

**Széchenyi István Szakkollégium**

Kurzusbeszámoló dolgozat

2020/21/ősz

Generating cycling networks

Scraping from procyclingstats

Author: András Meleg

Network Science course

Chief supervisor: Sándor Juhász

Professors: Balázs Lengyel, Zoltán Elekes, Gergő Tóth, Eszter Bokányi, Sándor Juhász,

# Abstract

The assignment with its appendix provides an overview of professional cycling, a brief understanding of the uniqueness of cycling as a sport,furthermore a code and the data to generate a network of professional cyclists. The data consist of nodes and weighted links, where nodes are riders, and the links are common participations on the same professional races, all scraped down from the Procyclingstats.com website. The data consist of 894 randomly selected riders from an initial sample of 1444, who have been selected based on participations on top level cycling races between 2015 and 2019. The data can be reproduced for essentially all the riders and races after 2000. The data is scraped in python 3.6 language and after generating the dataframe the graph and the construction of the network matrix is executed in the R programming language. The generated networks can be used for standard network science analysis, with a detailed overview given for possible and necessary improvements.

Contents

[1. Introduction 1](#_Toc58195964)

[2. The essence of cycling 2](#_Toc58195965)

[2.1. Medical revolution 3](#_Toc58195966)

[2.2. Performance output revolution 4](#_Toc58195967)

[2.3. State of professional cycling 5](#_Toc58195968)

[3. Literature overview of networks 7](#_Toc58195969)

[4. Setting up the network 8](#_Toc58195970)

[5. Further issues of discussion 10](#_Toc58195971)

[6. Summary 10](#_Toc58195972)

# Table of figures

Figure 1: market size of the global sports market from 2011 to 2018[[1]](#footnote-1) (page 3)

Figure 2: [[2]](#footnote-2)The worldwide sports events market, defined as all ticketing, media and marketing revenues for sports (page 5)

Figure 3: the webpage of Primoz Roglic, showing his 2019 race results (page 8)

„The important thing in life is not to triumph but to compete.

The most important thing in the Olympic Games is not winning but taking part; the essential thing in life is not conquering but fighting well.

All sports for all people”[[3]](#footnote-3).

# Introduction

The spread of globalisation comes together with the worldwide growth of the global sports industry – as a product. This paper does not root itself to these merits, but is not independent from it. The modern technology and the modern sport together created the modern sports analysis. 18 years ago a young physicist data scientist revolutionised baseball, and probably modern sports, with developing an algorithm to help the decision-making process of the Oakland Athletics. The problem is, there are sports, which have different tools and resources – at despite cycling did not miss its own revolutions in the last two or three decades – it certainly has a less sophisticated literature of data analysis, than some other major sports – especially at the academic field.

This work sets up the goal to set up a new framework to analyse professional cycling – a network based approach. The goal is to understand to be able to understand the unique characteristics of cycling, and give system related answers, and also to understand cycling a bit more. The work is not groundbreaking in its current state, but it provides the tools for it. With the attached codes it is easy to reproduce the database and the network, and it can give some ground of analysis to understand how the world of cycling behaves, and how does it change with each passing year, while it tries to clean the dirt of the early 2000’s, an era that is historically linked for everyone to systemathic and illegal doping.

In the first part of the work, I introduce cycling. I set up a context for sports, and also differentiate cycling from other sports in characteristics, while explaining the revolutions in the professional scene in the last two decades. In the second part, I give a brief overview about academic network science literature, while in the next part I explain the logic of my code that generates the network, and my code does some very simple network analysis based on what I was taught during the network science course.

In the final chapter I lay out further ways of how to work with the provided data, highlighting possible avenues that could all be relevant research topics for tuhe future. Among these are comparing the network characteristics in different time, for example 2015-2017 and 2017-2019. There are also ways to measure different clusterisation of different type of riders, for instance how the network of some riders with certain specialities differ from each other.

# The essence of cycling

Sport always had its important role in the human civilisation, although this role might have changed time to time. There are different academic approaches about the role and context of sports in human culture, but I would like to share my own personal input as well. One could say that sport is a mirror of society, that has its origins back with the ancinet humans, when sports are suspected to have been a form of practice for hunting and therefore surviving. One such theory is from Lombardo, (2012). He argues that even todays sports keep very similar characteristics, such as the fact that men are stronger, and most sports require qualities that hunting did in the ancient age.

