Measuring influence of coaches on NBA players' development
Term Project

Bence Szabó
Central European University, Statistical Methods in Network Science and Data Analysis

April 1, 2017

Introduction

The early period of any professional's career impacts heavily the potential trajectories one can have in a lifetime. Studies show for instance that intervention in early childhood health habits leads to adulthood improvements in several aspects (Conti, Heckman and Pinto, 2016). We can generally expect the same effects for professional performance as well, in this case: basketball. Having good coaches and veteran players around can help through several channels: building habits, helping with finances and trivially with the knowledge of basketball.

In my term project I look to measure the influence of coaches on NBA players' development based on the players' improvement measured in win shares. Using scraping methods I collect data of 17 seasons of NBA statistics (1999-2016), all coaches' and assistant coaches' teams of occupation to create network dataset with the following framework in mind. There are two types of nodes: mentors and mentees. Mentees are young players who are potentially influenced by coaches, assistant coaches and older players (the mentors). Controlling for their initial skill level I want to measure their prime basketball "production" and link it to previous mentors.

The mentors are allowed to have direct and indirect effects, the latter resulting from spillovers through their mentees becoming mentors later in their careers. For instance if there is a coach that was a good mentor, and his players later become good mentors as well, I want to account for that spillover effect to the initial coach. The question is interesting as it differs from the question of being the most successful coach: developing younger players can even decrease the team success in the short-run. I also include assistant coaches: it can easily happen that an assistant coach is responsible for the real effect of the development of a player, not the head coach. Naturally this analysis abstracts away from many things: the real dynamics of a player's development, effect of training staffs or that more talented players might need personalized trainings. These are limitations that I am not going to address. I also cannot address the bias coming from young coaches not having spent enough time in the league: they simply cannot develop players showing up in this setting.

This term paper is organized as follows: first I discuss the definitions of certain concepts of the NBA and the measurement process, and then look at the algorithm creating the dataset and the network. Afterwards I look at the results, examine their robustness and discuss the findings. Lastly I am going to look at the possible improvements and extensions of the project. All data and computations are attached as a separate file.

Context and definitions: what is the NBA and the statistics used

The National Basketball Association is the professional basketball league of the United States and Canada for male players, one of the largest sports businesses of the world with 30 teams and thousands of players having played over the last decades. An NBA season has two parts: a regular season which determines seedings between teams, and a playoffs period, where depending on the seedings teams are matched up one-to-one in a best-of-seven series by western and eastern "conferences", and in the Finals the best of the East competes against the best of the West. In this analysis I only look at regular seasons, consisting of 82 games.

Every team in the NBA has a coaching staff with a head coach and 3-4 assistant coaches, and a maximum roster of 15 players. Player's are traded within the season and coaches are fired within the year. Players and coaches are linked through being on the same team: in this analysis I am considering only those connections that happened at the start of the season. My reasoning is the following: every season begins with a training camp, where the mentoring of younger players can actually occur, while during the season this has more limitations due to frequent travels of the teams.

In general young players arrive to a team through the Draft, which is a lottery-based selection from a pool of players eligible to participate. In the Draft those teams are favored to pick first who had worse seasons in order to help them to get better. After being drafted, talented players however can get either under bad coaches and bad influences or good ones. Veteran players normally end up on teams via trade or "free agency" which is a market where players and teams match up. Coaches have a similar market.

