UNIVERSITY OF CALIFORNIA

Los Angeles

A Social Network Analysis of NBA Players

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Biancheng Wang

2019

© Copyright by
Biancheng Wang
2019

ABSTRACT OF THE THESIS

A Social Network Analysis of NBA Players

by

Biancheng Wang
Master of Science in Statistics
University of California, Los Angeles, 2019
Professor Mark S. Handcock, Chair

In this paper, we focus on the relationship of public figures using statistical network analysis methods. Specifically, we analyze the network of Instagram following relationship of NBA All-Stars in the most recent 5 seasons. Based on Latent Space Position Modeling, we find that there might be an International gang and a veteran "gang" within this network although it is not so clear. Triad Census Model implies that this network is more likely to be transitive. With the help of Exponential-family Random Graph Models (ERGM), we analyze this network from 4 aspects, including Demographical Characteristics, Social Media Characteristics, Business Factor and Basketball Factor. We find strong Nationality and Team homophily effects. As for Business Factors, it does show some Brand homophily effect. Also, players who have more NBA experience, more honors and are more active on social media are more likely to form a tie with others.

The thesis of Biancheng Wang is approved.

Ying Nian Wu

Hongquan Xu

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2019

To my parents

without whom I cannot be here

TABLE OF CONTENTS

1 Introduction **1**

1.1 Background 1

1.2 Data Description 2

2 Methods **4**

2.1 Introduction to Social Network 4

2.2 Measurements of Graphs 5

2.2.1 Density and Degree 5

2.2.2 Centrality 6

2.2.3 Social Balance and Transitivity 7

2.3 ERGM, ERNM, and TERNM 8

2.3.1 ERGM 8

2.3.2 ERNM 9

2.3.3 TERNM 9

3 Analysis and Results **10**

3.1 Network Description 10

3.2 Degree and Centrality Analysis 12

3.3 Latent Space Models 16

3.4 Triad Census Model 20

4 Further Analysis by ERGM **28**

4.1 Effects by Demographical Characteristics 29

4.2 Effects by Social Media Characteristics 31

4.3 Effects by Business Factor 32

4.4 Effects by Basketball Factor 33

5 Conclusion **37**

References **38**

LIST OF FIGURES

2.1 Transitivities of Triads[6] 7

3.1 Instagram Following Network of NBA All-Stars 10

3.2 Teammates Network of NBA All-Stars 11

3.3 Correlation Plot 15

3.4 Instagram Following Relationship of NBA All-Stars (International Players marked) 17

3.5 Instagram Following Relationship of NBA All-Stars (Frequent NBA All-Star Game Attendants marked) 18

3.6 Instagram Following Relationship of NBA All-Stars (3,4,5 and 6 sub-groups Bayesian posterior clusters marked) 20

3.7 Goodness-of-fit Diagnostics for the Model in Column (1) of Table 3.5 24

3.8 Goodness-of-fit Diagnostics for the Model in Column (2) of Table 3.5 25

3.9 MCMC Diagnostics for the Model in Column (1) of Table 3.5 26

3.10 MCMC Diagnostics for the Model in Column (2) of Table 3.5 27

4.1 Correlation Plot for all Numerical Variables 28

LIST OF TABLES

1.1 Node Attributes of Network 3

3.1 Degree and Centrality of NBA All-Stars 12

3.2 Statistical Summary of Degree and Centrality 15

3.3 Triad Census 21

3.4 Social Balance and Transitivity Analysis with transitive and cyclic triads 21

3.5 Social Balance and Transitivity Analysis using Triad Census Model 22

4.1 Effects by Demographical Characteristics conditional on Team homophily effect 29

4.2 Effects by Social Media Characteristics conditional on both Team and Nationality homophily effect 31

4.3 Effects by Business Factor conditional on both Team and Nationality homophily effect 32

4.4 Effects by Basketball Factor (Capacity and Experience) conditional on both Team and Nationality homophily effect 34

4.5 Effects by Basketball Factor (Others) conditional on both Team and Nationality homophily effect 36

CHAPTER 1

Introduction

1.1 Background

Much previous research has analyzed friendship networks, e.g., adolescent friendship[3] using The National Longitudinal Study of Adolescent Health data (Add Health)[1] and the relationships between monks using Sampson's monastery data. However, those relationships are not so complex when comparing with the network of public figures. Indeed, one reason they may have been studied so much is their relative simplicity. Adolescents networks often have few covariates while the grouping of monks is quite clear. One might be curious about what kind of factors may influence the relationship network among public figures.

In this paper, we choose professional basketball players, specifically, NBA All-Stars from the most recent 5 years, as our research objects. As the best athletes in the world in this field, it is obvious that they know each other. We often hear news about some of NBA players practicing together during the off-season, which might imply that there exist some small groups within these talented individuals. Since it is hard to figure out the true friendship between each other, so we focus on their relationships as presented on social media. For example, who do they follow on Instagram?

Since this is not an actual friendship network, more aspects should be considered in this case. We think the relationships on social media might be related to the following four aspects: Demographical Characteristics, Social Media Characteristics, Basketball Factors and Business Factors. We mainly use Exponential-family Random Graph Models (ERGM), a probability model for adjacency matrices, to answer how they influence the formation of

[1] www.cpc.unc.edu/projects/addhealth

ties in a network between public figures.

In the next section, we describe our dataset and give visual and numeric overview of our network. Chapter 2 will give more technical details of network analysis models. Chapter 3 and Chapter 4 are dedicated to the social network analysis and we try to figure out some patterns within this network and find characteristics which may impact the relationship. Finally, we conclude and give more discussion in Chapter 5.

1.2 Data Description

The dataset we work on in this paper is mostly collected from Wikipedia and Instagram on April 2nd, 2019. There are 50 NBA All-Stars from 2014-2019 seasons. However, Kawhi Leonard and Nikola Vucevic hardly use social media and do not have verified Instagram accounts while Nikola Jokic deleted his account last year. Chris Bosh, Tim Duncan and Kobe Bryant have already retired before this season. Therefore, we only collect data of 44 NBA All-Stars in our network.

Our dataset contains two networks. The first one is a network describing the following relationship between those players. In the next following sections, we will give some specific analysis regarding this network.

The second one is a network of teammates relationship. Not only this season but also previous experience of being teammates in NBA are included in this network. Although we mainly focus on the first network, we still think plotting Teammates network might be helpful to give us an additional information and better understanding about the relationship among those players.

---

[2]Years in NBA.

[3]The Number of Draft Pick which means the players Draft Order.

[4]Data Source: https://hoopshype.com/reps/

[5]Data Source: Ibid.

[6]There are 5 types of position in basketball, including Point Guard(PG), Shooting Guard(SG), Small

Table 1.1: Node Attributes of Network

Demographical Characteristics	Race	Basketball Factor	University
	International		Experience_NBA[2]
	Nationality		Num_of_Team_Change
	State		Draft[3]
	Age		Num_of_MVP
Social Media Characteristics	Ins_Post		Num_of_All_Star
	Ins_Follower		Num_of_Champion
	Ins_Follows		Team
Business Factor	Agent[4]		Division
	Agency[5]		Conference
	Brand		Position_detailed[6]

Table 1.1 gives us all 22 original variables we have, which are regarded as node attributes in network analysis. We can divide them into 4 categories. Within the Basketball Factor, there still exists different types of covariates. University, years in NBA and the number of team changed can stand for his overall basketball experience while the number of draft pick, MVP, All-Star and Champion can be a set of representative variables related to their capacity. We may also define some new variables based on these original variables in the next chapter.

Forward(SF), Power Forward(PF), and Center(C).

CHAPTER 2

Methods

2.1 Introduction to Social Network

Network is a collection of entities, together with a set of relations on those entities. Entities are also named as nodes or vertices.

Relations are also known as edges or ties. These relations can be characterized as directed or undirected. An undirected relation has only one value per pair while a directed relation has two values per pair representing the the perspective of each pair member. Also, relations can be characterized as binary or valued. A binary relations takes only two values while a valued relation can have more then two values.

There are some features of social relations. Individual heterogeneity in the propensity to form or receive ties, standing for the popularity or attractiveness of entities. Homophily by actor attributes means there might be higher propensity to form ties between actors with similar attributes. Transitivity of relationships is a phenomenon that friends of friends have a higher propensity to be friends. Besides, sometimes people establish and maintain balance in their relationships. We would like to test the existence of those features with the help of some statistical measures of graphs introduced below.

2.2 Measurements of Graphs

2.2.1 Density and Degree

For a n node undirected graph, it has $n(n-1)/2$ values. The density of this graph is the mean of the tie values, which is

$$\overline{Y} = \frac{1}{n(n-1)} \sum_{i \neq j} y_{ij} \tag{2.1}$$

where y_{ij} is the tie value between node i and node j.

For a n node directed graph, the density of the graph will be

$$\overline{Y} = \frac{1}{n(n-1)} \sum_{i,j} y_{ij} \tag{2.2}$$

The density is a global measure of the sociability of the graph. When focusing on node level summarizing, the density of a node is the mean of the tie values measuring the sociability of a node. Some nodes are more social than others.

For an undirected graph, the density of node i is

$$\overline{Y}_i = \frac{1}{n-1} \sum_{j:j \neq i} y_{ij} \tag{2.3}$$

For a directed graph, we have

$$\begin{aligned} \overline{Y}_i^{\circ} &= \tfrac{1}{n-1} \textstyle\sum_{j:j \neq i} y_{ij} \\ \overline{Y}_j^{\mathrm{i}} &= \tfrac{1}{n-1} \textstyle\sum_{i:j \neq i} y_{ij} \end{aligned} \tag{2.4}$$

For binary relations, nodal heterogeneity can be described by nodal degrees. High degree positions are influential, but also may be subject to a great deal of influence or stress from others.

For undirected relations, the degree of a node is the node's number of ties.