It would be tempting to take this for granted, but I would like to envision another perspectives either. For instance, very fresh studies tried to compare the gender gap in chess, a sport that you would not necessarily expect to require skills of hunting mammuts and tigers in the Ice-Age. (Wei, 2020). He argues that a probable performance-gap between top male and woman players in professional chess are due to participation gap – there are more professional male chess players in India than professional female chess players. Due to the natural distribution of skill among these players, it is explainable why top male players perform better than top female players. Although the article originally were a big hit recently, others pointed out that India is an outlier data point (just as Hungary) compared to other countries. (Camacho Collados, 2020)

The goal of this research is to set up a network about cycling, and do analyse that. It is fair to raise the question why chess, or gender gap, or the cultural role of sports is relevant? The point I would like to make here is the fact that despite major advances in data science worldwide, sports and data driven science is still just an emerging area of academic research, with limited resources, and numerous approaches still available for those who want to push their path on this field. It also sets up certain limitations and options further on, for instance why this research restricts itself for professional male cycling.

One thing we should be certain though: we should analyse sports on academic levels, it should not be a neglected area of research. It should be evident that the market concept of sport is currently emerging globally, sports become more infuluental in global economics, and it is strongly underregulated. (Nauright & Zipp, 2018)

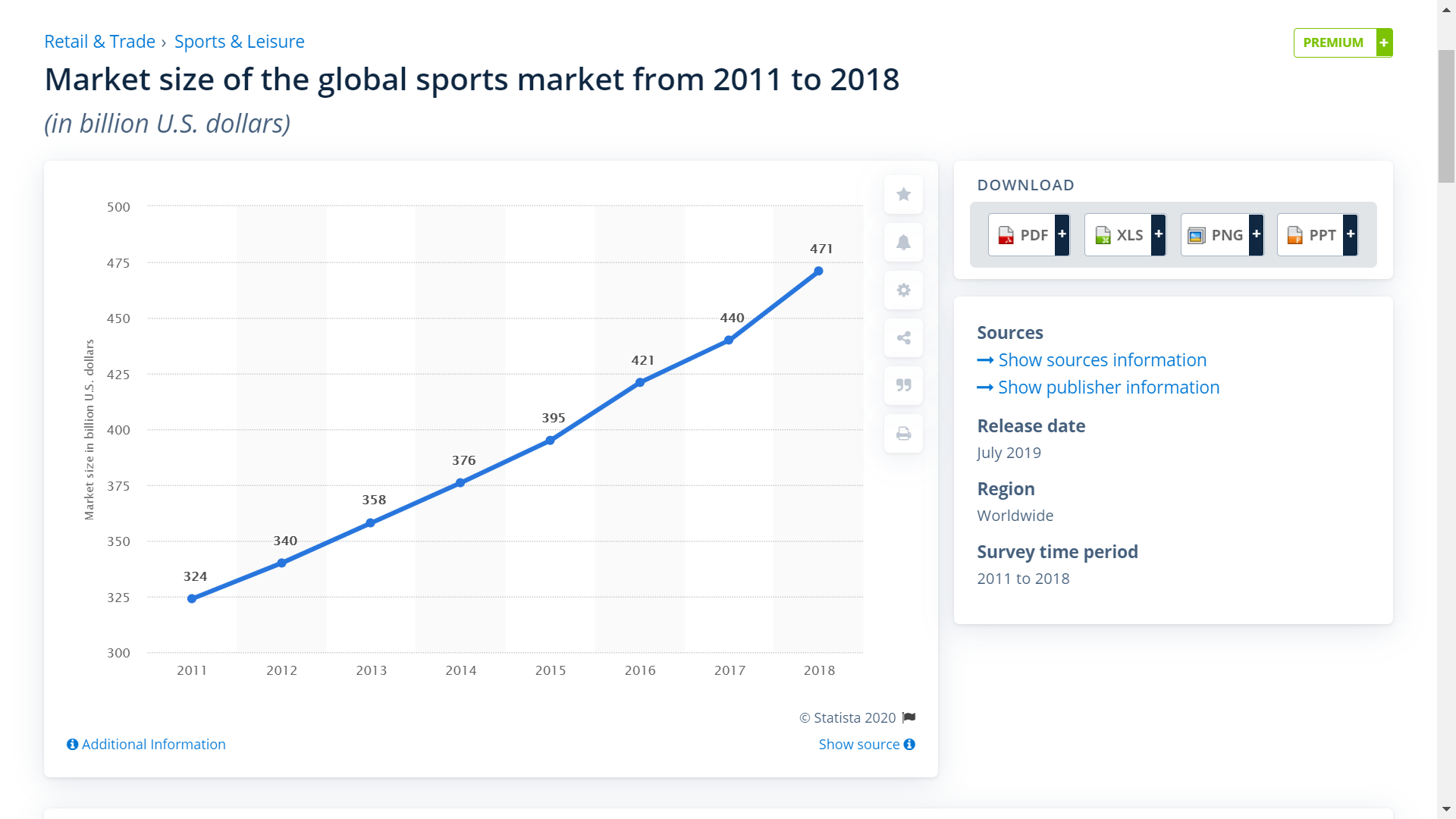


Figure 1: market size of the global sports market from 2011 to 2018[[4]](#footnote-4)

While private sport companies are already pushing data driven analysis for around two decades (Lewis, 2003), cycling has also had its on data revolution. Probably cycling is most notoriously famous about its doping scandals for the unexperienced eye, but there actually have been two major scientific revolutions in cycling since the end of the cold war, the *medical revolution* and the *performance output revolution.*

## Medical revolution

Arguably the most infamous cyclist is still Lance Armstrong – at the dusk of the previous century he was diagnosed with fatal testicular cancer. He managed to survive it, and when he restarted cycling, he managed to achieve outstanding results – results, that noone ever could achieve before – and nor could he. As it turned out, the late 1990’s and the early 2000’s when Armstrong was in his prime and seemingly won the Tour de France[[5]](#footnote-5) between 1999-2005 7 times were part of a major fraud, as his team, US Postal systemathically doped during those years. Despite the fact that journalists raised certain suspicions, Armstrong and co. systhematically kept refusing all allegations, and intentionally tried to destroy the reputation of those who crossed their lines. After more than a decade of investigation, finally Armstrong camouflage got torn apart in 2013, and he was forced to admit that he did doping. (Wall Street Journal, 2013) But already, the ideology of his era have already did their harm to cycling and for the reputation of cycling.