In the NBA almost every decision a player makes on a basketball court is tracked, creating rich databases for analysts. I am going to use two widely used statistics for this analysis, offensive and defensive win shares. These are composite indices proxying the marginal contribution of a player to wins in a season based on the offensive and defensive production. The offensive win share is calculated the following way (via `http://www.basketball-reference.com/about/ws.html`):

1. Calculate points produced for each player
2. Calculate total offensive possessions for each player
3. Calculate marginal offense = total points - 0.92 * league points per possession * offensive possessions
4. Calculate marginal points per win = 0.32 * league points per possession * $\frac{\text{team pace}}{\text{league pace}}$
5. Offensive win shares: $OWS_i = \frac{\text{marginal offense}}{\text{marginal points per win}}$

Defensive win shares are computed parellel to this in the following way (again via `http://www.basketball-reference.com/about/ws.html`):

1. Calculate the defensive rating for each player (this is a metric of Dean Oliver, estimating the player's points allowed per 100 possessions)
2. Calculate marginal defense for each player = $\frac{\text{player minutes played}}{\text{team minutes played}}*$team defensive possessions*(1.08*league points per poss
3. Calculate marginal points per win = $0.32*$ league points per game * $\frac{\text{marginal defense}}{\text{marginal points per win}}$
4. Defensive win shares: $DWS_i = \frac{\text{marginal defense}}{\text{marginal points per win}}$

These metrics are somewhat far from the original possessions in the game, however they account for individual effects, pace of the game and “stat stuffing”. This means that players who selfishly play for statistics only are penaltied, and it controls for the fact that some teams play more slowly, while also taking into account the effect of teammates.

Data and network formation

My data source was basketball-reference.com. In this analysis I use 17 seasons of data from 1999/2000 until 2015/2016. These data consist of around 1000 players and 100 coaches. The algorithm of the scraping had the following steps:

1. Use BeautifulSoup objects to store yearly statistics for the given seasons
2. Pre-process the data by cleaning of duplicates and account for name changes
3. Collect data on draft dates
4. Go through each coach and frame career into coaching destinations, and collect their win/loss record for further examinations
5. Create separate datasets for older players, younger players and coaches

The exact codes are attached. Using these data I define the following relationships:

- A coach is a mentor to a player if the coach and the player (mentee) were on the same team at the beginning of a season in the first 3 years of the player’s career
- A player is a mentor to another player (mentee) if the former one is in at least his 4th season while the latter is in his first 3 years of his career

These relationships create the base for links. The network has some attributes of a bipartite network: coaches can be separated from players, however some players can be mentees in their younger periods, while mentors in their older periods.

In this inquiry we are interested in the influence of players and coaches which can be modelled as indegrees: this results in links being directed such that the head of the link is the mentee, and the tail of the link is the mentor. In this setting the directed link means “being influenced by” a certain player.

As I want to define influence measured by the effect in developing a player, I need to attribute weights to the links being connected to this basketball production measure. First I want to account for the differences in baselines: players who are more talented arrive in the league with a greater skillset, so instead of raw win shares I base the measure on the following metric. Let us define OWS_{i1} and DWS_{i1} as the first season offensive and defensive win shares of player i, while also define OWS_{ip} and DWS_{ip} as the average offensive and defensive win shares during the player’s prime, which I defined as production from year 4 until year 8 (or until active). These intervals are somewhat ad hoc, but should be close proxies of what we are looking for regarding prime and initial productivity. So the weights are based on: $dOWS_i = OWS_{ip} - OWS_{i1}$.

So incorporating all this information I define weights of links between i and j as:

$$w_{ij} = \frac{N_{ij}}{N_i} dOWS_i$$

where N_{ij} is the number of links between mentee i and mentor j. So if they played together for two years in the first three years of the mentee's career while the mentor was already at least in his 4th year, $N_{ij} = 2$. N_i denotes the total number of "player-years", which is the total outdegrees of mentee i. So as an example: if a mentee played for 3 years for a coach in his first 3 years and had an old player for 2 years besides him, and had a change in win shares of 5, the weight of the relation between the coach and the mentee is $w_{ij} = \frac{3}{5} * 5 = 3$. This is the baseline configuration. As robustness checks I also used different definitions: standardized values of the change in win shares (value is: subtract mean and divide by standard deviation) and eliminating small weighted links, meaning if they are within 2 standard deviations of the mean change in win shares. I am going to call these configurations "raw", "standard" and "cleared".

