



Scientific Session: BIG DATA

uRos 2021: “Network Visualization of Multi-data Sources using R”

Rui ALVES



Shirley ORTEGA-AZURDUY



Christina PIERRAKOU



Chairman: Alexander KOWARIK, Statistics Austria

Thank you Alexander.

Good morning everyone, it's great to be here. We're going to make a presentation on "Network Visualization of Multi-data Sources using R"



Scientific Session: BIG DATA

Visit the github page and download the R script or the html file

<https://github.com/ruialv/VizNet-uRos2021>

Both the R script and the rendered result in html format are available for download in the link to the github account you have on your screen

If you'd like you can download it and run it on your Rstudio session or open the html file in any browser.

And now Christina is going to give you the context in which this "Network Visualization of Multi-Data Sources" emerged.

Christina, the floor is yours...

- [ESSnet2018] ESSnet on Big Data 2018 -2020 - Eurostat grant ESTAT-PA11-2018-8 Multipurpose statistics and efficiency gains in production.

https://ec.europa.eu/eurostat/cros/content/essnet-big-data-i_en#WP7_Multiple_domains

- ESSnet Big Data II WPJ – Innovative Tourism Statistics - Task 1C

https://ec.europa.eu/eurostat/cros/content/WPJ_Innovative_tourism_statistics_en

ESSnet Big Data II was a project within the [European statistical system \(ESS\)](#) jointly undertaken by 28 [partners which was run from November 2018 until December 2020](#). Its objective was the integration of big data in the regular production of official statistics, through pilots exploring the potential of selected big data sources, and through building and implementing concrete applications. It was a continuation of [ESSnet Big Data I](#) (from February 2016 until May 2018) and consists of 12 workpackages.

One of them was WPJ Innovative Tourism Statistics.



Objective

Develop a conceptual framework and setting up a prototype of Tourism Information and Monitoring System

Combining data

- Multi-purpose data (administrative)
- Survey data
- Web data (webscraping)

❑ The main objective of this package is to address the need of a conceptual framework and setting up a smart pilot Tourism Information System that will support statistical production in the field of tourism by integrating various big data sources with administrative registers and statistical databases using innovative statistical methods (data discrepancies, incoherent concepts, indirect relations between sources, redundancies of information).

The use of new information sources (including Big Data sources) in official statistics opens up completely new possibilities of enriching and improving the system of tourism statistics. Thanks to external sources of information, the data provided by official statistics can be more current and reflect the needs of users.

- How to get rapid overview of the data sources types
- How to trace back and forth where are sources used and to which end
 - Identify/Visualize which other countries use the same (or similar) sources
 - Understand the different purposes leading to the use of an specific source
 - Be able to directly browse into the external sources and get new insights
- Need to support production process to assess potential efficiency gains

#Thank you Christina.#

Why do we need a net work tool?

Official statistics rely strongly on monitoring properly and timely changes on economic activities. Economic activities can often cross territorial borders as well as knowledge fields. Tourism statistics is one of this kind because the hospitality industry requires and incorporates information of various fields like hotels and accommodations, but also about catering, housing & building markets, culture, sport, etc. Surveys and Registers are commonly used to produce official statistics. However, it is also known that the majority of tourist accommodations is booked online, that holds also for large leisure events. These sort of Big Data is become an important source of information for NSI's"

Hence, getting a rapid overview and intelligence of the data sources types, data sources used and outputs produced by other NSI's are of crucial importance.

Ideally, one should be able to trace back and forth the data sources and statistical output. It means:

- Identifying and visualizing which countries use the same (or similar) sources

- Understanding the purpose of use of an specific source
- Being capable to browse directly into external sources and get news insights.

