



Scientific Session: BIG DATA

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

Rui ALVES



Shirley ORTEGA-AZURDUY



Christina PIERRAKOU





Scientific Session: BIG DATA

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

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

The context

- [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.

The context



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.

The Problem

- 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

Why do we need a net work tool?

Why do we need a network tool?

Official statistics rely strongly on monitoring properly and timely changes on economic activities. Economic activities can often cross territorial borders and also cross 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. It is known that the majority of tourist accommodations is booked online, that holds also for large leisure events.

Hence, getting a rapid overview and intelligence of the data sources types, data sources used and outputs of other NSI's is 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 externa sources and get news insights.

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

The Problem

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. However, it was not easy to get the full picture. You can see that the heterogeneity is big across country workflows.

The Solution: Interactive network visualisation

Mention-worthy when it serves the demonstration:

From a user point-of-view, this interactive network visualisation has two main selection criteria: a single node and a group (topic).

As you can observe:

- Important nodes are centered: Green-dots are used for the data source types and Orange-dots represent the data sources.
- Bigger nodes relate to more countries.
- On the left, the user has green buttons for navigation , zoom-in and -out, re-center of the visualization.
- One can also refresh the page in the browser which re-builds the network
- The purple squares stand for the outputs.
- Moreover, hovering the mouse over a node shows additional information
- One has also the option to visit “Live” an “external link” (url) depicted as triangle-shaped nodes. These contain external links that can be html-, PDF-, image- or Excel-files, etc...
- One can also execute a multi-selection using “ctrl+click”
- Last, there is an option the visualization using the black-button “Export as PNG”

What's the importance of web scraped data?

Let say that one want to select a group or topic. Let us choose “web scraped data”.
(Notice: the bigger a node, the more countries involved using these data. Moreover look that the colour of the lines depict the predominant connection to a node.)

1. We can see “which sources” were scraped, namely hotels.com, booking and airbnb.
2. If you wonder “How many countries use booking?”, just hover on the node. There are 7 countries collecting hotels.com data.
(To what results?)
3. You may guess the purpose of use of these data in terms of official statistics, i.e. improving the accommodation inventory but also to learn more about visitors

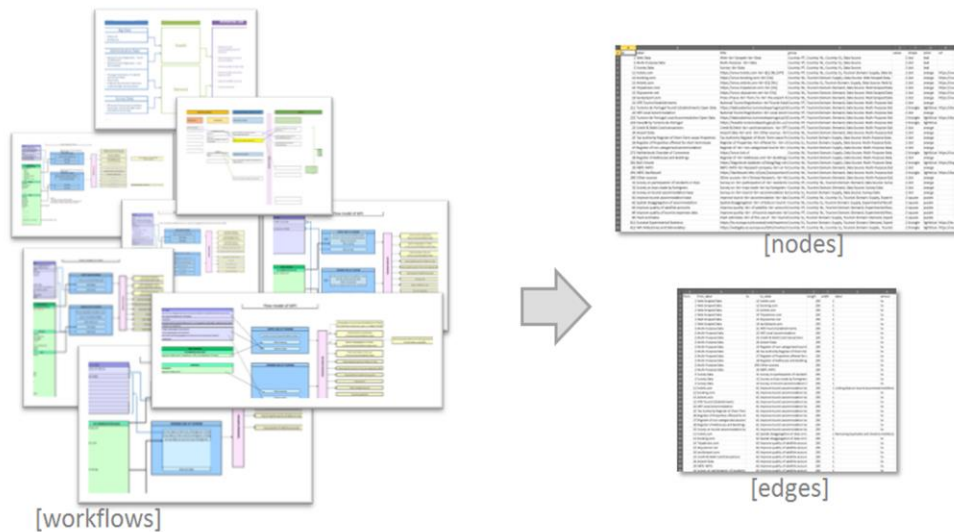
How many data sources are used to improve quality of tourist expenses data?

Let us now select an output node: Improve the quality of tourist expenses data. (Notice again: the bigger a node, the more countries involved using these data. All countries have a survey on visitor expenditures and a survey on participation of residents in trips.)

Using this 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

How does it work?