These conditions define the network where links are "flows of development influence". We can consider two different approaches to measure influence. First is to look at direct influence: add up all the mentees' development attributed to given mentor which is the weighted in-degree in this setting:

$$wd_j = \sum_i w_{ij}$$

But we aim to measure the direct and indirect influences of a mentor, which implies that for each mentor j:

$$D_j = \frac{1}{\lambda} \sum_{i \in N_j} D_i \implies \lambda D = gD$$

so we need to compute the eigenvector centrality of the nodes. I refer to this value as "Development Influence Index", DII to reflect upon the meaning of the metric.

Figure 1 and Figure 2 depicts the networks based on Offensive and Defensive Win Shares respectively. The intensity of the color is based on weighted in-degrees, while the size of the nodes are based on the eigenvector centralities, the Development Influence Indices. We can see that these two do not necessarily overlap: some nodes have low direct influence, while they have larger indirect influence and also the reverse as well.

Figure 1: **Network of Development Influence, Offensive Win Shares**

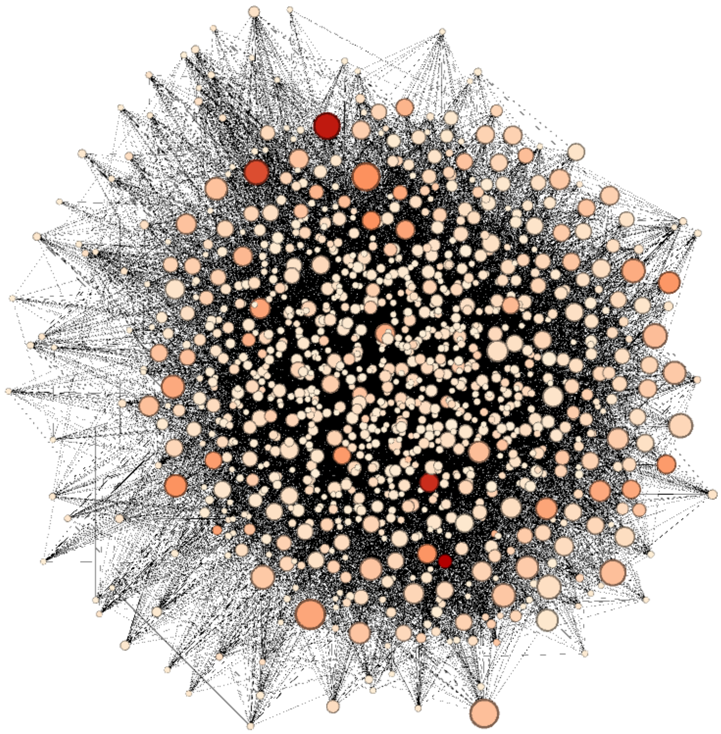


Figure 2: **Network of Development Influence, Defensive Win Shares**

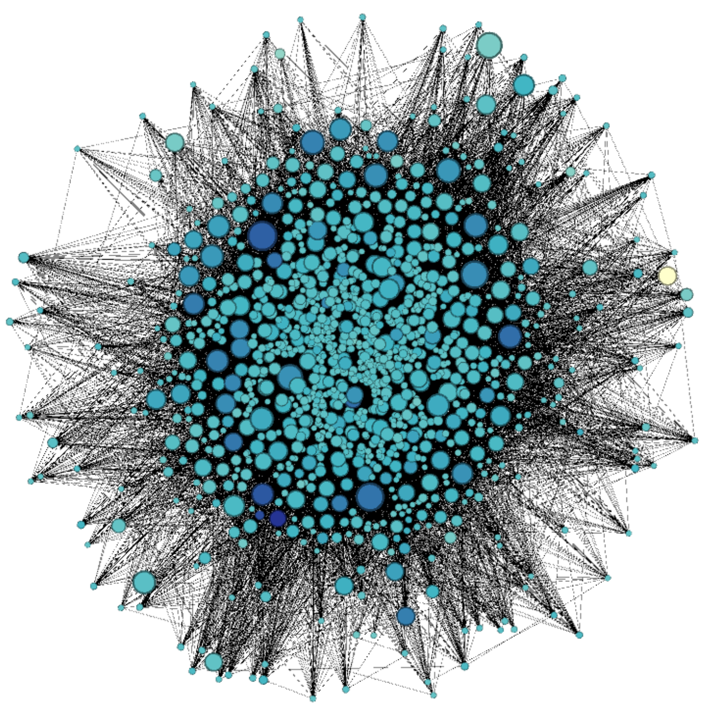
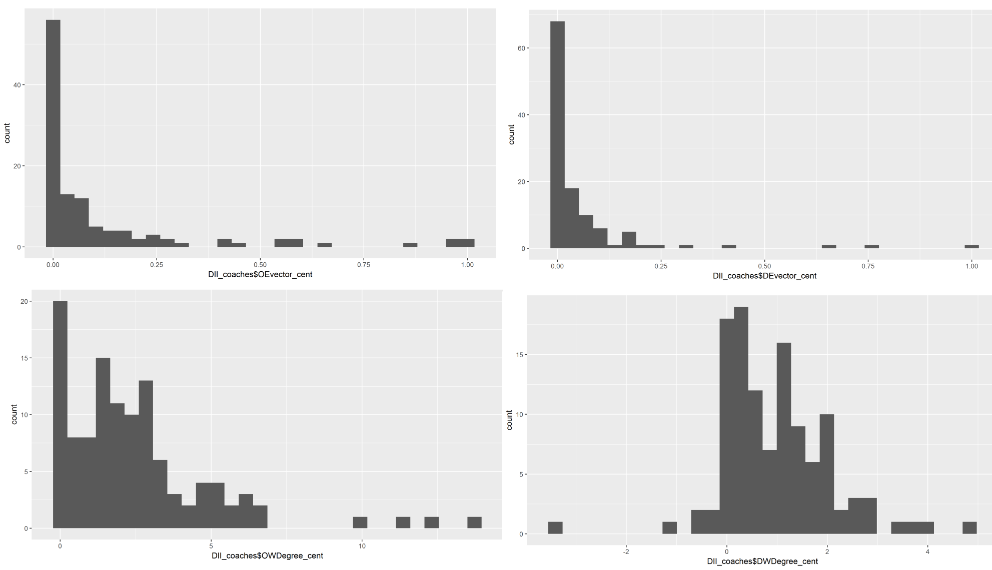


Figure 3 shows the distribution of DII (upper) and the weighted in-degrees (lower) for coaches, offensive are on the left handsight, while defensive is on the right handsight. We can see that the distributions are skewed: most coaches have little to no effect on the the development of players, while some have large effects comparatively. We can compare that to weighted degrees: on the offensive end we can see a slightly less skewed distribution, but on the defensive end we can see that negative values are more common. This means that actually there seem to be coaches whose influence is negative on players' defense.

Figure 3: **Densities of offensive and defensive Development Influence Index and weighted in-degrees of coaches**



