Chapter 8 / Capítulo 8

Applied bibliometrics. From data to publication (English Edition)
ISBN: 978-9915-9680-7-0
DOI: 10.62486/978-9915-9680-7-0.ch08

©2025 The authors. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Sankey Diagrams: Flows And Relationships Between Concepts / Mapas De Sankey: Flujos Y Relaciones Entre Conceptos

Sankey maps are a distinctive visual tool that captures flows of varying intensity between different categories or states in a bibliometric system. Unlike traditional graphs that show adjacency or connectivity relationships, Sankey diagrams visualize quantified transfers of resources, research attention, or intellectual influence using arrows whose width is proportional to the magnitude of the flow. This representation is compelling for analyzing temporal trajectories, interdisciplinary transfers, and redistributions of research focus over time.

8.1. Interpretation

Proper interpretation of these maps requires understanding that the width of the connections encodes specific quantitative values. At the same time, the spatial arrangement reveals the structure of the system's relationships.

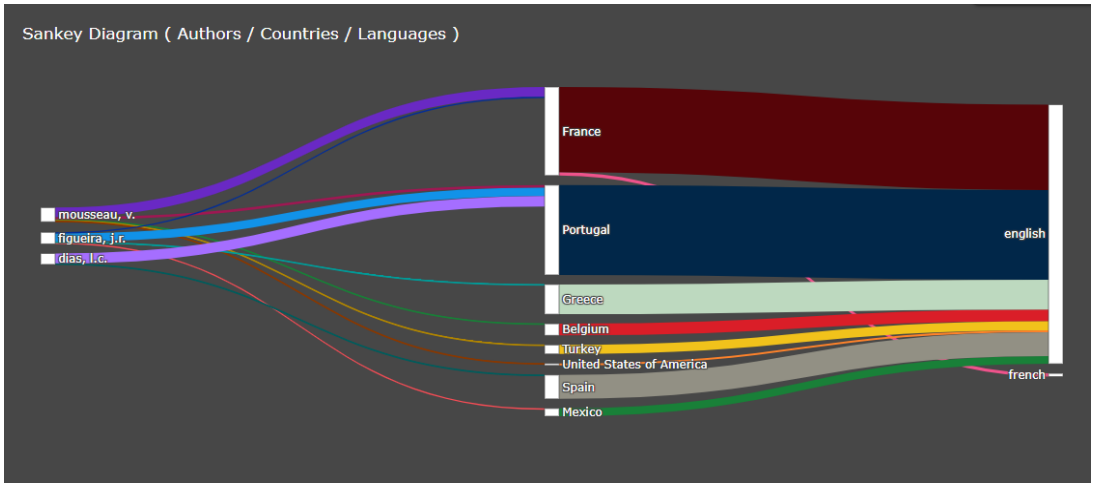

Figure 8.1. Example of a Sankey diagram

Reading a bibliometric Sankey diagram involves breaking it down into three main structural components: the nodes representing discrete categories (disciplinary fields, institutions, countries), the flows connecting these nodes indicating transfers or relationships, and the direction indicating the temporal or causal orientation of these transfers. The interpretation begins by identifying the dominant flows, those wider connections that represent the most intense relationships, to understand the main channels of intellectual exchange or institutional collaboration. Subsequently, the analysis shifts to secondary flows, which, although less intense, may reveal emerging connections or specialized niches with potential for future development. Finally, the identification of isolated nodes or those with few connections to areas disconnected from mainstream research, or to opportunities to establish new collaborations.

The temporal dimension in Sankey maps adds analytical depth by allowing the tracking of diachronic changes in flow patterns. When the diagram represents successive periods, changes in the thickness of the connections between different temporal segments reveal dynamics of consolidation, diversification, or reorientation of lines of research.

The progressive strengthening of specific flows indicates the emergence of new specialties or the intensification of established collaborations. In contrast, the thinning or disappearance of

others suggests the decline of paradigms or the breakdown of research alliances. This ability to visualize trajectories makes Sankey maps an exceptional tool for analyzing scientific mobility, the evolution of research interests, and the reconfiguration of disciplinary boundaries.

Contextualized interpretation transcends immediate quantitative analysis by integrating explanatory factors that give meaning to the observed patterns. An intense flow between two disciplinary fields may reflect both fruitful integration of conceptual frameworks and mere instrumental appropriation of methodologies, without proper intellectual synthesis. The concentration of flows around specific nodes may indicate genuine intellectual leadership or scale effects derived from the size of research communities.

The expert analyst distinguishes among these possibilities by triangulating with qualitative evidence and specialized domain knowledge, transforming the Sankey diagram from a descriptive representation into a diagnostic tool for understanding the dynamics of change in the scientific landscape.

