A Network Approach to High School Career and Technical Education in Chile

Pablo Geraldo Bastias

April 5, 2018

Abstract

I propose a network representation of the Career tracks offered during the High School in the Chilean context, in order to understand the relationship between tracks, study some structural properties of the network, and explore the Careers distribution in different regional contexts. Using administrative data from the Chilean Ministry of Education (2017), I first constructed a bipartite directed network in which schools has an edge pointing to the careers they offer, and the edges are weighted by the school enrollment in such careers. Then, I projected such network onto the groups, i.e., onto the careers, giving place to a cocitation network of careers connected by an edge if there is a school offering both careers, and the edges are weighted by the total number of schools that co-offers such career tracks. Using this network, I run a community detection algorithm, in order to identify clusters of careers being co-offered more frequently. I use those results to make some conjectures on the distance of different career tracks from the academic track and, therefore, to provide some hypothesis for future research on the consequences of different tracks for future academic attainment.

Career and Technical Education in Chile

In Chile, almost 40% of students attend Career and Technical Education (CTE) during the last two years of their high school. Previous research have shown that, although students are not officially sorted into the academic or vocational track by external means, nor they have to decide what track to follow until 11th grade, usually their parents make the decision for their children during their 8th grade, i.e., at least three years in advance to the formal beginning of the CTE, in a way that reproduces the socioeconomic inequality in the educational system (Farias, 2013; Geraldo, 2015). Students attending CTE come from families with lower income, less formal education, and lower academic expectations. Being exposed to CTE schools further decreases their expectactions and academic performance.

In addition, it is not only about attending CTE in general, because being enrolled in a particular school means to be attached to a limited set of options in the CTE track, because most of the schools are specialized not only on academic or CTE education, but also in specific training areas within the CTE. This could further affect students in relation to their future labor and educational plans. On the other hand, CTE is usually presented as a locally and economically pertinent type of education, in close relation with the local economic environtment, which would signify for the students immediate and advantageous working conditions in case of early insertion on the labor market. However, the scant research available on the Chilean case shows that this is not always the case; only a limited number of industrial specializations would offer labor market benefits to students (Bucarey and Urzua, 2013).

In this paper, I propose a network representation of the Career tracks offered during the High School in the Chilean context, in order to understand the relationship between tracks, study some structural properties of the network, and explore the Careers distribution in different regional contexts. I use administrative data from the Chilean Ministry of Education (2017) to construct a bipartite directed network in which schools has an edge pointing to the careers they offer, and the edges are weighted by the school enrollment in such careers. Then, I projected such network onto the groups, i.e., onto the careers, giving place to a cocitation network of careers connected by an edge if there is a school offering both careers, and the edges are weighted by the total number of schools that co-offers such career tracks.

The idea underlying the use of this one-mode projection onto the careers is inspired by the development of the "product space" (Hidalgo et al., 2007). I understand two Career tracks as connected if they are offered by the same school, and using this model I study the distance between different tracks and economic sectors. One could expect that two Careers could be offered in the same school either because the school resources and capabilities needed to offer both are somewhat related, or because one of the Careers is in general easy or cheap to implement. Trying to rule out this last possibility, and in line with previous research, in a subsequent analysis I prune network edges keeping only those that pass a certain treshold in terms of Revealed Comparative Advantage.

Using this network, I run a community detection algorithm, in order to identify clusters of careers being co-offered more frequently. I use those results to make some conjectures on the distance of different career tracks from the academic track and, therefore, to provide some hypothesis for future research on the consequences of different tracks for future academic attainment. Overall, my aim is to suggest that understand the structure of the Career and Technical Education network of training is really important in order to propose policies that, retaining the advantages of the model in terms of labor market outcomes, do not reproduces inequality in subsequent study trajectories of the youth exposed to CTE.

Data and Network Construction

Data Sources

The data from this project comes from administrative records of the Chilean Ministry of Education[1]. The entire dataset contains the total number of students enrolled in Chilean schools during 2017 (3,558,394), but I only used the cases of high school students in grades 11th and 12th, when the Career and Technical Education is offered. This produces a dataset of 410,494 students disrtributed among 2,903 schools: 62% of the students attend academic education, while the remaining 38% are students attending one of the 57 career tracks (belonging to five different economic sectors) offered in Chilean schools.

Network Construction

After filtering the relevant observations, I constructed a bipartite directed network as follows: each school s has an edge pointing to a particular career track c, if that school has students enrolled in the career during the year 2017. For some visualizations and analysis, I used a weighted version in which the edges were weighted by the total enrollment that the school has in that career. Formally, the incidence matrix ($\mathbf{B}$) of this network corresponds to an $c \times s$ matrix, where c is the number of careers (57 plus the academic track), and s is the number of schools (2,903):

$$B_{c,s} = \begin{cases} 1 & \text{if school } s \text{ offers career } c \\ 0 & \text{Otherwise} \end{cases}$$

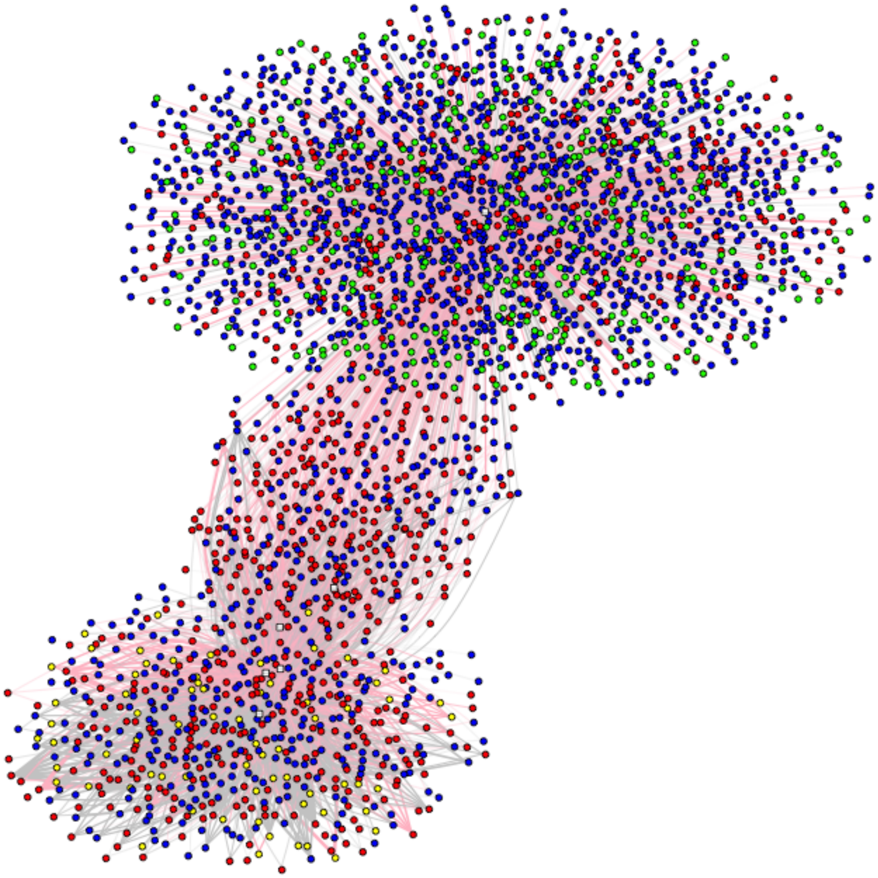
I constructed a second network following the same procedure described above, but this time the connections are not between schools and careers, but between schools and economic sectors. Recall that each career belongs to a particular economic sector, either Business, Industry, Service, Farming, or Maritime. This network converys similar information than the full career by schools network, but is more useful for visualizing and some analysis, so I will use alternatively the aggregate (sector) or desaggregate (careers) graph.

The resulting school by sector network is visualized in the Figure 1. One can infer, from this representation, a serie of characteristics of the high school system in Chile. First, the largest component corresponds to the academic education, which as I said provides education for more than 60% of students during grades 11th and 12th. Among the schools offering the academic track, there are Public schools (red ones), Subsidized schools (similar to the charter schools in the US, in blue), and notoriously all the Private Schools are in this component (depicted in green).