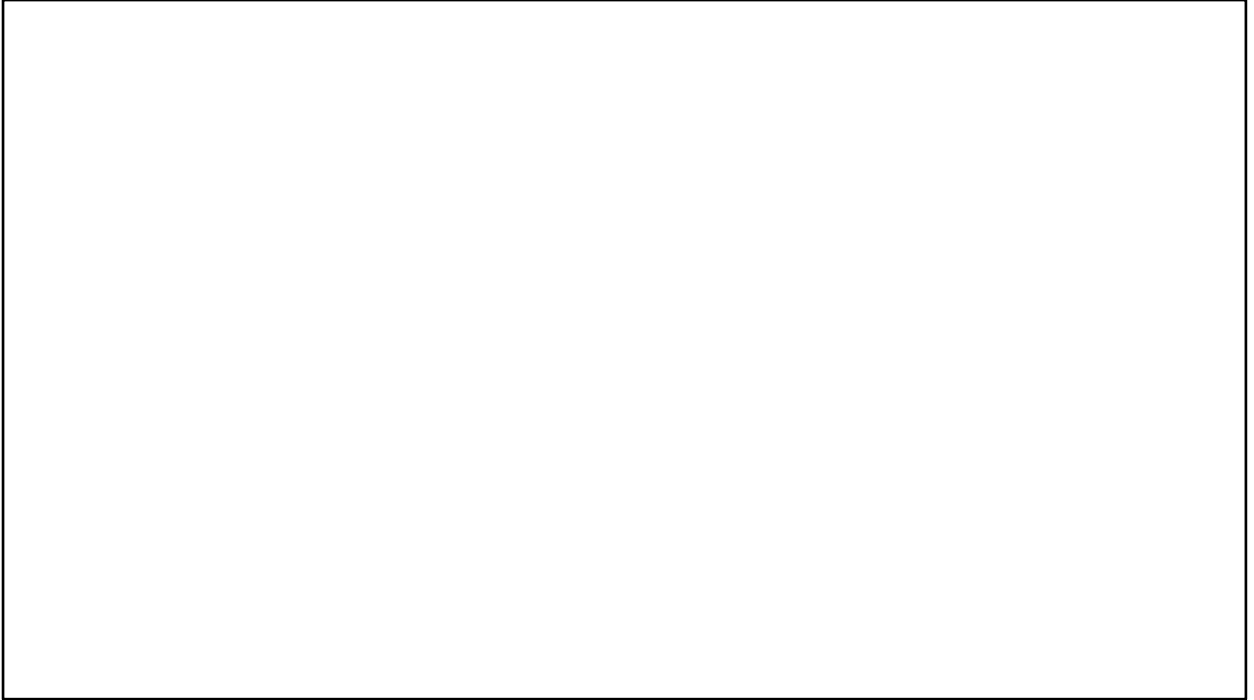
These all stages have the target of supporting statistical production processes in a way that potential efficiency gains can be assessed.

Multiple individual workflows do not give the full picture



A starting point of the WPJ was developing an inventory of current and potential data sources that can (or could) be used to produce fast and reliable statistical output.

A multiple number of individual workflows was collected. On this slide you can see that the heterogeneity across country workflows is large. Workflows are schemes or models showing the input and output of NSI's statistical processes to produce e.g. tourism statistics. You can notice that it was not easy to get the full picture.



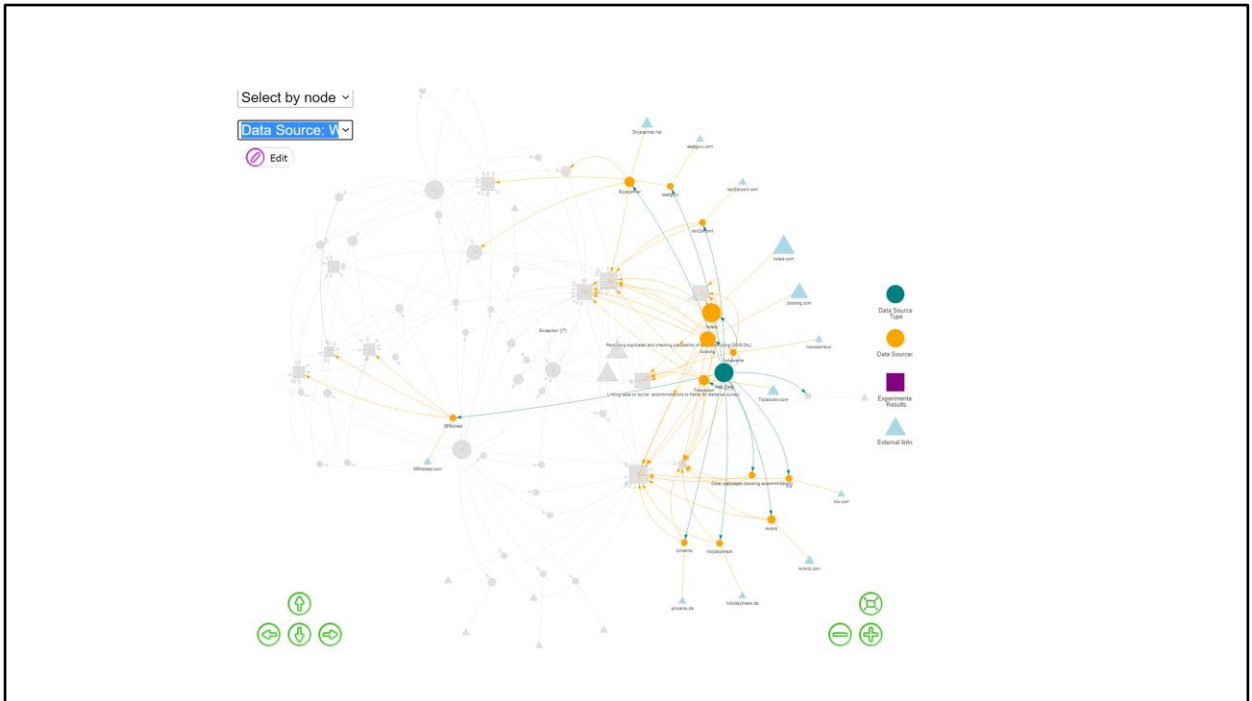
Mention-worthy when it serves the demonstration:

From a user point-of-view, the solution to this problem was developing an interactive network visualisation which has two main selection criteria: a single node and a group (or topic).

As you can observe:

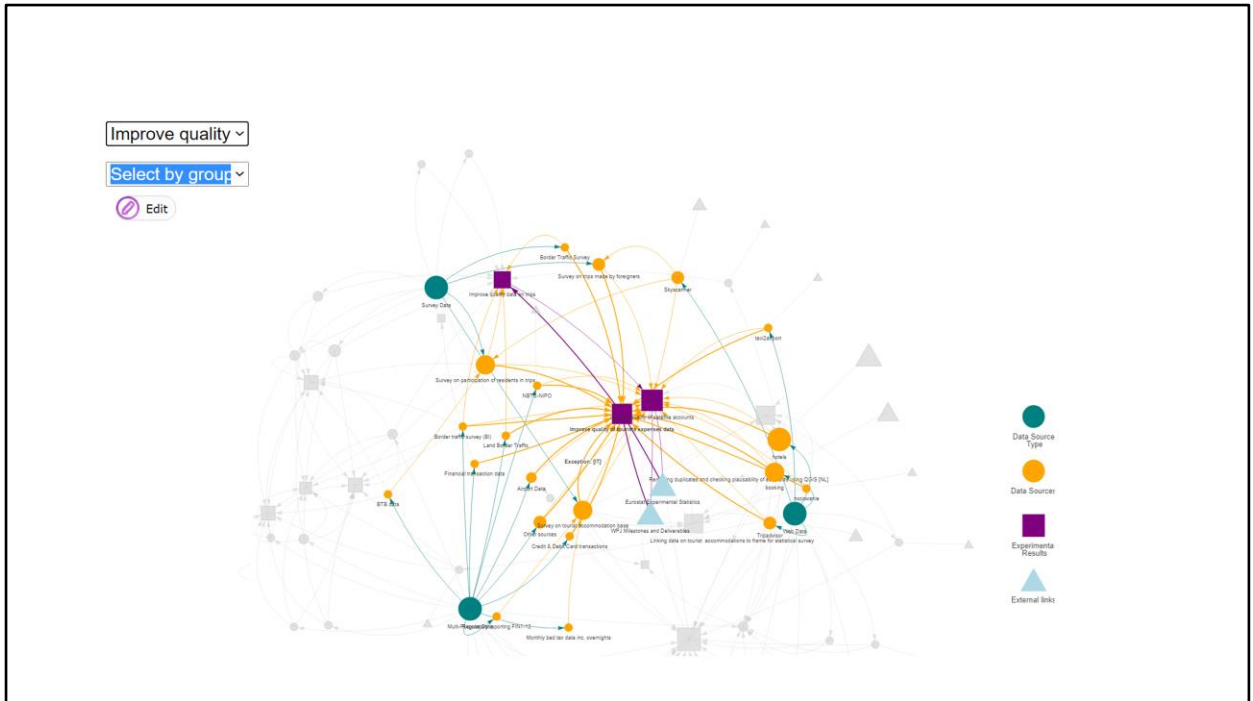
- Important nodes are centered. Green-dots are used to represent the data source types. Orange-dots represent the data sources and the purple squares stand for outputs.
- Notice that the bigger nodes relate to more countries.
- On the bottom, the user has green buttons: on the left for navigation and on the right to zoom-in and -out and to re-center the visualization.
- One can also refresh the page in the browser which re-builds the network and [**@Rui, PULL NETWORK**] one can pull the network (left or right or up or down) and this the software will rearrange the layout of the whole network.
- Further, by hovering the mouse over a node enable us to show additional information...
- To visit "Live" an "external link" (url), press on the triangle-shaped nodes. These external links can be either html-, PDF-, image- or Excel-files, etc...
- One can also execute a multi-selection using "ctrl+click"

- Last, there is an option to export a chosen layout of this visualization using the black-button: “Export as PNG”.



