Formula One in Network Science

J. Bou Karam Author, P. Abi Chacra Author, A. Assaf Author

# *Formula One is a motor racing championship that has gained popularity over the years, starting from the first launched race in 1946 up until today. Millions of people impatiently wait for every racing season because with it come the adrenaline and excitement that only few racing games have. Although this game is highly admired by people from all around the world, studies on this sport have been primitive and little. On the other hand, network science has played a major role in several topics, ranging from social media all the way to sports. In this case, it studies how drivers pass others in races and how strong the relationship among the drivers is. For this reason, viewing this topic from a network analysis point of view gives the ability to study patterns and trends followed by racers through analyzing the number of times one overtook another and interpreting such results for predicting future wins.*

# Introduction

Grand Prix motor racing, also known as GP, has existed since the year 1894. It only took place in France where cars would race across towns from the start to end points. However, over the years, GP got broken down into several kinds of races and ended with its being called Formula One after the year 1946 where it replaced the name of Formula A. The Formula One Championship is a sport millions of people watch, ranging from a very small age to an old one. It is considered one of the best industries in the world and among the top 10 sports businesses to exist. This sport, also known as F1, despite its being less popular than other sport activities like football, basketball, and baseball, is considered the greatest and most exciting type of racing in the world. It has increasingly gained the interest of its audience over the years and is currently the most watched racing game.

Before going deeper into the subject, it is important to understand the game. To begin, the races take place in several countries such as The United States of America, Canada, Bahrain, and other locations. Each racetrack varies approximately between 2.5 and 4.7 miles depending on the country in which the game is taking place with 12 rounds for every race. Each car’s speed can reach about 220 miles an hour, but of course, its strength and confidence in winning the race depends on the driver and constructor team.

On the other hand, network science is one of the most optimal and efficient ways in studying relationships among sets of data known as nodes. For this reason, studying Formula One from a network science perspective improves one’s understanding of the topic at hand. Throughout this report, several measures are studied to discover patterns followed by drivers in passing others, to measure the importance of a driver based on several centrality measures, to identify communities of drivers through differentiating the most from the least number of overtaking, and to study how this sport evolved over the years. All of the aforementioned points, along with several others that are discussed throughout this paper will improve ongoing predictions for future games.

# Idea

Formula One is a car sports game that, despite its being watched by millions of people, does not come close to the hype witnessed by other sports. Moreover, the major part of the audience tends to be a fan of the most famous drivers such as Hamilton, Vettel, or Schumacher without getting to know other drivers who can be just as good if given the proper attention by the media and the audience.

For the reason of familiarizing the audience with all the players, which in a way widens their knowledge about the good, the underestimated, and the over-hyped drivers, this project represents the number of trespasses made by each driver and the number of times he was passed between the years 1981 and 2020 by building a network that represents the relationships among the drivers.

# Literature Review

Very few papers exist that study the science behind Formula One let alone the network generated by such a game. However, some were taken into account in a way that widened our knowledge of the game and resulted in seeing it from a different and more developed perspective. Note that the project has little dependency on the literature since the game of Formula One is taken from a whole new perspective.

Bothner, Kim, and Smith, in their article ‘*How does status affect performance?’* show that a driver’s status affects his competence and productivity in the race. A driver’s status depends on his position at the end of the race and the derived statistics that show the number of overtaking, the constructor, standing, and other attributes. The literature discusses two opposing points of view where the first argues that a driver’s performance solely depends on his status, while the opposing opinion claims that prestige has no relation with the driver’s performance. For the reason of confirming one of the two views, this paper is taken as reference throughout the project to come up with a final conclusion. As a result, the network science part of this project deals with the overtaking figures and constructor and the machine learning part with the rest to finally confirm the significance of relationships among the drivers and to predict possible wins in future races.