On the bottom of the figure, we have the cluster of schools offering vocational education, mainly Public and Subsidized schools, but it is also notorious the presence of Corporation schools (in yellow), leaded by firms

[1] Available here.

Figure 1: Bipartite Graph of Schools and Career Tracks

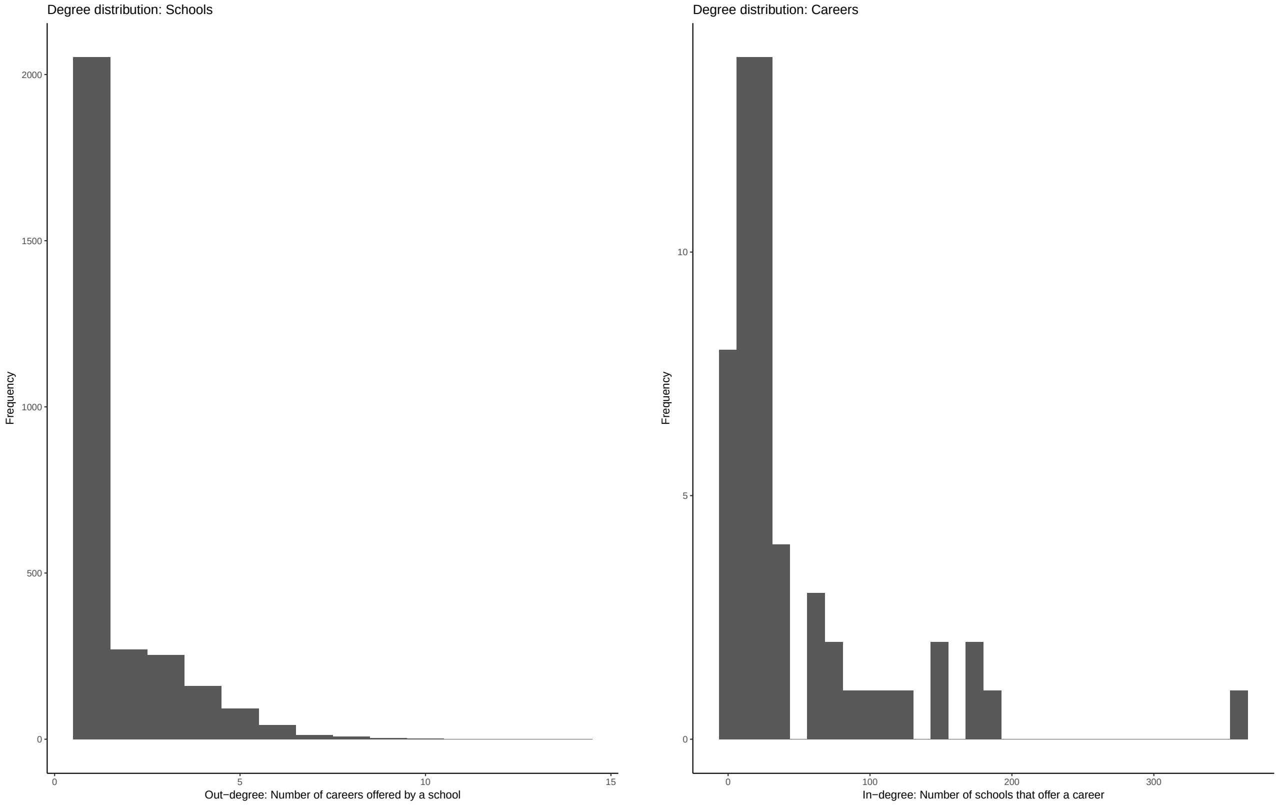


Note: Circle nodes represent schools, with the color denoting ownership or administrative dependence (red: public, blue: subsidized, green: private, yellow: corporation), while the white boxes correspond to career tracks. Edges are weighted by the school enrollment in the particular track, and colored according to if the proportion of women enrolled in the track is more than 0.5 (pink edges), or less (grey edges).

and bussiness associations offering training with strong connections to the workplace. In the middle, finally, there are some schools, Public and Subsidized, that offers both types of education, academic and CTE. I will further refer to this schools as "mixed schools". Therefore, in general, Public and Subsidized schools are track-diverse, i.e., there are some offering exclusively academic education, others offering exclusively CTE, and also some mixed schools. On the other hand, Private schools are completely in the academic track component, while Corporation schools belong completely to the CTE track component.

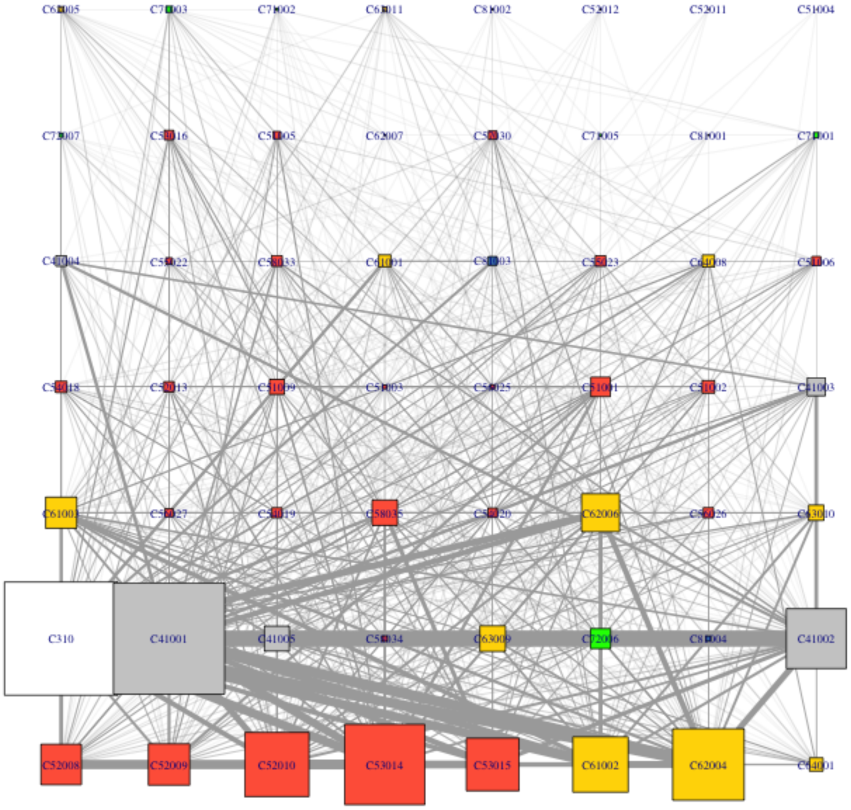
Another important aspect of this network is encoded in the edge color. In Figure 1, edges are pink if more than 50% of students in a given school and career are women, and grey otherwise. Most importantly, it is clear that within the CTE cluster, there are some economic sectors that concentrate more proportion of women enrollment, and others with more proportion of men enrollment. In the following sections, I will further analyze the characteristics of the network, more precisely quantifying some of the intuitions just developed, such the distribution of sectors and careers by school ownership and gender enrollment.

Figure 2: Degree distribution of Bipartite Projection



Figures 2 report the in-degree and the out-degree of the resulting directed bipartite networks. In the case of out-degree (upper panel), this refers to the number of careers offered by a school. As we can see, most schools offer at most five tracks, including the academic one, but there are cases in which some schools appear offering more than ten different career tracks. Certainly, there are reasons to be worried about the capacity of a school to offer high quality training in so many different areas, in this would also explain why most school offer a few options. On the student side, however, this shows the lack of alternatives that a student faces when enroll in a certain school. Regarding the in-degree (bottom panel), i.e., in relation to careers offered by schools, it is possible to see that some career tracks are offered by less than 50 schools in the country, probably (as we will further explore below) in a geographic-specific way. There are some careers, however, that have presence in more schools, even more than 300. In this plot, the academic track was removed, because it has an in-degree of more than 2000.

Figure 3: Projection onto the Careers



Note: Nodes represent CTE tracks (careers), and are colored according to the economic sector that they belong to (White: academic, grey: business, red: industry, gold: services, green: farming, blue: maritime). The node size and edge width are weighted to represent the amount of connection shared, i.e., the number of schools that offer each career and they in conjuction to others.

