Digital Topics in Science and Industry - Dashboard

G. Burdloff, C. Lobet, R. Mondjehi, V. Schott

Faculté des **sciences économiques** et de **gestion** Université de Strasbourg

Dashboard Purpose

We aim through this dashboard to make available some insights on the propagation of digital technologies in research - both in science and in the industry.

1 Deliverables

1.1 Code and Plots in the Github repo

In the root folder, excepting the file *app.R*, code files are numbered. We thereafter reference these codes by their number. Because of storage constraints, the original databases are not provided. However, we provide the extracted data that were then used to create most of the plots. These are located in the *data/* folder in addition to three more required files that are called in the code files. Finally, the *app autarcy/* folder contains all the files necessary to launch and host the Shiny application. As a result, all the plots are provided in the sub-directory *plots/*. The user can download this folder and run the app in his/her local RStudio session.

1.2 The Shiny App

The application is hosted on *shinyapps.io* in a free tier virtual server. For this reason, this web-hosted version is not highly responsive. The user is provided with the source code of the app so as to make his/her choice of which version to use (online or local). Note that, as discussed in section 5, using the dashboard could require to zoom out the webpage (to approximately 80%).

1.3 Report

A brief report on the dashboard is proposed hereafter.

2 Presentation

This work has been performed with two datasets.[1] The first is composed of scientific articles metadata that include digital keywords, either as a research target or as a tool for applied research. The second

[1] Data have been provided by M. Müller and S. Bianchini - BETA Strasbourg. Scientific papers have been obtained from Clarivate (Web of Science).

is the corollary for industrial research i.e. made of patents metadata.

Because we used a set of digital topics keywords different from the one used by the data providers, we end up with a slightly smaller amount of data, namely about 280,000 papers and about 24,000 patents. Both share a similar time window from 2000 to 2020. The final datasets containing the extracted topics can be obtained out of the code files 1, 2 and 3. These same codes also produce several small datasets then used to create the following plots (code files): wordclouds (4), digital topics networks (5), evolution in science fields (6), authors & journal contributions (7) and causality heatmaps (9). Additionally, we provide visualizations (networks) of the scientific research areas over time and across countries.

3 Visualizations

3.1 Digital topics relationships and their Contributors (C. Lobet)

Most visualizations, even further those discussed in this section, have required a *basis* data processing that can be found in code files 1, 2 and 3. Since our vision was to create all the plots outside of the dashboard code - mainly because of the computational constraints encountered in a free tier virtual machine hosting solution - small datasets have been derived from the main database, so as to make the plot creation easier for everyone. All the small datasets are made available in the *data/* folder. The main metric in these datasets is the number of occurrences of the digital topics. In the case of scientific articles, an alternative metric that can be retrieved in several visualizations as *popularity* is the number of citations obtained by these digital topics.

From these datasets have been created - among other visualizations - the three first tabs (after *Home*). The first one, *Overview*, is aimed at providing through simple wordclouds insights on the most important digital topics in science, industry and both of these fields, for the whole period 2000-2020 and for single years in this range. Another visualization in this tab is there to inform about the relationships between these digital topics. The topics are represented by nodes which sizes depend on topics occurrences or popularity. The networks edges indicate co-occurrences of the topics and edges widths are weighted by the corresponding amount. Corresponding code files are 4 and 5.

We propose a very casual visualization of contributions to digital topics research - or use in research - from authors (science), publishers (science) and inventors (science) in the tab *Contributions*. Code file 6. In the picker menu, the entities (author / publisher) are sorted according to their overall contribution.

Finally, it seemed interesting to observe the evolution of digital topics popularity inside a scientific research area. In the *Evolutions* tab, the user will find simple visualizations of this question with the possibility to select the desired topic (sorted by overall popularity in the scroll menu). We make available the opposite visualizations i.e. the evolution of research areas interested in one specific digital topic. Corresponding code file is 7.

3.2 Causality (R. Mondjehi)

The Granger causality test to test the causality between the series and the vector autoregressive model (VAR) to measure the direction of causality between our series.

To do this, we had to construct series of the most studied topics in a general way for the documents and patents that we will study separately. And construct a cross-tabulation between topics and scientific research areas.