For directed relations, the indegree of a node is the nodes number of incoming ties while the outdegree of a node is the nodes number of outgoing ties.

2.2.2 Centrality

The core question here is how do individual positions vary? Some nodes seem to be more "central" than others in the network. Therefore, we need a measure of centrality.

Koschtzki D. et al. (2005) [4] attempted to classify different centrality measures into 4 types, including Reach centrality, Flow Mediation centrality, Vitality centrality and Feedback centrality.

Reach measures centrality based on ability of ego to reach other vertices. Flow Mediation centrality is based on quantity of walks passing through ego. Vitality centrality focus on effect of removing ego from the network. Feedback is a centrality of ego defined as a recursive function of alter centralities.

Four ways we used to measure centrality in this research include degree centrality, eigenvector centrality, closeness centrality and betweenness centrality. The things they are measuring are not the same.

Degree centrality measures how connected an entity is by counting the number of directed links it has to others in the network.

$$C_D\left(n_i\right) = \frac{\sum_{j:j\neq i} y_{ij}}{n-1} \tag{2.5}$$

Betweenness centrality counts the number of paths passing through each entity in the network. Therefore, it quantifies extent to which position serves as a bridge.

$$C_B\left(n_i\right) = \sum_{j<k} \frac{g_{jk}\left(n_i\right)}{g_{jk}} \tag{2.6}$$

where $g_{jk}$is the number of j, k geodesics, $g_{jk}\left(n_i\right)$ is the number of j, k geodesics including i.

Closeness centrality measures the proximity of an entity to others.

$$C_C\left(n_i\right) = \sum_{j=1}^{n} \frac{1}{d\left(n_i, n_j\right)} \tag{2.7}$$

where $d\left(n_i, n_j\right)$ is the geodesic distance from i to j.

Eigenvector centrality measures how connected an entity is and how much direct influence

it might have over other connected entities.

$$c_i^e = \frac{1}{\lambda} \sum_{jj \neq i} y_{i,j} c_j^e \tag{2.8}$$

2.2.3 Social Balance and Transitivity

In a balanced triads, a friend of a friend is a friend. For directed relations, actors seek out transitive relations and avoid intransitive relations. A triple is transitive if $i \to k$ when $i \to j$ and $j \to k$.

Directed dyads have three states, including mutual, asymmetric and null. Every triad is composed of 3 dyads. We can identify traids based on the number of each type. Then we have 16 types of triads which are shown in Figure 2.1. Figure 2.1 also gives us corresponding types of transitivity for all 16 triads. This can help us have a better understanding on the transitivity of network.

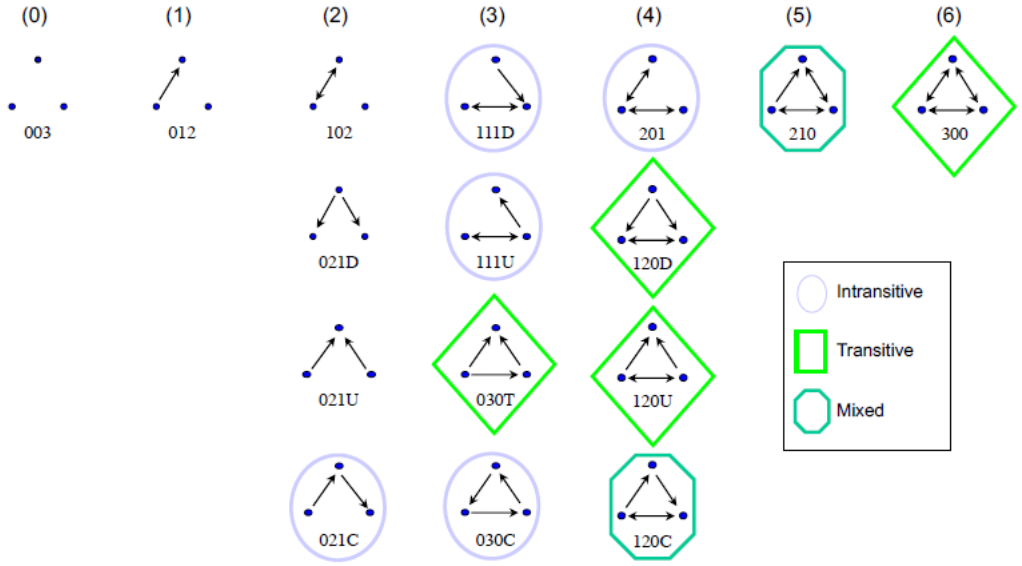


Figure 2.1: Transitivities of Triads[6]

2.3 ERGM, ERNM, and TERNM

2.3.1 ERGM

Random graphs, where the connections between nodes are considered random variables, have wide applicability in the social sciences. Exponential-family Random Graph Models (ERGM) have already shown themselves to be a useful class of models for representing complex social phenomenon.

When we try to construct statistical models for social network, we regard network ties as the outcome of unobserved social processes which tend to be local and interactive.

Therefore, we assume network ties are variable and nodes are fixed and the observed network is generated by a stochastic process. We can estimate parameters of the process and do some inference.

As for goodness of fit issue, we can use some traditional diagnostics, including Akaike information criterion (AIC), Bayesian information criterion (BIC) and estimation diagnostics looking at MCMC performance. Besides, once we have estimated the parameters of probability distribution, we can draw graphs at random and compare their characteristics with those of the observed network. If the model is good, then sampled graphs will resemble observed network[2].

The Exponential-family Random Graph Models (ERGM) can be represented as

$$P(Y = y) = \frac{\exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\}}{c(\theta)} \tag{2.9}$$

where $\theta_{1,2\ldots k}$ are parameters, $g_{1,2\ldots k}(y)$ are statistics, and $c(\theta)$ is a normalizing constant:

$$c(\theta) = \sum_{y \in \mathcal{Y}} \exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\} \tag{2.10}$$

Also, we can rewrite it as follows in a log-odds form:

$$\log \frac{P\left(Y_{ij} = 1 | y_{ij}^c\right)}{P\left(Y_{ij} = 0 | y_{ij}^c\right)} = \sum_{k=1}^{K} \theta_k \left[g_k\left(y_{ij}^+\right) - g_k\left(y_{ij}^-\right)\right] \tag{2.11}$$

where y_{ij}^c is the rest of a graph excluding y_{ij}, y_{ij}^+ is the graph with $Y_{ij} = 1$ and y_{ij}^- is the graph with $Y_{ij} = 0$.

Then we can get some useful implications. Each unit change in g_k for tie (i, j) present increases the conditional log-odds of (i, j) by θ_k. The statistics g_k used here can be some summary measurement of networks, e.g., triad census, and node attributes.

2.3.2 ERNM

We can also model the attributes of the social actors as random variables, thus creating a random model, which we call Exponential-family Random Network Models (ERNM).

ERNM is a joint modeling of Y and X. Let $\mathcal{N}$ be the sample space of Y and X. We have

$$P_\eta(Y = y, X = x) = \frac{\exp\{\theta \cdot g(y, x)\}}{c(\theta, \mathcal{N})} \quad y, x \in \mathcal{N} \tag{2.12}$$

where θ is q-vector of parameters, $g(y, x)$ is q-vetor of network statistics, and $c(\theta, \mathcal{N})$ is a normalizing constant.

Then we have latent space model[9] using ERNM to find the latent class for each nodes.

2.3.3 TERNM

The model degeneracy[5] means only small range of graphs covered as parameters vary and it may hamper ERGM and ERNM. Tapered Exponential-family Random Network Models (TERNM) can retain the advantages of ERGM and ERNM but remove the degeneracy properties. Thus we can have a better goodness of fit.

These family of models can be represented as:

$$P(Y = y|\theta, \beta) = \frac{\exp\left\{\sum_{k=1}^K \theta_k g_k(y) - \sum_{k=1}^K \beta_k^{-2} \left(\mu_k(\theta, \beta) - g_k(y)\right)^2\right\}}{c(\theta, \beta)} \tag{2.13}$$

where $\beta > 0$ are vectors of hyper parameter and mean value parameters $\mu(\theta, \beta) = E_P(g(Y))$. The augmented term tapers the likelihood of configurations far from the mean μ.

CHAPTER 3

Analysis and Results

3.1 Network Description

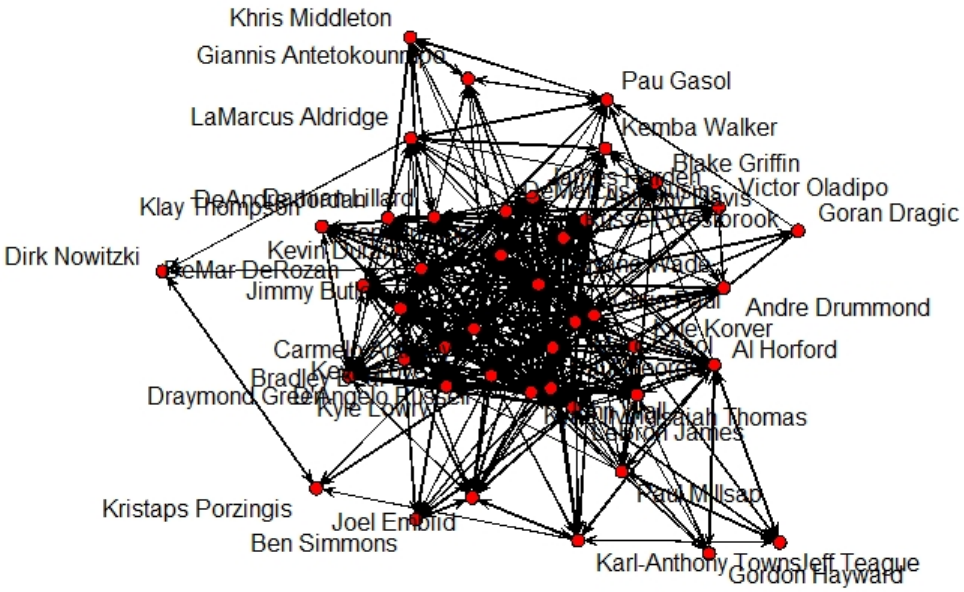


Figure 3.1: Instagram Following Network of NBA All-Stars

Instagram Following Network is a fully observed binary network, which is shown in Figure 3.1. This network is a directed network with 44 vertices and 624 edges. Since there are 1936

possible edges, the density of ties in this network is 0.322. There is only one component which means there is no one in this network who does not follow anyone else and is not followed by others. The average of geodesic distances is 1.747. This is a measure of the average length of the shortest path between two players. There is a large cloud with dense ties in the center of the Figure 3.1 while some players seem to have less ties.

