Higher Education Courses Unemployment - Portugal

| Mário Reis  70969  Lisbon, Portugal  mario.fb.reis@gmail.com | Artur Fonseca  75456  Lisbon, Portugal  artur.fonseca@tecnico.ulisboa.pt | André Pires  76046  Alenquer, Portugal  pardal.pires@gmail.com |
| --- | --- | --- |

# ABSTRACT

The purpose of the project is to learn not only about the unemployment among universities graduates, colleges’ graduates, courses graduates and courses unemployment by area/sub-areas across the years in Portugal, but also the “entry grade-unemployment” relation of the public courses in 2016. This report describes the work developed during the first semester of 2016/2017 in the Information Visualization course at Instituto Superior Técnico. It summarizes what we proposed to answer with our visualization, what we achieved, how we achieved the goals, the problems during the development and future work we could do to improve the solution or to add new functionalities. We will describe every step we took, what we learned from this project and what we pretend to represent.

## Keywords

VIS; Information Visualization; Unemployment; Higher Education; Public; Private; Portugal; University; College; Course; d3; Pentaho; Dataset; Data; Derived Measure

# INTRODUCTION

Our domain for the project is the employment/unemployment of higher education courses graduates from all the public and private universities of Portugal during the period of 2007-2015.

We chose to analyze this domain because, as students, it is a very important topic to us and we are concern about which courses/areas have more or less employment/unemployment, especially to the ones that want to pursue a high level academic life and are undecided to what area want to pursue or what course inside a specific area. With the unemployment graduates analysis across these entire dimensions public/private sector, universities, colleges, courses and course’s areas we hope to achieve some interesting conclusions and help the future freshmen choose a course. Beyond the future freshmen students people (college chairman’s, public education government department…) interested in analyze the state or success/failure of universities, colleges, specific courses or certain course’s areas in terms of unemployment in Portugal, can use this visualization to extract conclusions.

In the beginning we decided to focus on creating a VIS so that we could accomplish **5 tasks (aka Originals)** but as we developed our solution we added more tasks **(aka Expanded)**:

1. Query->Compare - Compare the unemployment (%) of different courses (regardless of course conclusion year of the graduates). **[Original]**
2. Analyze->Consume->Present – Present the information about unemployment (%) from a specific course graduates across time. **[Original]**
3. Query->Identify – Identify the university with more unemployment (%).**[Original]**
4. Consume->Present – Relation between minimum entry grade and unemployment (%).**[Original]**
5. Query->Summarize – Summarize the employment/unemployment by graduation areas. **[Original]**

Initially, we set **6 representative questions (aka Originals Questions)** and later on with the development we set more questions that we wanted to answer **(aka Expanded Questions)** (for the tasks above):

1. Does the Computer Science graduate in IST have more unemployment, in 2015, than Computer Science in ISEL? And in 2007? **(Task 1) [Original]**
2. Is Computer Science in IST having less unemployed graduates in last years? **(Task 2) [Original]**
3. What was the year which had less unemployed people from Computer Science in IST? **(Task 2) [Original]**
4. What is the university with more unemployment? **(Task 3) [Original]**
5. Where the unemployment will be higher? In a course with 14 minimum entry grade or one with 17? **(Task 4) [Original]**
6. What is the graduation area with less/more unemployment? **(Task 5) [Original]**

We did manage to answer all the tasks and questions we proposed ourselves in the first checkpoint; while some of them are answered by only one idiom, others are answered by interacting with two or more idioms, but we made it in an intuitive way in order to improve the user experience.

The main framework we used to develop the graphs for this project was **d3**, we used **bootstrap** JavaScript/CSS library to do the HTML layout and **Pentaho (PDI)** for analyze, treat, clean and integrate data in order to come up with good “cleaned” data to be easily consumed by our visualization.

# RELATED WORK

We didn’t find much work like what we proposed to do.

We have seen this one initially <http://infocursos.mec.pt/> that allow to see all the higher education courses but it presents only basic statistic and information related to the courses like the gender/age/grade/unemployment distribution etc with poor interaction and a not very good use of idioms (too much and wrongly used pie charts). In the end it only presented the unemployment for each course isolated and without the possibility to navigate through time or compare courses for example.

We found many articles in journals; a few examples are listed below:

* <http://www.dn.pt/portugal/interior/quais-os-cursos-superiores-que-dao-emprego-veja-a-lista-5233967.html>
* <http://observador.pt/2016/06/17/veja-quais-os-cursos-superiores-com-mais-e-menos-emprego/>

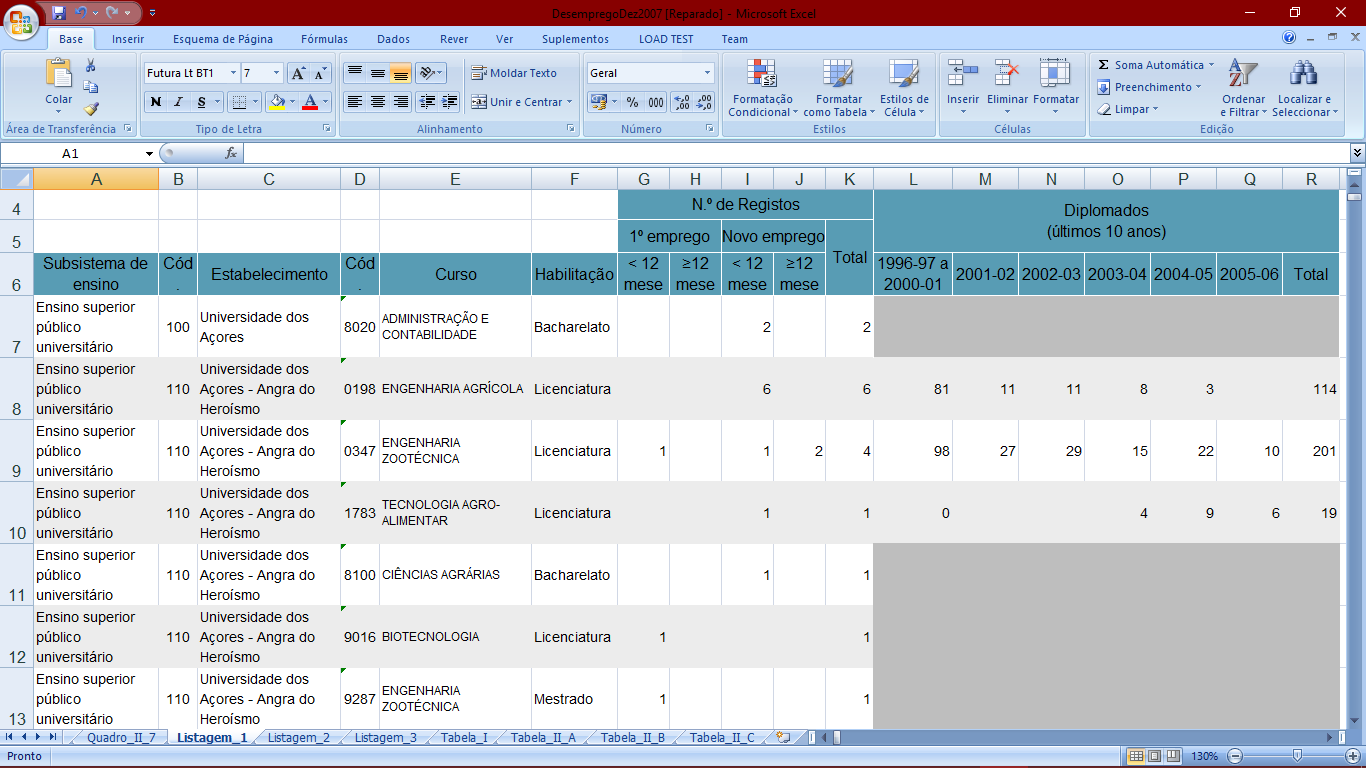
The journals have no interaction with the user, limited space to present the data and no possibility to a depth analysis or a more accurate one.