If you are interested in more depth about how doping became a part of cycling, why it became so serious, who were the first victims, what moral questions does doping raise, you can always ask the author of the research for some extra knowledge – but this work does not allow us to go into such lengths. What it is necessary though is to understand how this could happen, and why should it make analysing cyclist networks an interesting tool regarding this. And I would argue that the reason for this is that cycling is a relatively simple sport.

There are many sports that require much more complex movements, much more technique and coordination, compared to road cycling. Cycling is essentially from a very simple movement: two pedals powered by legs rotate around an axis, and this is almost everything the human does in cycling. This is a large simplification, but also true on the other hand – cycling is seemingly simple. But its also competitive, and therefore it would seem natural that everyone tries to find and edge. But like in other sports these edges can first materialise in different tactics, lineups (team-based ballgames), or probably other kind of equipment, or a careful data analysis of players (Baseball, etc.) a sport such endurance based as cycling probably more naturally turns toward medical innovations in the first place. The good thing is, this does not necessarily means pure harm, because competition produces development, and a development of health and antidoping measures may prove as a fair price for society, for the vandalism of the doping era. *But can we ever be certain that the sports are clean, anymore?*

## Performance output revolution

If you can’t dope your riders illegally, better do it legally. This could have been the motto of the 2010’s, the era, that seemingly is an era of cleaner professional cycling than its predecessor. The previous 10 years were mostly dominiated by Team Sky / Ineos, a Brittish cycling team, and also their thechnical approach, the *wattage.* To put it simply, rather than focusing on speed or heart rate as measurements of performance, they focused on power output – utilising the technical concept of power meters. (Global Cycling Network, 2016)

Team Sky utilised the boundaries of human physiology and cycling’s existing strategy. The concept is fairly simple: you win races, if you are at the front of the race. Staying in the front of the race costs more energy, than going slower at the back, but human physiology has two very important threshold in competitive racing, the *lactate threshold* and the *burnout threshold.* If you perform over a certain threshold, your body starts to use carbonhydrates instead of fat, and human generally tend to be capable to perform at the edge of this threshold for an hour.