The second article ‘*The top 10 F1 drivers of 2020 - as chosen by the drivers. F1 - The Official Home of Formula 1® Racing’* written by Barretto lists the best drivers of the 2020 season, which are studied throughout the machine learning process in predicting further conclusions. Moreover, this piece is also taken as reference throughout the whole analysis process for the purpose of either confirming or disproving the hypotheses. The results are compared with the list of drivers presented in the paper along with the evaluation made by the author to increase the network’s credibility and reliability.

Therefore, both papers are crucial in making the network a credible one when analyzing by confirming the results and for making the machine learning results a reliable source of prediction.

# Research Questions

Since there exists a small number of papers dealing with the sports of Formula One, the research questions raised in this project have neither been discussed not studied previously.

To determine the objective of the study, it is crucial to raise questions that cover every point mentioned throughout the paper.

This project inclines to answer one main question that is broken down into several others: are drivers fairly judged by the audience and media?

For this reason, we have come up with a few hypotheses that will either be proved or refuted throughout the coming sections.

For instance, not all good drivers are equally well-known and cheered for by the public due to a number of factors that range from political or cultural all the way to publicity and race rankings. In other words, some drivers tend to be less famous than others regardless of their statistics. As a result, some players are underestimated and overshadowed by the media.

Second, some remarkable players tend to fall behind due to problems with their cars or the constructor itself, which, in turn, leads to the driver’s losing his market value and being less noticed by the audience. For this reason, this project also focuses on both the underestimated and overestimated drivers throughout all the years.

Third of all, even though the constructors play a big role in maintaining the car throughout the race and making sure it operates at full capacity, a driver’s constructor guarantees a higher chance of winning the race or being within the top 3.

Last but not least, a driver’s chance at overtaking others increases as years go by since the championship rapidly adapts to evolution in technology, which makes the participants more competent and more threatening to each other.

In conclusion, this paper sheds the light on the dominating communities and probable problems associated with cars overtaking each other, which in turn play a part in predicting future races’ possible winners or losers.

# Dataset

The Formula One overtaking data was gathered from [www.cliptheapex.com](http://www.cliptheapex.com) that required a payment for enabled access. This website is made up of discussions where every forum contains the overtaking data of the years ranging from 1981 up until 2021.

However, most years demanded work on gathering their data together. In other words, most pages provided information on every race instead of every season and in text form which was not the target information to work on. For this reason, all the data from every race was copied into a text file where every text file merged all the races within the same season together. Afterwards, some data preprocessing was made to sum every driver’s overtaking and place the result in matrix form to match the format of the other data that existed in picture form.

On the other hand, the data that existed in picture form was transformed into an actual csv file using a website called https://online2pdf.com and remained in matrix form.

As a result, all the target data was made into csv format and then merged together into two files. The first file, consisting of two columns (id and label), contains the nodes, which are the names of all drivers, and the second file represents the edges and is made up of several columns such as the passer, overtaken, race start year, and end year. The machine learning dataset was initially obtained from Kaggle where two datasets were used: “Formula 1 Race Data” and “Formula 1 World Championship (1950 - 2021)”. Then, they were combined with the resulting data to generate the final csv file that will be worked upon throughout the machine learning process.

The code is discussed in the Code section.

# Code

## As previously mentioned in the Dataset section, the data needed to be further worked upon to be able to generate a network.

## Since no forum presented the targeted data directly, some preprocessing was made to solve the two main issues: incomplete data and data in picture form instead of text.

## Note that all the lines of code showing in the below figures are only snippets of the whole code corresponding to an appropriate process and not all codes were mentioned below due to the large number of python files.

## First, the information was incomplete since the data that existed was in the form of a text and specified who passed whom. For this reason, after going over 21 races for every season and placing them inside their corresponding text file, the latter underwent some python code using the “pandas” library. Every word was placed under a column and a corresponding csv file was generated for every year. Then came the cleaning of the unused columns where the only columns that remained were “Pass” and “Passed” that represented the driver who overtook another and the one that was overtaken respectively. After this procedure was completed, the previously mentioned csv file was also processed using the same python library where a JSON object was generated. This object was made up of a nested dictionary within every key in the form of key: {key: value}. The former key represents the driver in question while the second the name of all racers that may have been overtaken by this driver during a certain race, whereas the value represents the number of times the latter was passed.