We wanted something that had the possibility to dig all courses (in Portugal) unemployment data as well the relation with the course’s area and the universities.

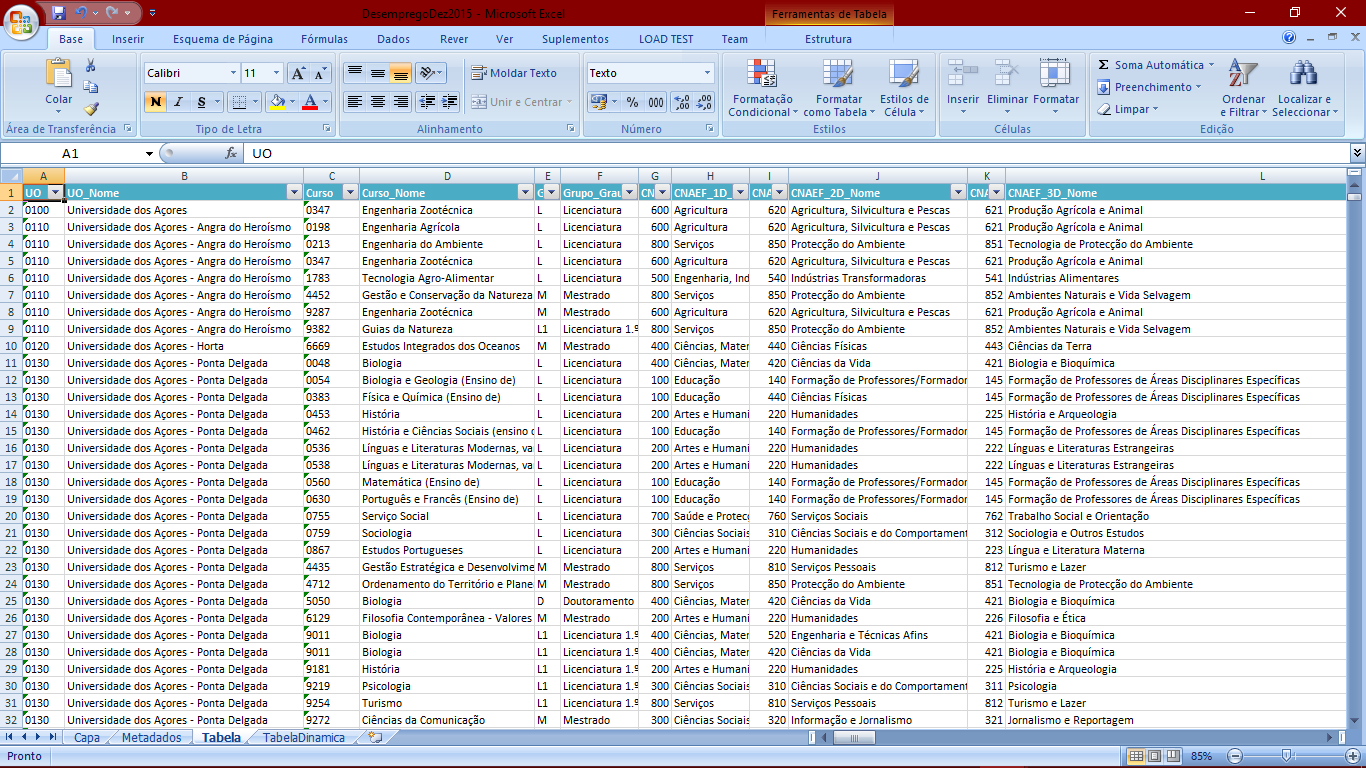
# THE DATA

**Initially (II-Checkpoint)**

The data we used was a set of documents (excel format) provided by DGEEC (“Direção-Geral de Estatísticas da Educação e Ciência”) with the information about all the courses and the respective year. This data is available at DGEEC website (<http://www.dgeec.mec.pt/np4/92/>).

Our initial dataset was one excel file for each year (2007-2015) about **unemployment from all the higher education courses registered in “Centro de Desemprego”** and other file with the entry grades of 2016 for all the public higher education courses since the private ones doesn’t have it. The files from 2007–2015 had different layouts and different tables (very heterogeneous design) as seen in Figure 1 and Figure 2, which is truncated because it has more than 50 columns/attributes but in the end all of them had the same information.

**Figure 1 – Courses 2007**



**Figure 2 – Courses 2015**

We started to select the following attributes: Year, Course Name, Course Code, University Name, University Code, Degree Level, Total Unemployed, Total Graduates, Course Area Code, Course Area Name and Entry Grade.

The course’s areas deserve a little explanation. There are 3 hierarchical course’s area levels. Top level has multiple roots (one for each big area) “it is a forest of trees”, the figure 3 summarizes the idea showing only one tree (leaves are the courses).

**Figure 3 – Course’s Areas Levels**

We calculated the following **derived and aggregation measures** for each year:

* **Derived: % Unemployment by Course** (100 \* Total Unemployed of Course / Total Graduates of Course) to compare the unemployment (%) of different courses (regardless of course conclusion year of the graduates) and to present the information about unemployment (%) from a specific course graduates across time.
* **Aggregation: %** **Unemployment by Each University** (100 \* Total unemployed graduates of university courses / Total graduates of university courses) to identify the university with more unemployment (%).
* **Aggregation: %** **Unemployment by Each Area Level** (100 \* Total unemployed of all courses from the area / Total graduates of all the courses from the area) to summarize the employment/unemployment by course areas.

We calculated these percentages (**normalizing the data**) instead of using the total unemployed because the totals were meaningless; courses that produce 400 graduates per year probably have more unemployed than one that produce only 30.

We used the table from 2015 to obtain the courses area relationship, because it was the only one with that information and Merge Join [by Course Code and University Code] it with all the other tables from the other years. Some records were lost due to the extinction of courses from “Bolonha” and restructurings (e.g. 2012 dataset had 5100 entries, resulting in 4933 entries). We crossed our data with a table that have information about what sector the university belong (private or public). Some courses didn’t have information about total graduates, we dropped those.

All of these data calculations and treatments we explained above were made using the **Pentaho (PDI) tool,** in the end this process produced Json files as output.