Now, let say we want to know the importance of web scraped data. So, let us choose the **group “Data Source: Web scraped data”**.

1. We can see “which sources” were scraped, namely hotels.com(8), booking(6), tripadvisor(3) and airbnb(2).
(Notice that the bigger a node, the more countries are involved using these data.)
2. If you wonder “How many countries use booking?” Just hoover on the node. There are 6 countries collecting booking.com data.
(To what results?)
3. You may guess the purpose of using these data in terms of official statistics, i.e. improving the accommodation inventory but also to learn more about aggregated information of visitors on expenditures levels and visited areas. (Look at the colour of the lines. They show the predominant connection of a node.)

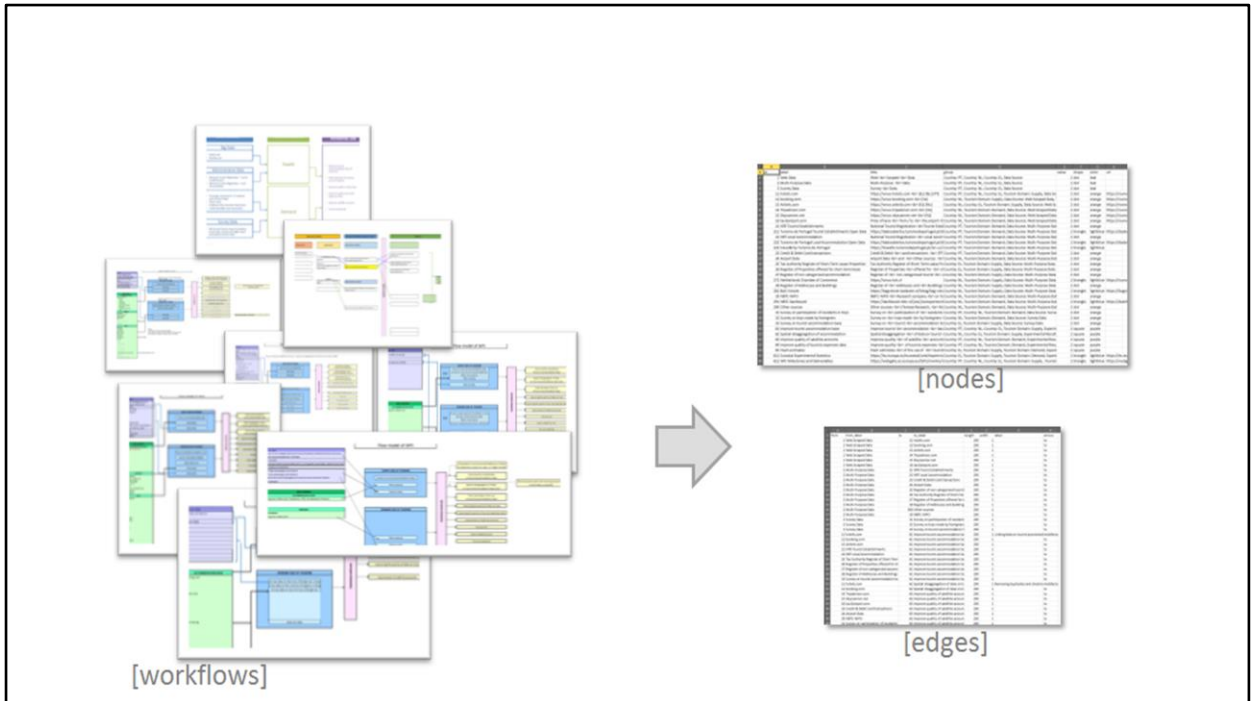


Let us now **select a node**. The output node: **“Improve the quality of tourist expenses data.”** (Notice again: the bigger a node, the more countries involved using these data.) We observe that 6 out of 8 countries have a Survey on participation of residents in trips and also 6 countries have a Survey on tourism accommodations base. Two countries used the “Airport registers” and (as shown before) also webscraped data of hotels, booking and other booking websites.