To conduct some of the analysis, I also work with the network projections onto the groups (careers) and onto the people (schools). The projection onto the groups, i.e., onto the sectors or careers, produces a cocitation matrix, in which there is a link between two careers when the same school “cite” (provide training on) those careers. For the projection onto the people, i.e., onto the schools, two schools are connected if they offer the same CTE tracks, giving place to a bibliographic coupling matrix, in the sense that two schools cite the same careers.

Figure 3 shows the projection onto the careers. I used a grid layout to prevent nodes overlap and to highlight the weights of the edges. It is clear, from this visualization, that in addition to the academic track, the careers that are most frequently offered by schools in conjuction to other careers are from the business sector, followed by industry and services. With far less connections, we observe farming and maritime careers. It is also noticeable that, although many career are cocited with others from the same economic sector, there is also an important cross-sector connection, such as schools offering both industry and business, or business and service careers.

Methods and Analysis

In this section I describe and apply several methods that permit to test more formally some of the hypothesis provided when describing the network constructed. In particular, I am interested in looking for different types of assortativity of schools and careers in relation to ownership, economic sector, and gender enrollment; in evaluating the proximity of different careers in what could be called the “training space”, eventually detecting communities of careers; and finally, in identify some local variation at the regional level in the type of careers offered.

Assortativity in the CTE Network

It is a established regularity in social networks that vertex of the same type tends to connect each other. One could think many questions for the CTE careers network in terms of assortativity, or the tendence to observe links between simmilar vertex. For example, one could expect that schools offering more careers would be offering careers from the same economic sector, because they would be more similar to implement; on the other hand, for different reasons one could exactly the opposite, because in a market-oriented educational systems, schools would try to diversify their offer to attract more students. In this sections, I will briefly confront three different questions that could be thought in terms of network assortativity:

- Are schools of the same ownership more likely to offer the same careers?

Recall that, in the projection onto the schools, two schools would be connected if they offer the same career. Given the information of the school ownership type (Public, Private, Subsidized, Corporation), I can explore if there is ownership specialization (in which case we would observed assortativity) or diversification (in which case we would observe dissortativity).

- Are careers jointly offered more likely to be from the same economic sector?

This question can the answered in the projection onto careers. As above, careers are connected if they are jointly offered in a school. Does this connections happen more frequently between careers of the same economic sector (specialization), or not (diversification)?

- Are careers jointly offered more likely to have a similar gender-orientation?

To answer this question, we can use the projection onto groups. Recall that two careers or economic sectors are connected if they are offered in the same school. We also now the overall proportion of men and women enrolled in such tracks, what we can use as a proxy of gender orientation. For example, Industry careers are prevalently male-enrolled, while Service careers are more female in enrollment.

The first two questions corresponds to assortative mixing on discrete characteristics, because there is a small number of categories to which each node could belong (4 ownerships, and 5 economic sectors plus the academic track), while the thrid question is better understood as a question on assortative mixing on continuous or scalar characteristics (because gender enrollment moves continuously from 0 to 100 %).

Newman (2003) define the following *assortative coefficient* as a measure of assortative mixing on discrete characteristics:

$$r = \frac{\sum_i e_{ii} - \sum_i aib_i}{1 - \sum_i a_i b_i}$$

Where a_i and b_i are the fraction of edges starting and ending in nodes of type i respectively (both are the same in undirected network as our projections, but in the directed case the distinction is relevant). This measure is bounded between -1 and 1, being 1 for perfect assortativity and 0 for random mixing. Negative values indicate dissortative mixing.

We can use this measure to evaluate our first two question. Are schools of the same ownership regime (Public, Subsidized, Corporation, or Private) more or less likely to offer the same careers, or careers from the same sector? Results of Table 1 show a contrasting scenario. If we consider all the schools in the dataset, in fact there is assortativity in both the sector and the careers offered in schools of the same ownership (signed by the positive values). Nevertheless, one question emerges: could be this assortativity being driven by the presence of Private schools? In fact, Private schools are all connected to each other, because they offer *only one* type of education, the academic track. This could be increasing our assortativity measure.

Table 1: Assortative mixing in the CTE network

School Ownership	All Schools	Excluding Privates
by Sector	0.015	-0.393
by Career	0.008	-0.412
Economic Sector		
by Career	0.066	0.093
Women Enrollment	All Schools	Excluding Privates
by Sector	-0.200	-0.200
by Career	-0.027	-0.039

If we remove from the sample Private schools, in fact, the result reverses. Now, and to a much higher degree, schools seem to be dissortative by ownership, in terms careers and sectors covered. Although this results deserve more attention, a potential explication of the pattern would be the fact that, at least at the local level, schools could be trying to differentiate their educational offer from one another. One could expect that the need for more differentiation in relation to careers happes precisely among schools of the same ownership, as they probably share a common space in the education market.

Regarding question two, the result reveals some degree of assortativity in relation to economic sector, but not as much as one could expect given the hypothesis of similarity in their implementation. This means that schools tend to offer jointly careers from the same economic sector, but there is also a fair amount of jointly offered careers across economic sectors. This results holds even if we remove the Private schools from the sample.

Question three, on the other hand, require the use of the assortative mixing measure for continuos characteristics. The coefficient is defined analogously, with the further implication of being equivalent to the Pearson correlation coefficient. Recall that we want to know if a school that offers a CTE track that, in the population, has higher proportion of women enrollment, is more likely to offer another career (or from a sector) that also has higher women enrollment, or not.

Our results in Table 2 show that school's careers are in fact dissortative in terms of women enrollment and, in this case, the answer does not depend on the inclusion of the Private schools. In other words, this means that, once a school offers a career (or an economic sector) highly associated with women's enrollment, it is less likely to offer another similar career, but more likely to offer a career associated with men's enrollment. This results intuitive, and is in line with our exploratory analysis of the whole network. In substantive terms, this means that schools tend to diversify their offer within a school in gender-specific terms, which makes sense considering that, in the Chilean system, most schools have mixed-gender enrollment, and only a few are exclusively for men or women.

Proximity and Communities

A next question I would like to addres is to understand if there is some underlying structure in the way that careers are cocited by schools. In doing this, I am inspired by the "product space" analysis from Hidalgo et al. (2007). In fact, we can imagine a "training space" in which some training programs, our CTE tracks, are closer to others in terms of how likely are they jointly offered by schools. This co-citation could be reflecting similarities in the capabilities that a school need to have in order to implement such programs.

However, following the argument made by the product space, it is possible that a school offers two careers not because their intrinsic similarity or complementarity, but because some randomness, or simple because one of those careers is highly prevalent in the entire network of schools. In order to control for that possibility, we can use the Revealed Comparative Advantage (RCA), which measures if a school enrollment in a career, as a proportion of its total enrollment, is higher than the overall proportion of that career in the total enrollment of the system. Formally:

$$RCA_{s,c}\frac{X(s,c)/\sum_c X(s,c)}{\sum_s X(s,c)/\sum_{s,c} X(s,c)}$$

Where s stands for school, and c for career. $RCA_{s,c} > 1$ if the proportion of school s enrollment in career c is higher than the overall proportion of career c in the system-wise enrollment.