**After II Checkpoint**

To make the following data treatments and calculation we wrote some JavaScript scripts that received the Json files from **Pentaho (PDI)** and produced new ones that were directly consumed by the visualization.

In the first prototype that we developed we noticed that there was too many colleges (private ones + public ones = 400) and that would be hard to represent all of them to make a good visualization (**Scalability Issue**). In reality when we inspect our data carefully we notice that we didn’t have universities per course but colleges. So we aggregate our courses into colleges and our colleges into universities and shrunk our data set due to aggregation. In the end the logical view of our data was University<-College<-Course.

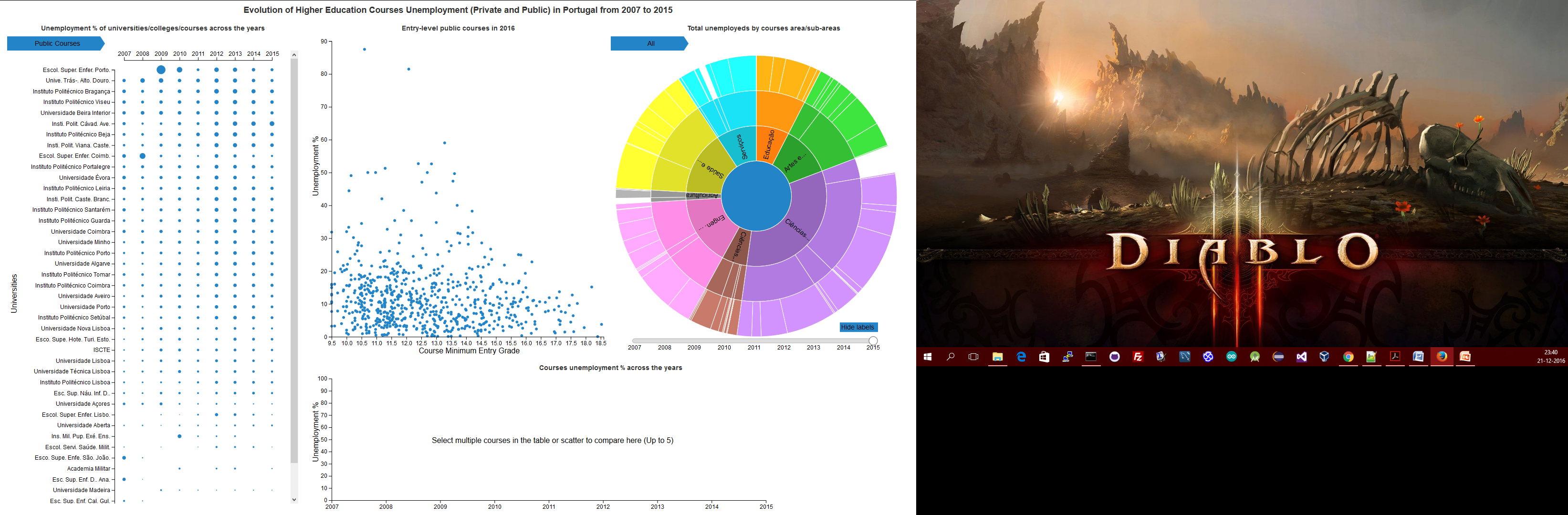
In the developing of our prototype we concluded that we needed a way to sort universities, colleges and courses by unemployment, example to support the **Task 3**. So we calculated **other derived measure** that we called **UnemploymentAverage** for each university, college and course. We did this calculation using the previous unemployment % measures calculated and making a simple average across all the years (2007-2015). Some courses didn’t have data in some years probably because there was none graduate registered in “Centro de Desemprego” in that year so we made the average counting that year as 0 unemployed benefiting the course, with the idea that if there are no one registered so all graduates are employed, this is not true because can exist people that are not registered in “Centro de Desemprego” because drop the search for job but was a decision we had to make.

We preprocessed our data using **d3 nest component** to make our data in a hierarchical object graph that can be easily consumed by visualization. This way we don’t need to do real time calculation and the search for an item is fast making visualization light in terms of performance and smooth to the user. We did this transformation because our data had strong hierarchies:

* University<-College<-Course
* AreaLevelTop<-AreaLevel1<-AreaLevel2

# VISUALIZATION

**Overall Description**

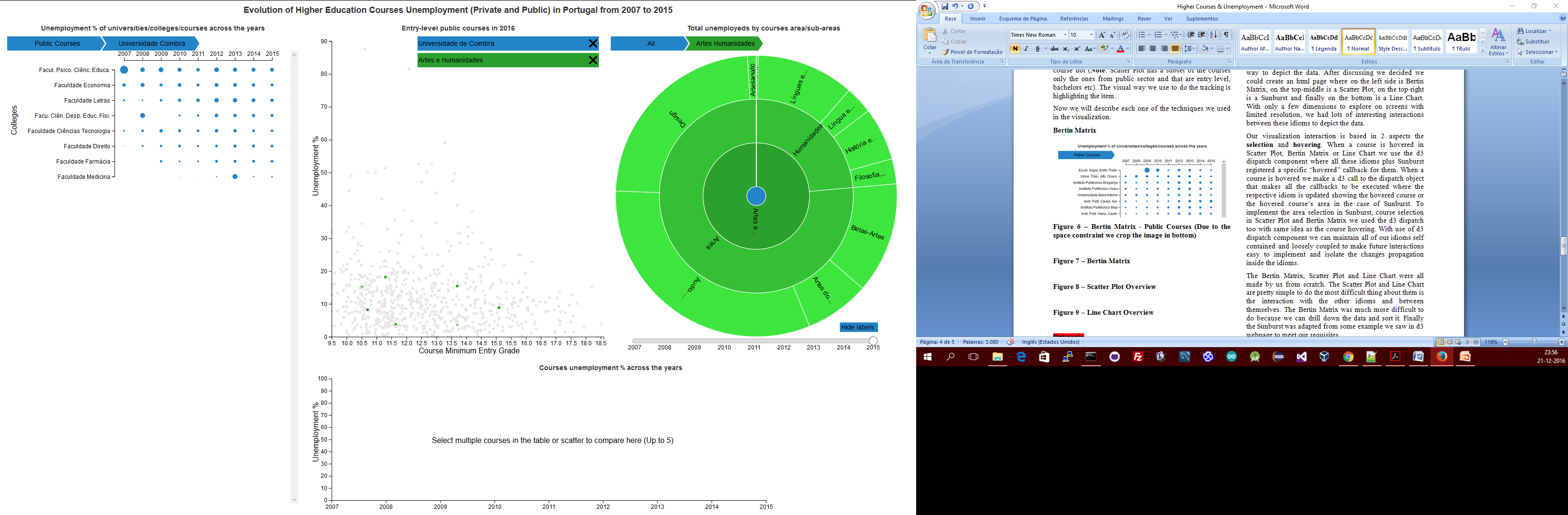
**Figure 4 – Layout overview: Bertin Matrix (Left), Scatter Plot (Top-Middle), Sunburst (Top-Right) and Line Chart (Bottom).**

* Bertin Matrix has the unemployment percentage for the University<-College<-Course hierarchy for each year (2007-2015)
* Scatter Plot has the courses from the public sector for 2016 with unemployment % and minimum entry grade relation.
* Sunburst represents the total of unemployed distributed by courses areas for each year (2007-2015)
* Line chart is used to visualize the courses unemployment % across years (2007-2015) to a more detailed analysis.