Without wanting to make 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, is to “translate the static” workflow into 2 tables (or dataframes): nodes and edges. These are the core of a network

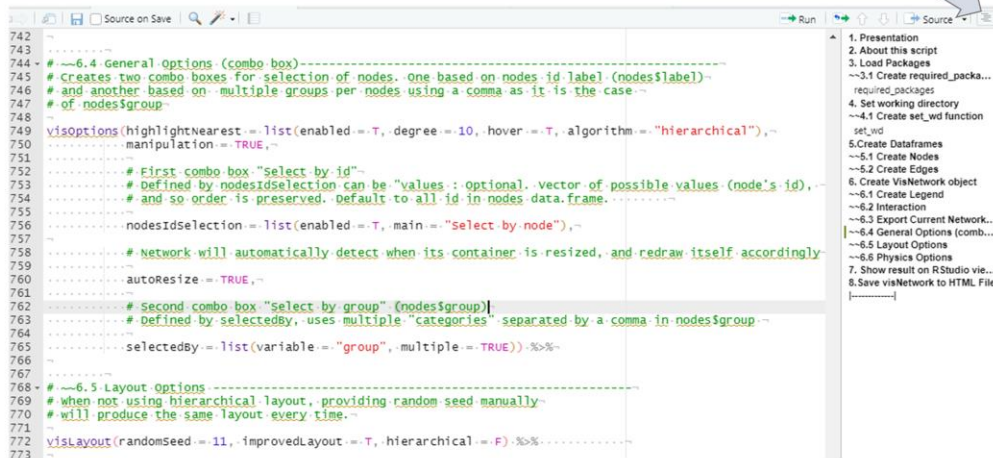
Nodes table has numeric ID of the nodes and their properties: size, shape, colour. It also has label, title (provide the tooltip info when hovering over a node) and group that have the data to use in the combo boxes. In this case, size indicates the number of countries that use that particular data source, for example.

Edges table has information on the connections (“from” and “to”) 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 ~
744 # ---6.4 General Options (combo box)-----
745 # Creates two combo-boxes for selection of nodes. One based on nodes id label (nodes$label)~
746 # and another based on multiple groups per nodes using a comma as it is the case~
747 # of nodes$group~
748 visoptions(highlightNearest = list(enabled = T, degree = 10, hover = T, algorithm = "hierarchical"),~
749 ~
750 ~
751 ~
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 ~
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 ~
769 # ---6.5 Layout Options-----
770 # when not using hierarchical layout, providing random seed manually~
771 # will produce the same layout every time~
772 ~
773 vislayout(randomSeed = 11, improvedLayout = T, hierarchical = F) %>%~
774 ~
```

1. Presentation
2. About this script
3. Load Packages
~~3.1 Create required_packages
required_packages
4. Set working directory
~~4.1 Create set_wd function
set_wd
5. Create Dataframes
~~5.1 Create Nodes
~~5.2 Create Edges
6. Create VisNetwork object
~~6.1 Create Legend
~~6.2 Interaction
~~6.3 Export Current Network...
---6.4 General Options (comb...
~~6.5 Layout Options
~~6.6 Physics Options
7. Show result on RStudio vie...
8. Save visNetwork to HTML File

The next thing was to create a script to produce a dynamic and interactive network. We used the visNetwork package visNetwork package3, 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 some functionality such as action related events.

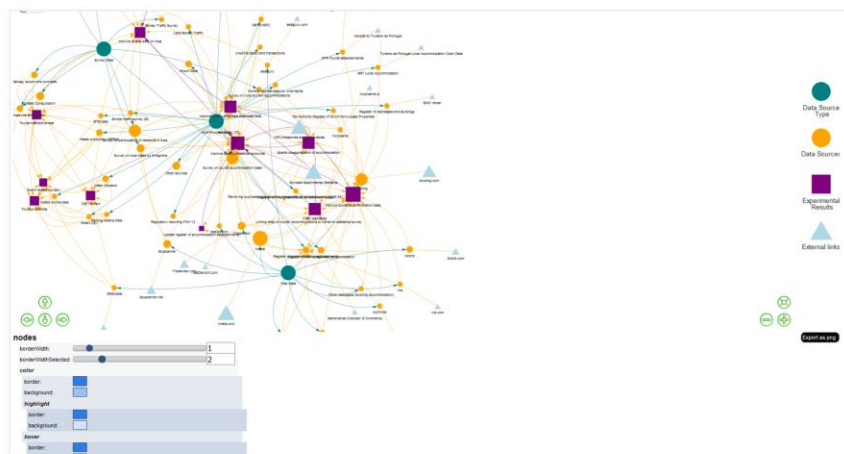
The visNetwork is very flexible and accessible not only because it is based on open-source software, but also because it works on any modern browser for up to a few thousand nodes and edges.

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. We can do this with dput {base} R command to recreate a dataframe
- Commented, useful to understand what the script is doing and changing parameters
- Organized in an outline layout (useful to navigate the code), Show Document Outline: Ctrl+Shift+O

How does it work? `visNetworkEditor{visNetwork}`



Now, I'd like to share with you this very useful command (`visNetworkEditor`) that needs shiny package and lets you configure and view options directly and immediately on your network .

And without writing a single line of code.

LIVE DEMO:

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

R Packages

- **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 use it

Visnetwork: it's what makes the heavy work

Rstudioapi: needed for the set_wd funtion

Shiny: to run 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