## To go in details, the first part of the code loops over the “Pass” and “Passed” columns to gather the names of all the drivers. Second, another loop was made over the “Passed” column again to sum up all the times a player was overtaken by a certain driver. If the driver’s name corresponds to the one that overtook him in the “Pass” column, the sum was incremented by 1. This process was repeated for all years from 1981 until 2020. Figure91 displays the code generated to sum the number of times a certain driver was overtaken. Finally, after storing the JSON object in a file, it was converted to csv format using http://convertcsv.com

for j,i in enumerate(pd['Pass']):  
  
 if overtakes[i][pd['Passed'][j]] == '':  
 overtakes[i][pd['Passed'][j]] = 0  
 overtakes[i].update({pd['Passed'][j]: overtakes[i][pd['Passed'][j]]+1})

Figure 9 Python code that calculates the number of times a Formula One driver was overtaken by another in the same season.

On the other hand, the second issue that was addressed was converting the documents from image form to text while making sure no data was lost through a website called https://online2pdf.com. The file was then checked for missing or wrong data by comparing the resulting csv with the origin image.

After gathering all data from all races and all seasons, the last phase was applied in which two steps were followed after merging all the files that contained the sums into one huge csv file. First, the names of all racers were merged into one file representing the nodes. The first loop also iterated over the “Pass” column while the second went over the “Passed” one to make sure no name was lost. Figure 10 shows a snippet of the code where the name of the drivers was added to a list along with the range of years in which they played.

for i in pd['pass']:  
 if len(overtakes) == 0:  
 overtakes.append(i)  
 else:  
 for j in overtakes:  
 if j == i:  
 break  
 else:  
 if overtakes.index(j) == len(overtakes) - 1:  
 overtakes.append(i)

Figure 10 Python code that saves the names of all drivers in a list.

The second and final step consisted of improving the file with the sums further where the name of the driver that passed another, the overtaken, the sum, the year, and end year were only kept and represented the edge list.

As a result, all the needed information became available and feasible to work with in creating and studying the network.

On the other hand, the machine learning part of the project required additional attributes like the race, constructor, driver, position, and others that contribute to a more accurate prediction. The data dealing with the overtaking was gathered from ClipTheApex website for each year between 1981 and 2020 and for each driver that has participated in the races.

It was then cross referenced with a dataset of approximately 23,000 rows from Kaggle to be able to give the machine learning model the number of overtakes of each driver by year. In addition to that, the same thing was done with 3 other datasets also collected from Kaggle to get the during every race the driver’s age, nationality, country, circuit in which he drove, his final position, and points earned at the end of the race. This process of going over the 23,000 records and rerunning the whole code bu t on a whole different dataset was repeated on the constructors of every driver in every race, too.

The “results” csv was taken as a reference to all other files since it contains the foreign key of every attribute, which in turn was combined with the keys in every file. As a result, a new file was generated for the machine learning. For instance, when adding the driver name to the new file, the unique ID in the drivers.csv file is compared with that of the results.csv file. If the values match, then the driver’s name is added to the new csv file. Figure 11 shows a snippet of the whole code where the driver is compared and added to the array, which is in turn added to the final file used in machine learning.

for index, result in results.iterrows():  
 print(index)  
 for driverIndex, driver in drivers.iterrows():  
 if(int(result['driverId']) == int(driver["driverId"])):  
 driverName.append(driver['driverRef'])  
 break;

Figure 11 Python code that compares the drivers in both files and saves them in a list.

## Network

## Network Creation

## The Formula One network is characterized by being both a dense and sparse depending on the year, directed, and weighted structure having approximately 23 nodes for every year or season. Note that since a range of about 40 years was gathered, most drivers’ names do not exist in all seasons. The nodes represent the drivers whereas the edges the connection between two drivers. In other words, since the edges are directed, the direction to which an arrow points portrays who overtook whom. Moreover, every edge has a weight representing the number of times the person got passed.

