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, which are the nodes, pass others in races where the act of overtaking is represented by the edges. 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 racer overtook the other 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.

# 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

All IBM authors must obtain clearance from IBM before submitting the final manuscript. The “[Publication Clearance](https://apps.na.collabserv.com/wikis/home?lang=en-us#!/wiki/W18e544042a85_4b63_915a_1d1ed2cf8338/page/Publication%20clearance)” wiki provides details about the procedure. IBM Research employees must use the [Research Publication Clearance System](https://watapp.watson.ibm.com/publication/) to obtain publication approval. Non-Research IBMers should follow the procedure outlined in the “Publication Clearance” wiki.

Non-IBM authors must transfer the copyright to IBM by addressing this question during the submission process in ScholarOne. For any questions about initial or final submission requirements, please contact the [journal editor](mailto:rdannuc@us.ibm.com?subject=IBM%20Journal%20of%20Research%20&%20Development:%20Early%20Access%20Constraint).

# Literature Review

Sections following the introduction should present your results and findings. The body of the paper should be approximately 6,000 words. The manuscript should evolve so that each sentence, equation, figure, and table flow smoothly and logically from whatever precedes it. Relevant work by others, as well as relevant products from other companies, should be adequately and accurately cited. Sufficient support should be provided (or cited) for the assertions made and conclusions drawn.

Headings may be numbered or unnumbered (“1 Introduction” and “1.2 Numbered level 2 head”), with no ending punctuation. As demonstrated in this document, the initial paragraph after a heading is not indented.

# Research Questions

Use American English when writing your paper. The serial comma should be used (“a, b, and c” not “a, b and c”). In American English, periods and commas are within quotation marks, like “this period.” Other punctuation is “outside”! The use of technical jargon, slang, and vague or informal English should be avoided. Generic technical terms should instead be used.

# Dataset

The Formula One overtaking data was gathered from www.cliptheapex.com. 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 code is discussed in the Code section.

## Network

## Network Creation

## The Formula One network is characterized by being a dense, 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.

## 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 columns.

Table

Description automatically generated

Figure 1 Nodes list where the id and label represent the drivers’ names.

## A new column was then added in which the time interval was created and in turn, the dynamic network.

## 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. Figure 3 shows the whole network when all years are used, while figure 4 displays only one season, 2018, with its corresponding color.

