NETS 150 Homework 5 - Project

Team Members

Ankit Das - Yoni Nachmany - Evan Tao

Project Categories

- Document Search (aka Information Retrieval)
- Graph and Graph Algorithms

Project Type

Empirical Analysis with IR Code

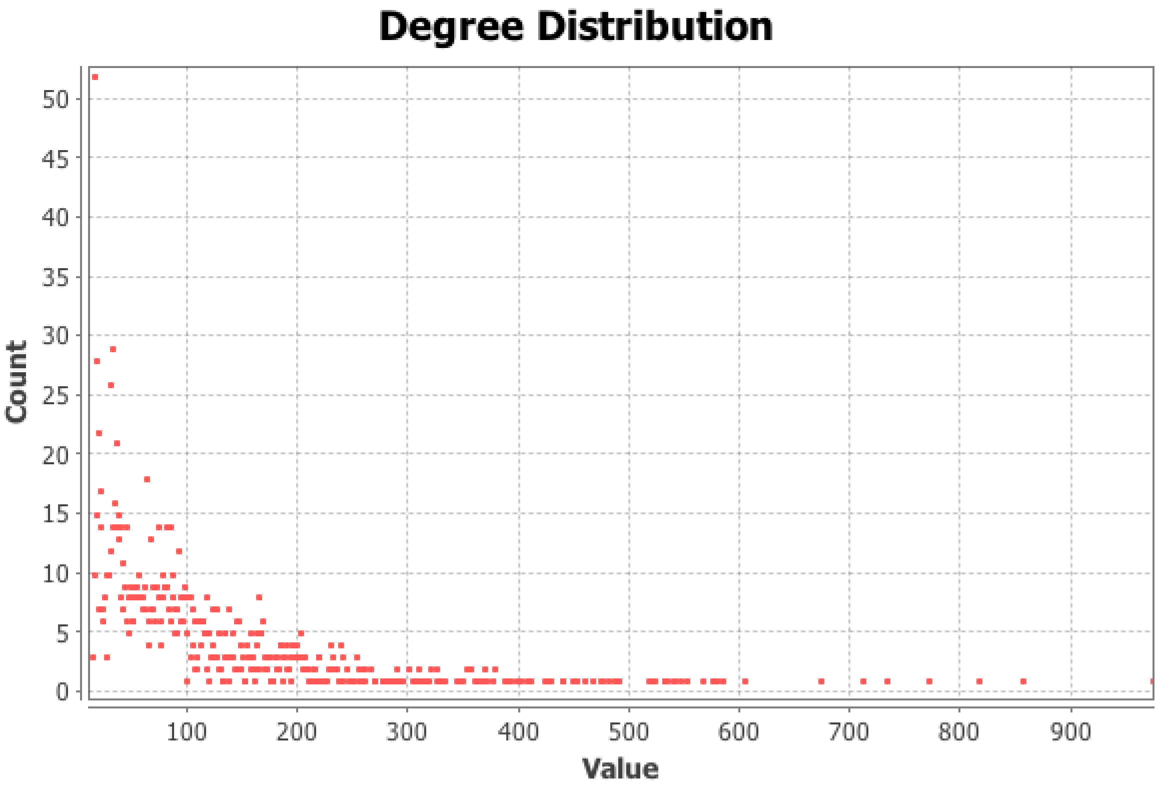
In our project, we analyzed the network of NBA players, where there is an edge between two players (who represent vertices) if they have played with each other in the last 10 years.

We hypothesized that the graph exhibited properties of large-scale networks, namely a small diameter, one giant connected component, a high clustering coefficient, and a heavy-tailed degree distribution. We further hypothesized a correlation between degree and number of teams played for as well as degree and career duration.

Our project required document search (aka information retrieval) to get the data required from an online website of rosters, and then graphs and graph algorithms to analyze the data and test our hypotheses. By feeding in our data in an adjacency list representation to Gephi, the open graph viz platform, we were able to both visualize the graph and obtain certain statistics like diameter and clustering coefficient.