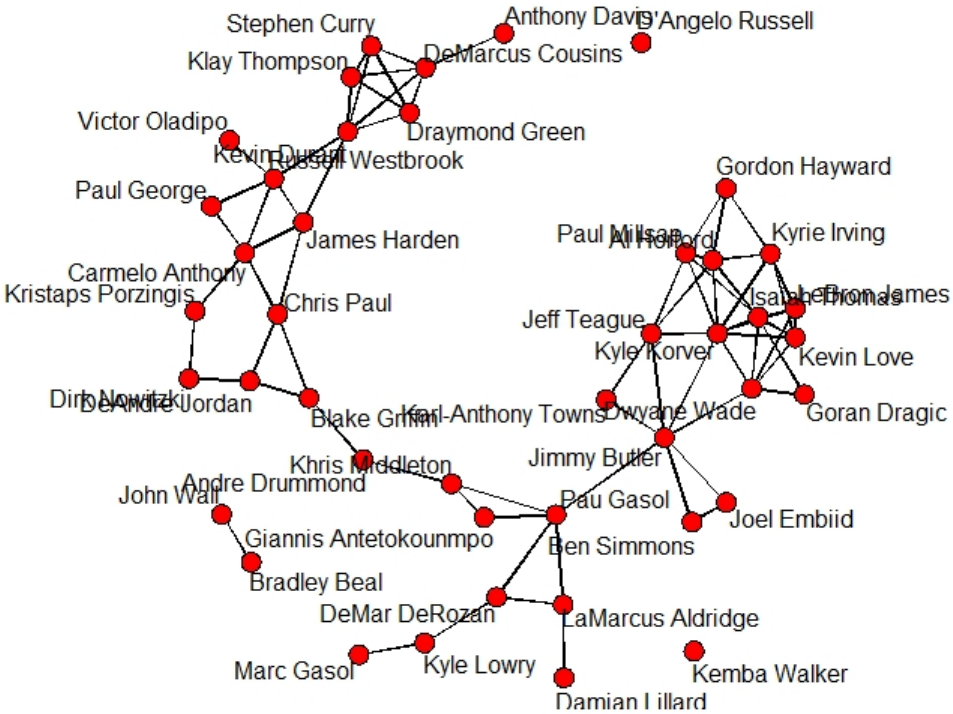


Figure 3.2: Teammates Network of NBA All-Stars

Teammates Network is an undirected network with 44 vertices and 75 edges. Since there are 968 possible edges for this undirected network, the density of ties is 0.077 which is much sparser than the first network. From Figure 3.2, we can find that there are 1 large component and 3 small components including 2 isolates, which are nodes not adjacent to any other nodes. Kemba Walker and D'Angelo Russell do not work with any other players

in this network.

3.2 Degree and Centrality Analysis

To capture the sociality of the network, we calculate the indegree and outdegree for each player in 2nd and 3rd columns of Table 3.1. In our network, the indegree of a player stands for how many players in this network follows him on Instagram while the outdegree stands for how many players he follows. Therefore, we think indegree can tell us how popular this player is or at least how popular he seems to be on the social media.

Table 3.1: Degree and Centrality of NBA All-Stars

	Player name	Indegree	Outdegree	Eigenvector Centrality	Closeness Centrality	Betweenness Centrality
1	LeBron James	25	19	0.194	0.642	61.400
2	DeMarcus Cousins	22	20	0.193	0.642	28.707
3	Anthony Davis	19	16	0.159	0.597	12.501
4	Kevin Durant	20	20	0.183	0.642	76.519
5	Kyrie Irving	27	17	0.159	0.623	90.817
6	LaMarcus Aldridge	5	10	0.069	0.518	12.775
7	Bradley Beal	16	20	0.204	0.652	15.141
8	Goran Dragic	2	4	0.039	0.506	0.349
9	Andre Drummond	7	9	0.0.85	0.531	3.799
10	Paul George	28	28	0.260	0.741	82.088
11	Kevin Love	18	24	0.234	0.694	44.475
12	Victor Oladipo	9	7	0.069	0.494	4.969
13	Kristaps Porzingis	5	3	0.021	0.448	46.953
14	Kemba Walker	13	4	0.038	0.462	3.571
15	John Wall	24	24	0.222	0.683	90.932
16	Russell Westbrook	26	18	0.169	0.597	38.682

	Player name	Indegree	Outdegree	Eigenvector Centrality	Closeness Centrality	Betweenness Centrality
17	Stephen Curry	23	23	0.211	0.672	61.408
18	Giannis Antetokounmpo	7	2	0.005	0.358	3.132
19	DeMar DeRozan	17	17	0.183	0.614	16.789
20	Joel Embiid	13	10	0.096	0.551	14.385
21	James Harden	19	11	0.123	0.538	5.572
22	Jimmy Butler	19	17	0.157	0.597	44.136
23	Draymond Green	16	16	0.165	0.597	9.642
24	Al Horford	9	14	0.107	0.581	33.413
25	Damian Lillard	22	18	0.180	0.623	36.461
26	Kyle Lowry	17	20	0.206	0.652	10.379
27	Klay Thompson	9	8	0.092	0.524	1.219
28	Karl-Anthony Towns	8	7	0.054	0.538	16.643
29	Carmelo Anthony	25	21	0.207	0.652	89.005
30	Isaiah Thomas	18	15	0.162	0.597	50.946
31	Paul Millsap	4	14	0.134	0.597	9.425
32	Marc Gasol	5	29	0.203	0.754	62.868
33	Gordon Hayward	4	4	0.035	0.473	0.177
34	DeAndre Jordan	15	14	0.156	0.573	11.403
35	Pau Gasol	9	5	0.028	0.500	31.348
36	Chris Paul	22	21	0.199	0.662	41.676
37	Dwyane Wade	28	34	0.272	0.827	202.774
38	Ben Simmons	9	6	0.067	0.500	1.272
39	Blake Griffin	11	9	0.086	0.500	6.717
40	D'Angelo Russell	12	19	0.156	0.642	61.642
41	Dirk Nowitzki	3	1	0.001	0.312	1.591

	Player name	Indegree	Outdegree	Eigenvector Centrality	Closeness Centrality	Betweenness Centrality
42	Khris Middleton	4	7	0.059	0.506	27.629
43	Kyle Korver	5	14	0.131	0.581	13.279
44	Jeff Teague	5	5	0.037	0.483	5.389

We can find that the Top 5 players with largest indegree are Paul George, Dwyane Wade, Kyrie Irving, Russell Westbrook, Carmelo Anthony and Lebron James (tie with Carmelo Anthony) while Goran Dragic, Dirk Nowitzki, Paul Millsap, Gordon Hayward, and Khris Middleton has the least number of indegree. It is nature to think about what kind of factors will influence the popularity of a player. Carmelo Anthony, Dwyane Wade and Lebron James are from 2003 Draft Class which might be the best class in the NBA history and most people believe that they will be a part of NBA Hall of Fame without any doubt. However, some of those players seem to be not so popular have relatively less NBA All-Star experience. Besides, two of them are also international players. The mean of indegree is 14.18. It seems that those isolates in Figure 3.2 tend to have less indegree when comparing with others. This might imply that teammates are more likely to follow each other, which is very understandable.

To better understand the relationship between variables and popularity, a correlation plot is shown as below in the 1st row of Figure 3.3. We define some new variables here. Top_30, Top_14 and Top_3 respectively stand for whether a player is picked at the first-round pick, the lottery pick and the first 3 picks. Generally, gifted player will be picked early during NBA Draft. Change is whether the player have changed team through all his career. MVP and Champion are dummy variables indicating whether he has got those titles or not. We set the median value 4 as a threshold to identify if the player is frequently selected to attend NBA All-Star Game. Since most of players sign their shoes contract with Nike or Air Jordan which is a brand produced by Nike company, we add Nike variable which is whether his sponsor is Nike company.

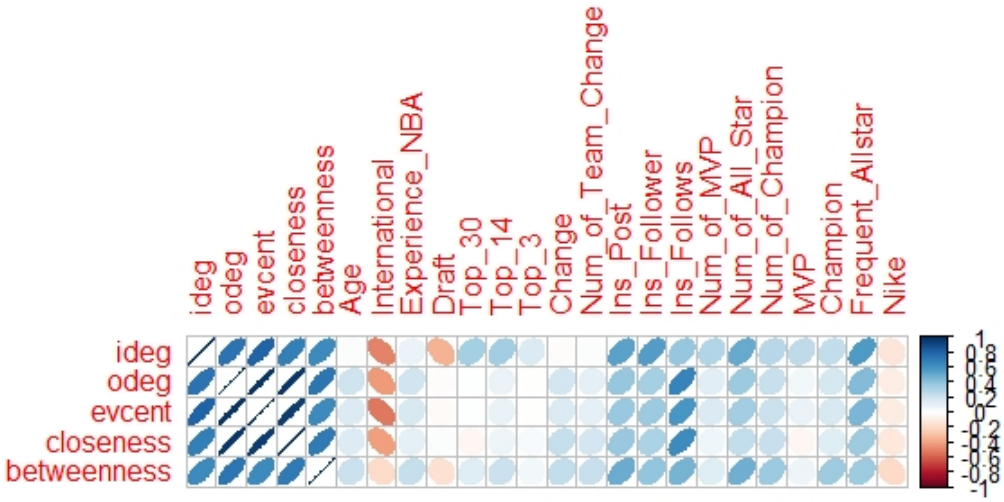


Figure 3.3: Correlation Plot

From Figure 3.3, we can find that International players tend to be less popular in the players network. From the aspect of social media, those with more Instagram followers, which means they are popular all over the world, are also popular among other players. Besides, players with more honors and more gift has more player followers. Those findings can help us to fit ERGM for further analysis.