## Since the goal of this project to study trends over the years, the network was turned into a dynamic one where the interval of every season depended on the start and end date. Every driver’s constructor was also taken into account.

## First, the nodes list was imported and modified to a directed graph and then came the edges list’s turn that was appended to the same workspace. Figure 1 shows the nodes list format with the id and label, dates, and constructor names columns. A new column was then added in which the time interval was created and in turn, the dynamic network.

Table

Description automatically generated

Figure 1 Nodes list where the id and label represent the attributes.

## In order to differentiate among the years, a unique color was assigned to every subnetwork, which is a way of turning the graph into communities based on the driver’s active years in the championship. Figure 2 shows the whole network with a selection of all the existing years, while figure 3 displays only one season, 2019, with its corresponding color.

A picture containing text, map, flock, several

Description automatically generated

Figure 2 Whole network where every color represents a season and the texts display the drivers’ names.

A picture containing sky, outdoor, day, several

Description automatically generated

Figure 3 2019 subgraph that shows the overtaking data in edge form.

## After creating the needed network, the next stage of analyzing and interpreting it began.

The network consists of 159 nodes and 4602 directed and weighted edges that vary according to the selected interval.

## Network Analysis

Before going deeper into the network analysis, the layout of the graph was modified using the Force Atlas layout that pushes connected edges to each other and repulses non-connected ones. As a result, figure 4 represents the new network layout.

Map

Description automatically generated

Figure 4 New network layout

In order to come up with conclusions, it is necessary to study some important measures that help in understanding the network structure and goal.   
To determine the most significant metrics to use, it is vital to differentiate among the various methods. For instance, studying the weighted degree distribution is more reliable than the ordinary degree distribution in this case since the key feature behind this study is the number of overtaking done by every driver, and since the weights represent the total sum of overtakes performed, the weighted degree distribution is the most optimal metric to use.

First and most importantly, the weighted degree distribution is a crucial metric that helps in understanding the network structure. To begin, in this case, this distribution, which has an average of 67, specifies the number of times one driver overtook or was overtaken by another. For this reason, it is crucial to focus on the two main mentioned types where the weighted in-degree represents the number of times a driver was overtaken and the weighted out-degree the number of times this driver passed another. On the level of the former, the network was filtered by node size where the larger the node, the more this driver was passed. Figure 5 shows the network the weighed in-degree distribution filtered by node size where the driver who was mostly overtaken is Grosjean with a value of 370, and the one with the lowest number is Tarquini with a value of 0. Note than Grosjean is a very well-known driver regardless of being the most overtaken. On the other hand, the out-degree is equally as crucial since it represents the number of times this driver overtook another. Figure 6 shows the weighted out-degree network distribution where the node size is proportional to the sum of passes made by every player. It is observed that Perez, the most famous driver, has the highest value (508) of overtaking, while Sato is among the lowest with a value of 1 trespass, Even if the most famous drivers have the highest number of overtaking, better-known drivers tend to have lower distribution than others who are less famous than them, like Perez and Hamilton who has a number of 450, but this does not apply to the drivers having the highest number of in-degree. In other words, even though Grosjean has the highest in-degree distribution, he is very well known due to the threat he imposes on other drivers. Finally, figure 7 displays the distribution in a graph that shows the weighted degree distribution of every node on the axis. For instance, 10 nodes have degree 1, which means that they either overtook or were overtaken only once, while one other node has a degree of 500. This type is not as reliable as the weighted in and out degree distributions since it shows a general view rather than a specific one, which makes it ambiguous to understand.

Diagram

Description automatically generated

Figure 5 Network representing the drivers’ in-degree distribution throughout all the years.

# Diagram Description automatically generated with medium confidence

Figure 6 Network representing the drivers’ out-degree distribution throughout all the years.