Results and robustness checks

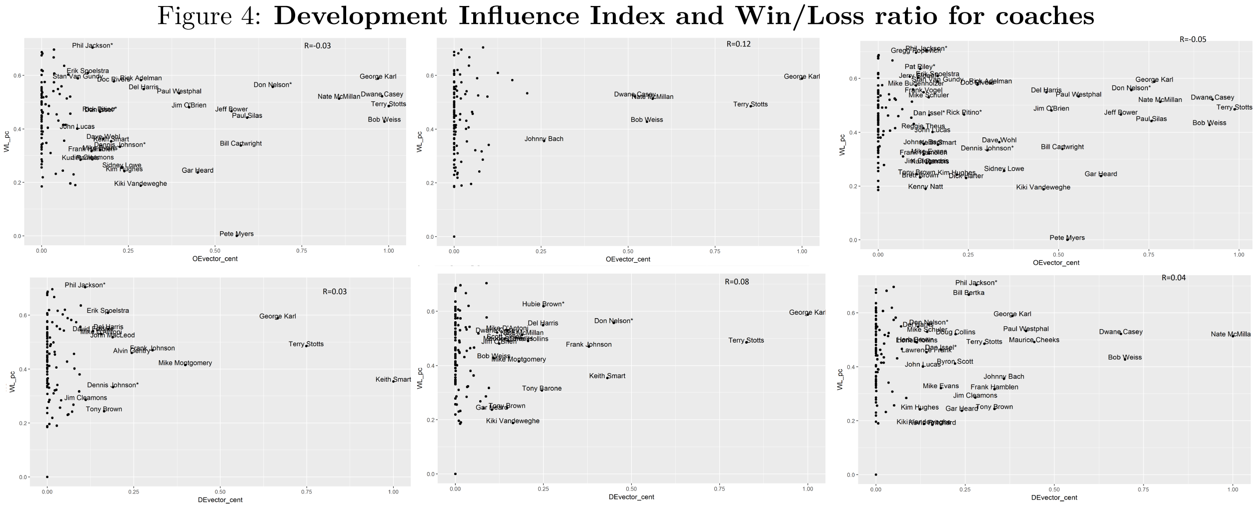
The term project aims to evaluate and rank coaches based on their influence on players' development which we defined in the last sections. I used R and igraph package during these computations. First I display in Table 1 the top 5 coaches based on the two main categories and the 3 different definitions (raw, standard and cleared, mentioned in network formation part). We can see that the different definitions might alter the exact ranking, but the names seem to be stable within the categories. Let us remember that this ranking is not based on winning percentage or general success but on development in player's production who were in contact with a coach in the player's first 3 years in the NBA. With that in mind we can see that well-regarded coaches are in the top 5: George Karl is a former Coach of the Year, Terry Stotts was one of the best assistants and also Coach of the Year nominee in the recent years, while Don Nelson is a Hall of Fame coach. It is easily verifyable for the reader by a Google search that these coaches are highly appreciated by their peers and players. So this ranking seems to be corroborated by common sense, even though it was based on a relatively distant and abstract measures and methods.

Table 1: **Top 5 coaches based on their Development Influence Index**

	Offense			Defense		
Rank	Baseline	Standard	Cleared	Baseline	Standard	Cleared
1	Terry Stotts	George Karl	Terry Stotts	Keith Smart	George Karl	Nate McMillan
2	Bob Weiss	Terry Stotts	Dwane Casey	Terry Stotts	Terry Stotts	Bob Weiss
3	Dwane Casey	Nate McMillan	Bob Weiss	George Karl	Don Nelson*	Dwane Casey
4	George Karl	Bob Weiss	Nate McMillan	Mike Montgomery	Keith Smart	Maurice Cheeks
5	Nate McMillan	Dwane Casey	George Karl	Frank Johnson	Frank Johnson	Paul Westphal