We are also curious about how their positions in the network vary, in other words, the extent to which they are central in the network.

Table 3.2: Statistical Summary of Degree and Centrality

	Min	Median	Mean	Max	Overall
Indegree	2	14	14.18	28	0.4097

Outdegree	1	14.5	14.18	34	
Eigenvector centrality	0.00117	0.15606	0.13207	0.272	0.1432
Closeness centrality	0.3116	0.5972	0.579	0.8269	0.2596
Betweenness centrality	0.1769	16.7162	33.7273	202.7741	0.09578

Table 3.2 gives a summary of those different measures of centrality. We also include the overall centrality of the network in Table 3.2.

Besides, more details are given in Table 3.1. It has shown the indegree, outdegree and different centrality measures for each player. Also, we plot the correlation between centrality measures and other variables in Figure 3.3. We can find those different measures are highly correlated with each other. Compared with popularity, a quite similar conclusion can be made for Centrality. American players with more honors and more Instagram activities, including the number of Instagram Posts and Instagram Follows, are more centered in this network.

3.3 Latent Space Models

In this section, we want to find out whether there exist some sub-groups within this network. For example, is there any International gang or OG gang among those talented athletes? A latent space model [9] is the best way to capture the cluster of nodes. We can use it to represent network in a lower dimension and preserve their relative distance at the same time.

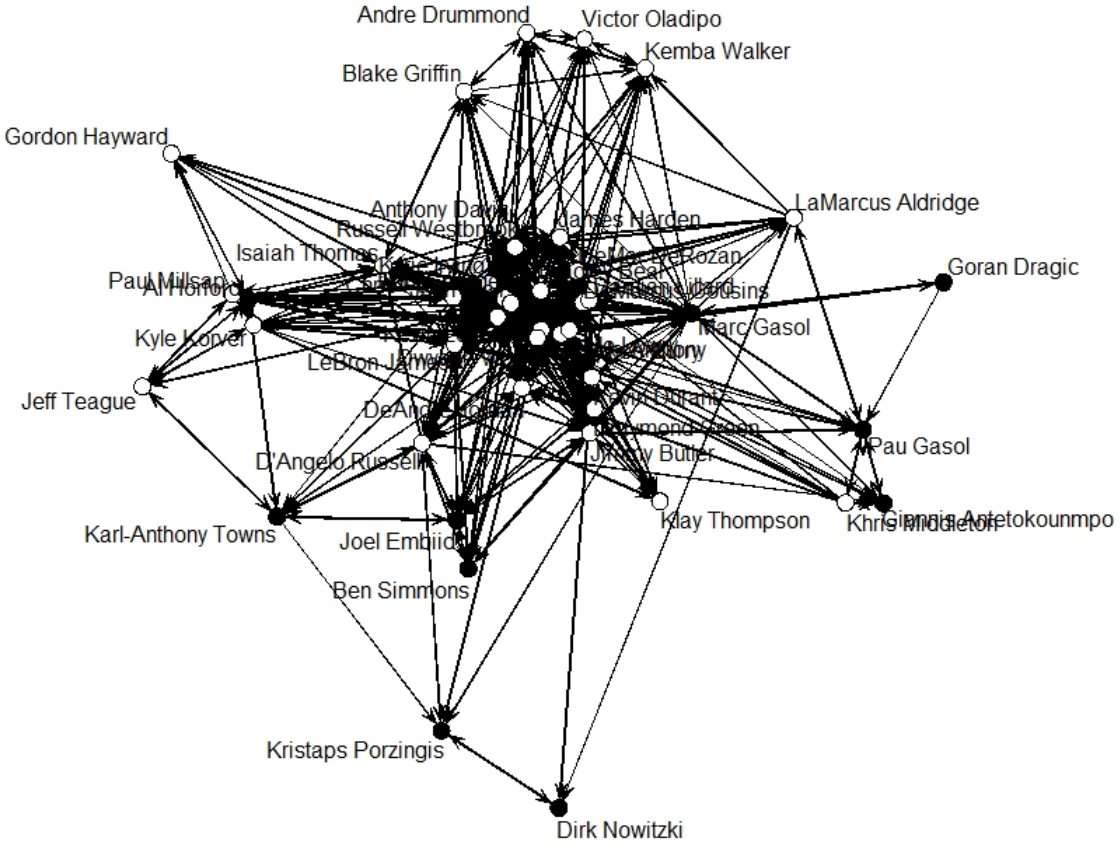


Figure 3.4: Instagram Following Relationship of NBA All-Stars (International Players marked)

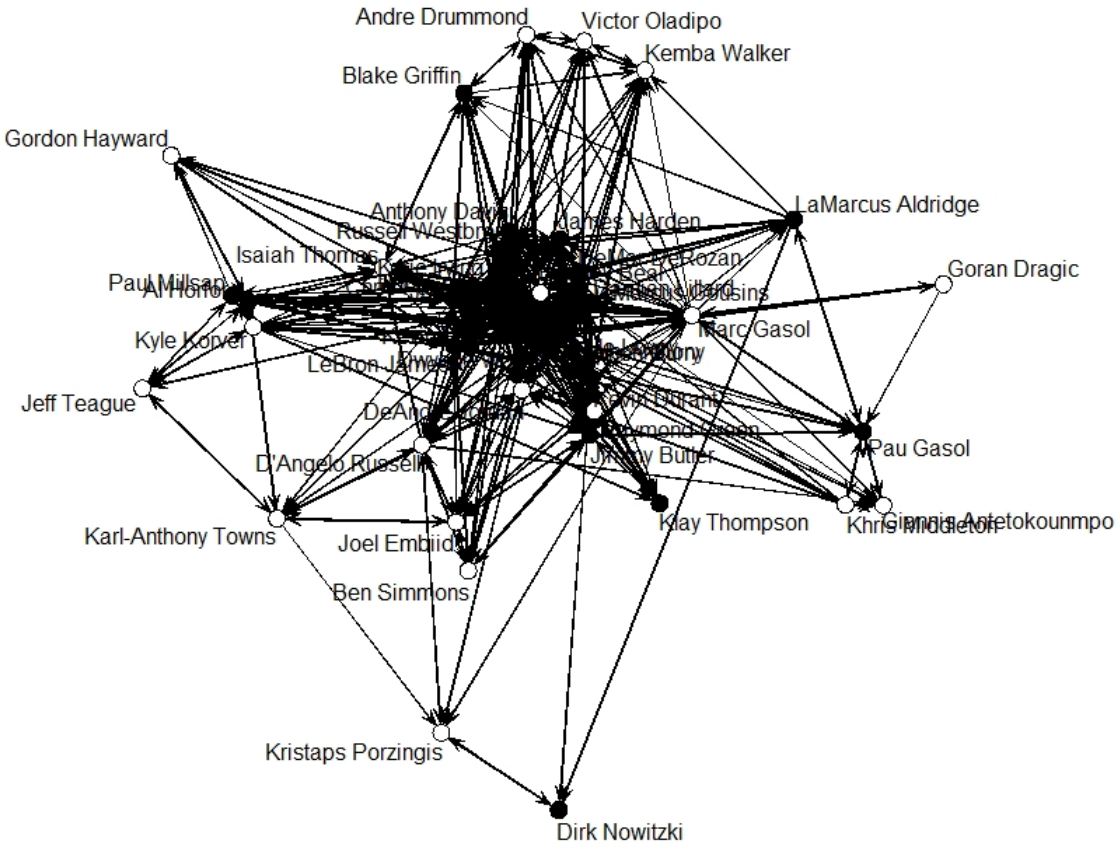


Figure 3.5: Instagram Following Relationship of NBA All-Stars (Frequent NBA All-Star Game Attendants marked)

Therefore, we fit a two-dimensional latent space model first. Since we are more interested in the International players group versus American players group and veteran groups versus rookie groups, we plot results of models with International players marked black in Figure 3.4 and with frequent NBA All-Star Game attendants marked black in Figure 3.5.

From these two figures, it is hard to find clear groups in this network. However, we can still find some patterns with the help of marks. Many players are clustered at the middle of the plot. Two isolates in Figure 3.2 are relatively far from the center.

International players which has black marks in Figure 3.4 are mainly located at the lower part and right part of the figure. But we can find LaMarcus Aldridge is closer to those International players. That might be caused by the culture of his team San Antonio Spurs

where there are many International players in this decade. Khris Middleton from Milwaukee Bucks locates very close to his two International All-star teammates. Also, Al Horford, a Dominican professional basketball player seems to be close to the American group. That is also reasonable since he moved to the United States with his family when he was in the middle school and played his high school and university games in this country. Spanish player Marc Gasol also moved as teenager with his parents to the United States and his position is also more "central" than other International players. There is a common characteristic among International players who are very far from the American group, e.g., Goran Dragic, Pau Gasol, Kristaps Porzingis and Giannis Antetokounmpo. They did not have either high school or university level basketball experience in the United before entering NBA. Thus, although the group is not so clear, we can find some homophily effects based on the plots.

From Figure 3.5, the clustering is not as clear as what is shown in Figure 3.4. There are still some patterns can be found as follows. Most of players in the middle have more NBA All-Stars experience. The upper groups of players are all Eastern Conference players who just have limited NBA All-Stars experience and has no superstar teammates. The lower left part are all International rookies. Besides, an interesting fact is that the far left group of players have played for Atlanta Hawks before together.