# Graphical user interface, chart, application, table, Excel Description automatically generated

Figure 7 Scatterplot representing the drivers’ degree distribution throughout all the years.

# Second, in order to study how fully connected the network is, it is crucial to run the graph density measure over the directed network. In other words, the more connected the nodes, the higher the density where a higher density displays a great number in overtakes done by every driver on every other driver. However, in the case of the Formula One championship over the years, the graph will never be fully connected since a majority of drivers have never played together in any race and season or never trespassed each other due to the huge discrepancy in their positions. In a way, it is impossible to have a dense graph as a general view, but the chances of having more connected nodes on a smaller scale is high due to the limited number of players and the number of passes done by each one of them. This proves the aforementioned hypothesis where, after running the graph density algorithm on Gephi, a value of 0.18 was generated. However, when observing it on a smaller scale, the density tends to be greater. For instance, the year of 1986 until 1987 has a value of 0.3, which is also sparse but denser than the overall graph density. This value gradually increases from 0.7 in 2001 to 0.9 in 2007 and continues on rising to reach 2.5 in 2016 and 2.8 in 2018. It can be perceived that with every passing year, the graph density increases to reach its highest in 2018. In other words, every year witnesses more trespasses than the year before. This can rely on a number of reasons such as fair competitiveness among drivers, closely competent constructors, and an advancement in technology with the championship’s fast adaptation to it in comparison to the previous years.

The third measure is known as centrality, most specifically the betweenness centrality that is the most relevant metric in this case. To begin, the node with the highest betweenness centrality is the one that lies on the shortest path between two nodes. In this case, the betweenness centrality is represented by the node that falls between two drivers. For instance, throughout 2017, the node with the highest centrality measure is Verstappen with a value of 1419 whereas the driver with the lowest value of 2 is Hartley. In other words, Verstappen often falls between two nodes that tend to fight for a position, whereas Hartley is rarely fought with. Figure 8 represents the network structure in 2017 where the nodes’ size is directly proportional to the betweenness value. This measure plays a major role in determining who the riskiest driver in every race is due to his position within the track that tends to be challenging for players with close positions as him.