Performing on the edge of this threshold allows endurance athletes to produce a stable amount of power for an extended amount of time, but makes it very hard to overreach, and certainly easier to reach burnout threshold, where you can only perform for seconds, before you empty the tank. The innovation is power meters is the fact that they realised that the body and the heart reacts much lower to altering power, therefore it gives less accurate results of performance, especially because BPM is also a dependent on other variables as well, for instance how much coffee the rider drank in the morning.

As power meters allowed to measure the power output during every turn of the pedal, Team Sky got more efficient in training, and they could also race more on the lactate threshold with a more strategic approach than other teams did, especially because they had the strongest riders in their team. As a result, they sort of dominated the early years of the 2010’s, until very recently.

## State of professional cycling

As it has shown above, cycling is in constant motion and evolution, but it is probably still could be argued to be less developed compared to other sports in the world. This is clearly illustrated by Figure 2.

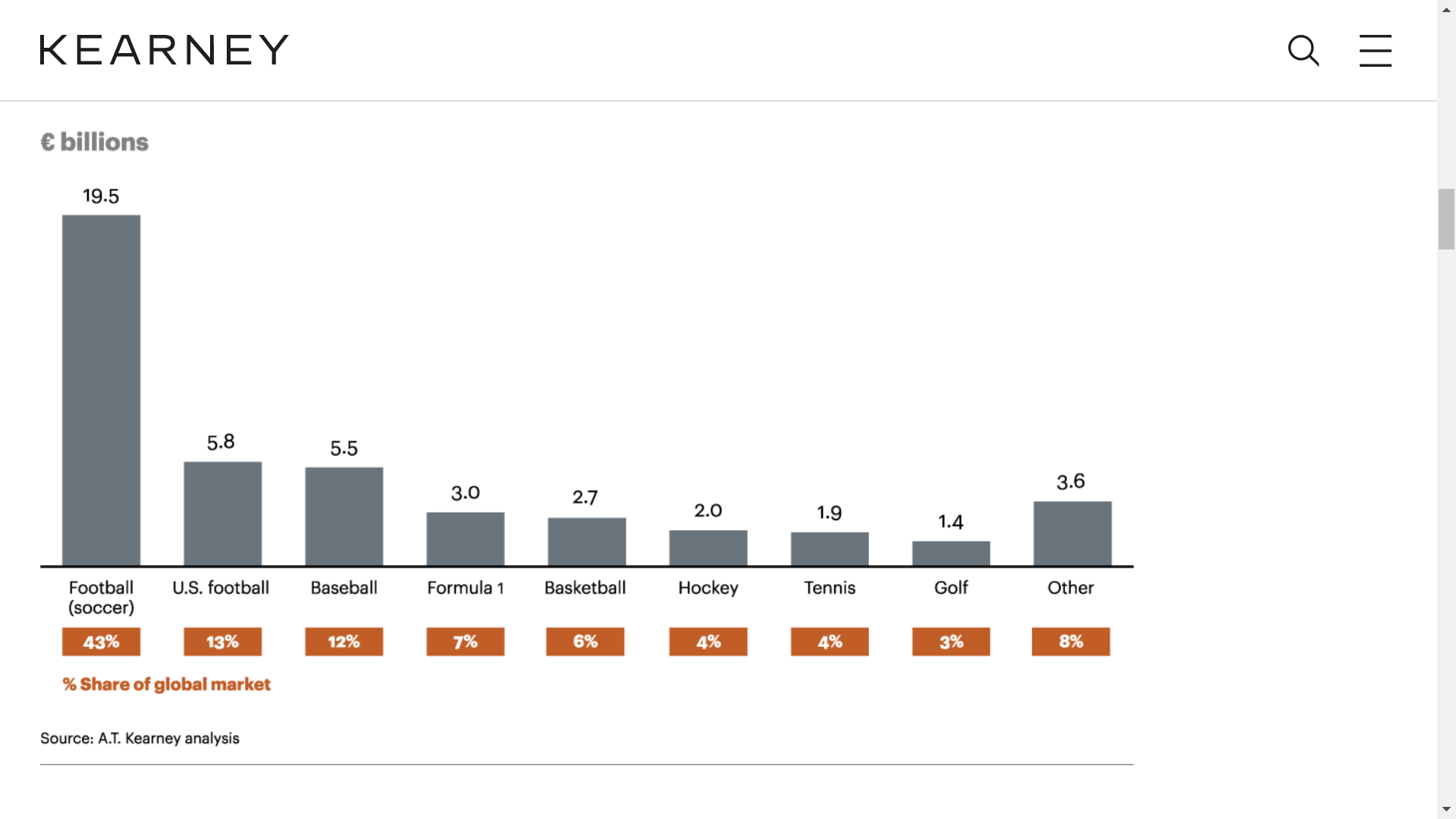


Figure 2: [[6]](#footnote-6)The worldwide sports events market, defined as all ticketing, media and marketing revenues for sports

Therefore, cycling is supposedly using less resources than those traditionally superpopular sports. Top level professional cycling teams has to deal with certain issues, that limit the research about cycling, and before moving on, I would like to point out a few factors, so it becomes clear why cycling networks wasn’t yet analysed in specific depths yet. Obviously, if more people watch soccer than cycling, the revenue becomes smaller, and teams will have less resources, but there is more about that.

First of all, top level professional cycling (World Tour level) has a fixed calendar every year, and teams are obligated to participate. This means that professional teams has to allocate resources of travelling around the world, compared to ball-sports, where teams are usually located in the same country, or at least the same continent. This system is probably similar to Formula 1, but F1 has a much larger revenue.

Second on that, cycling has no income for participation, therefore teams has no access to income from participation fees, which is probably very specific to most of the sports around the world. Even other endurance sports such as athletics or skiing can generate ticket income, but cycling does not. This means that only reliable income of teams come from sponsorhips. But this can be as reliable as much the sponsors are committed, and its extremely hard to find sponsors since the doping scandals.

There is a compensation system involved, as race competitions give race rewards based on performance, but the teams have no input on these, and this is the third issue – in cycling, the race organisers has much larger effect on the sport than in other sports, for instance the ASO company organises the two of the three biggest races (Tour de France and La Vuelta), and just this gives company a strong enough position in the cycling world to have control over things, and when there is less competition, there is less progress.