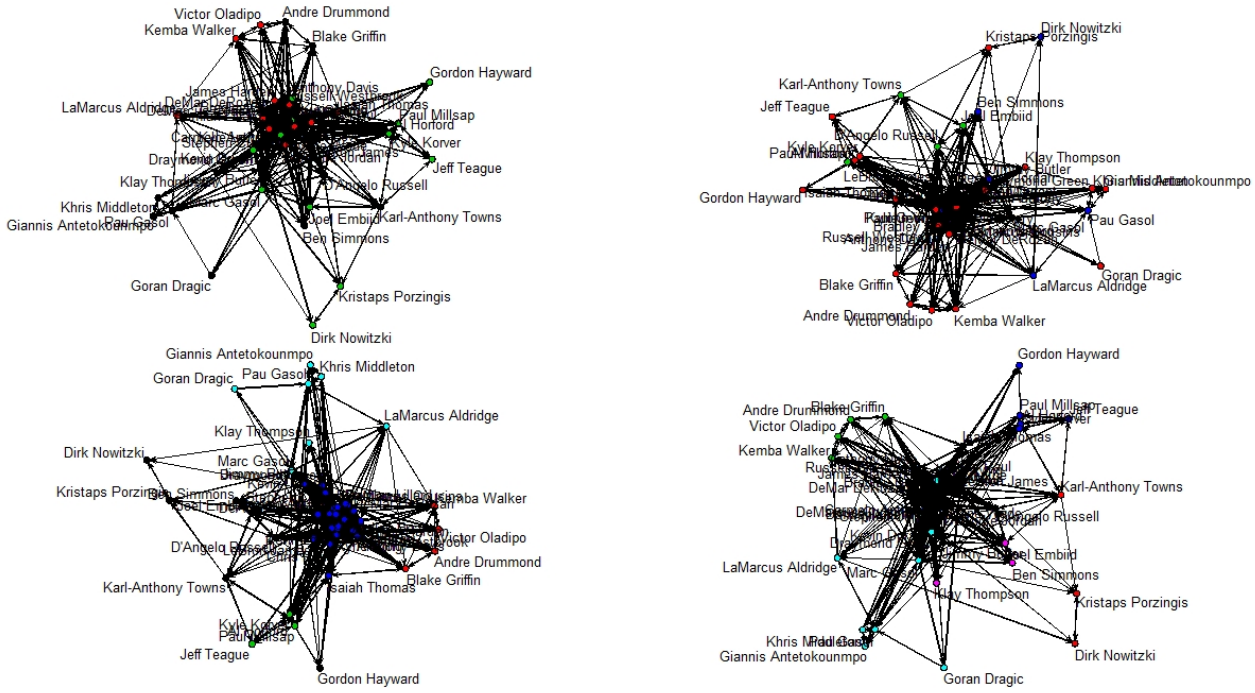


Figure 3.6: Instagram Following Relationship of NBA All-Stars (3,4,5 and 6 sub-groups Bayesian posterior clusters marked)

For further analysis, we try to fit two-dimensional latent cluster model with different number of sub-groups from 3 to 6 and plot them in Figure 3.6. But the result is not so convincing especially when looking at the posterior probability. Some of players are quite ambiguous to be classified into any group using all these models.

Therefore, we think the network of public figures showed on the social media is too complex and we cannot simply divide them into clear sub-groups. But the findings we get from the visualization of network by latent space model do help.

3.4 Triad Census Model

Transitivity has been found in many friendship networks. How about this relationship among public figures? Since it is just a kind of relationship shown on the social media, which might be related to business and other factors instead of purely true friendship, the results may be different. In this section, we want to answer this question regarding social balance and

transitivity in this network using ERGM.

Table 3.3: Triad Census

Triad	003	012	102	021D	021U	021C	111D	111U
Count	3504	1725	3467	274	122	92	700	646
Triad	030T	030C	201	120D	120U	120C	210	300
Count	70	1	1210	194	77	44	486	632

Table 3.3 shows the number of all 15 types of triad appearing in this network. It is obvious that this network is not balanced since a graph is balanced if all subsets of 3 nodes are balanced.

Table 3.4: Social Balance and Transitivity Analysis with transitive and cyclic triads

	(1)	(2)
mutual	1.547^{***}	1.437^{***}
	(0.140)	(0.145)
nodematch.Team		3.730^{***}
		(1.002)
ttriple	0.020^{**}	0.021^{**}
	(0.007)	(0.007)
ctriple	-0.199^{***}	-0.203^{***}
	(0.026)	(0.025)
AIC	2424	2351
BIC	2440	2373
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 3.4 above gives us the results of fitting ERGM including triad statistics. The model in Column (2) is conditional on the Team homophily effects. The second model has

a better fit based on AIC and BIC. The significant positive coefficient of nodematch.Team implies that players in the same team are more likely to follow each other. Considering the momentum of the team, players, especially the superstar, need to tell the public that they are a consolidated team even though their relationships are not such close. Following other superstars in the same team can be regarded as a way to show their solid and unified relationships.

In both two models, since the coefficient of the transitive triad statistics is significantly positive while the coefficient of the cyclic triads is significantly negative, it appears to be a general preference for balanced friendship ties.

Table 3.5: Social Balance and Transitivity Analysis using Triad Census Model

	(1)	(2)
edges	-0.426^{***}	-0.569^{***}
	(0.048)	(0.091)
nodematch.Team	3.309^{***}	2.998^{***}
	(0.916)	(0.888)
triadcensus.012	-0.075^{***}	
	(0.022)	
triadcensus.102	-0.083^{***}	
	(0.014)	
triadcensus.021D	-0.015	
	(0.041)	
triadcensus.021U	0.039	
	(0.073)	
triadcensus.021C	-0.390^{***}	-0.523^{***}
	(0.102)	(0.101)
triadcensus.111D	-0.144^{***}	-0.082^{**}
	(0.032)	(0.031)

	(1)	(2)
triadcensus.111U	-0.152^{***}	-0.090^{**}
	(0.032)	(0.032)
triadcensus.201		-0.082^{***}
		(0.022)
triadcensus.300		0.219^{***}
		(0.034)
AIC	2233	2198
BIC	2283	2236
Note:	*p<0.05; **p<0.01; ***p<0.001	

To add more evidence, we also use Triad Census model shown in Table 3.5. Since we find a strong Team homophily effect, we still fit the model conditional on such effect. Tapered ERGM is used here for a better fit.

First, we fit ERGM with 7 types of triad census which contributes to the large proportion of triangles in the network. According to the results in Column (1) of Table 3.5, Team homophily effect is still significantly positive. It seems that there appear to be a general preference for transitive friendship ties since the coefficients of the number of intransitive triads, e.g., 021C, 111D and 111U, are significantly negative.

Next, we include all triad census into the model and then select the most significant ones to improve the fitting of models. The AIC and BIC of the model in Column (2) of Table 3.5 are better than the previous one.

From the goodness-of-fit diagnostics in Figure 3.7 and Figure 3.8, we can find that the second model does perform better in edge-wise shared partners and minimum geodesic distance.

According to Figure 3.9 and Figure 3.10, MCMC diagnostics also look pretty good and they are converged.

Besides what we found in the previous model, the number of transitive triad 300 has

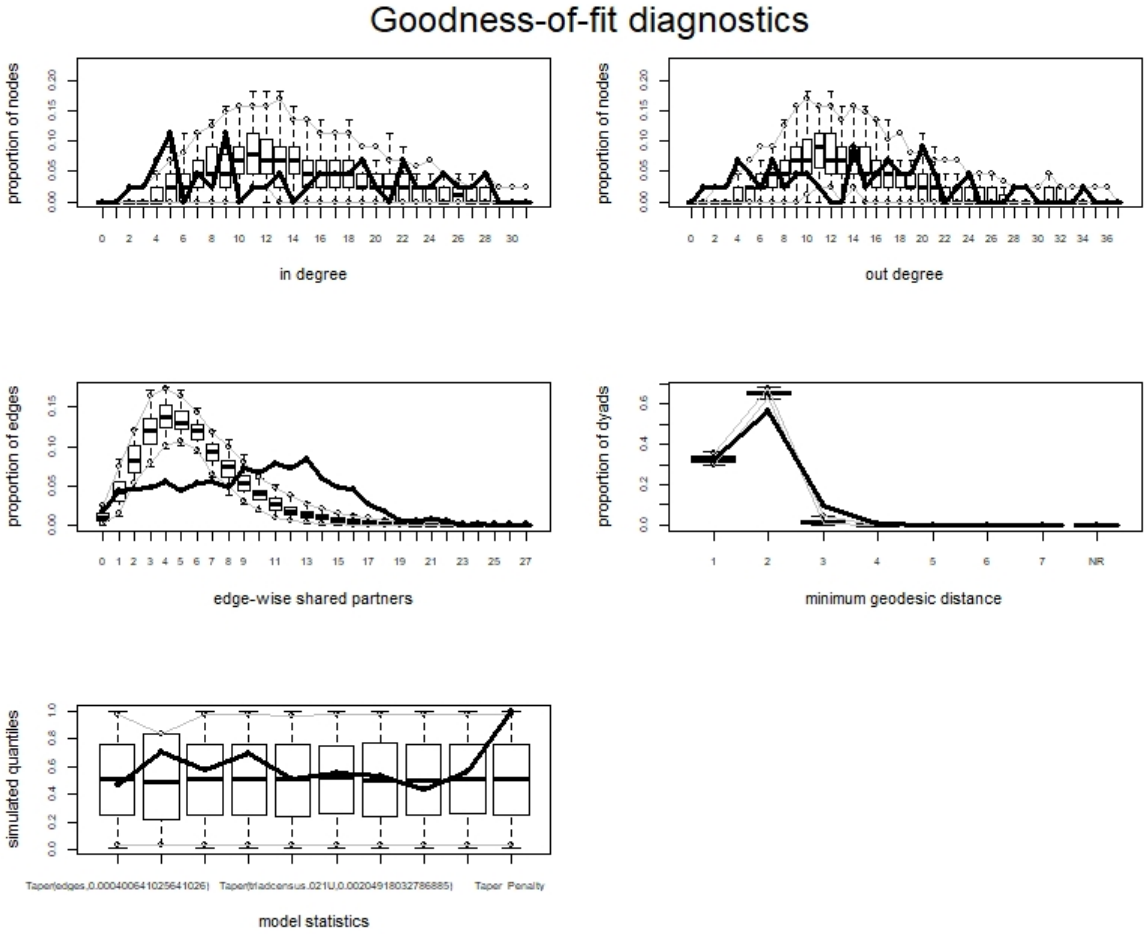


Figure 3.7: Goodness-of-fit Diagnostics for the Model in Column (1) of Table 3.5

a significantly positive coefficient while the number of another intransitive triad 201 has a significantly negative coefficient, implying that this network is more likely to be transitive. This means that there is a tendency that a friend of your friends is also your friend.