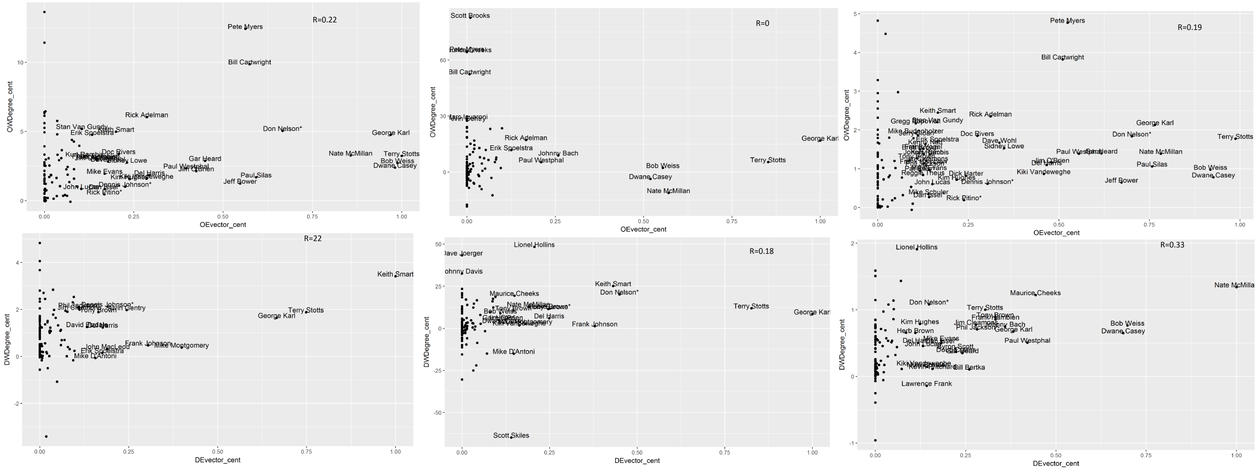
We can compare the computed DII values of the coaches' to their career win/loss ratios, which describes their success with respect to winning games in the regular seasons during their careers as head coaches (so we cannot attach this to their assistant coach tenure). Figure 4 displays the following: in the first row we can see the offensive DII-s displayed vs. win/loss ratios, in the second row we can see the defensive DII-s in the same manner. The three columns are the baseline, the standard and the cleared configurations. We can see that there is no strong relationship between winning and developing players: Phil Jackson is the "winningest" coach in the sample, however his development points are quite low, while the most successful developers have an around 0.5-0.6 winning ratio. This observation is valid across all definitions and categories. This is expected: having good players on the team and being a good coach leads to success, however developing young players does not necessarily translate to wins for the coach himself; it might be rewarded only in the long run.

Figure 4: **Development Influence Index and Win/Loss ratio for coaches**



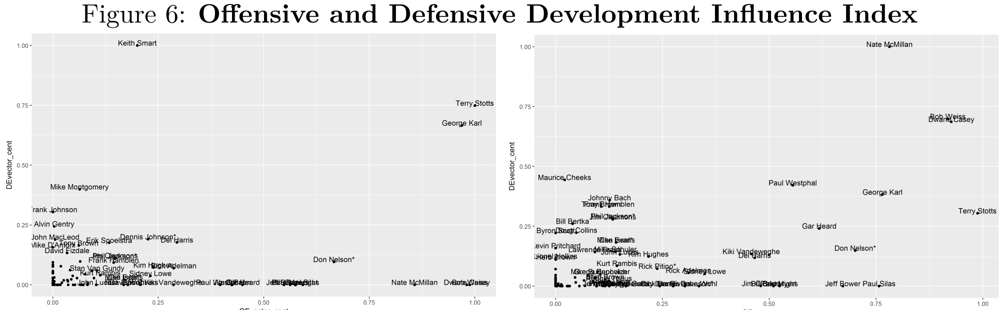
The following Figure 5 displays in similar manner the relationship between weighted in-degrees (direct influence) and DII (direct+indirect influence). We can observe a weak positive correlation (~0.3 at most) which is low compared to the fact that DII is derived from the weighted in-degrees. It shows that direct influence does not measure the overall influence of a coach well, so accounting for the spillovers is necessary for proper measurement.

Figure 5: **Development Influence Index and weighted in-degree for coaches**



We can inspect both offensive and defensive DII by looking at Figure 6. I only consider the baseline and the cleared version as the standard version is almost the same as the baseline. We can see that dismissal of small weighted links influences the results here for some coaches (Nate McMillan), but in general only a few coach could develop players effectively in both dimensions: most coaches have an advantage either on offense or defense, which is not surprising.