Finally, the proximity ϕ of a pair of careers, c_i, c_j in the training space is computed as:

$$\phi_{c_i,c_j} = \min\{P(RCA_{c_i}|RCA_{c_j}), P(RCA_{c_j}|RCA_{c_i})\}$$

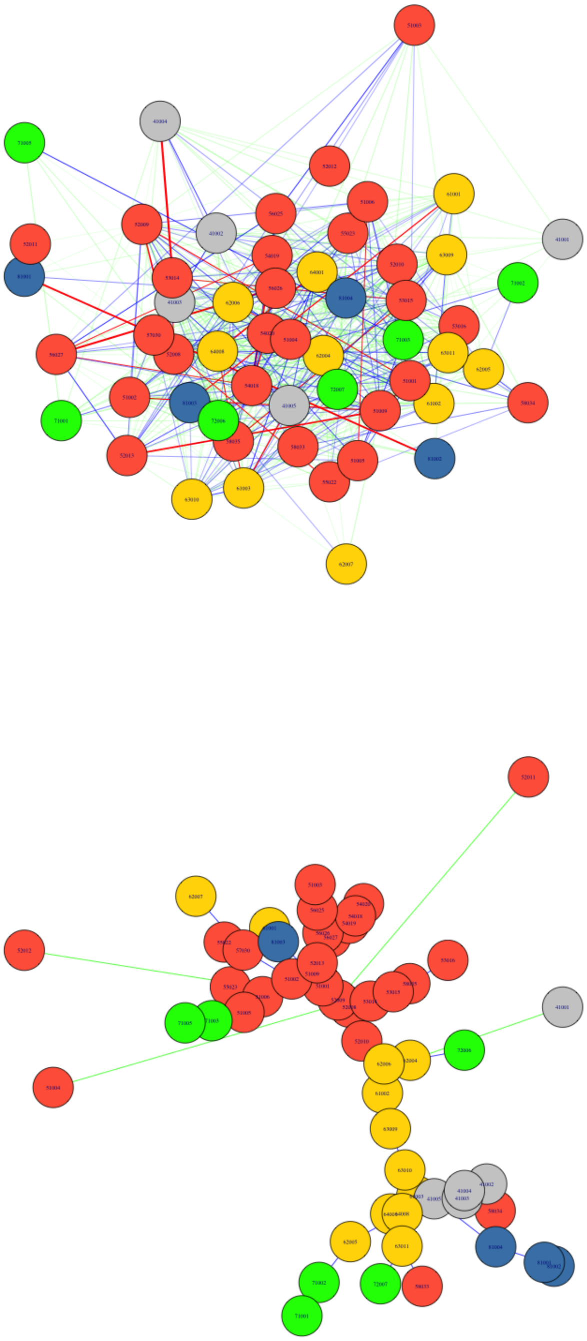
where the *min* argument is for making symmetric the measure.

A first step, then, would be filtering the network, keeping only those combination with $RCA_c > 1$. This result is depicted in the upper panel of Figure 3. We can see that there is a core of highly connected careers, although with varying levels of proximity, and then a few careers much less connected, in general, which are in the periphery of the graph. We can further filter the edges by taking the maximum spanning tree, which facilitates the visualization of the resulting network. Figures 3 shows the resulting graphs after implementing this transformations in our network, including only the edges with $RCA > 1$, and then only those in the Maximum Spanning Tree.

Looking at this networks, we can identify some characteristics of the training space of the Chilean CTE. First, careers belonging to the same economic sector tend to be clustered, which makes sense because they probably requiere similar conditions of implementation (teachers, infrastructure, firm support, etc.). This is also consistent with our previous results, in the sense of career-to-career connections being assortative by economic sector.

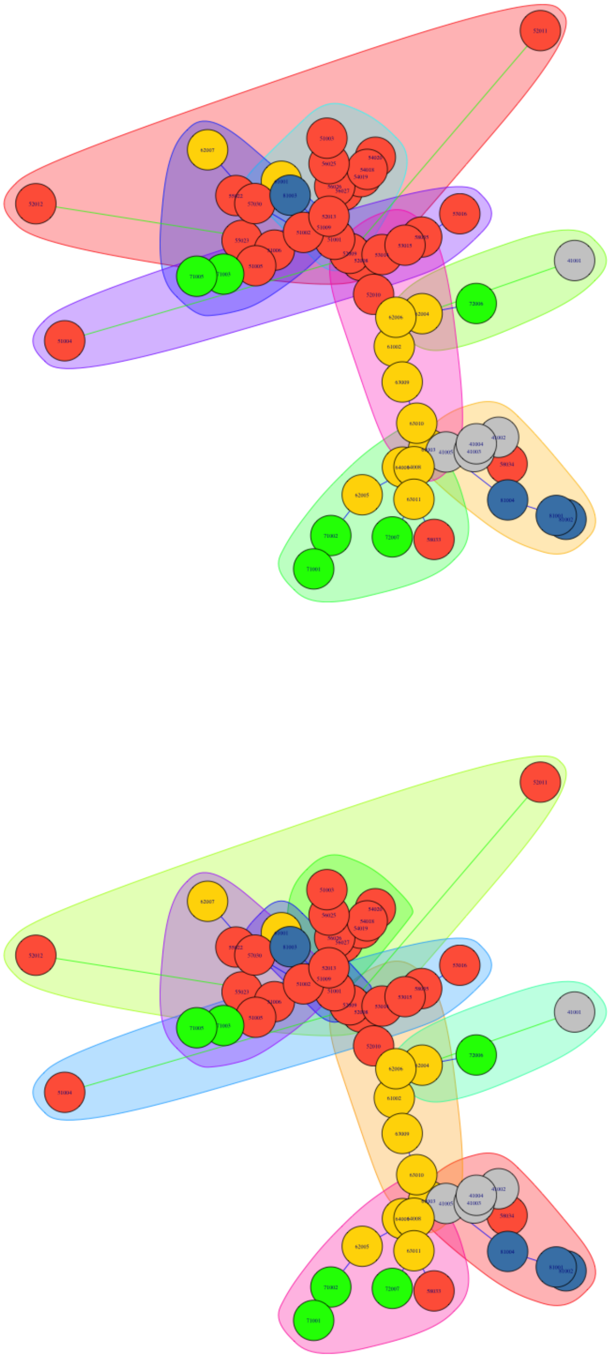
In addition, as I said above, it is possible to identify some careers that are in the "periphery" of the network. This is clearer in the bottom panel. There are noticeably some industrial, maritime, and farming careers that are weakly connected with the core of the training space. Those are, for example, aeronautical mechanics among industry, wood processing among farming, and fishery among maritime. It is likely that those careers are either highly difficult to be implemented at the high school level (as aeronautical mechanics), or in fact closer to an unskilled job rather than to a specialized training (wood processing and fishery). Either because of the lack or excess of specialization, those CTE tracks seem to be in the periphery of the training space.

Figure 4: Training Space (RCA > 1) and MST representation



Note: Nodes represent CTE tracks (careers), and color identify the economic sectos to which the careers belong (grey: Business, red: Industry, gold: Services, green: Farming, blue: Maritime). Edge color maps to the proximity, where higher proximity is depicted in red, followed by blue, and finally green (for products connected but being far in the Training Space).

Figure 5: Career Communities in the Training Space



Note: Nodes represent CTE tracks (careers), and color identify the economic sectos to which the careers belong (grey: Business, red: Industry, gold: Services, green: Farming, blue: Maritime). Edge color maps to the proximity, where higher proximity is depicted in red, followed by blue, and finally green (for products connected but being far in the Training Space). In addition, groups from the Louvain clustering (upper) and the optimal clustering (bottom) algorithms are included.

Finally, in order to explore if there is a underlying agglomerative structure in the careers in the training space not reductible to the economic sector, I used the two clustering algorithms to identify the underlying communities: Louvain method, and optimal method. Both algorithms try to maximize the modularity in the network, in the sense of having the partition into groups that produces the higher connectivity within, but being weakly connected to other groups. The search stops when it is not possible to increase the modularity by adding more nodes to a group. The difference is that the Louvain method is greedy, in the sense that follow the path that increase the modularity in each step. The optimal method, on the other hand, explores a wider range of options, giving an overall best result (at computational costs). Figure 5 shows the result of this excercise, that seems to follow closely our previous description. Both algorithms identify practically the same structure.

Regional Variations

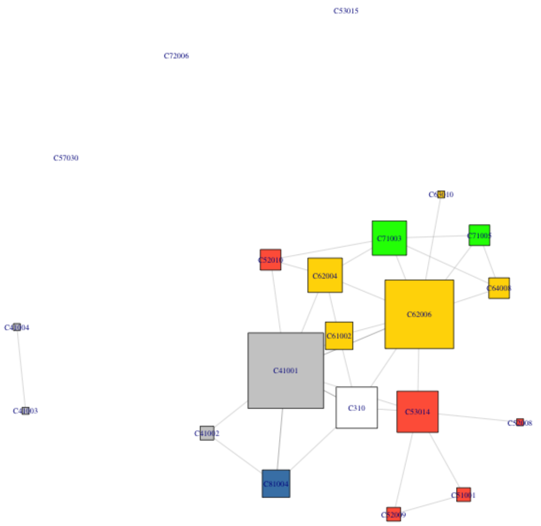
One of the main purposes of the CTE is to be strongly connected with the local labor market. A first exploration of this feature could be made by looking at some regional subgraphs, considering the careers offered in regions of Chile with particular productive structure. Figures 6, 7, and 8 show these graph for Tarapaca (North), Metropolitan (Center), and Araucania (South) regions.