After extracting the series of topics and their scientific application areas, we wanted to investigate whether there was a causal relationship between the topics studied on the one hand and the impact of these topics on the different scientific application areas on the other.

This method will thus allow us to establish causal connections not only between the topics and but also between the topics and the relative scientific domains.

The granger causality tests the null hypothesis according to which a series X causes another series Y when the p-value is lower than the significance threshold (0.05). If p-value is greater than 0.05 then there is no causality in the granger sense. The result obtained is contained in a causality matrix that we will visualize in a heat map.

Then the VAR model is used to determine the direction of influence of the series on each other. A correlation matrix is obtained as a result that we will also visualize in a heat map.

Causality tests and correlation matrices have only statistically proven our apprehension about how pervasive data science and AI are in most research fields today. Furthermore, one could extend the VAR model to predict the research topic.

3.3 Research fields and Countries (V. Schott)

We decided to provide these insights about papers publication to highlight main collaborations between most popular research areas, such as Engineering, Computer Science, Robotics, Chemistry, Surgery, etc. We also filtered for each year to allow analysing the evolution of those links. The idea for the country analysis is basically the same.

To setup these features, we first had to process our original data. In fact, for country analysis we used for each paper the affiliated country for each author. Thus, a paper wrote by an author from a Chinese University and two authors from a French University, we'll count this paper as coming from China and from France. We found around 160 countries. For visualization purpose, we choose to keep only 25 most important countries, ranked by number of papers.

For the research area analysis, we conducted a similar process, and kept the 25 most import research areas, based on number of links in order to show the most collaboratives ones.

In addition, we decided to provide another feature, a combination of the topic and the country analysis. In fact, we provide several networks, one for each country highlighting most popular research areas and most linked research areas. This is an interesting feature because shows the main differences between countries in term of papers publication.

For this part of the network, we used python for the entire process (from data pre-processing to network displaying). For data processing we mainly used pandas, then we built our networks using the networkx package, finally to create very smooth and dynamic networks, we "send" the networkx networks in pyvis.

The main issues encountered doing this part of the project was the visualization using Networkx combined with Pyvis. In fact, these two packages are not entirely compatible, because most features that can be added using Networkx are added after displaying the main graph, and only this main graph can be send to Pyvis. For example, I was not able to increase the font size of the node name in Pyvis, neither to make the size of the edges related to the intensity of the actual link, even thought I had all the needed information.

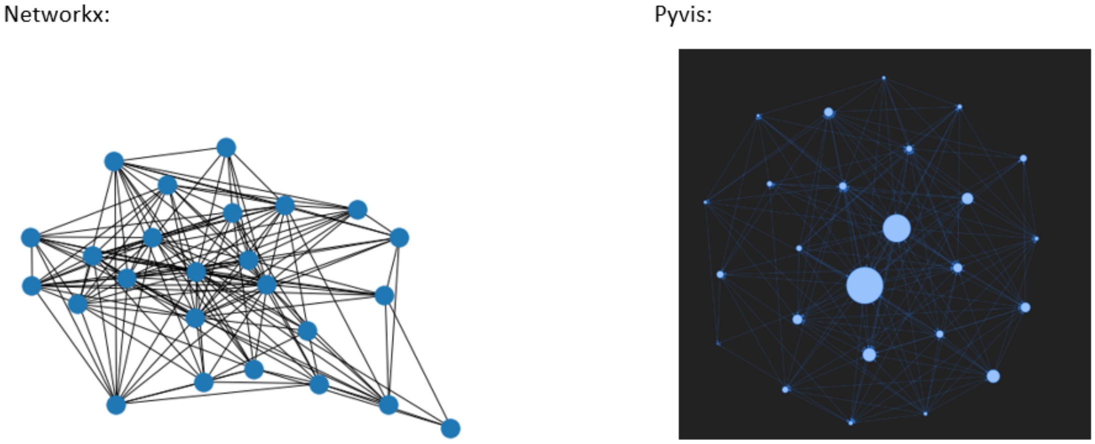

Figure 1: Networkx and Pyvis

4 Dashboard features (G. Burdloff)

The dashboard is constituted of several tabs dedicated to a specific type of graph in which the user can set some parameters to visualize the graph he wants. In some tabs the UI is dynamic which means that by choosing some specific parameters in the tuner, some options will appear (or disappear) depending on previous parameters which avoid issues of data unavailability. The user can access to six different tabs:

- The *Home* page: which is empty but could have replace some part of this report.
- *Overview*: the user can visualize some network through a dynamic interface about digital topics in science, industry or both fields, according to the box he chose to tick in the tuner, and filter them by year. There is also a panel dedicated to word clouds which basically works in the same way.
- *Contributions*: In this tab the user can choose the type of contributor he wants by ticking the associated box, which will update the first dropdown menu that contains the name of all contributors which belongs to the chosen category. After choosing a contributor, the second dropdown menu will contain all the available plot associated to the contributor's name that the user chose before. Notice that the second dropdown menu display the real file name which is not very user friendly and can be improved by using other potentially more complex solutions.
- *Evolutions*: This tab contains two panels which respectively shows the evolution of digital topics popularity cross time and digital topics popularity in specific scientific domain cross time. In both panels the user can choose between four levels of popularity, specify a digital topic or a scientific domain and finally select an available plot to display in the same way as the contributions tab.
- *Causality*: A very basic tab that contains heatmaps displayed as images, corresponding to science, industry or both fields depending on what the user chose in the tuner. We advise the user to enlarge the panel in which heatmaps are displayed (with the dedicated button on top right corner) and zoom in it to read them in detail (zoom with the navigator tool).
- *Countries*: As in the Overview tab, the user can visualize networks and filter them by year or countries (depending on the selected panel) to examine relations between countries in the first panel, links between topics in a chosen country in the second and finally the link between topics in the world, filtered by year.

5 Limitations / known issues

- In *Contribution* and *Evolution* tabs, make the plot picker more user-friendly, i.e. get rid of choosing the plot by its filename. This has been done properly in all other tabs.
- Include the light versions of *Overview* networks (these versions exist and can be found in the Github repository). See if it loads significantly faster compared to original networks.
- Manage plots dimensions. Right now the dashboard is well displayed on a 1080p format but without borders (the browser takes some place so we lack space). Potentially requires some *javascript* skills. For now this issue can be solved by zooming out the web page to about 80%.
- In the *Fields and Countries* tab - in the sub-tabs *Topics* and *World* - the user has to manually select an option for the visualization to be loaded. The default plot does not load automatically.
- In the same tab, nodes names are quite small so the user has to zoom in to see them.
- In the *Contributions* tab, for *inventors* plots, several are not working because of path specification issues.

6 Contributions of authors

G. Burdloff I was in charge to design entirely the dashboard, the global structure of it, basically where the elements are placed, but also the script running in backend using the shiny reactive components which allows to create some dynamic and interactive features in the dashboard. I'm used to work with shiny but not with a lot of external elements as we did for this project, so I had to find a way to integrate all the external HTML and JS files used to render the different plot designed by my team. I also did a bit of data manipulation to handle some issues in the text files used to fill the different menus for example. The code of the Shiny app is named *app.R* in the Gthub repository.

R. Mondjehi Data pre-processing, causality and correlation test, matrix, and heat map. The corresponding code is the one labeled 9 on Github.

V. Schott Work on the research fields and countries collaborations. More details in section 3.3. The corresponding code is the one labeled 10 on Github.

C. Lobet My technical missions were to pre-process the data, extract the digital topics and produce the plots of the first three tabs. Corresponding codes are 1 to 8 in the repository and the extracted datasets can be found in the folder *data/*. I was also in charge of testing the app and hosting it on *shinyapps.io*. Finally, I had the opportunity to supervise to some extent this project and follow and bring together the work of other members (exception is V. Schott who made his own initiatives in proposing visualizations of the countries collaborations and of the evolution and relationships of research areas in science). Consequently I gathered members works and compiled them in the Github repo as well as in this report. As an endnote, members all were implicated and feedbacks revealed a good learning curve for everyone in this project, which is obviously the main purpose of the attached course.