Figure 6: **Offensive and Defensive Development Influence Index**



It is also possible to group coaches based on their Development Influence Indices. We can define the following Euclidean type distance:

$$dist_{j1,j2} = \sqrt{(ODII_{j1} - ODII_{j2})^2 + (DDII_{j1} - DDII_{j2})^2}$$

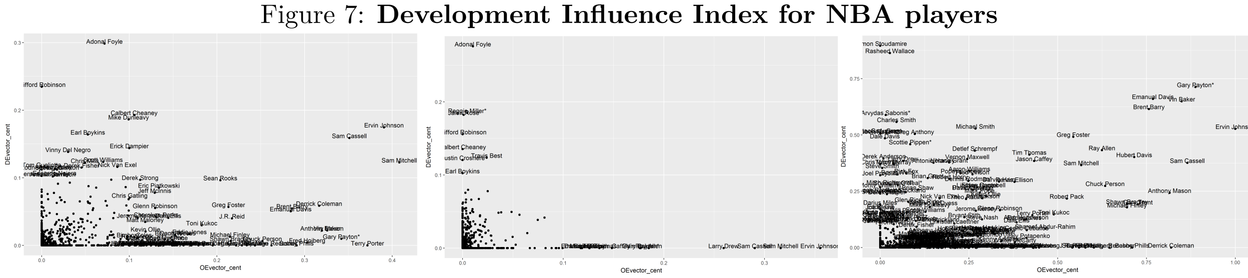
based on the offensive and defensive influence of coach $j1$ and $j2$. Based on this distance, we can cluster coaches using the K-means algorithm. The top tier coaches for the three different definitions of baseline, standard and cleard weights are displayed in Table 2. We can see that zeroing out small weights actually decreases the distance between top coaches and the rest. There are 5 coaches appearing under all definitions showing the robustness of the results.

Table 2: **Top tier coaches based on their Development Influence Index and K-means**

Baseline	Standard	Cleared
Bob Weiss	Bob Weiss	Bob Weiss
Dwane Casey	Dwane Casey	Dwane Casey
George Karl	George Karl	George Karl
Keith Smart	Keith Smart	Nate McMillan
Nate McMillan	Nate McMillan	Terry Stotts
Terry Stotts	Terry Stotts	Don Nelson
		Gar Heard
		Jeff Bower
		Paul Silas
		Paul Westphal

I do not discuss the DII values computed for players in detail: the most important aspect is that there are too many players to meaningfully evaluate the validity of the results. We can also see that these results are less robust across the three weight definitions. However if we accept the methodology that I described previously, we can derive some information from these as well. Figure 7 displays offensive and defensive DII for NBA players in case the reader wants to look at it in more depth.

Figure 7: **Development Influence Index for NBA players**



Discussion and conclusion

There are certain problems with the approach which I could not handle partially due to methodological limitations. In earlier attempts it became obvious that the results were very sensitive to number of total links of the mentors, which I controlled for by including the total number of links as normalizing factors in the weights. This solved the problem that players who were traded often and coaches who coached a lot of teams appeared to be overly influential, but it mechanically makes it impossible that teams would hire people just to teach young players on the team (e.g. case of Kevin Ollie).

There are also some other aspects to consider: it could happen that certain teams are more willing to sacrifice player development for present winning. This would result in creating bias such that coaches hired here would not have the autonomy to sacrifice sufficient amount of time to develop players. This could however open up a new research line to find out teams' ability to develop players.

Also it is quite plausible that player development is not linear: very talented players might need special coaching. I do not account for that fact. And although I intend to control for initial talent and prime production, there can be more sophisticated ways of measuring player development.

The greatest added value in my term paper is the creation of the dataset and the network, while creating a method that can be universally applied across sports (and even some other fields). Although the method has some issues, we can see that it produces some results which seem to be plausibly describing the phenomenon at hand.

References

Conti, G., Heckman, J. J., & Pinto, R. (2016). The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour. *The Economic Journal*, 126(596).

Definition of win shares: http://www.basketball-reference.com/about/ws.html. Access: 04.01.2017.

Source of data: http://www.basketball-reference.com