Our visualization idioms are highly interactive. We pointed to visualization with many interactions between the idioms and many redundant ways to perform some of the interactions so the user is not limited to the system workflow. With these guidelines in mind we tried design the visualization in a way that the user doesn’t feel forced to going one place only to select something when he could do it where we was looking.

**Selecting a university** (left click in the row) in the Bertin Matrix updates the Scatter Plot highlighting the courses from all the Colleges of that university maintaining all the others to give the context to the user. After we select a university in Bertin Matrix we can select a college from that university, if we do this the Scatter Plot is updated in similar way of the university selection but this time only the courses from that college are highlighted. In end the matrix selection filter what we see in Scatter Plot.

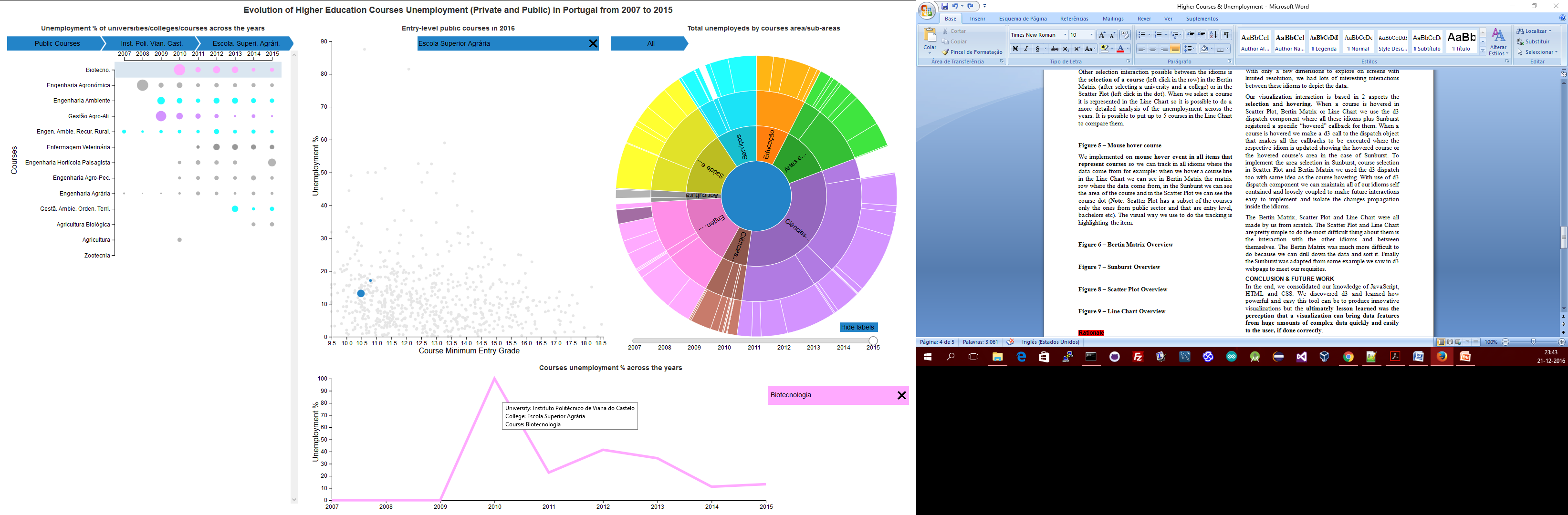
**Selecting an area or sub-area** (left click in the area section) in Sunburst affects the Scatter Plot in a similar way of the Bertin Matrix selection but in this case highlighting all the courses that belong to that area. In the end the Sunburst filter the courses in Scatter Plot.



**Figure 5 – Example of the composition of filters in Scatter Plot imposed by Sunburst and Bertin Matrix selection (In the top are the active filters)**

These two filters explained above work at same time so it is possible to combine them to see something more advanced in the Scatter Plot for example: it is possible to see all the courses from a specific college that belong to a specific area and see the disposition of them in the Scatter Plot (See Figure 5).

Other selection interaction possible between the idioms is the **selection of a course** (left click in the row) in the Bertin Matrix (after selecting a university and a college) or in the Scatter Plot (left click in the dot). When we select a course it is represented in the Line Chart so it is possible to do a more detailed analysis of the unemployment across the years. It is possible to put up to 5 courses in the Line Chart to compare them.

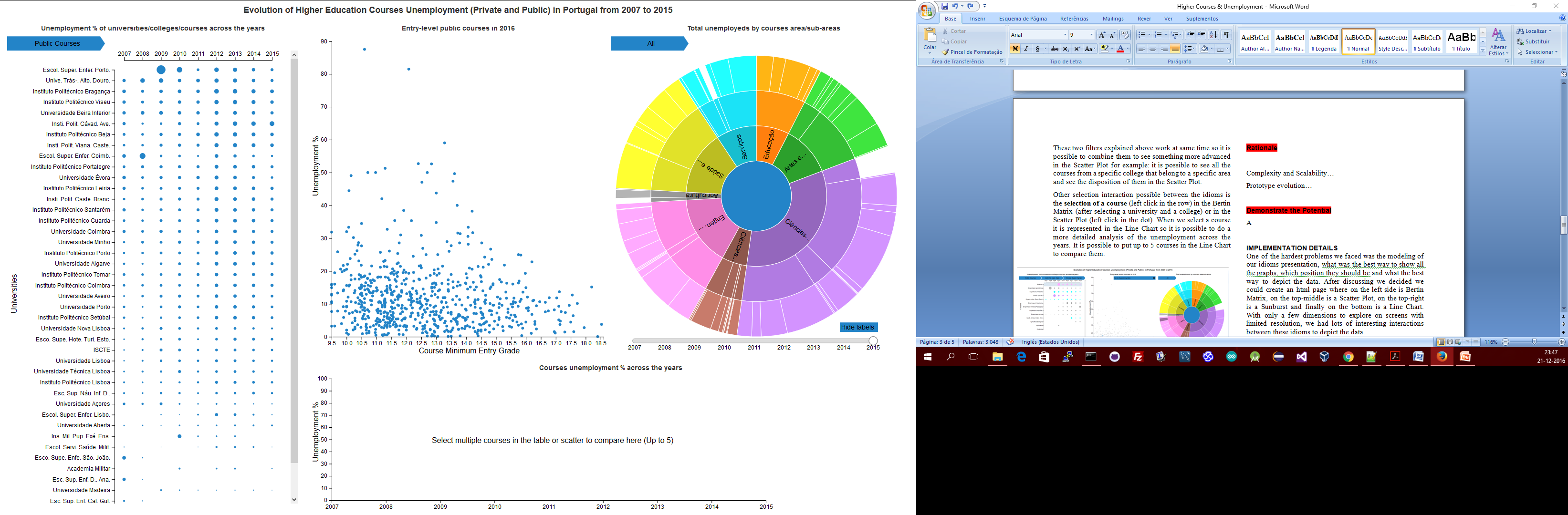


**Figure 6 – Mouse hover course in Line Chart and the effect in the other idioms**

We implemented on **mouse hover event in all items that represent courses** so we can track in all idioms where the data come from for example: when we hover a course line in the Line Chart we can see in Bertin Matrix the row where the data come from, in the Sunburst we can see the area of the course and in the Scatter Plot we can see the course dot (**Note**: Scatter Plot has a subset of the courses only the ones from public sector and that are entry level, bachelors etc). The visual way we use to do the tracking is highlighting the item (See Figure 6).