For all the following reasons I think it is important not to turn the back on cycling, and do good quaility empirical research, because it is a beautiful sport, which is very popular, and accessible for everyone. But probably the motivation of this paper comes from two other reasons as well. I think cycling is a great thing, and I would like people to know that how much is to adore about it.

But also, cycling has a unique nature, that it requires constant cooperation by its participants, therefore people are very interconnected all the time. In cycling, it is very hard to reach advantage, due to the wind resistance, and riders usually form a *big bunch[[7]](#footnote-7)* and cooperate, even when they are racing against each other. What this means is that the relationship of riders and teams may have just as important in formulating tactics and strategies, as the strength of the riders. It is also important to point out that this makes cycling a very complex network, because the individual strength of riders in a team, or in different subsets of riders may vary based on how much they can cooperate together, or how much they can rely on each other.

**Professional cycling is basically a multidimensional prisoner’s dilemma, in constant motion, and it makes understanding and researching cycling an incredibly unique challenge.**

# Literature overview of networks

Network Science is unique combination of traditional data science, quantitative economy, sociology, and arts. It uses the tools of data science, such as programming languages and linear algebra basics. It uses quantitative economy, when trying to find correlation between values of and correlation of hidden measures in networks. It also roots in sociology, not just historically, but in logic as well. (Borgatti et al., 2009) The inspiration of network science is to understand networks, and use this information to align it with our knowledge about the world. (Hidalgo, 2016) And finally, networks and graphs are more than just doing some algebra on matrixes, and the representation should catch the heart and the eye, so it is safe to say it is closer to arts than the author of this work will ever be.

Networks generally have two different type of elements: nodes, and links. Nodes are the elements in the network, whose behaviour we want to understand, and mathemathically they are the points of the graph. Links are connections between these nodes, that can either be directed or non-directed, and can have quantitative or qualitative attributes too. A qualitative attribute is a discrete, cathegoric differentiation between links, for instance when we are connecting cyclists, it can be teammate-teammate link, or can be teammate-nonteammate link. Quanitative link attributes are different, they give different weight to each link, instead of defining a connection with a logical boolean (True or False) it gives it a numeric value ( 0.5, 2, 130, etc.)