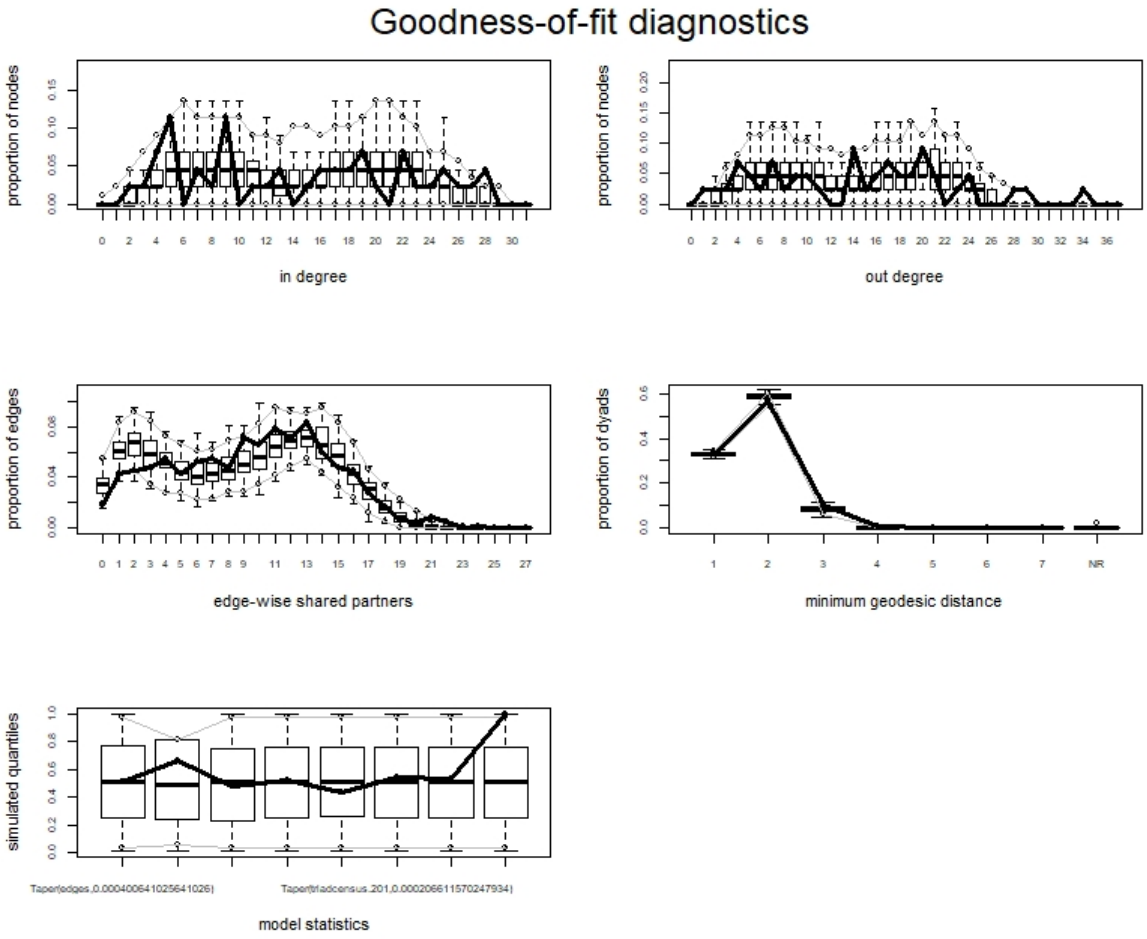


Figure 3.8: Goodness-of-fit Diagnostics for the Model in Column (2) of Table 3.5

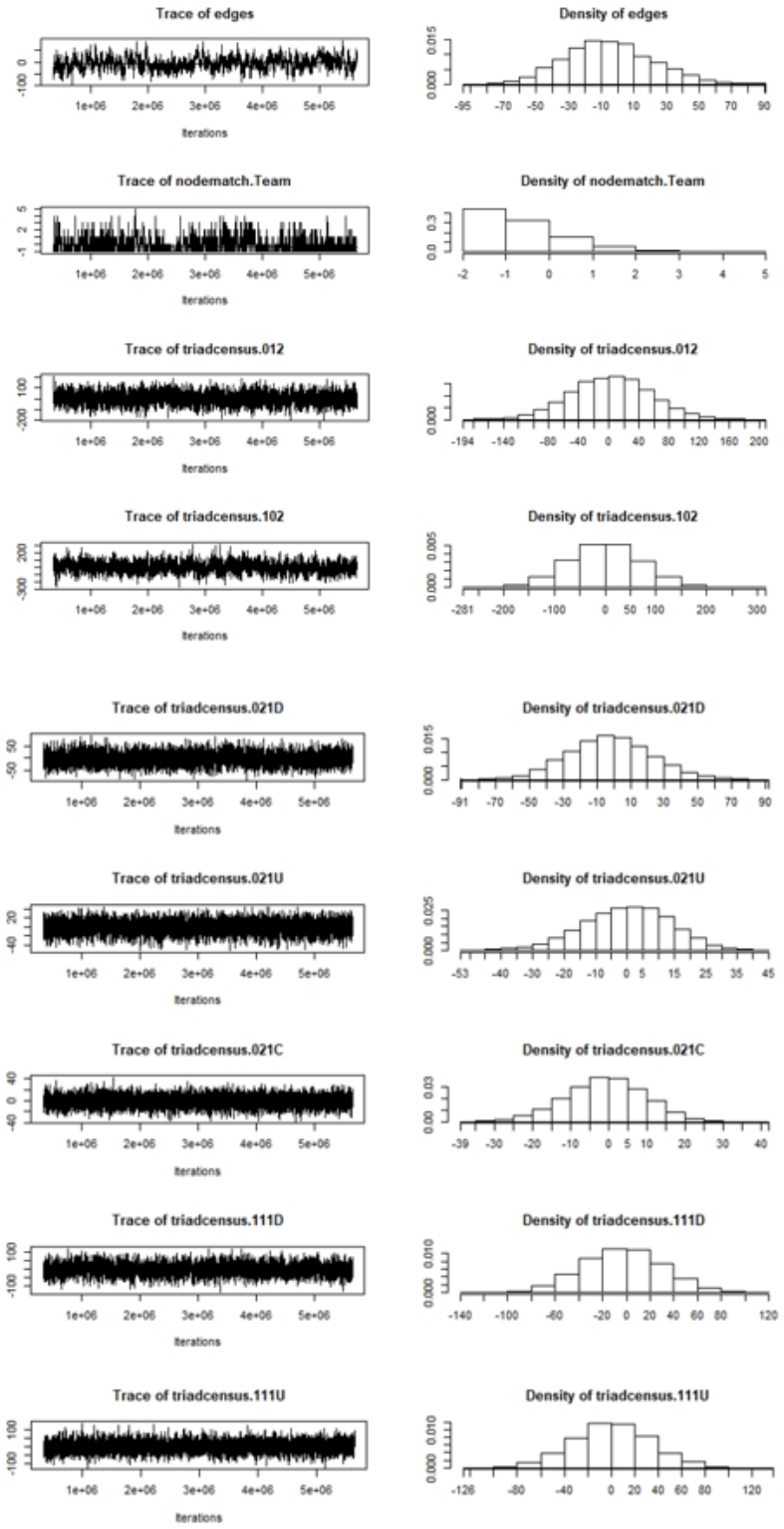


Figure 3.9: MCMC Diagnostics for the Model in Column (1) of Table 3.5

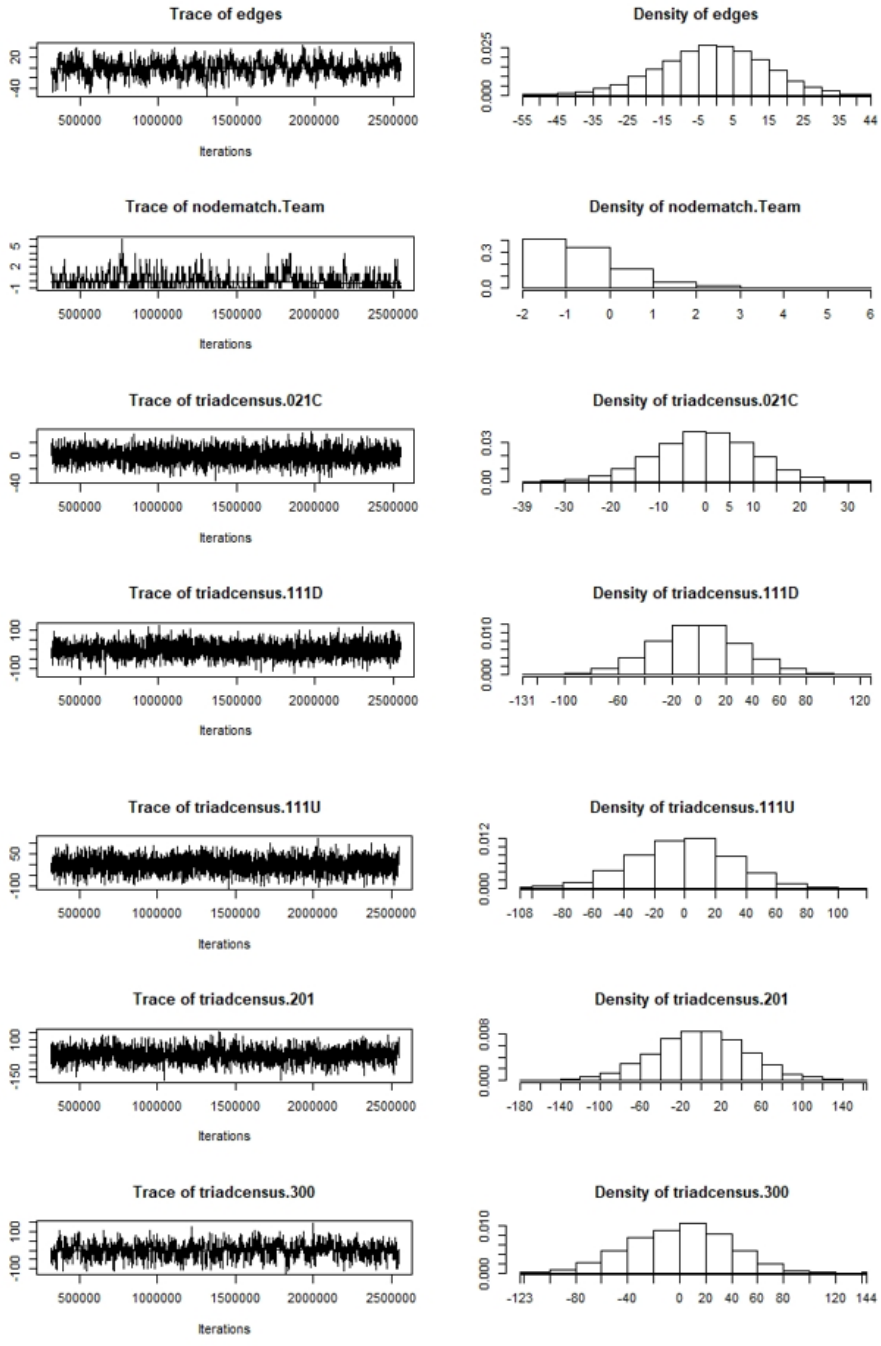


Figure 3.10: MCMC Diagnostics for the Model in Column (2) of Table 3.5

CHAPTER 4

Further Analysis by ERGM

In this chapter, we will find the effects of those characteristics by ERGM.

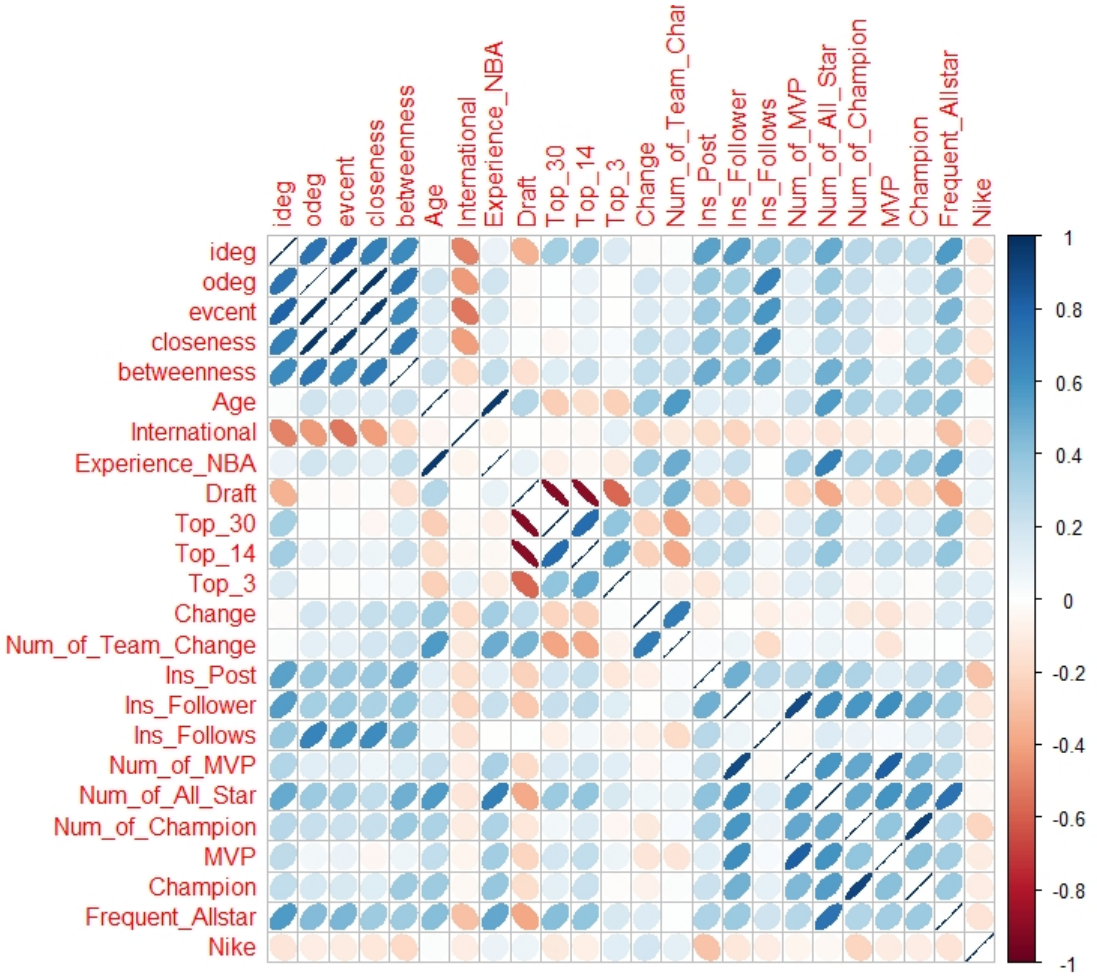


Figure 4.1: Correlation Plot for all Numerical Variables

We have 4 categories of variables and we plot the correlation among those numerical variables in Figure 4.1. From the figure above, some highly correlated variables, especially within the same category, can be easily found. Also, some variables cross the category can

also show high correlation, for example, Age and some Basketball Factor standing for a player's honor.

4.1 Effects by Demographical Characteristics

First, we start with Demographical Characteristics. We also fit ERGM conditional on Team homophily effect as before. For Demographical Characteristics, we think Race, International, Nationality and State[1] are measuring a quite similar latent variable. Since there are too many Nationalities and States, we only use nodematch instead of nodemix which is used for International and Race.

Table 4.1: Effects by Demographical Characteristics conditional on Team homophily effect

	(1)	(2)	(3)	(4)	(5)	(6)
edges	-2.254^{***}	-1.291^{***}	-1.907^{***}	-1.862^{***}	-0.869^{***}	-1.542^{***}
	(0.517)	(0.264)	(0.116)	(0.108)	(0.052)	(0.110)
nodematch.Team	4.941^{***}	5.257^{***}	5.246^{***}	5.301^{***}	4.934^{***}	4.957^{***}
	(1.008)	(1.016)	(1.016)	(1.016)	(1.010)	(1.011)
nodecov.Age	0.024^{**}					
	(0.009)					
mix.International.0.0		0.984^{***}				
		(0.271)				
mix.International.1.0		-0.476				
		(0.307)				
mix.International.0.1		-0.771^{*}				
		(0.316)				
mix.International.1.1						

[1]For International players, if he played high school basketball in the United States, we use the state his high school is located. Otherwise, we just use the country they come from.

	(1)	(2)	(3)	(4)	(5)	(6)
		(0.000)				
nodematch.International.0			1.599***			
			(0.131)			
nodematch.International.1			0.616*			