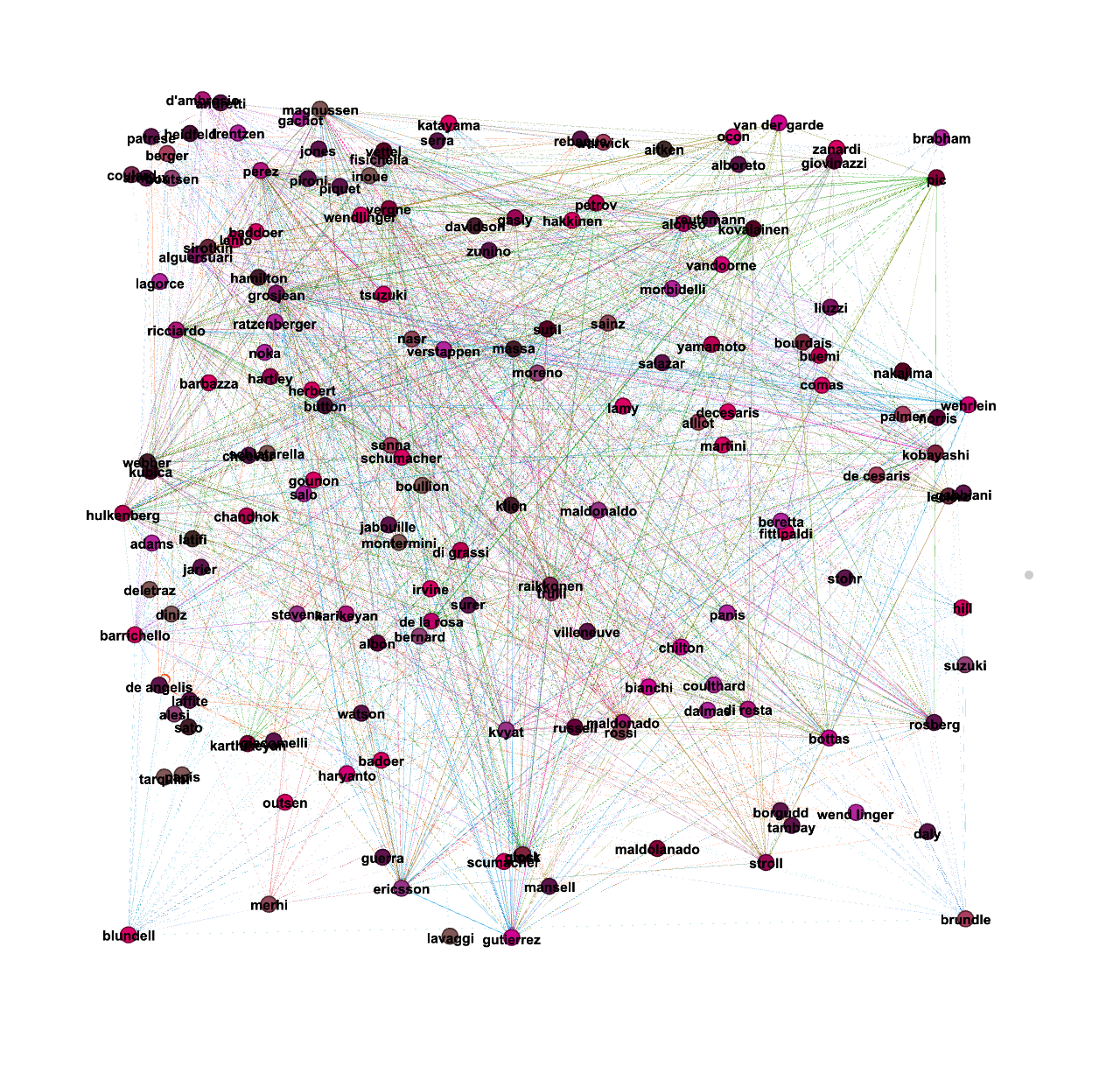


Figure 3 Whole network where every color represents a season and the texts display the drivers’ names.

A picture containing sky, day

Description automatically generated

Figure 4 2018 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

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

First and most importantly, the weighted degree distribution is a crucial metric that clarifies the network. To begin, in this case, the degree specifies the number of times one driver overtook another or was passed by someone else. To begin, the average weighted degree, although it shows the total number of passed, is unreliable due to its summing up the weighted in and out-degrees. For this reason, it is crucial to focus on the two main degree mentioned measures where the former represents the number of times a driver was overtaken and the latter the number of times this driver passed another. On the level of the in-degree, the network was filtered by node size where the larger the node, the more this driver was passed in comparison to his opponents. Figure 5 shows the network representing the in-degree 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 famous 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 out-degree network 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 a lower distribution than others who are less famous than them, like Perez and Hamilton, but this does not apply to the drivers having the highest number of out-degree. In other words, even though Grosjean has the highest out-degree distribution, he is very well known due to his being a risky and courageous player. Finally, after summing the in and out degrees of all drivers throughout the years, figure 7 displays the distribution in a graph that shows the value 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 can lead to unreliability if the in and out-degree distributions were missing.

Diagram, map

Description automatically generated

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

# Diagram, map Description automatically generated

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

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

Figure 6 Scatterplot representing the drivers’ degree distribution throughout all the years.

# 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.

## 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. Figure 1 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 1 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 2 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 2 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 was now available and feasible to work with in creating and studying the network.

## Findings and Contribution

## Findings

All acronyms should be defined at first mention in the abstract and in the main text. Define in figures, tables, and footnotes only if not defined in the discussion of the figure/table. Acronyms consist of capital letters (except where salted with lowercase), but the terms they represent need not be given initial caps unless a proper name is involved (“central processing unit” [CPU] but “Fourier transform” [FT]). Use of “e.g.” and “i.e.” okay, but refrain from using “etc.” It is preferable to use these abbreviations only in parentheses (e.g., like this).

Abbreviate units of time (s, min, hr, day, mo, yr) only in virgule constructions (10 μg/hr) and in artwork; otherwise, spell out, e.g., 10 days, 3 months, 25 minutes. Units of measure (Kb, MB, kWh, etc.) should always be abbreviated when used with a numeral. If used alone, spell out (“16 MB of RAM” but “these values are measured in micrometers”).

## Contribution

The project was completed by three members: Joy Bou Karam, Pierre Abi Chacra, and Ali Assaf. Every member contributed in his/her own way.

* Joy Bou Karam:
  + Data Preprocessing and Processing from years 2010 to 2020.
  + Converting data from picture form to text form.
  + Report writing (sections Literature Review, Dataset, Code, Network Creation, Analysis).
  + Creation of dynamic graph.
* Pierre Abi Chacra:
  + Data Preprocessing and Processing from years 1999 to 2009.
  + Report writing (sections Introduction, Idea, Research Question, Analysis, Conclusion).
  + Further data processing.
  + Machine Learning.
* Ali Assaf:
  + Data Preprocessing and Processing from years 1981 to 1998.
  + Report writing (section Analysis, Findings).
  + Machine Learning.
  + Creation of dynamic graph.

# Conclusion

# 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. IBM Corporation, IBM Knowledge Center - IBM Secure Service Container (Secure Service Container). [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/HW11R/com.ibm.hwmca.kc\_se.doc/introductiontotheconsole/wn2131zaci.html (URL)

3. J. Williams, “Narrow-band analyzer,” PhD dissertation, Dept. of Electrical Eng., Harvard Univ., Cambridge, MA: 1993. (Thesis or dissertation)

4. J. M. P. Martinez, R. B. Llavori, M. J. A. Cabo, et al., "Integrating data warehouses with web data: A survey," *IEEE Trans. Knowledge and Data Eng.*, preprint, Dec. 21, 2007, doi:10.1109/TKDE.2007.190746. (PrePrint)

5. W.-K. Chen, *Linear Networks and Systems,* Belmont, CA: Wadsworth, pp. 123–135, 1993. (book)

6. S. P. Bingulac, “On the compatibility of adaptive controllers,” *Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory,* pp. 8–16, 1994. (conference proceedings)

7. K. Elissa, “An overview of decision theory," unpublished. (Unpublished manuscript)

8. R. Nicole, "The last word on decision theory," *J. Computer Vision,* submitted for publication. (Pending publication)

9. C. J. Smith and J. S. Smith, Rocky Mountain Research Laboratories, Boulder, CO, personal communication, 1992. (Personal communication)