There were 1411 nodes and 31700 edges in our undirected graph. The diameter was a low 3, the clustering coefficient was a relatively high 0.547, and the degree distribution was heavy-tailed (as seen in the visualization on the following page). All of this matches the attributes of a large-scale network, thus verifying our initial hypothesis.

These statistics of the NBA player graph make sense. Since there are only 30 teams and players do move around teams, we would expect to have a diameter representing that of a “small world”. Additionally, neighbors of a given node are expected to have an edge between them because everyone on one team in one year forms a completely connected clique. Furthermore, we would expect there to be a small fraction of people that had played with many more players than average over the course of their careers -- be it through longevity in the league or frequently switching teams.

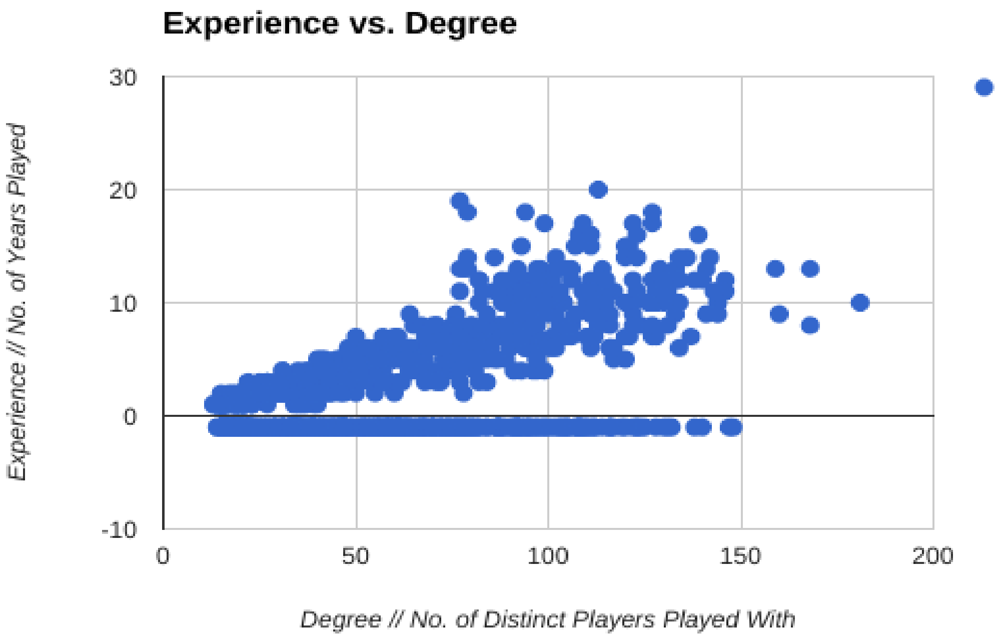


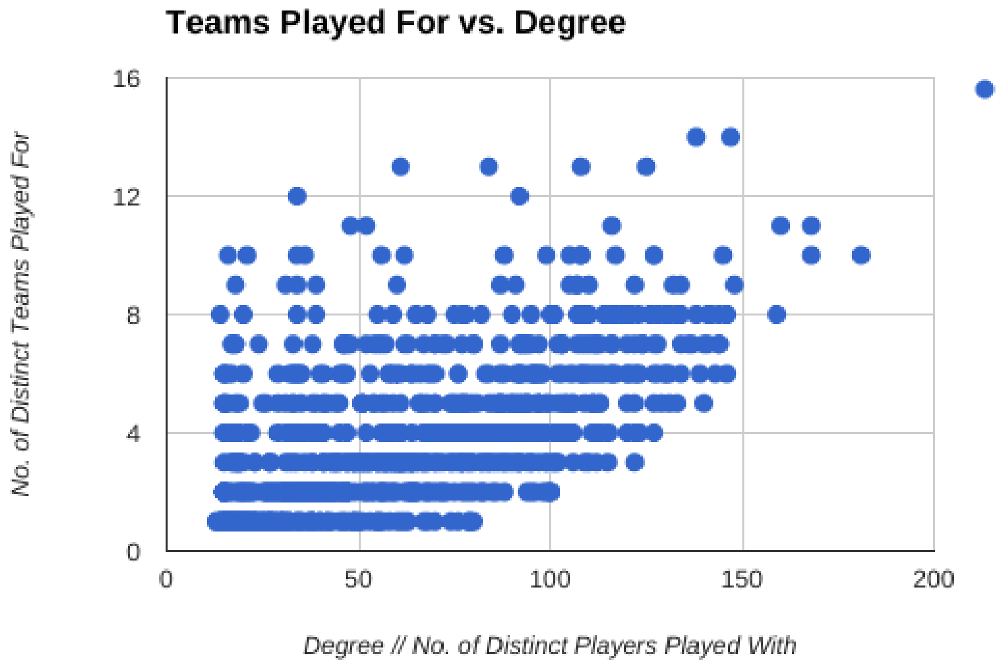
We collected several measures of data among NBA players, including name, position played, number of teams played for[1], number of players played with[2] and years of experience[3], and we collected these statistics over the last ten years (2005-2006 season to 2014-2015 season). Our first objective was to extract this data and visualize relationships between measures. Extraction was simple: the program was run, the necessary data was written to a file, and that data was copied to a spreadsheet. On the other hand, the relationships we pinpointed were less predefined: they were either personal curiosities that we wanted to investigate or trends in the actual NBA that we wanted to verify. The following graphs display the two relationships that we analyzed:

[1] With regard to the website we used for our data, extracting the number of teams played for was a difficult task. Thus, we substituted it with a closely related measure: the number of jerseys worn. This measure was much more straightforward to retrieve.

[2] Playing with "oneself" is included; thus, all players have a self-edge. This means that a player who played for one season will have a degree of 15, not 14.

[3] Retired players, by our definition, have -1 years of experience.


Experience vs. Degree
30
20
10
0
-10
Experience // No. of Years Played
0
50
100
150
200
Degree // No. of Distinct Players Played With


Teams Played For vs. Degree
16
12
8
4
0
No. of Distinct Teams Played For