As expected, the CTE training networks present some particularities in relation to the regions they belong. First, however, it is necessary to note some strinking similarities. In fact, in all the three regions analyzed, one of the more important careers is Administration (Business), which appears as highly prevalent irrespective of place. In a lower level, also Electricity seems to be transversal. But there are also noticeable differences. In particular, in Tarapaca Region, in the North of the country, Port Operation appears as important, in consonance with the productive structure[2]. In Arauco, on the other hand, one of the careers that appear as relevant is Agricultural Technician, again in close relationship with the local economic environment.

This analysis, although brief, suggest that there is, in fact, regional specializations in the Career and Technical Education offered by schools. Nevertheless, this regional-specific careers are far from being the most prevalent ones even in those regions, and the ones that have higher enrollment are similar across the country.

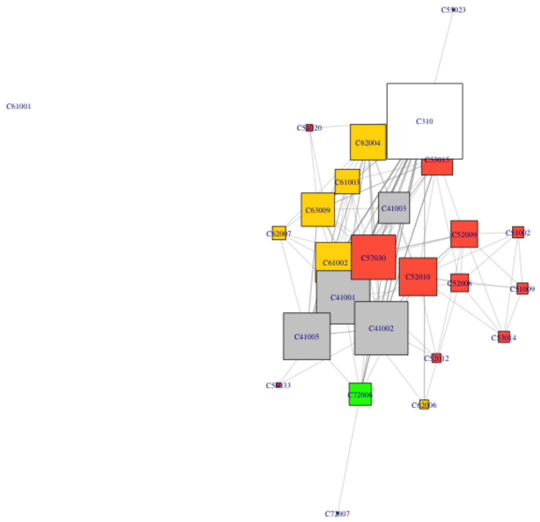
[2] The principal city of Tarapaca region, Iquique, is one of the largest international ports of Chile.

Figure 6: Subgraph of Tarapaca Region



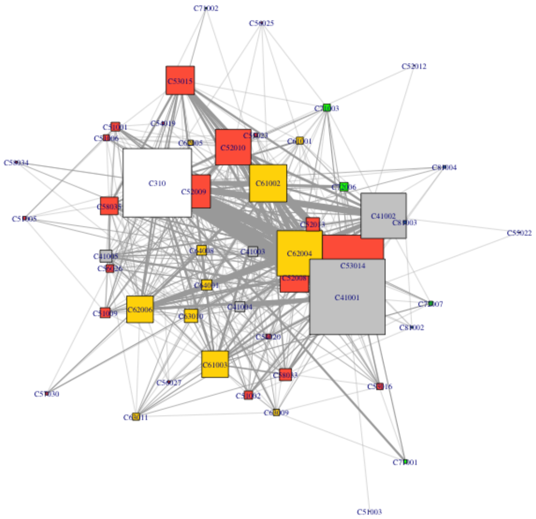
Note: Nodes represent CTE tracks (careers), and are colored according to the economic sector that they belong to (White: academic, grey: business, red: industry, gold: services, green: farming, blue: maritime). The node size and edge width are weighted to represent the amount of connection shared, i.e., the number of schools that offer each career and they in conjuction to others.

Figure 7: Subgraph of Metropolitan Region



Note: Nodes represent CTE tracks (careers), and are colored according to the economic sector that they belong to (White: academic, grey: business, red: industry, gold: services, green: farming, blue: maritime). The node size and edge width are weighted to represent the amount of connection shared, i.e., the number of schools that offer each career and they in conjuction to others.

Figure 8: Subgraph of Araucania



Note: Nodes represent CTE tracks (careers), and are colored according to the economic sector that they belong to (White: academic, grey: business, red: industry, gold: services, green: farming, blue: maritime). The node size and edge width are weighted to represent the amount of connection shared, i.e., the number of schools that offer each career and they in conjuction to others.

Discussion and Conclusion

In this brief project, I proposed a network approach to high school Career and Technical Education (CTE) in Chile. After constructing the basic network, and based on both into the groups (careers), and into the people (schools) projections, I explored some features that appear clearly due to this network representation. Further, I adapted the product space to a training space in order to estimate the proximity of different careers between each other, and try to identify clusters of careers in the space. Finally, I provide a glimpse into the regional variability that this network representation permits to explore, findings some evidence of the adaptation of the training opportunities to the local economic environment.

Surely, however, this is only a first attempt to use network analysis to understand the structure of Technical and Vocational Training in Chile, and has several limitations both in terms of methods and substance. In methodological terms, it would be valuable to develop some analysis considering the directed nature of this network in the first place, before using the projections, as we know that every projection loose some information. Also, it is important to note that the results of the community detection are sensible to the algorithm selection, and therefore would be necessary to explore different approaches and evaluate how robuste the training clusters are to this methodological choice. Finally, it would be valuable to characterize more formally, using different centrality measures, the role that different schools, careers, and economic sectors, play in the training network, as well as study the macro-properties of the network itself.

In substantive terms, there are many questions that I could not take into account in this paper, but that are crucial to better understand the consequences of the CTE for Chilean students. For example, is there a relationship between distance from the academic track in the network, measured as the proximity (ϕ) in the training space, and the performance of students attending certain careers? In the same line, are there careers that have lower trade-off fore students in terms of labor market and educational outcomes? Finally, is the relationship between local network characteristics and local economic structure predictive of labor market outcomes of CTE students? Being relevant questions for some of the most disadvantages Chilean students, further research on this area is warranted, and the training space approach could be a valuable contribution to CTE studies.

References

Anderson, K. (2017) Skill Networks and Measures of Complex Human Capital. PNAS

De la Cruz, J. & Riker, D. (2012) Product Space Analysis of the Exports of Brazil. US ITC Working Paper

Farías, M. (2013). Effects of Early Career Decisions on Future Opportunities: The Case of Vocational Education in Chile. Stanford University.

Geraldo, P. (2015) Efectos de la educación media técnico profesional en la reproducción de la desigualdad educativa. Tesis para optar al grado de magister en Sociología. Pontificia Universidad Católica de Chile.

Hidalgo, C., Klinger, B., Barabasi, L., & Haussmann, R. (2007) The Product Space Conditions of the Development of Nations. Science

Newman (2003) Mixing patterns in networks. https://arxiv.org/pdf/cond-mat/0209450.pdf