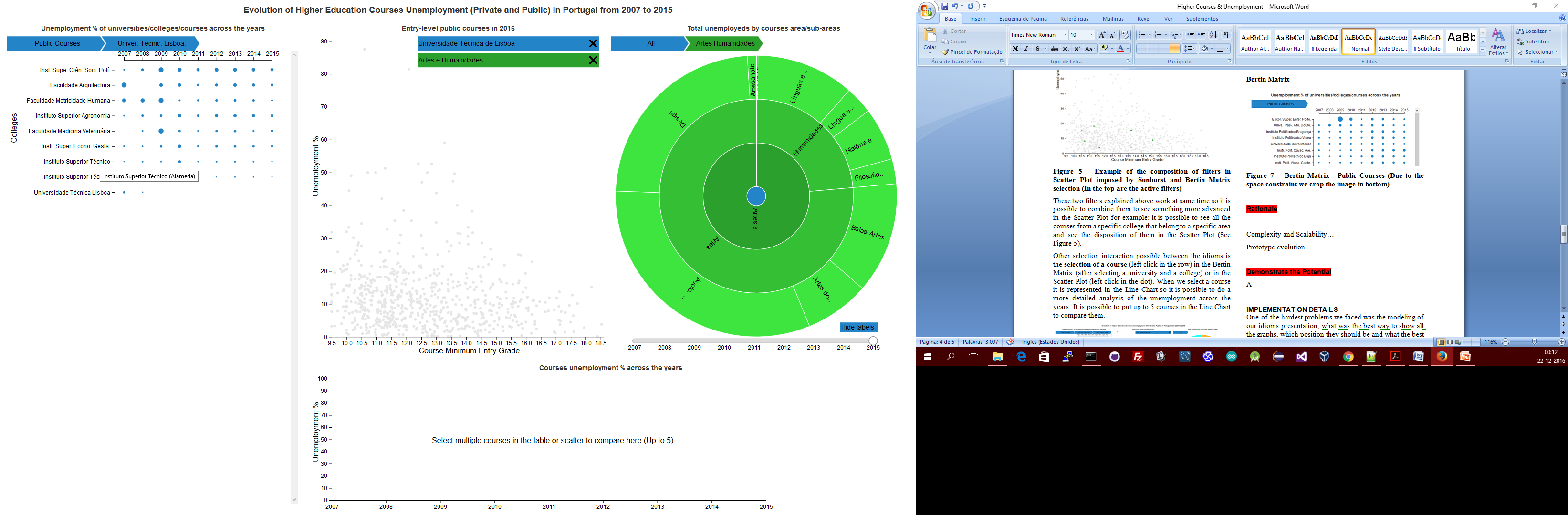
Now we will describe each one of the techniques we used in the visualization.

**Bertin Matrix**



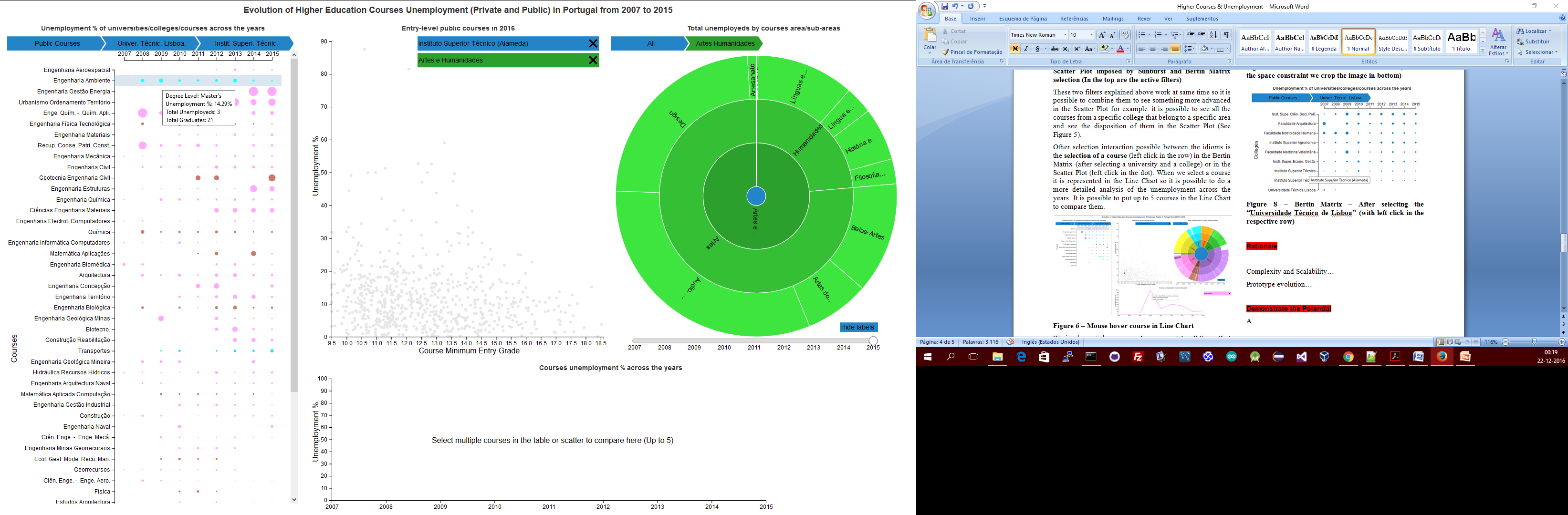
**Figure 7 – Bertin Matrix - Public Universities (Due to the space constraint we crop the image in bottom)**

In figure 7 we can see part of the Bertin Matrix with all the public universities represented. The circles are proportional to the unemployment % of all the graduates of that university. In the top we can see “Breadcrumbs” that indicate where the user are in the hierarchy (recall instead of remember).