In the network science course we learned several measurements that can help to describe the weight of networks, among many, there are a few important ones, such as edgebetweenness (e), average distance (L), average number of links (k), node-link ration (d) fragmentation (F), clusters , coreness (c).

But above all measurments, the most important element of network science is the network itself, therefore, we should understand, what kind of networks are we looking for, and how can we use it. After careful evaluation, as a first framework approach to set up professional cycling’s network analysis, it seemed like races as connecting links, and riders as nodes, could be a sufficient network.

# Setting up the network

In order to construct a network, I had to collect data, and this turned out to be a hard challenge. I have decided to scrape the data from the internet, using the data available at procyclingstats.com.

Procyclingstats is a website, that collects the results of professional cycling road races, even with some archived results back to the 90’s. Today, every race result is online on procyclingstats. Each rider has a webpage, that contains their individual results on each race, and also some rankings. You can see it in a screenshot from procyclingstats at figure 3.

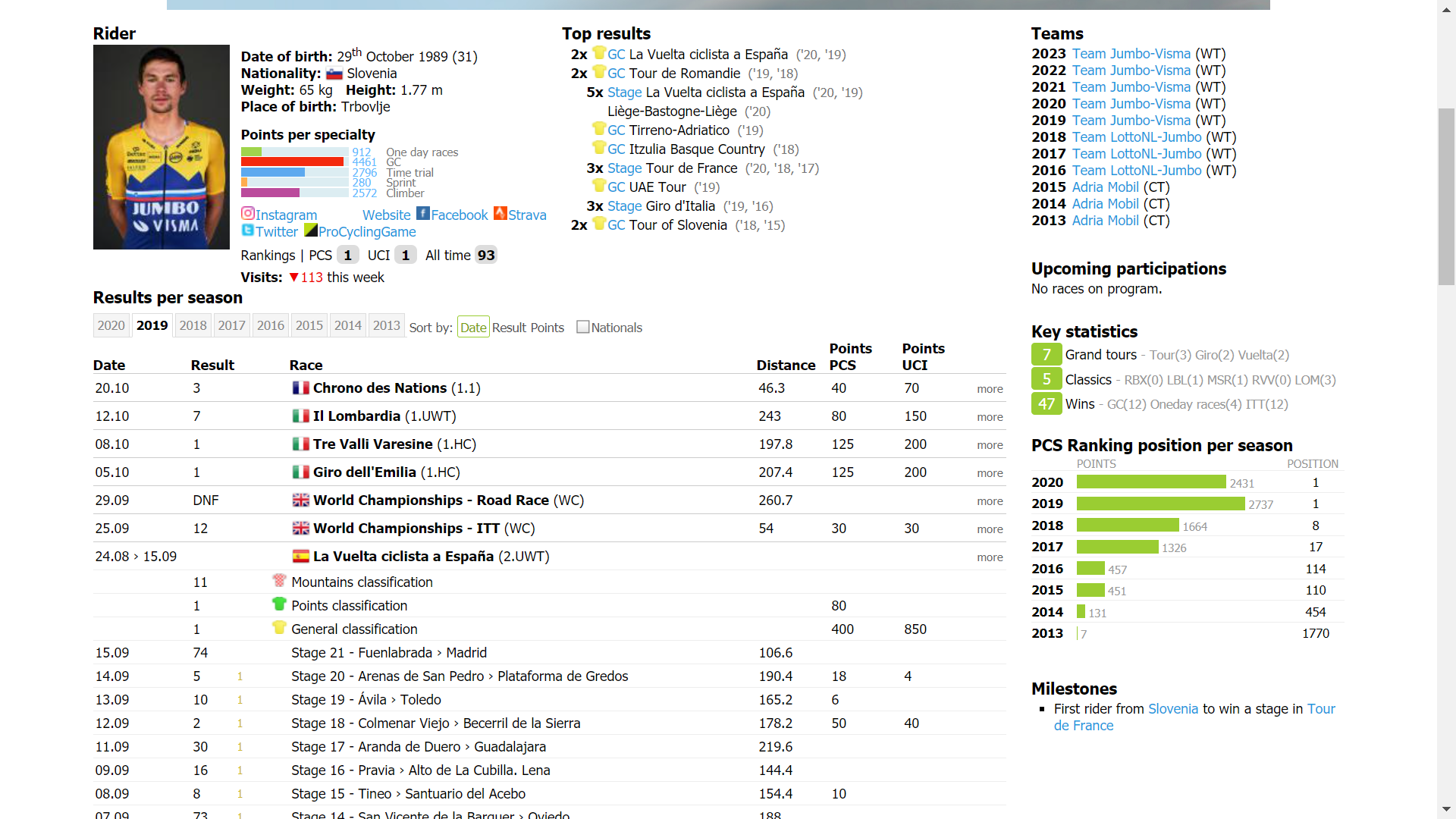


Figure 3: the webpage of Primoz Roglic, showing his 2019 race results

The first goal was to scrape down the data of the individual racers. After some research I’ve decided to use the work of arjundahir,( 2018) from Github. His project, *ProCyclingStats* *Points Prediction* aimed to use a test-train-split method in order to predict how successfull in the site’s own poing ranking system each rider will become according to previous results. He reached a good correlation rate, and his method was sufficiently general in terms of data collection, that I could use his original code, but using it for my own purposes.

His idea was to manually choose some high-prestige races (such as Tour de France, etc) and download to startlist of these competitions to separate lists. After downloading the startlists, the code would filter out the duplicate participants. Afterwards, each individual riders html profile is downloaded, for all the seasons the rider has participated. Finally, after downloading the html files, the code uses a BeautifulSoup html parser to collect all the race results from the individiual seasons of the riders, all the races that all the riders have ever participated.

I had to make a few modifications on the code, but I did not alter the logic of this method. First of all, I have decided to increase the initial race sample size. [[8]](#footnote-8)In the end, I have collected 1444 professional World Tour Riders, who have participated on any of the mentioned races. Afterwards, I have downloaded the individual files for each rider, but this took a while. (In fact, it would take an extra day of computational scraping to download all the files, so at this moment I only have the data of 894 of the 1444 riders.) I created my own code to get the necessary informations from this point, because for the network I didn’t need that much data, and the environment of my 2020 python libraries are probably outdating the 3 year old versions. (There is always the option to set up older environments, but probably its also a better way of learning if I do write my own scripts.)