0
50
100
150
200
Degree // No. of Distinct Players Played With

Our second objective was to verify our hypothesis regarding the correlation between players' attributes, and thereafter make reason of trends. In the first graph, the long stream of data points at -1 years of experience suggests that retired NBA players have degrees (number of teammates) that vary across the board; that is, NBA players may retire after playing with any number of players. This makes sense as injuries or other mishaps can cut a career off at any point.

Next, there was a strong linear trend among active NBA players between numbers of years played and number of players played with. This confirms what we had hypothesized: the longer a player has been in the league, the more teammates he will have had. This could be for two reasons: a player sticks with one team, hence seeing many players come and go alongside him, or a player jumps around many teams, hence playing with many new faces.

The second visualization also showed a roughly linear trend between number of teams played for and number of teammates played with. This makes sense: the more teams one has played for, the more teammates he will have played with. However, there were many more outliers -- data points that did not fit the linear trend -- in the second graph than in the first. Particularly, there were many players that, based on the number of players played with, had been on more teams than we would expect. These are players whose data points appear above where a line of best fit would be placed. The primary reason for this is as follows: these are players that had to swap jerseys multiple times while playing for the same team; thus, it appears that they switched teams when in reality they didn't. Hence, they have actually played for less teams than our data indicates, and they would actually be closer to the line of best fit.

Overall, we learned a fair amount from our project and had fun doing so. From information retrieval to graph analysis to data visualization, our project involved many different aspects and touched on various things we learnt in class. We would like to thank the entire NETS 150 team for a great semester.

See compiled data here: https://docs.google.com/a/seas.upenn.edu/spreadsheets/d/1w5Gtr4S_0MNPEu4bIWgfU4Ty0fOTv5GNA-yHwc2Ztrw/edit#gid=777911147

See attached code in AnswerGetter.java and Main.java. Raw data comes from here: http://www.basketball-reference.com/teams/.