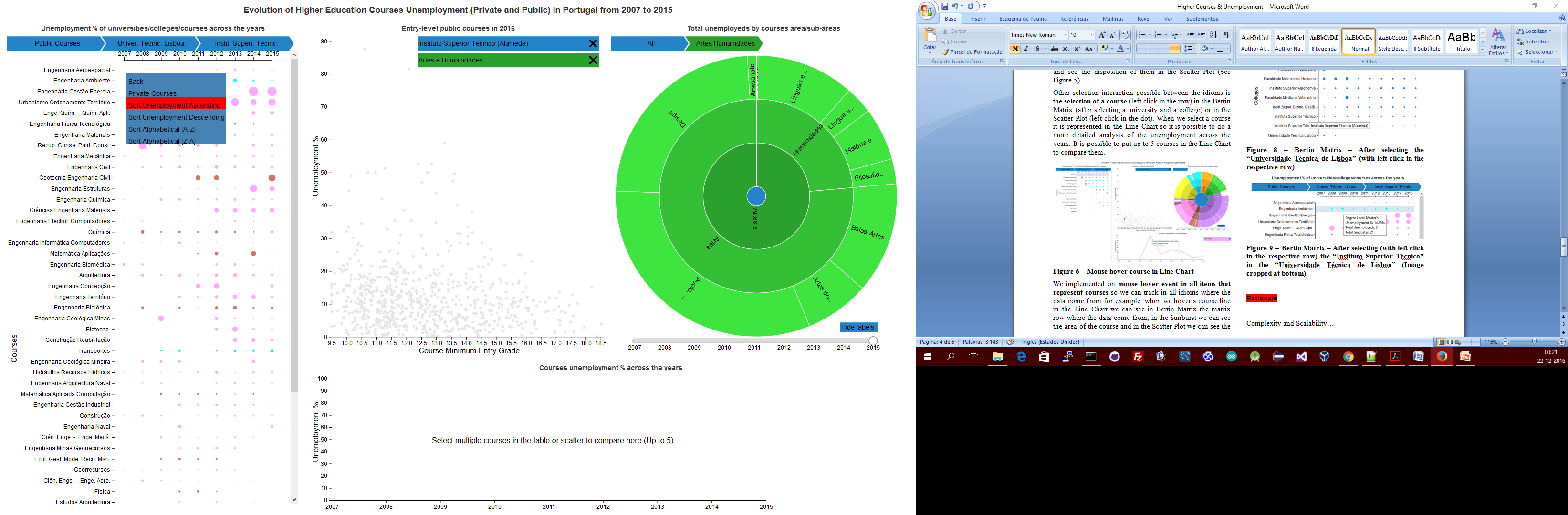
**Figure 8 – Bertin Matrix – After selecting the “Universidade Técnica de Lisboa” (with left click in the respective row)**

In figure 8 we can see the Bertin Matrix in a new state where we navigate from all the public universities to the “Universidade Técnica de Lisboa”, now we have all the colleges that belong to it listed. Now circles are proportional to the unemployment % of all the graduates of this specific college. The breadcrumbs in top were updated. We can navigate back clicking in the breadcrumbs.



**Figure 9 – Bertin Matrix – After selecting (with left click in the respective row) the “Instituto Superior Técnico” in the “Universidade Técnica de Lisboa” (Image cropped at bottom).**

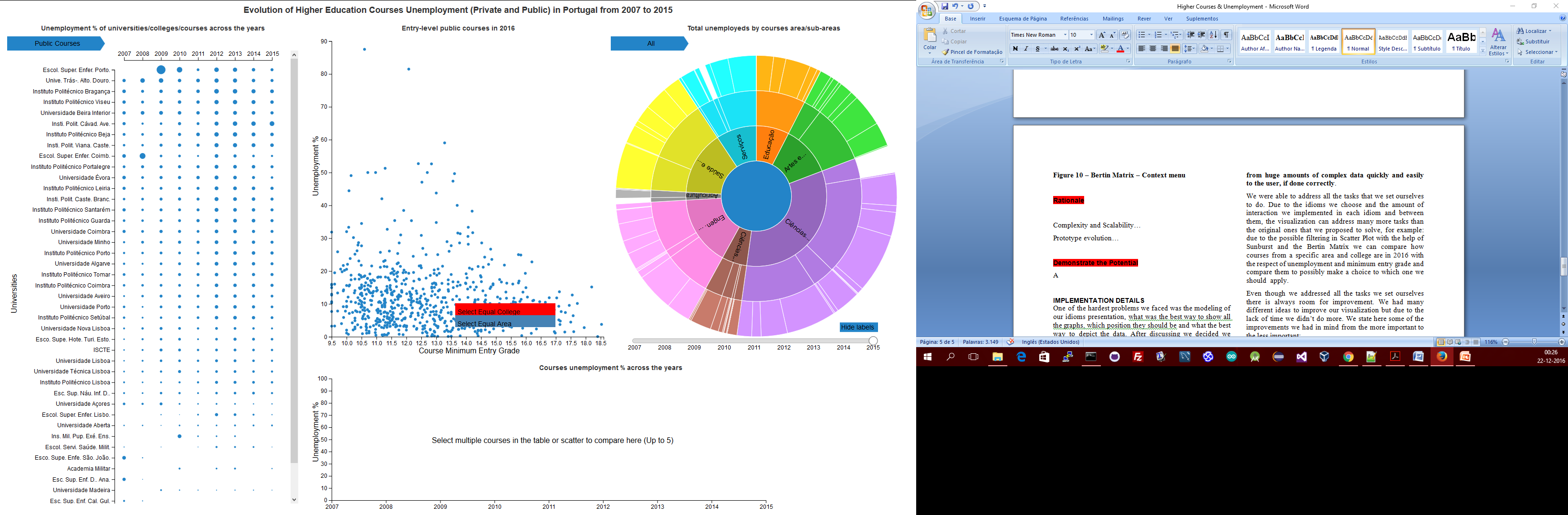
In figure 9 we can see all the courses from “Instituto Superior Técnico”. The circles here have different colors from the previous ones (See Figure 8 and 7) because these ones represent courses and the color of the circles match the course’s area color in the Sunburst. The blue we see in Figure 8 and 7 is the color of the root (See Figure 12) so we used it to represent aggregation items or items where we don’t want to expose the area. We can hover a circle with the mouse to see a tooltip with more detailed data.



**Figure 10 – Bertin Matrix – Context menu (right click in the matrix to make it appear)**

In figure 10 we can see the context menu where we can navigate back in the hierarchy. We can sort the current the data in the matrix by unemployment and alphabetical. We have the possibility to change the data in the matrix to the private education sector dataset too.