Appendix

Code

```
# [Section 1]
########################
### DATA PREPARATION ###
########################

# Data available from Chilean Ministry of Education
stu_2017 <- read.csv("MatriculaAlumno_2017_03212018.csv", header=TRUE, sep=";")
colnames(stu_2017) <- tolower(names(stu_2017))

# Keeping observations in grades 11th and 12th, when CTE is offered
stu_2017 <- stu_2017 %>% filter(cod_ense==310 | cod_ense==410 | cod_ense==510
                                | cod_ense==610 | cod_ense==710 | cod_ense==810,
                                cod_grado==3 | cod_grado==4)
# Assign Career number to students in the academic track
# Equal to their sector number
stu_2017$cod_espe[stu_2017$cod_ense==310] <- 310

stu_net <- stu_2017 %>%
  # Add "S" (school) or "C" (career) to the id, so we can combine safely after
  mutate(school_id=paste0("S", rbd), # School
         career=paste0("C", cod_espe), # Career
         sector=paste0("ES", cod_ense)) %>% # Economic Sector
  # Measures of total and by gender enrollment by school
  group_by(rbd) %>%
  mutate(school_enroll=n(), school_womprop=mean(ifelse(gen_alu==2,1,0))) %>%
  # Measures of total and by gender enrollment by sector
  ungroup() %>% group_by(cod_ense) %>%
  mutate(sector_enroll=n(), sector_womprop=mean(ifelse(gen_alu==2,1,0))) %>%
  # Measures of total and by gender enrollment by career
  ungroup() %>% group_by(cod_ense, cod_espe) %>%
  mutate(career_enroll=n(), career_womprop=mean(ifelse(gen_alu==2,1,0))) %>%
  # Measures of total and by gender enrollment by (school x sector)
  ungroup() %>% group_by(rbd, cod_ense) %>%
  mutate(sch_sec_enroll=n(), sch_sec_womprop=mean(ifelse(gen_alu==2,1,0))) %>%
  # Measures of total and by gender enrollment by (school x career)
  ungroup() %>% group_by(rbd, cod_espe) %>%
  mutate(sch_car_enroll=n(), sch_car_womprop=mean(ifelse(gen_alu==2,1,0))) %>%
  # Select variables of interest
  select(school_id, rural_rbd, cod_depe2,
         cod_reg_rbd, cod_com_rbd, nom_com_rbd,
         career, sector,
         school_enroll, school_womprop,
         sector_enroll, sector_womprop,
         career_enroll, career_womprop,
         sch_sec_enroll, sch_sec_womprop,
         sch_car_enroll, sch_car_womprop) %>%
  unique() %>% ungroup()
stu_net$rbd <- NULL
stu_net$cod_espe <- NULL
```

```
# Rename variables
colnames(stu_net) <- c("school_id", "rural", "owner",
                       "region", "city_code", "city_name",
                       "career", "sector",
                       "school_enroll", "school_womprop",
                       "sector_enroll", "sector_womprop",
                       "career_enroll", "career_womprop",
                       "sch_sec_enroll", "sch_sec_womprop",
                       "sch_car_enroll", "sch_car_womprop")
# Save the database
write.csv(stu_net, file="student_career_2017.csv", row.names=FALSE)

# Data set previously filtered to keep only last two years of high schools
# (When the Career and Technical Education is offered)
stu_net <- read.csv("student_career_2017.csv", header=TRUE)

# Creating the edges list, edges attributes, nodes attributes

# Sector vertex attributes
sector_list <- stu_net %>% group_by(sector) %>%
  mutate(id=sector,
         enrollment=max(sector_enroll), wom_prop=max(sector_womprop)) %>%
  select(id, enrollment, wom_prop) %>%
  mutate(owner=5) %>% unique() %>% ungroup()
sector_list$sector <- NULL

# Career vertex attributes
career_list <- stu_net %>% group_by(career) %>%
  mutate(id=career,
         enrollment=max(career_enroll), wom_prop=max(career_womprop)) %>%
  select(id, enrollment, wom_prop, sector) %>%
  mutate(owner=5) %>% unique() %>% ungroup()
career_list$career <- NULL

# Schools vertex attributes
school_list <- stu_net %>% group_by(school_id) %>%
  mutate(id=school_id,
         enrollment=max(school_enroll), wom_prop=max(school_womprop)) %>%
  select(id, region, city_code, city_name, owner, rural, enrollment, wom_prop) %>%
  unique() %>% ungroup()
school_list$school_id <- NULL

# Create the node list with attributes
# At the school by sector level
nodes_school_sector <- bind_rows(school_list, sector_list)
write.csv(nodes_school_sector, file="school_sector_nodes.csv", row.names = FALSE)

# At the school by career level
nodes_school_career <- bind_rows(school_list, career_list)
write.csv(nodes_school_career, file="school_career_nodes.csv", row.names = FALSE)

# Create the edge list with links attributes
edges_school_sector <- stu_net %>% select(school_id, sector,
                                          sch_sec_enroll, sch_sec_womprop)
write.csv(edges_school_sector, file="school_sector_edges.csv", row.names = FALSE)

edges_school_career <- stu_net %>% select(school_id, career,
                                          sch_car_enroll, sch_car_womprop)
write.csv(edges_school_career, file="school_career_edges.csv", row.names = FALSE)





# [Section 2]
##############################
### NETWORK VISUALIZATIONS ###
##############################

### SECTOR LEVEL
library(igraph)
# Descriptios at the school by sector level
stu_net <- read.csv("student_career_2017.csv", header=TRUE)
node_sector <- read.csv("school_sector_nodes.csv", header=TRUE)
edge_sector <- read.csv("school_sector_edges.csv", header=TRUE)

# Turn into an igraph network object
# vertices=node_list does not work with NAs in attributes
# I have to make the round through package:network

net <- graph_from_data_frame(edge_sector, directed=FALSE, vertices=NULL)
# Add Nodes attributes using package:network
library(intergraph)
library(network)
net2 <- asNetwork(net)
net2 %v% "owner" <- node_sector$owner
net2 %v% "region" <- node_sector$region
net2 %v% "city" <- node_sector$city_code
net2 %v% "rural" <- node_sector$rural
net2 %v% "enrollment" <- node_sector$enrollment
net2 %v% "wom_prop" <- node_sector$wom_prop
list.vertex.attributes(net2)

# Turn it back into igraph
net_igraph <- asIgraph(net2)
detach("package:intergraph", unload=TRUE)
detach("package:network", unload=TRUE)

# Make it bipartite
V(net_igraph)$type <- FALSE
V(net_igraph)$type[V(net_igraph)$vertex.names %in% stu_net$school_id] <- TRUE

# Add shape and color to the nodes
V(net_igraph)$shape <- ifelse(V(net_igraph)$type == FALSE, "square", "circle")
V(net_igraph)$color <- recode(V(net_igraph)$owner,
                              `1`="red",
                              `2`="blue",
                              `3`="green",
                              `4`="yellow",
                              `5`="white")
V(net_igraph)$labels <- recode(V(net_igraph)$vertex.names,
                               `ES310`="Academic",
                               `ES410`="Business",
                               `ES510`="Industry",
                               `ES610`="Services",
                               `ES710`="Farming",
                               `ES810`="Maritime")

# VISUALIZATIONS
# Bipartite plot
png("sch_sec_biplot.png", 800, 800)
plot(net_igraph,
     vertex.label=ifelse(V(net_igraph)$type==FALSE,
                         V(net_igraph)$labels,NA),
     vertex.size=1.5,
     vertex.label.cex=0.1,
     vertex.label.color="black",
#     vertex.size=V(net_igraph)$enrollment*0.0001,
     edge.color=ifelse(E(net_igraph)$sch_sec_womprop>0.5, "pink","grey80"),
     edge.width=E(net_igraph)$sch_sec_enroll*0.005)
dev.off()

# PROJECTIONS
# Projection onto sector
proj_sec <- bipartite.projection(net_igraph, which="false")
# Projection onto schools
proj_sec_sch <- bipartite.projection(net_igraph, which="true")

# Projection plots
png("proj_sec.png", 800, 800)
plot(proj_sec, layout=layout_on_grid,
     vertex.label.cex=0.7,
     vertex.size=45*strength(proj_sec)/max(strength(proj_sec)),
     edge.width=E(proj_sec)$weight*0.025,
     vertex.label=V(proj_sec)$vertex.names)
dev.off()

#png("proj_sec_sch.png", 800, 800)