			(0.288)			
nodematch.Nationality				1.563***		
				(0.124)		
nodematch.State					1.173***	
					(0.232)	
nodematch.Race.Black						1.028***
						(0.125)
nodematch.Race.White						0.068
						(0.365)
AIC	2267	2095	2093	2086	2250	2198
BIC	2284	2128	2115	2102	2266	2220
Note:						*p<0.05; **p<0.01; ***p<0.001

From the results in Table 4.1, we can find that a strong International, Nationality and State homophily effect from their significant positive coefficients. As the Age of player increases, he tends to have more relationship with other players.

The most interesting finding is about the attributes International and Race. Since the coefficient of mix.International.0.0 is larger than mix.International.0.1 and both of them are significant, it seems that American players are more willing to follow American players. The coefficient of mix.International.1.0 is negative, which implies that International players prefer following other International players. But this phenomenon is not statistical significant. However, it still seems that the homophily does exist. Besides, from the results of using differential homophily term, we can find that the homophily effect among American players is larger than the one among International players. As for Race, there is a strong homophily

effect among Black athletes, but such effect is not statistical significant when applying to White athletes. The difference of homophily effect of International and Race might be the reason why those American and Black players are more clustered than those International and White players in Figure 3.4.

In all those models presented in Table 4.1, Model in Column (4) seems to have the best fit based on AIC and BIC. Therefore, we will then fit ERGM conditional on both Team and Nationality homophily effects in the next few sections.

4.2 Effects by Social Media Characteristics

Social Media Characteristics can tell us some social behavior of the players. Here are some of assumptions we want to test in this part of analysis. The number of followers may indicate whether he is popular on the social media. The player who has more follows might result in a higher probability that he follows other players while the one who has more post might result in a higher probability of being followed.