I created a loop that iterates through each downloaded html file. The loop first collects the name, then creates an empty list to the racename, racetype, and raceday length for each race, and when the loop iterates through a race, appends the racename to the races list, the type of race (1 day race or stage race) to racetype, and the total number of stages of the race to the racedays list. After this, still in the same loop mentioned at the start of this paragraph, the algorithm puts the rows into a dataframe, as each row has the rider’s name, the name of the race, the racelength, and the racetype. (Although racetype ’gc’ or ’lt’ code basically means a logical connection of True if lt = 1, and False if gc <1, so it does not contain specfic information. After the iteration, I have saved this pandas dataframe into a csv file, and opened an R notebook.

In the R notebook, I have paraphrased a work from one of the courses, when my mentor, Sándor showed us how to generate a network from a CSV file, with given very similar data – but that network was not weighted. I got stuck for a while here, I didn’t really know, how to combine the racelength into account, but I think it is correct to expect that if you ride together on a 5 raceday long race with someone, you will be more connected than riding together on a 1 raceday long race. I would probably argue that the connection will be even stronger – after all, if you go for a one day race, you only show up around the start of the race, do the race (when there is probably limited amount of talking), and then leave the location. If you travel together with a peloton for many days, there is more chance for connection. So I have decided to go back to the pandas dataframe, and duplicated each stage race row for the number of racedays the given race had. After this, I once again generated a network in the R notebook, that now weighted races according to their length in racedays. The network is ready.

# Further issues of discussion

This has only been a staring point, but the work is far from over. There are further ways for improvement, and I hope that those who evaluate this work may give me some useful feedback regarding which directions to pursue.

It would be probable a good idea to set up certain thresholds in the network, and only evaluate those edges that are above the threshold, as unweighted networks. It could give some interesting results, to see how the network is changing, and trying to find some patterns in the future. What will be the connection between L, k, and transitivity? Another approach could be to try different methods to see the clusterisation of the network, as clusters of teams are expected, but also clusters based on the terrain speciality of the riders, such as one day specialist or general classification specialist.