#plot(proj_sec_sch, layout=layout_nicely,
#     edge.width=E(proj_sec_sch)$weight*0.025,
#     vertex.size=1.5,
#     vertex.label=NA)
#dev.off()


### CAREER LEVEL
# Descriptios at the school by sector level
# Filter a case with missing career (23 obs, sector 610)
node_career <- read.csv("school_career_nodes.csv", header=TRUE) %>% filter(id!="C0")
```

```
edge_career <- read.csv("school_career_edges.csv", header=TRUE) %>% filter(career!="C0")

# Degree distributions of the bipartite network
net <- graph_from_data_frame(edge_career, directed=TRUE, vertices=NULL)

# Out degree: Number of career track by school
# Removing 0 out-degree (careers)
out_degree <- ifelse(degree(net, mode="out")>0, degree(net, mode="out"), NA)
png("degree_dist_out.png", 800, 800)
qplot(out_degree, geom="histogram") +
  theme_classic() + stat_bin(binwidth = 1) +
  labs(title="Degree distribution: Schools",
       y="Frequency", x="Out-degree: Number of careers offered by a school")
dev.off()

# In degree: Number of schools by career track
# Removing 0 in-degree (schools)
# Removing more than degree 2000 (academic track)
in_degree <- ifelse(degree(net, mode="in")>0 & degree(net, mode="in")<2000
                    , degree(net, mode="in"), NA)
png("degree_dist_in.png", 800, 800)
qplot(in_degree, geom="histogram") +
  theme_classic() + stat_bin(binwidth = 1) +
  labs(title="Degree distribution: Careers",
       y="Frequency", x="In-degree: Number of schools that offer a career")
dev.off()

# Turn into an igraph network object
# vertices=node_list does not work with NAs in attributes
# I have to make the round through package:network
net <- graph_from_data_frame(edge_career, directed=FALSE, vertices=NULL)

# Add Nodes attributes using package:network
library(intergraph)
library(network)
net2 <- asNetwork(net)
net2 %v% "owner" <- node_career$owner
net2 %v% "region" <- node_career$region
net2 %v% "city" <- node_career$city_code
net2 %v% "rural" <- node_career$rural
net2 %v% "enrollment" <- node_career$enrollment
net2 %v% "wom_prop" <- node_career$wom_prop
list.vertex.attributes(net2)

# Turn it back into igraph
net_sch_car <- asIgraph(net2)
detach("package:intergraph", unload=TRUE)
detach("package:network", unload=TRUE)

# Make it bipartite
V(net_sch_car)$type <- FALSE
V(net_sch_car)$type[V(net_sch_car)$vertex.names %in% stu_net$school_id] <- TRUE
```

```r
# Add shape and color to the nodes
V(net_sch_car)$shape <- ifelse(V(net_sch_car)$type == FALSE, "square", "circle")
# Weird: one vertex changes the owner
#V(net_sch_car)$vertex.names[V(net_sch_car)$owner==2 & V(net_sch_car)$type==FALSE]
V(net_sch_car)$owner[V(net_sch_car)$vertex.names=="C52008"] <- 5
V(net_sch_car)$color <- recode(V(net_sch_car)$owner,
                              `1`="red",
                              `2`="blue",
                              `3`="green",
                              `4`="yellow",
                              `5`="white")

### VISUALIZATIONS
# Bipartite plot
png("sch_car_biplot.png", 800, 800)
plot(net_sch_car,
     vertex.label=ifelse(V(net_sch_car)$type==FALSE,
                         V(net_sch_car)$vertex.names,NA),
     vertex.size=1.5,
     vertex.label.cex=0.1,
     vertex.label.color="black",
     edge.color=ifelse(E(net_sch_car)$sch_car_womprop>0.5, "pink","grey80"),
     edge.width=E(net_sch_car)$sch_car_enroll*0.005)
dev.off()

### PROJECTIONS
# Projection onto career
proj_career <- bipartite.projection(net_sch_car, which="false")
V(proj_career)$sector <- as.numeric(str_sub(V(proj_career)$vertex.names, start=2, end=2))
V(proj_career)$color <- recode(V(proj_career)$sector,
                              `3`="white",
                              `4`="grey80",
                              `5`="tomato",
                              `6`="gold",
                              `7`="green",
                              `8`="steel blue")
# Projection onto schools
proj_car_sch <- bipartite.projection(net_sch_car, which="true")

# Projection plots
png("proj_car.png", 800, 800)
plot(proj_career, layout=layout_on_grid,
     vertex.size=30*degree(proj_career)/max(degree(proj_career)),
     vertex.label.cex=0.7,
     edge.width=E(proj_career)$weight*0.15,
     vertex.label=V(proj_career)$vertex.names)
dev.off()

#png("proj_car_sch.png", 800, 800)
#plot(proj_car_sch, layout=layout_nicely,
#     edge.width=E(proj_car_sch)$weight*0.025,
#     vertex.size=1.5,
#     vertex.label=NA)
```

```
#dev.off()

# Macroproperties: Degree Distribution




# [Section 3]
###############################
### Assortativity Measures ###
###############################

# 1) Are schools of the same ownership more likely to offer the same careers?
#assortativity.nominal(net_igraph, type=V(net_igraph)$owner, directed=FALSE)
# This doesn't make sense, because this is a bipartite network.
# We need the projections!
# Projection onto schools: by sector
assortativity.nominal(proj_sec_sch, type=V(proj_sec_sch)$owner, directed=FALSE)
# 0.01538797
# Projection onto schools: by career
assortativity.nominal(proj_car_sch, type=V(proj_car_sch)$owner, directed=FALSE)
# 0.007764737 Much less, but still positive
# But are these measure only driven by private academic schools?
# By sector
proj_sec_sch_sub <-
  subgraph.edges(proj_sec_sch, eids=V(proj_sec_sch)[owner!=3])
assortativity.nominal(proj_sec_sch_sub, type=V(proj_sec_sch_sub)$owner, directed=FALSE)
# -0.3925975 Completely the opposite!!!
# By career
proj_car_sch_sub <-
  subgraph.edges(proj_car_sch, eids=V(proj_car_sch)[owner!=3])
assortativity.nominal(proj_car_sch_sub, type=V(proj_car_sch_sub)$owner, directed=FALSE)
# -0.4119342 Same!!!


# 2) Are careers offered in the same schools more likely to be from the same sector?
assortativity.nominal(proj_career, type=V(proj_career)$sector, directed=FALSE)
# 0.06626649
# Same as above, what if the assortativity is driven by private schools?
sch_car_sub <- subgraph.edges(net_sch_car, eids=V(net_sch_car)[owner!=3])
proj_career_sub <- bipartite.projection(sch_car_sub, which="false")
V(proj_career_sub)$sector <-
  as.numeric(str_sub(V(proj_career_sub)$vertex.names, start=2, end=2))
assortativity.nominal(proj_career_sub, type=V(proj_career_sub)$sector, directed=FALSE)
# 0.09324824 Not in this case (the result holds)


# 3) Are sectors and careers assortative on wom_prop
# Sector (including privates)
assortativity(proj_sec, types1=V(proj_sec)$wom_prop, directed=FALSE)
# -0.2
# Sector (excluding privates)
sch_sec_sub <- subgraph.edges(net_igraph, eids=V(net_igraph)[owner!=3])
```

```
proj_sec_sub <- bipartite.projection(sch_sec_sub, which="false")
assortativity(proj_sec_sub, types1=V(proj_sec_sub)$wom_prop, directed=FALSE)
# The resulting graph is identical, why???
# Career (including privates)
assortativity(proj_career, types1=V(proj_career)$wom_prop, directed=FALSE)
# -0.02704507
# Career (excluding privates)
assortativity(proj_career_sub, types1=V(proj_career_sub)$wom_prop, directed=FALSE)
# -0.03880667






# [Section 4]
##########################
### Proximity Network ###
##########################

# Filtering the Network by the RCA
# (Methodology from the Product Space)
Proximity <- stu_net %>%
  mutate(RCA_sch_sec = (sch_sec_enroll/school_enroll)/(sector_enroll/41049),
         RCA_sch_car = (sch_car_enroll/school_enroll)/(career_enroll/41049),
         RCA_sec = ifelse(RCA_sch_sec>1,1,0),
         RCA_car = ifelse(RCA_sch_car>1,1,0)) %>%
  group_by(sector) %>% mutate(ubiq_sec = sum(RCA_sec)) %>%
  group_by(career) %>% mutate(ubiq_car = sum(RCA_car)) %>% ungroup() %>%
  group_by(school_id) %>% mutate(RCA_tot = sum(RCA_car)) %>%
  filter(RCA_tot>1, RCA_car==1) %>% ungroup()

Proximity$career <- as.numeric(str_sub(Proximity$career, start = 2))
career_number <- unique(Proximity$career)
RCA_net <- expand.grid(career_number, career_number)
RCA_net$weight <- NA

