall rating.

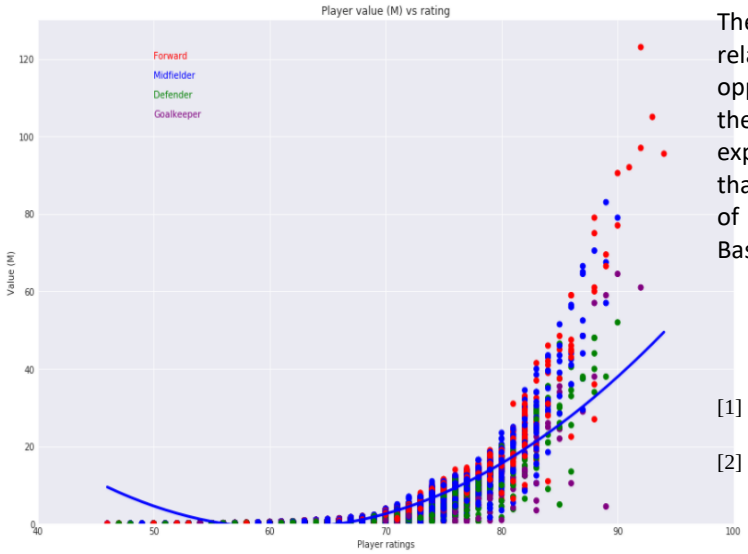


Fig 6.2. OVERVALUED PLAYERS VS UNDERVALUED PLAYERS

The blue line represents the gradient that differentiates the overpriced ones from the highly undervalued ones. The players below the ones are the diamonds in the rough, these players can bring in high value to the team at a very low cost rather than the highly valued players who pay back the investment but will cost the team a fortune.

This plot is just an improvement on the previous plot. It gives a detailed information about the undervalued players, the position they play.

VII. CONCLUSION

Through this study report, I have learnt about some of the commonly used methodologies for SNA and correlated them with one of the least researched topics like player scouting. There were some challenges in building the model as I was relatively new to Python, but this gave me the perfect opportunity to learn something new. The results indicate only the preliminary implementation of the algorithm. It can be expanded to fit any of the aforementioned scouting models that can be beneficial for the club in the long run. The limits of its usage can be further expanded into sports like NFL, Baseball etc.

REFERENCES

[1] Petek Askar, “Social network analysis for e-learning environments,” Izmir University of Economics, 35330, Turkey.

[2] Lukas Zenk, Christoph Stadtfeld, “Dynamic organizations. How to measure evolution and change in organizations by analyzing email communication networks”, 6th Conference on Applications of Social Network Analysis.

[3] Bruce Cronin, “A window on emergent European social network analysis,” , University of Greenwich, Park Row, London SE10 9LS, UK, 4th & 5th UK Social Networks Conference.

[4] Donghyuk Shin, Suleyman Cetintas, Kuang-Chih Lee, Inderjit S. Dhillon, “Tumblr Blog Recommendation with Boosted Inductive Matrix Completion”.

[5] David Ediger, Karl Jiang, Jason Riedy, David A. Bader, Courtney Corley, Rob Farber, William N. Reynolds, “Massive Social Network Analysis: Mining Twitter for Social Good,” 2010 39th International Conference on Parallel Processing.

[6] Silvio Lattanzi, “Algorithms and models for social networks,” Sapienza, Università di Roma, Dottorato di Ricerca in Computer Science, XXIII Ciclo – 2010.

[7] Mohsen Jamali and Hassan Abolhassani ,“Different Aspects of Social Network Analysis”, Sharif University of Technology, Tehran, Iran.

[8] C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh, “Text and structural data mining of in uenza mentions in web and social media,” Public Health Informatics special issue in International Journal of Environmental Research and Public Health, vol. 7, 2010.

[9] Steve Belichick , “ Football Scouting Methods”.

[10] James Tippet, “Breaking the Football Code”.