In short, using this vizNetwork tool, we can see back and forth the data sources and the data types involved.

Besides we can also read out that there are two other output connected to the expenditures, namely “Improve the survey data on trips” and “Improving the quality of satellite accounts”.

Thank you for your attention. I am very glad to leave the floor to Rui from Statistics Portugal (the engine behind this innovative tool).



Thank you Shirley, you're too kind

Without wanting to turn this presentation into a tutorial, we'd like to share with you how we did this.

This first step required a significant work of re-conceptualization in order to make the different workflows compatible and consistent, without tampering with their substance.

The majority of the countries used a similar template but its implementations and content varied slightly.

Making everything compatible was quite challenging.

The first thing to do, at least in this particular case, was to “translate” the static workflows into 2 tables (or dataframes): nodes and edges. These are the core of a network.

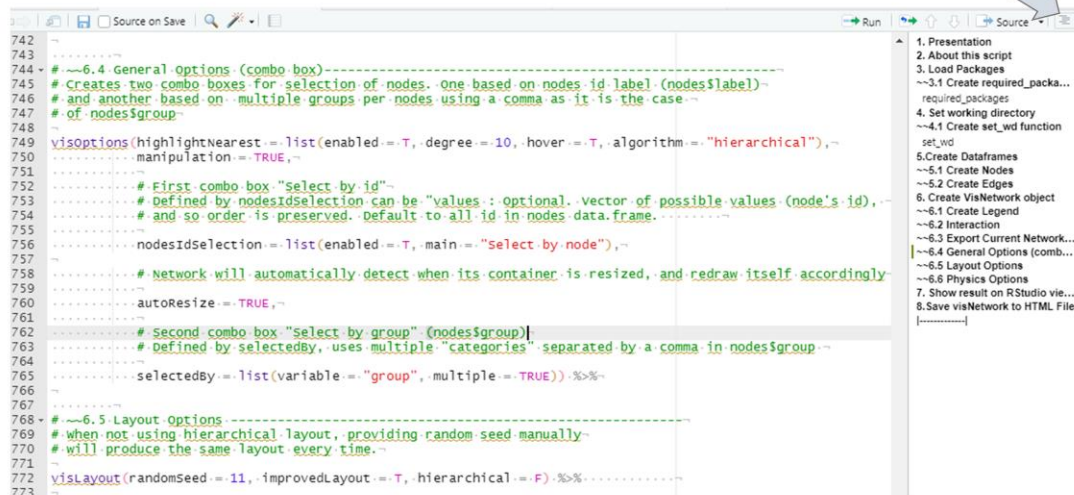
Nodes dataframe has numeric ID of the nodes and their properties: size, shape and colour. It also has **label**, **title** (provide the tooltip info when hovering over a node) and **group** that have the data for the combo boxes.

Edges table has information on the connections and also on length, width and

arrow type. The ID's for the "from" and "to" are the same as node's ID.

<https://github.com/ruialv/VizNet-uRos2021>

Ctrl+Shift+O



```
742 ~~~~~
743 # ~6.4 General options (combo box)~~~~~
744 # Creates two combo boxes for selection of nodes. One based on nodes id label (nodes$idlabel)~
745 # and another based on multiple groups per nodes using a comma as it is the case~
746 # of nodes$group
747 ~~~~~
748 visoptions(highlightNearest=list(enabled=T, degree=10, hover=T, algorithm="hierarchical"),~
749 ~~~~~
750 ~~~~~
751 manipulation=TRUE,~
752 ~~~~~
753 # First combo box "Select by id"~
754 # Defined by nodesidselection can be "values: Optional. Vector of possible values (node's id),~
755 # and so order is preserved. Default to all id in nodes data.frame.~
756 ~~~~~
757 nodesidselection=list(enabled=T, main="Select by node"),~
758 ~~~~~
759 # Network will automatically detect when its container is resized, and redraw itself accordingly~
760 ~~~~~
761 autoresize=TRUE,~
762 ~~~~~
763 # Second combo box "Select by group" (nodes$group)~
764 # Defined by selectedBy, uses multiple "categories" separated by a comma in nodes$group~
765 ~~~~~
766 selectedBy=list(variable="group", multiple=TRUE)) %>%~
767 ~~~~~
768 # ~6.5 Layout Options~~~~~
769 # When not using hierarchical layout, providing random seed manually~
770 # will produce the same layout every time.~
771 ~~~~~
772 vislayout(randomSeed=11, improvedLayout=T, hierarchical=F) %>%~
773 ~~~~~
```

The next thing was to create a script to produce a dynamic and interactive network. From several available options, we used the visNetwork package, an R interface to “vis.js” JavaScript library.