The specific bibliometric applications of Sankey maps range from the analysis of academic mobility, tracing the trajectories of researchers between institutions or countries, to the study of thematic evolution, showing how concepts migrate between different disciplinary fields. In scientific policy evaluation, they enable visualization of funding flows among agencies, thematic areas, and research teams. In interdisciplinary studies, they reveal citation patterns between distant fields or the integration of diverse methodologies.

The versatility of this representation makes it a particularly valuable tool for communicating complex findings to non-specialist audiences, translating abstract research dynamics into intuitive visual narratives about the flows that shape contemporary knowledge.

8.2. Types of Sankey diagrams

Sankey maps of thematic evolution visualize how research concepts and topics migrate across disciplinary fields or transform over time. These diagrams show the emergence, consolidation, and decline of lines of research, revealing processes of specialization, paradigm fusion, or the emergence of interdisciplinary areas. Their construction is based on the diachronic analysis of terms extracted from titles, abstracts, and keywords, segmented into successive time periods, during which the most significant terminological transitions are identified using measures of semantic similarity and evolutionary co-occurrence analysis.

Citation flow diagrams depict the transfer of intellectual influence across fields, institutions, or research groups. Unlike conventional citation networks that show specific connections, citation Sankey diagrams capture net flows of influence, showing how certain areas act as net exporters of knowledge while others function primarily as importers.

This approach is particularly valuable for studying interdisciplinary relationships, where citation flows can reveal asymmetries in intellectual exchange between established and emerging fields, or between different epistemological traditions.

Scientific collaboration Sankey maps visualize patterns of cooperation between institutions, countries, or disciplines, showing not only the existence of collaborations but also their intensity and evolution over time. These diagrams enable the identification of stable research consortia, emerging patterns of South-South or North-South cooperation, and the dynamics of collaborative network formation and dissolution.

They are typically constructed using co-authorship data segmented by period, where the flows represent the volume of joint publications between different entities, allowing analysis of how institutional collaboration strategies evolve in response to funding programs, scientific policies, or technological developments.

Funding flow diagrams depict the distribution of economic resources across different subject areas, institutions, or research types, showing how funds are allocated, transferred, and concentrated within the scientific ecosystem. These maps are handy for evaluating the impact of funding policies, identifying mismatches between stated priorities and actual allocation patterns, and analyzing the efficiency of research resource distribution. Their construction requires integrating data from calls for proposals, awarded projects, and scientific outputs, ensuring complete traceability from funding sources to the research products generated.

Sankey diagrams of scientific production by sector visualize the distribution of research outputs among different types of organizations (universities, research centers, companies, hospitals) and their evolution over time.

These diagrams capture trends in sectoral specialization in research, in knowledge transfer from academia to industry, and in the relative roles of different actors in the innovation system. Their implementation requires consistently classifying institutional affiliations by sector and analyzing their evolution over time, revealing patterns of diversification or specialization in the scientific production of different types of organizations.

Each type of bibliometric Sankey diagram answers specific research questions and requires particular strategies for data preparation, processing, and validation. The analytical objectives should guide the selection of the appropriate type, the availability of data with the necessary granularity and temporal coverage, and the study's communication needs. Regardless of the specific type, they all share the ability to transform complex bibliometric relationship data into intuitive visual narratives about the flows that structure the dynamics of contemporary scientific knowledge.

8.3. Construction

8.3.1 Implementation in R Bibliometrix

Sankey maps in Bibliometrix are constructed using the *sankeyPlot()* function, which transforms transition matrices between bibliometric categories into interactive flowcharts. The process begins with preparing a transition matrix, where the rows represent the source states, and the columns represent the destination states, with cell values indicating the frequency or intensity of the flow. For temporal thematic analysis, this matrix captures the migration of concepts across consecutive periods, while in collaboration studies, it can represent institutional mobility or changes in co-authorship patterns.

Technical implementation first requires processing bibliographic data using *biblioAnalysis()* to obtain the basic structure, then using *coupling()* or *cocMatrix()* to generate specific association matrices. The *sankeyPlot()* function accepts customization parameters such as *node. Width* to adjust the width of the nodes, *node.pad* for the spacing between elements, and *units* to specify the metric represented in the flows. Bibliometrix integrates the *networkD3* library to render interactive visualizations that allow users to drag nodes, *hover* over flows to see exact values, and filter connections by intensity thresholds.