**Scatter Plot**



**Figure 11 – Scatter Plot – Context menu**

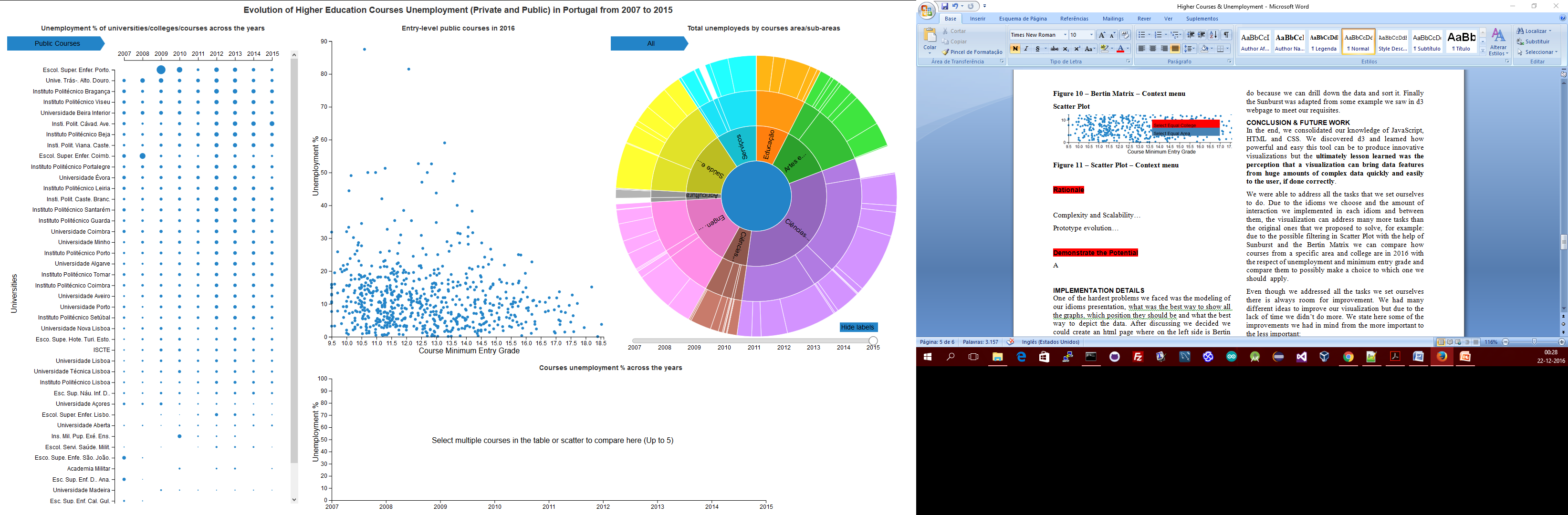
In figure 11 we can see a context that appear when we right click in a dot (course) that allow the user to filter all the courses in Scatter Plot that belong to the same college or to the same area. These interactions have the same effect as if we make the same selections in the Bertin Matrix and in the Sunburst allowing redundancy in interaction.

We can remove the filters in the “Breadcrumbs” of the Sunburst and Bertin Matrix but we can remove it in Scatter Plot too (See Figure 5 – Top of the figure) allowing redundancy in interaction too

The colors of the dots change if we have an area filtered (See Figure 5). We use the blue color of the Sunburst root when we don’t have any area selected, making the same decision we made in the Bertin Matrix to maintain the consistency in our visualization.

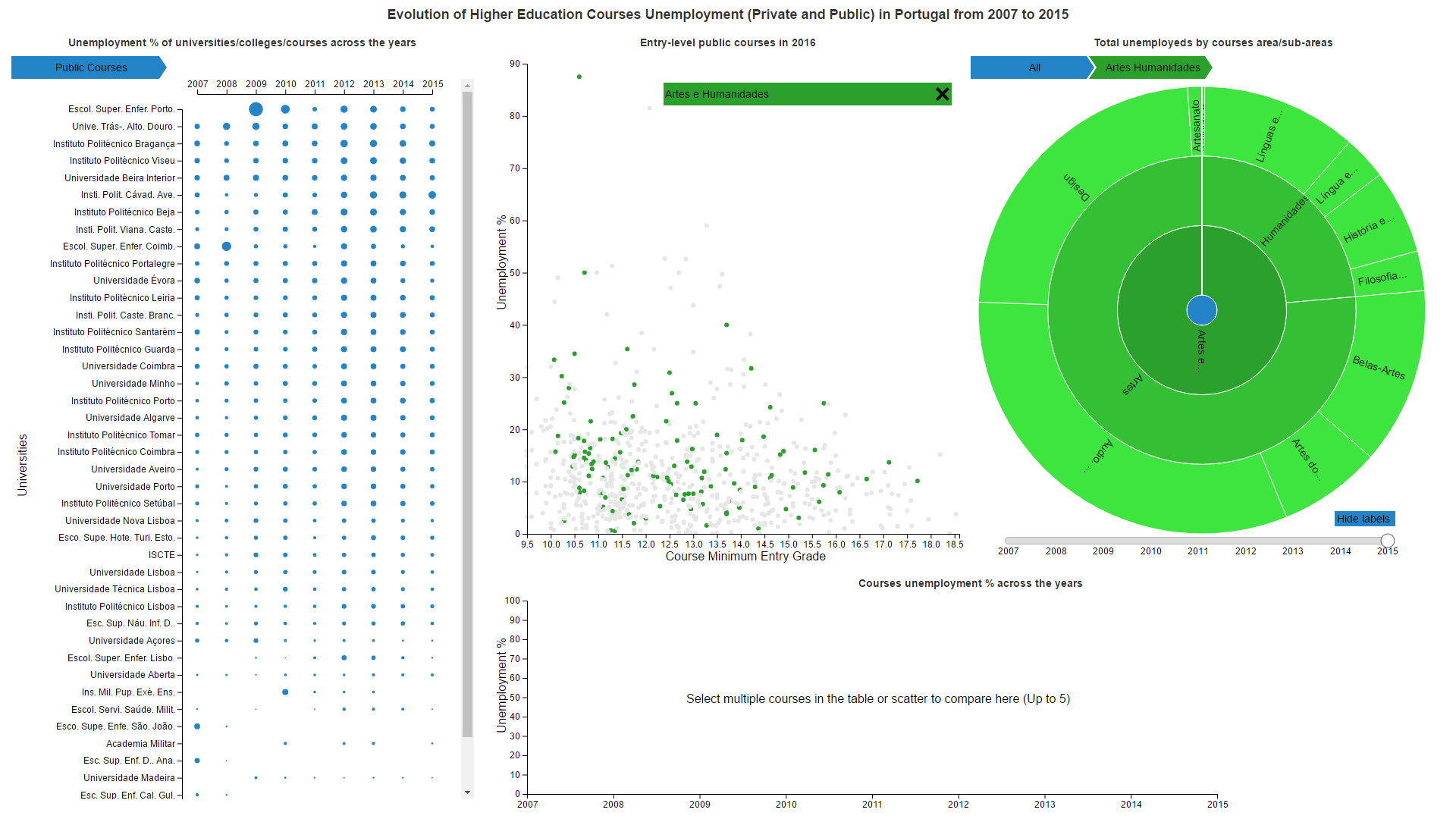
Hovering the dot with the mouse allow the user to see more details of that course with the help of a tooltip.