Table 4.2: Effects by Social Media Characteristics conditional on both Team and Nationality homophily effect

	(1)	(2)	(3)	(4)
edges	-1.826^{***}	-1.801^{***}	-1.989^{***}	-1.898^{***}
	(0.109)	(0.109)	(0.119)	(0.118)
nodematch.Team	5.423^{***}	5.383^{***}	5.784^{***}	5.861^{***}
	(1.016)	(1.017)	(1.024)	(1.025)
nodematch.Nationality	1.455^{***}	1.435^{***}	1.603^{***}	1.421^{***}
	(0.126)	(0.126)	(0.133)	(0.134)
nodecov.Ins_Post	0.360^{***}			0.142^{**}
	(0.039)			(0.045)
nodecov.Ins_Follower		0.317^{***}		0.254^{***}
		(0.039)		(0.044)

	(1)	(2)	(3)	(4)
nodecov.Ins_Follows			0.479***	0.451***
			(0.040)	(0.042)
AIC	1995	2016	1938	1864
BIC	2017	2039	1960	1897
Note:			*p<0.05; **p<0.01; ***p<0.001	

Since the numbers of Instagram posts, followers and follows are too large, we scale them first in the data pre-processing step. Then we fit ERGM and present the results in Table 4.2. All variables have significant positive coefficients which correspond well to our assumptions above. With the help of AIC and BIC, the model in Column (4) has a best fit using Social Media Characteristics conditional on both Team and Nationality homophily effect.

4.3 Effects by Business Factor

Just like other industry, many Business Factor may influence the decision of insiders. Therefore, we cannot just ignore the impact from the business world. Still we fit ERGM conditional on both Team and Nationality homophily effect.

Table 4.3: Effects by Business Factor conditional on both Team and Nationality homophily effect

	(1)	(2)	(3)
edges	−1.864***	−1.914***	−1.928***
	(0.108)	(0.109)	(0.117)
nodematch.Team	5.304***	5.342***	5.290***
	(1.016)	(1.016)	(1.016)
nodematch.Nationality	1.561***	1.559***	1.576***
	(0.124)	(0.124)	(0.124)
nodematch.Agent	0.075		

	(1)	(2)	(3)
	(0.254)		
nodematch.Agency		0.535**	
		(0.174)	
nodematch.Brand			0.165
			(0.112)
AIC	2088	2078	2086
BIC	2110	2101	2108
Note:	*p<0.05; **p<0.01; ***p<0.001		

From Table 4.3, it seems that the Agent s not so important. However, when we use Agency instead, it becomes significant. It makes sense since we define two players do not have the same agent if they have more than one different agents but, in fact, some players have more than one agent in the same agency company. Also, in NBA, a superstar, e.g., LeBron James, can have some influence on the decision of his team. It might be useful for a player to join a team if he can build a solid relationship with the No.1 player in that team. Therefore, as for the agency, in order to help players sign larger contracts with NBA teams, some agency companies might serve as bridges among their own players.

Besides, we can find that signing the same shoes brand do not increase the possibility of this following relationship significantly.

4.4 Effects by Basketball Factor

There are two sub-categories within Basketball Factor. One is more related to the capacity a player showed in basketball as well as his experience. The other one includes the position he plays in the game, which NBA Division and Conference he comes from as well as which University[2] he played for before going to NBA.

[2]For some International players who did not have University basketball experience in the United States, we use the country they come from instead. For those American players who did not have University

Table 4.4: Effects by Basketball Factor (Capacity and Experience) conditional on both Team and Nationality homophily effect

	(1)	(2)	(3)	(4)	(5)
edges	-2.571^{***}	-1.602^{***}	-1.352^{**}	-1.076^{***}	-1.071^{***}
	(0.240)	(0.116)	(0.415)	(0.256)	(0.134)
nodematch.Team	5.357^{***}	5.465^{***}	5.343^{***}	5.307^{***}	5.453^{***}
	(1.017)	(1.021)	(1.017)	(1.018)	(1.020)
nodematch.Nationality	1.563^{***}	1.590^{***}	1.565^{***}	1.576^{***}	1.350^{***}
	(0.124)	(0.125)	(0.124)	(0.125)	(0.127)
nodecov.Experience_NBA	0.035^{***}				
	(0.010)				
mix.Top_14.0.0		-0.649^{**}			
		(0.231)			
mix.Top_14.1.0		-0.950^{***}			
		(0.153)			
mix.Top_14.0.1		-0.319^{*}			
		(0.138)			
mix.Top_14.1.1					
		(0.000)			
mix.MVP.0.0			−0.609		
			(0.410)		
mix.MVP.1.0			−0.531		
			(0.433)		
mix.MVP.0.1			0.0001		
			(0.431)		

basketball experience, we use the high school he graduated from instead.

	(1)	(2)	(3)	(4)	(5)
mix.MVP.1.1					
			(0.000)		
mix.Champion.0.0				-0.966^{***}	
				(0.255)	
mix.Champion.1.0				-0.714^{**}	
				(0.274)	
mix.Champion.0.1				-0.550^{*}	
				(0.273)	
mix.Champion.1.1					
				(0.000)	
mix.Frequent_All_Star.0.0					-1.426^{***}
					(0.184)
mix.Frequent_All_Star.1.0					-1.033^{***}
					(0.144)
mix.Frequent_All_Star.0.1					-0.760^{***}
					(0.139)
mix.Frequent_All_Star.1.1					
					(0.000)
AIC	2076	2048	2078	2073	2001
BIC	2098	2087	2116	2112	2040
Note:					*p<0.05; **p<0.01; ***p<0.001

All the following analysis are the effects conditional on both Team and Nationality homophily effect. From the Column (1) of Table 4.4, we can find that Year in NBA has a positive influence on the formation of a tie. This is the same as Age in Table 4.1 since most players enter NBA at quite similar ages. The Column (2) seems that players with high picks are less likely to follow those with low picks while players with low picks are more likely to follow those with high picks. The homophily of NBA draft pick is not significant. As

shown in Column (3), it looks like whether a player have won MVP is not so important in this network. Players who have won Championship and are frequently selected to be NBA All-Stars are more likely to be followed. Therefore, players with more basketball skills will be a bonus for being followed by other players.

Table 4.5: Effects by Basketball Factor (Others) conditional on both Team and Nationality homophily effect

	(1)	(2)	(3)	(4)
edges	-1.894^{***}	-1.872^{***}	-1.876^{***}	-1.875^{***}
	(0.111)	(0.119)	(0.112)	(0.108)
nodematch.Team	5.122^{***}	5.291^{***}	5.300^{***}	5.315^{***}
	(1.025)	(1.018)	(1.016)	(1.016)
nodematch.Nationality	1.571^{***}	1.563^{***}	1.557^{***}	1.560^{***}
	(0.124)	(0.124)	(0.125)	(0.124)
nodematch.Division	0.210			
	(0.159)			
nodematch.Conference		0.020		
		(0.107)		
nodematch.Position			0.047	
			(0.110)	
nodematch.University				0.895^{**}
				(0.406)
AIC	2086	2088	2088	2083
BIC	2108	2110	2110	2105
Note:				*p<0.05; **p<0.01; ***p<0.001

In Table 4.5, we test the effect of other Basketball Factors. Most variables do not show the statistical significance. One interesting finding is that players from the same University are are likely to form a tie as we can see a significantly positive coefficient in Column (4).

CHAPTER 5

Conclusion

In this research, we analyze a relationship among public figures with many available variables. The part of Degree and Centrality analysis gives us some ideas about which factors may influence this Instagram Following relationship. The Latent Space Model does not indict clear sub-groups since this relationship is more complex than a simple clustering such as the friendship among monks. We find strong homophily effects of Team and Nationality and have done a lots of effect analysis conditional on such two homophily effects. Generally, Basketball Factors standing for Capacity and Experience of a player and Social Media Characteristics are important in forming relationship while Business Factors and some other Basketball Factors are not so important.

There are still more research we can do. We can use more community detection techniques besides Latent Space Model and then compare them with each other. Also, we focus more on statistical inference in this paper. We could try to use social network models to do some prediction and compare them with other machine learning techniques.

REFERENCES

[1] David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau and Martina Morris. 2008. *ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. Journal of Statistical Software*, 24(3).

[2] David R Hunter, Steven M Goodreau and Mark S. Handcock. 2008. *Goodness of Fit of Social Network Models. Journal of the American Statistical Association*, 103, 1, pp. 248-258.

[3] Goodreau, Steven M., James A. Kitts, and Martina Morris. 2009. *Birds of a Feather, Or Friend of a Friend?: Using Exponential Random Graph Models to Investigate Adolescent Social Networks. Demography*, 46.1: 103-26.

[4] Koschtzki, D., Lehmann, K.A., Peeters, L., Richter, S., Tenfelde-Podehl, D., Zlotowski, O. 2005. Centrality Indices, Chapter 3 in Brandes and Erlebach *Network Analysis: Methodological Foundations*. Volume 3418 of *LNCS Tutorial*, Springer.

[5] Mark S. Handcock. 2003. *Assessing Degeneracy in Statistical Models of Social Network. CSSS Working Paper*, No. 39.

[6] Mark S. Handcock. Fall 2018. *Social Balance and Transitivity*. lecture Notes. *STAT 218: Statistical Analysis of Networks*, UCLA, delivered 1 Nov 2018.

[7] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau and Martina Morris. 2008. *statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. Journal of Statistical Software*, 24(1).

[8] Pavel N. Krivitsky and Mark S. Handcock. 2008. *Fitting Latent Cluster Models for Social Networks with latentnet. Journal of Statistical Software*, 24(5).

[9] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. 2002. *Latent Space Approaches to Social Network Analysis. Journal of the American Statistical Association*, 97, 1090-1098.