for (i in career_number) {
  for(j in career_number){
    numerator <- Proximity %>% group_by(school_id) %>% filter(career==i | career==j) %>%
      summarise(tot=n()) %>% filter(tot==2) %>% summarise(num=n())

    denominator <- max(Proximity$ubiq_car[Proximity$career==i],
                       Proximity$ubiq_car[Proximity$career==j])

    RCA_net$weight[RCA_net$Var1==i & RCA_net$Var2==j
                   | RCA_net$Var1==j & RCA_net$Var2==i] <-
      numerator/denominator
  }
}

# Removing values 0, the relation of a career to itself
RCA_net <- RCA_net %>% filter(Var1!=Var2, weight>0) %>%
  mutate(weight=as.numeric(weight), weight_inv=1/weight)
```

```r
# Removing double counting a relationship
RCA_net <- RCA_net[!duplicated(t(apply(RCA_net[1:2], 1, sort))), ]

# Node characteristics
nodes_prox <- tibble(id=unique(RCA_net$Var1)) %>%
  mutate(sector=str_sub(id, end=1))

RCA_car_net <- graph_from_data_frame(RCA_net, directed=FALSE, vertices=NULL)
library(intergraph)
library(network)
net2 <- asNetwork(RCA_car_net)
net2 %v% "sector" <- nodes_prox$sector
# Turn it back into igprah
RCA_car_net <- asIgraph(net2)
V(RCA_car_net)$color <- recode(V(RCA_car_net)$sector,
                               `3`="white",
                               `4`="grey80",
                               `5`="tomato",
                               `6`="gold",
                               `7`="green",
                               `8`="steel blue")

# Complete network
png("career_RCA.png", 800, 800)
plot(RCA_car_net,
     layout=layout_with_kk,
     vertex.label.cex=0.5,
     vertex.label=V(RCA_car_net)$vertex.names,
     vertex.cex=0.5,
     edge.width=E(RCA_car_net)$weight*7.5,
     edge.color=ifelse(E(RCA_car_net)$weight>0.15, "red",
                       ifelse(E(RCA_car_net)$weight<0.05, "green", "blue")))
dev.off()

# Pruned network
tree_net <- RCA_car_net
E(tree_net)$weight <- E(tree_net)$weight_inv
mst_net <- mst(tree_net)

png("career_tree.png", 800, 800)
plot(mst_net,
     layout=layout_with_kk,
     vertex.label=V(mst_net)$vertex.names,
     vertex.label.cex=0.5,
     vertex.cex=0.1,
     edge.color=ifelse(1/E(mst_net)$weight>0.15, "red",
                       ifelse(1/E(mst_net)$weight<0.05, "green", "blue")))
dev.off()
```

```
# [Section 5]
###########################
### Community Detection ###
###########################

career_spinglass <- cluster_spinglass(mst_net, weights=1/E(mst_net)$weight)
career_louvain <- cluster_louvain(mst_net)
career_optmod <- cluster_optimal(mst_net)
career_optmod2 <- cluster_optimal(mst_net, weights=1/E(mst_net)$weight)

png("career_tree_louvain.png", 800, 800)
plot(mst_net,
     layout=layout_with_kk,
     vertex.label=V(mst_net)$vertex.names,
     vertex.label.cex=0.5,
     vertex.cex=0.1,
     edge.color=ifelse(1/E(mst_net)$weight>0.15, "red",
                       ifelse(1/E(mst_net)$weight<0.05, "green", "blue")),
     mark.groups = communities(career_louvain))
dev.off()

png("career_tree_optimal.png", 800, 800)
plot(mst_net,
     layout=layout_with_kk,
     vertex.label=V(mst_net)$vertex.names,
     vertex.label.cex=0.5,
     vertex.cex=0.1,
     edge.color=ifelse(1/E(mst_net)$weight>0.15, "red",
                       ifelse(1/E(mst_net)$weight<0.05, "green", "blue")),
     mark.groups = communities(career_optmod))
dev.off()




# [Section 6]
#########################
### Regional Analysis ###
#########################

tarap <- subgraph.edges(net_sch_car, eids=V(net_sch_car)[region %in% 1])
tarap_career <- bipartite.projection(tarap, which="false")
V(tarap_career)$sector <-
  as.numeric(str_sub(V(tarap_career)$vertex.names, start=2, end=2))
V(tarap_career)$color <- recode(V(tarap_career)$sector,
                              `3`="white",
                              `4`="grey80",
                              `5`="tomato",
                              `6`="gold",
                              `7`="green",
                              `8`="steel blue")
```

```r
png("tarap.png", 800, 800)
plot(tarap_career, layout=layout_with_kk,
     vertex.size=30*strength(tarap_career)/max(strength(tarap_career)),
     vertex.label.cex=0.7,
     edge.width=E(tarap_career)$weight*0.5,
     vertex.label=V(tarap_career)$vertex.names)
dev.off()


rm <- subgraph.edges(net_sch_car, eids=V(net_sch_car)[region %in% 13])
rm_career <- bipartite.projection(rm, which="false")
V(rm_career)$sector <-
  as.numeric(str_sub(V(rm_career)$vertex.names, start=2, end=2))
V(rm_career)$color <- recode(V(rm_career)$sector,
                              `3`="white",
                              `4`="grey80",
                              `5`="tomato",
                              `6`="gold",
                              `7`="green",
                              `8`="steel blue")

png("rm.png", 800, 800)
plot(rm_career, layout=layout_with_kk,
     vertex.size=30*strength(rm_career)/max(strength(rm_career)),
     vertex.label.cex=0.7,
     edge.width=E(rm_career)$weight*0.5,
     vertex.label=V(rm_career)$vertex.names)
dev.off()

arauco <- subgraph.edges(net_sch_car, eids=V(net_sch_car)[region %in% 9])
arauco_career <- bipartite.projection(arauco, which="false")
V(arauco_career)$sector <-
  as.numeric(str_sub(V(arauco_career)$vertex.names, start=2, end=2))
V(arauco_career)$color <- recode(V(arauco_career)$sector,
                              `3`="white",
                              `4`="grey80",
                              `5`="tomato",
                              `6`="gold",
                              `7`="green",
                              `8`="steel blue")

png("arauco.png", 800, 800)
plot(arauco_career, layout=layout_with_kk,
     vertex.size=30*strength(arauco_career)/max(strength(arauco_career)),
     vertex.label.cex=0.7,
     edge.width=E(arauco_career)$weight*0.5,
     vertex.label=V(arauco_career)$vertex.names)
dev.off()




proj_career2 <- bipartite.projection(net_sch_car, which="false", multiplicity=FALSE)
```

```r
V(proj_career2)$sector <- as.numeric(str_sub(V(proj_career2)$vertex.names, start=2, end=2))
V(proj_career2)$color <- recode(V(proj_career2)$sector,
                              `3`="white",
                              `4`="grey80",
                              `5`="tomato",
                              `6`="gold",
                              `7`="green",
                              `8`="steel blue")





net_sch_car_dir <-

hist(degree(net_sch_car))
hist(degree(proj_career2))

hist(degree(proj_career))
hist(degree(proj_car_sch))






net_2 <- graph_from_data_frame(edge_career, directed=TRUE, vertices=NULL)

hist(degree(net_2, mode="in"))

# Add Nodes attributes using package:network
library(intergraph)
library(network)
net2 <- asNetwork(net)
net2 %v% "owner" <- node_career$owner
net2 %v% "region" <- node_career$region
net2 %v% "city" <- node_career$city_code
net2 %v% "rural" <- node_career$rural
net2 %v% "enrollment" <- node_career$enrollment
net2 %v% "wom_prop" <- node_career$wom_prop
list.vertex.attributes(net2)

# Turn it back into igraph
net_sch_car <- asIgraph(net2)
detach("package:intergraph", unload=TRUE)
detach("package:network", unload=TRUE)

# Make it bipartite
V(net_sch_car)$type <- FALSE
V(net_sch_car)$type[V(net_sch_car)$vertex.names %in% stu_net$school_id] <- TRUE

# Add shape and color to the nodes
V(net_sch_car)$shape <- ifelse(V(net_sch_car)$type == FALSE, "square", "circle")
```

```r
# Weird: one vertex changes the owner
#V(net_sch_car)$vertex.names[V(net_sch_car)$owner==2 & V(net_sch_car)$type==FALSE]
V(net_sch_car)$owner[V(net_sch_car)$vertex.names=="C52008"] <- 5
V(net_sch_car)$color <- recode(V(net_sch_car)$owner,
                              `1`="red",
                              `2`="blue",
                              `3`="green",
                              `4`="yellow",
                              `5`="white")
```