**Sunburst**



**Figure 12 – Sunburst – Representing the unemployed graduates by area in 2015**

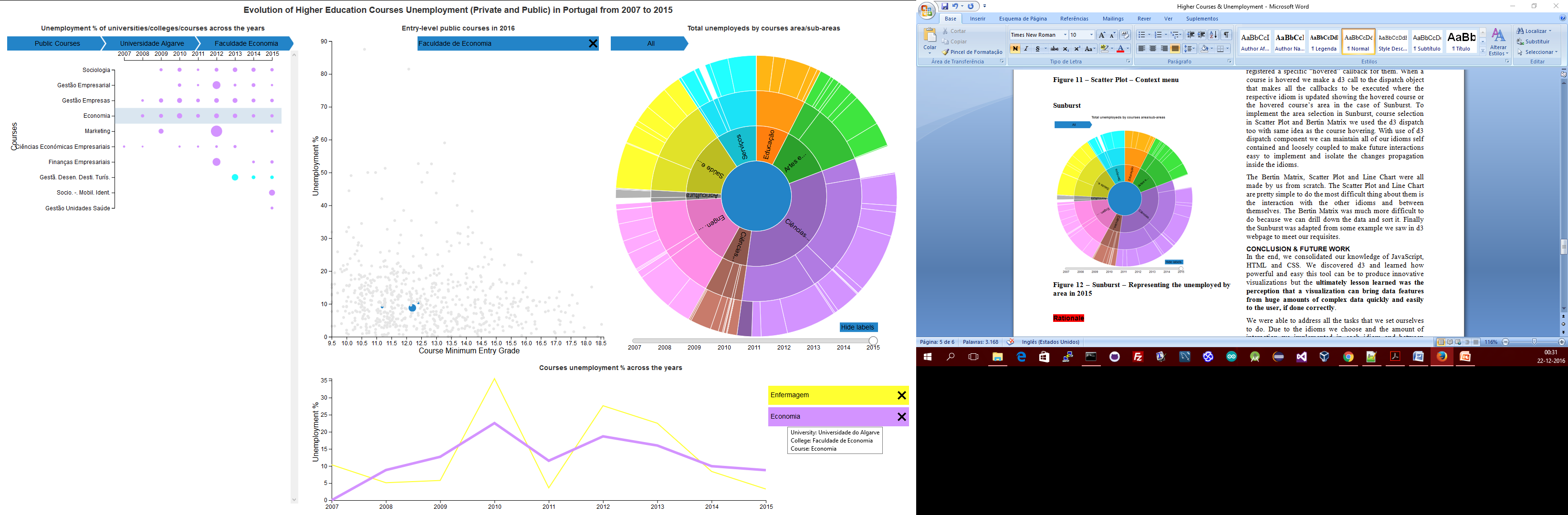
In figure 12 we can see the year selector in the bottom where is possible change the dataset of the sunburst. With a click in an area we can “zoom” the sunburst to that area. The button above the selector allows the user to hide/show the area labels in the sunburst. In the top is possible to see the “Breadcrumbs” that maintain the user navigation in the sunburst, exactly the same idea of the Bertin Matrix “Breadcrumbs” to maintain the consistency of the visualization.



**Figure 13 – Sunburst – Zoomed to the area of “Artes e Humanidades”**

In figure 13 we can see the sunburst state after we click in the area of “Artes e Humanidades”, the breadcrumbs updated accordingly and more labels appeared because we have more space to draw them.

**Line Chart**



**Figure 14 – Line Chart – Comparing the unemployment % of two courses (mouse hovered “Economia” course)**

In figure 14 we can see the Line Chart with two courses (selected in Bertin Matrix or Scatter Plot). The y-axis are dynamic and adapt to the maximum value of the data it contains so it is possible to compare courses with very low unemployment % making a good use of the space. The menu at right allows the user to remove a course from the Line Chart. Hovering the line or the menu item with mouse shows more details about the course.

**Rationale**

Complexity and Scalability…

Prototype evolution…

**Demonstrate the Potential**

A

# IMPLEMENTATION DETAILS

One of the hardest problems we faced was the modeling of our idioms layout disposition. What was the best way to show all the graphs? Which position they should be? And what the best way to depict the data? After discussing we decided we could create an html page where on the left side is Bertin Matrix, on the top-middle is a Scatter Plot, on the top-right is a Sunburst and finally on the bottom is a Line Chart. With only a few dimensions to explore on screens with limited resolution, we had lots of interesting interactions between these idioms to depict the data.

Our visualization interaction is based in 2 aspects the **selection** and **hovering**. When a course is hovered in Scatter Plot, Bertin Matrix or Line Chart we use the d3 dispatch component where all these idioms plus Sunburst registered a specific “hovered” callback for them. When a course is hovered we make a d3 call to the dispatch object that makes all the callbacks to be executed where the respective idiom is updated showing the hovered course or the hovered course’s area in the case of Sunburst. To implement the area selection in Sunburst, course selection in Scatter Plot and Bertin Matrix we used the d3 dispatch too with same idea as the course hovering. With use of d3 dispatch component we can maintain all of our idioms self contained and loosely coupled to make future interactions easy to implement and isolate the changes propagation inside the idioms.

The Bertin Matrix, Scatter Plot and Line Chart were all made by us from scratch. The Scatter Plot and Line Chart are pretty simple to do the most difficult thing about them is the interaction with the other idioms and between themselves. The Bertin Matrix was much more difficult to do because we can drill down the data and sort it. Finally the Sunburst was adapted from some example we saw in d3 webpage to meet our requisites.

# Conclusion & FUTURE WORK

In the end, we consolidated our knowledge of JavaScript, HTML and CSS. We discovered d3 and learned how powerful and easy this tool can be to produce innovative visualizations but the **ultimately lesson learned was the perception that a visualization can bring data features from huge amounts of complex data quickly and easily to the user, if done correctly**.

We were able to address all the tasks that we set ourselves to do. Due to the idioms we choose and the amount of interaction we implemented in each idiom and between them, the visualization can address many more tasks than the original ones that we proposed to solve, for example: due to the possible filtering in Scatter Plot with the help of Sunburst and the Bertin Matrix we can compare how courses from a specific area and college are in 2016 with the respect of unemployment and minimum entry grade and compare them to possibly make a choice to which one we should apply.

Even though we addressed all the tasks we set ourselves there is always room for improvement. We had many different ideas to improve our visualization but due to the lack of time we didn’t do more. We state here some of the improvements we had in mind from the more important to the less important:

1. Try to arrange a new color schema in the sunburst to distinguish areas in the same level specially areas that have the same parent area but are different. We should try get a color scheme that differentiate the sub-areas but maintain a visual relation with the parent and child ones.
2. The Sunburst idiom hides the total number of unemployed across time in 2007 were 9560 and 2015 were 23975, we wanted to implement a year selector that show this data using the idea of a “Scented Widget” to guide the user navigation/exploration.
3. Make animation transition between Sunburst states so the user can see which areas grown and which areas shrunk. We tried this one in the end but it had problems that we couldn’t fix so we drop it to have a solid visualization.
4. In the Scatter Plot we had some data points hide behind other data points and when filtered we couldn’t select the ones we wanted because they were behind due to SVG Z-index limitation. We should have reordered the SVG elements to bring them to front.
5. In the Sunburst the labels in the left side are with upside down and we should rotate them to be easy to read.
6. Use of buttons instead of the context menus in matrix and Sunburst to be more direct to the user perception and to the usability improvement.
7. Add filters in the Bertin Matrix to filter courses by degree level (Graduation, Bachelors, Masters, and PhD) and use some kind of visual representation to distinguish between them in Line Chart too.