For advanced diachronic analysis, Bibliometrix allows you to segment data into time periods

using *timeslice* and generate successive transition matrices that show the evolution of flows. The *Evolution ()* theme function complements this analysis by identifying specific thematic trajectories, which can then be visualized as specialized Sankey maps. The export of results includes HTML formats to preserve interactivity, as well as SVG vector images for academic publications, maintaining visual clarity at different scales and resolutions.

The code to generate a Sankey map is:

1. Install and load packages

```
install.packages("bibliometrix")
install.packages("plotly")
install.packages("networkD3")
library(bibliometrix)
library(plotly)
library(networkD3)
library(dplyr)
```

2. Load data

```
data(management, package = "bibliometrix")
M <- management
```

3. Create data for a Sankey chart showing evolution over time

```
years <- c(2015, 2020)  # Range of years
sankey_data <- thematicEvolution(M,
                        years = years,
                        n = 10,  # Number of terms per year
                        field = «ID")  # "ID" for Keywords Plus, "DE" for Author Keywords
```

4. Generate Sankey plot

```
sankey_plot <- plotThematicEvolution(sankey_data$Nodes, sankey_data$Edges)
sankey_plot
```

8.3.2. Implementation in PyBibX

PyBibX constructs Sankey maps using the SankeyDiagram class, which provides granular control over all aspects of the visualization. The basic implementation requires creating a diagram instance by specifying the flow data as a list of tuples or a nested dictionary, where each entry defines an individual flow with its source, destination, and magnitude. The *generate()* method processes this data and applies automatic positioning algorithms that minimize connection crossings and optimize the diagram's overall readability.

Advanced customization in PyBibX includes aesthetic adjustments using parameters such as *color_palette* to define color schemes consistent with the study domain, *node_alignment* to control the vertical arrangement of nodes, and *flow_opacity* to manage the visual overlap of multiple flows. Native integration with Plotly allows you to create interactive visualizations where users can explore specific flows, dynamically rearrange nodes, and apply real-time filters based on different magnitude or category criteria.

For specialized bibliometric analysis, PyBibX includes methods such as *author_mobility_sankey()*, which transforms institutional-affiliation data into academic mobility maps, and *concept_evolution_sankey()*, which visualizes terminological migration across research periods. The tool also offers preprocessing functions, such as *normalize_flows()*, to adjust magnitudes

based on different impact metrics, and *cluster_flows(),* to group minor connections and simplify complex visualizations without losing substantive information.

Recap

- Sankey diagrams show directed flows between categories, with the width of each band proportional to the flow magnitude.
- They allow for the visualization of thematic, citation, or funding transfers.
- Each link requires source-destination data and a magnitude.
- The directionality of the flow must be evident through arrows or layout.
- Normalization helps avoid visual saturation and improve comparability.
- Logarithmic scales help when flows are heterogeneous.
- Cleaning and unifying categories are essential before analysis.
- Hierarchical Sankeys show subcategories within larger flows.
- The order of nodes affects readability; minimize crossings.
- Colors should distinguish groups and have an explanatory legend.
- Sankeys should include verifiable units and sums.
- Popular software: R (ggalluvial), Python (plotly), D3.js.
- Interactivity improves the exploration of complex flows.
- In time series, panels by period or animations are used.
- Checking for outliers and aggregation errors increases validity.
- Grouping minor categories under “others” facilitates visual clarity.
- They complement graphs and heat maps by showing dynamic routes.
- Documenting aggregation criteria ensures reproducibility.
- Numerical tables should accompany Sankey diagrams.
- Interpreting flows in their disciplinary context prevents erroneous conclusions.

Self-assessment questions

1. What does the width of a band in a Sankey diagram represent?
2. What minimum data does a Sankey diagram require?
3. How would you show temporal evolution using a Sankey diagram?
4. Why are categories normalized before constructing them?
5. What advantages does it have over a static graph?
6. When would you use a logarithmic scale?
7. What effects does an incorrect node order produce?
8. How can visual overload be reduced in complex Sankey diagrams?
9. What tools can be used?
10. Why is it vital to document aggregation criteria?

BIBLIOGRAPHY

1. Tufte ER. The Visual Display of Quantitative Information. 2nd ed. Cheshire (CT): Graphics Press; 2001. ISBN: 9780961392147.

2. Healy K. Data Visualization: A Practical Introduction. Princeton (NJ): Princeton University Press; 2018. ISBN: 9780691181622.

3. Murray S. Interactive Data Visualization for the Web: An Introduction to Designing with D3. 2nd ed. Sebastopol (CA): O’Reilly Media; 2017. ISBN: 9781491921289.

4. Knaflic CN. Storytelling with Data: A Data Visualization Guide for Business Professionals. Hoboken (NJ): Wiley; 2015. DOI: 10.1002/9781119055259.