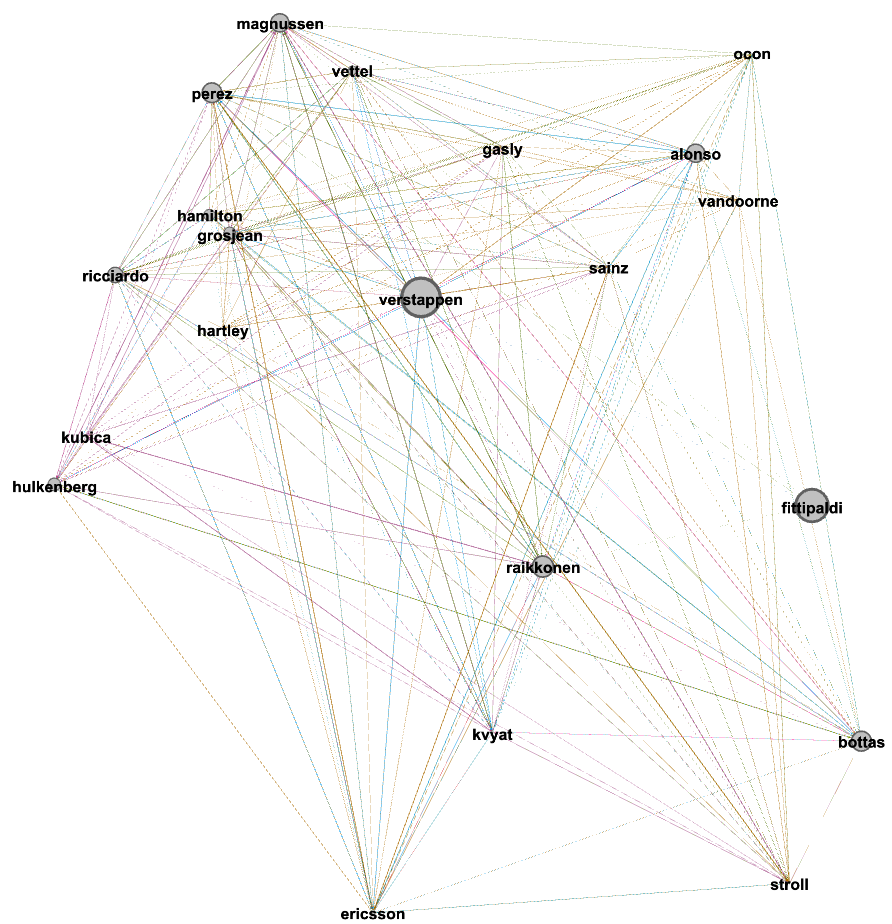


Figure 8 2017 Network representing the drivers’ betweenness centrality by node size.

Modularity is one of the most important metrics used due to its ability to detect communities within the network. In this case, when running the modularity over all the years, five main communities were detected based on shared similarities among them, such as the clusters to which they belong. Figure 9 displays the graph filtered based on every community along with the applied Force Directed layout where every color represents a different community and every node size represents the weighted in-degree of every node. It can be observed that nodes colored in blue tend to have the highest in-degree based on the size with a partition of 22% of the whole graph, the ones in green make up 23% of the whole partition and have a slightly less degree, followed by the ones in red and pink with values of 17% and 34%respectively ending with the light green nodes that represent only 1% of the whole graph. Moreover, every color also fit in a certain interval of time where the blue and light green colored nodes belong to recent years, whereas the other colors go way back from 1981 up until approximately 2000, which can be seen in figures 10 and 11.

Diagram

Description automatically generated

Figure 9 Network with five communities.

Chart

Description automatically generatedDiagram

Description automatically generated with medium confidence

Figure 10 Communities in 2015 vs communities in 1990.

After categorizing the network via the constructor names in the inter edges filter, it was noticed that the most popular constructors do not belong to the most popular drivers that have the highest overtaking statistics. In other words, there exist 12 racers belonging to the “Williams” constructor where the best two drivers do not fall within the top 3 positions in ranking. Moreover, the driver with the highest number of wins, Hamilton, is known to drive a Mercedes and has been a champion for a long time; yet the corresponding constructor does not fall within the top ones as it falls somewhere in the middle of the list.

## Findings and Contribution

## Findings

Based on the analysis mentioned previously, several conclusions were found that either approve or refute the hypotheses mentioned in the sections above.

First of all, drivers with the highest number of overtaking are rarely placed within the top 3 race champions regardless of their having better statistics. The weighted in-degree is a significant measure that shows how competitive the driver is. It also shows how threatening he is to other drivers since he has a high risk of passing others and causing them to lose their place.

In addition to that, due to technology’s evolution and proper driver training that makes them fairly competitive, a higher interaction among the drivers is established where every year witnesses an increase in overtaking. This is verified using the graph density that increases year after year; hence, making the game more thrilling.

A combination of both findings confirms the resulting network generated by modularity where a total of 5 communities were detected. This approves the fact that the more recent the years are, the higher the number of overtaking and the denser the network gets.

However, the final hypothesis has been refuted by the network analysis where the constructors do not necessarily affect the resulting position or rank of the driver.

As a result, the generated network has proven and refuted the aforementioned hypotheses suggest in the previous sections where several conclusions have been made. Studying this network and the derived results plays an important role in helping the audience improve their judgment and understanding of the game and its statistics.

## Contribution

Since the number of times a driver passes another or was passed by another is focused on throughout the seasons from 1981 until 2020, all findings, measures, and additional data hugely influence the direction of future research and discoveries such as on the level of machine learning.