Checking overlapping data could also be helpful, especially if our simple size is extended to the doping era as well. Hopefully we will see a little bit different network structure this time, which is not correlation with a cleaner sport, but maybe a point to the more just direction. But there are deeper questions to ask here – what if we compare the ego network of proved dopers with a random sample of riders? Are we going to see different characteristics? Is it likely for instance that a shrinkage of connectedness indicates a more likely transition towards doping? How does different clusters tend to have different ego networks? These are the question for future research.

# Summary

In the first part of the work I did explain the relevance of sports in human culture, and how there are no evident truths about sport. I did also explain how cycling is different from other kind of sports, why the resources of professional teams are limited, and why it prevents the sport from rapid development. I did point out how the sport has changed systemathically in the recent years, and why it gives great roots for scientific reasearch. I did pointt out the doping era and the widespread introduction as power meters, as relevant factors.

In the second part, I did explain how network science is different from other fields of science as an interdisciplinary science. I also gave a brief introduction to basic concepts and definitions. Then I provided the logic of my codes, what they do and how they work, in order to create the data for network analysis. I generated the data of 894 of 1444 selected riders from a simple to set up the network, and I collected every professional race these riders have ever participated. I stored the race day have participated, the year the race happened, the rider who participated, and also how long the race was. Altogether this data contains of 12000 years of individual race seasons. Afterwards, I calculated the edgebetweenness values for the generated network.

Further research is required to improve on this assignment, and I have mentioned a dozen of ways to start it, once all the data will be downloaded for it, not just 894 of the 1444 riders results.

## Bibliography

arjundahir. (2018). *ProcyclingStats Points Prediction*. GitHub. https://github.com/arjunsudhir/procyclingstats

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Corrected 24 april 2009; see last page. *Science*, *323*(April), 892–896.

Camacho Collados, J. (2020). *The gender gap in top level chess*. Josecamachocollados.Medium.Com/. https://josecamachocollados.medium.com/the-gender-gap-in-top-level-chess-15591d8990ba

Global Cycling Network. (2016). *How do powermeters work?* YouTube. https://www.youtube.com/watch?v=nzXZCNa6QLY

Hidalgo, C. A. (2016). Disconnected, fragmented, or united? a trans-disciplinary review of network science. *Applied Network Science*, *1*(1), 1–19. https://doi.org/10.1007/s41109-016-0010-3

Lewis, M. (2003). *Moneyball*. WW Norton & Co.

Lombardo, M. P. (2012). On the evolution of sport. *Evolutionary Psychology*, *10*(1), 1–28. https://doi.org/10.1177/147470491201000101

Nauright, J., & Zipp, S. (2018). The complex world of global sport. *Sport in Society*, *21*(8), 1113–1119. https://doi.org/10.1080/17430437.2018.1469846

Wall Street Journal. (2013). *Lance Armstrong Admits to Doping - Armstrong Confesses to Oprah*. YouTube. https://www.youtube.com/watch?v=u1qf9TlFTl4

Wei, J. M. (2020). *What gender gap in chess?* Chessbase.Com. https://en.chessbase.com/post/what-gender-gap-in-chess

1. Obtained from <https://www.statista.com/statistics/1087391/global-sports-market-size/> AD: 06/12/2020 [↑](#footnote-ref-1)
2. Obtained from <https://www.de.kearney.com/communications-media-technology/article?/a/the-sports-market> AD: 06/12/2020 [↑](#footnote-ref-2)
3. Pierre de Coubertin, founder of the International Olympic Committee [↑](#footnote-ref-3)
4. Obtained from <https://www.statista.com/statistics/1087391/global-sports-market-size/> AD: 06/12/2020 [↑](#footnote-ref-4)
5. The Tour de France is the most prestigious bike race of the world, and winning it is the highest achievement in cycling [↑](#footnote-ref-5)
6. Obtained from <https://www.de.kearney.com/communications-media-technology/article?/a/the-sports-market> AD: 06/12/2020 [↑](#footnote-ref-6)
7. Which is called peloton [↑](#footnote-ref-7)
8. My race sample size for startlists, between 2019-2015: Tour de France, Giro d’ Italia, Vuelta Espana, Strade Bianchi, Paris-Nice, Tirreno-Adrietico, Dauphine Libere, Milano Sanremo, Gent-Wevelgem, Ronde-van-Vlandeeren, Paris-Roubaix, Amstel Gold Race, Tour of California, Liege-Bastogne-Liege, Tour de Suisse, Il Lombardia, Tour de Romandie [↑](#footnote-ref-8)