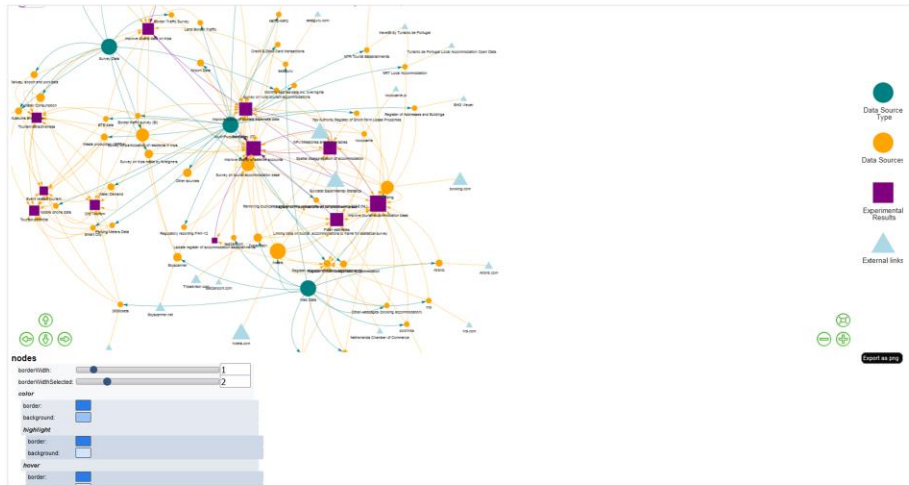
Although it is not mandatory to know JavaScript to use this package, some basic knowledge is helpful for things such as action related events.

The visNetwork package is very flexible and accessible, not only because it is based on open-source software, but also because it works on any modern browser for up to a few thousand nodes and edges.

It is based on html widgets, so it is compatible with Shiny, R Markdown documents, and RStudio viewer.

The script, which can be downloaded from that github account is:

- **“Self-contained”** in the sense data is embedded. No need to load or import data. Running the script will produce the dataframes. We can do this by using the dput {base} R command to recreate a dataframe.
- **Commented**, useful to understand what the script is doing if you want to change parameters. This is important for us because we really want to encourage everyone to try this with other data and so we made an effort to make this as easy as possible
- **Organized in an outline layout** (useful to navigate the code), and To Show Document Outline you can either use Ctrl+Shift+O or just click that button on the top right corner of your Rstudio window



Now, I'd like to share with you this very useful command (`visNetworkEditor`) that lets you configure and view options directly and immediately on your network. And without writing a single line of code or going through that tedious process of changing the parameters on the script, run the all thing and then finally observe the result in the Viewer Window.

LIVE DEMO:USE THE PHYSICS SLIDERS -> MAX VALUES

You just change some parameters (check boxes and sliders) and you immediately see the result. Once you're satisfied, just use it on your final script.

- **{dplyr}**
 - Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7 <https://CRAN.R-project.org/package=dplyr>
- **{visNetwork}**
 - Almende B.V. and Contributors, Benoit Thieurmél and Titouan Robert (2021). visNetwork: Network Visualization using 'vis.js' Library. Rpackage version 2.1.0. <https://CRAN.R-project.org/package=visNetwork>
- **{rstudioapi}**
 - Kevin Ushey, JJ Allaire, Hadley Wickham and Gary Ritchie (2020). rstudioapi: Safely Access the RStudio API. R package version 0.13. <https://CRAN.R-project.org/package=rstudioapi>
- **{shiny}**
 - Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges (2021). shiny: Web Application Framework for R. R package version 1.7.1. <https://CRAN.R-project.org/package=shiny>

And these are the citations for the packages mentioned in this presentation.

dplyr: always end up using it

Visnetwork: it's what makes the heavy work

Rstudioapi: needed for the set_wd function

Shiny: if you want to run the visNetworkEditor command



Scientific Session: BIG DATA

Thank you for your attention

<https://github.com/ruialv/VizNet-uRos2021>

 INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

 Statistics Netherlands

 Hellenic Statistical Authority

And... that's it for me. Thank you for your attention