All these studies led to one important part; machine learning. Predicting future wins and losses is a crucial way in hooking more people to the game, in making its market value grow, and in increasing the audience’s interest by placing bets based on real data and not just assumptions. The machine learning part of it, in turn, directs the audience’s analysis and judgment of the game into a new different perspective away from all the media and propaganda towards studying the statistics generated by every driver and therefore, encouraging the most deserving one.

Moreover, this project also studies motor racing matches from a new point of view never existed before, which can lead to further evolution in the game, in betting over drivers, and in fairly judging all players based on their expertise rather than the media.

Finally, and in addition to all points previously mentioned, a wider potential will be given to all kinds of sports and most importantly the racing types in further growing and becoming as popular as other athletic championships that have millions of people around the globe stuck to their television screens impatiently waiting for them.

# Machine Learning

After analyzing the network and gathering all necessary results, the data was placed in a file processed by Python for the machine learning process. For the purpose of creating the model, we relied on two Kaggle repositories, “Formula 1 Race Data” and “Formula 1 World Championship (1950 - 2021)”.

The datasets were manipulated in such a way to formulate a meaningful file that will be later used for the training of the model. This file contains many attributes of each driver such as age, constructor, points, position in a given race and some others.

In python, the file was then split into X and Y arrays, where the former consists of all attributes but the position which was allocated for the latter array only. Each of the two arrays was then split into train and test portions by 80-20 percentages, respectively. For this purpose, we used the “train\_test\_split” method from the famous “sklearn” library.

Next, we performed multiple models like SVM and linear regression and ended up with a modest accuracy of 0.439. We then proceeded with SVM to take advantage of the probability feature that allowed us to calculate the probability of every driver to win a certain race. After that, we prepared a sample data to put our model into a practical scenario. This data contains drivers from the current season (2021) that we intended to calculate their chances of winning the Spanish grand prix of this season. We imported this dataset into python, fed the model with it and got the corresponding estimated probability for our drivers.

# Conclusion

Formula One is a sport that requires lots of adrenaline and excitement despite its being just a racing championship. It has evolved over the years to become one of the most popular shows that people impatiently wait for year to year.

The weighted degree distribution, rather than the ordinary one, network density, centrality measures, and others show that even the most famous drivers tend to have lower overtaking statistics and vice versa. In other words, such drivers tend to be overshadowed by others.

This project, both on the level of the network and machine learning, are a huge step in changing the Formula One science and research from primitive to more advanced since it enabled the audience and the researchers to expand their knowledge, better judge the drivers based on statistics rather than just relying on the media or the fans, and predict future events. As a conclusion, network analysis plays an important part in studying the science behind Formula One and in widening the approach for further development and research both on the smaller and more general levels.

# References

1. Burns, A. (2012, March 1). F1 project - Stakeholders in Formula 1. Academia.edu - Share research. <https://www.academia.edu/10973191/F1_Project_Stakeholders_in_Formula_1>

2. Bothner, M., Kim, Y., & Smith, E. (2011, July 27). How does status affect performance? Status as an asset vs. status as a liability in the PGA and NASCAR. PubsOnLine. https://pubsonline.informs.org/doi/abs/10.1287/orsc.1110.06793.

3. Barretto, L. (2020, December 21). Exclusive: The top 10 F1 drivers of 2020 - as chosen by the drivers. F1 - The Official Home of Formula 1® Racing. <https://www.formula1.com/en/latest/article.exclusive-the-top-10-f1-drivers-of-2020-as-chosen-by-the-drivers.yhxUA1J590x2tFTOb24z5.html>

4. ClipTheApex for overtaking data <https://cliptheapex.com/forums/overtaking/>

5. Kaggle for machine learning data between 1950 and 2020 <https://www.kaggle.com/rohanrao/formula-1-world-championship-1950-2020>

6. Dataset of Races from Kaggle for machine learning data https://www.kaggle.com/cjgdev/formula-1-race-data-19502017

7. Machine Learning models https://machinelearningmastery.com/make-predictions-scikit-learn/