Herr II, Bruce W., Russell Jackson Duhon, Katy Börner, Elisha F. Hardy, Shashikant Penumarthy. 2008. 113 Years of Physical Review: Using Flow Maps to Show Temporal and Topical Citation Patters. Proceedings of the 12th Information Visualization Conference (IV2008), London, UK, July 9-11, 421-426. IEEE Computer Society Conference Publishing Services.

113 Years of Physical Review: Using Flow Maps to Show Temporal and Topical Citation Patterns

Bruce W. Herr II, Russell J. Duhon, Katy Börner, Elisha F. Hardy, Shashikant Penumarthy
Cyberinfrastructure for Network Science Center
School of Library and Information Science, Indiana University Bloomington, IN, USA
{bherr, rduhon, katy, efhardy, sprao}@indiana.edu

Abstract

We visualize 113 years of bibliographic data from the American Physical Society. The 389,899 documents are laid out in a two dimensional time-topic reference system. The citations from 2005 papers are overlaid as flow maps from each topic to the papers referenced by papers in the topic making intercitation patterns between topic areas visible. Paper locations of Nobel Prize predictions and winners are marked. Finally, though not possible to reproduce here, the visualization was rendered to, and is best viewed on, a 24" x 30" canvas at 300 dots per inch (DPI).

Keywords---network analysis, domain visualization, physical review

1. Introduction

This is the very first map of a 113-year scholarly dataset that captures the structure and evolution of an entire field: Physics. The visualization aggregates 389,899 papers published in 720 volumes of 11 journals between 1893 and 2005. Time runs horizontally. PACS codes run vertically and are labeled from '00-General' to '09-Geophysics, Astronomy and Astrophysics' on the right.

The 91,762 papers published from 1893 to 1976 take up the left third of the map. In 1977, the Physical Review introduced the Physics and Astronomy Classification Scheme (PACS) codes, and the visualization subdivides vertically into the top-level PACS codes. The 217,503 papers from 1977 to 2000, for which references but no citation data is available, occupy the middle third on the map. The 80,634 papers from 2001 to 2005, for which citation data is available, fill the right third of the map.

Each 'bar graph' is subdivided vertically into the journals that appear in it with height proportional to the number of papers. The area for each journal is subdivided horizontally into the volumes of the journal. Overlaid on this two-dimensional base map are all citations from every Physical Review paper published in 2005.

Small Nobel Prize medals are placed where the prize-winning paper is located on the map. Thomson Scientific predicts three Nobel Prize awardees in physics each year based on citation counts, high-impact papers, and discoveries or themes worthy of special recognition. Correct predictions by Thomson Scientific are indicated by larger medals.

2. Data

The Physical Review dataset was provided by the American Physical Society (APS). It comprises optical character recognition (OCR) of the full text, markup SGML/XML of the full text, and bibliographic metadata in XML format for 398,005 papers.

XSLT was used to parse the many files that reported the bibliographic metadata in XML format. Among others, it provides a unique Digital Object Identifier (DOI) for each paper; a listing of its authors; date(s) when the paper was received, revised, corrected and published; code(s) from the Physics and Astronomy Classification Scheme (PACS) [1, 2] and citation references.

PACS codes have been in use since 1975. They are assigned by the author in consultation with reviewers and editors. The average number of PACS codes per paper since then is about 1.5. There is no 'main' PACS code for a paper.

3. Technique

3.1 Reference System

How to layout 389,899 papers published over 113 years? Using the temporal and topical PACS organization of papers, a simple, two-dimensional reference system was used. Time runs from left to right, the PACS codes increase from bottom to top with the lowest line of bar graphs having no PACS codes assigned. The first 83 years take up the left third of the page, the next 25 years each occupy about 1cm in the middle third, and the last 5 years with about 20% of all PhysRev publications are plotted in the right third of the map.

The next question concerned the layout of papers within each 'PACS code by year' box. The original idea was to group papers published in the same year by bibliographic coupling [3]. That is, the more references two papers share the more similar they are. Hierarchical agglomerative clustering would then be applied to compute a partition of the paper set that has high within cluster similarity and low between-cluster similarity. Clusters would then be size coded by the number of papers they contain and laid out using self organizing maps [4] or a Dorling cartogram.

However, a layout using circles was not space efficient. Also, the same journals and volumes would be positioned in different places of the 113 x 11 boxes. Hence, we decided to group and display papers using their natural 'containers': journals and journal volumes.

The space was divided into horizontal slices for each journal proportional in height to the quantity of papers from that journal containing that PACS. Those slices were then subdivided into vertical slices for each volume in that journal, proportional in width to the quantity of papers from that volume containing that PACS.

Out of the 261,190 papers with PACS codes, 110,978 contain more than one top-level PACS code. They were placed in each of them, increasing the size of the encompassing volume and contributing their citations to the set of papers cited by that PACS, see Fig. 1 and Fig. 2.

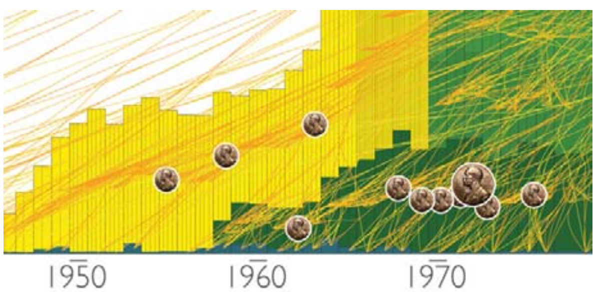

Figure 1 Reference system close-up

3.2 Data Overlays

The different journals are color coded, and they always appear in the same sequence, improving the readability of the map considerably. As mentioned previously, the height of each journal bar corresponds to the number of its papers published in this year. Journal volume subdivisions reflect the percentages of papers published in them.

Interested to show the cross-fertilization of different areas of research, we tried to find a way to show the citation linkages among the 389,899 papers. For 2005 alone, there are 958,913 citation linkages including DOIs from 42,516 papers published in 2005 to 145,665 papers within the PhysRev dataset and to 176,473 outside the set. Aggregated at the level of journal volumes, there are 17,343 tendrils from ten top-level PACS codes in 2005 to 2,691 volumes.

Metaphorically and visually, citation linkages might resemble 'blood vessels' that interconnect and feed different areas of science. The flow map layout [5] captures this metaphor algorithmically, see also section 5 on 'Related Work'. Papers were aggregated into journal volumes and volumes clustered into a tree according to their spatial arrangement. Lines are then drawn from the root (top level PACS code) to each volume, remaining merged into single lines until the bounding box for the upcoming cluster is approached, then splitting into lines proportional to the amounts that 'flow' to the respective volumes of the cluster. For this map, the flow map layout was independently applied to all 2005 citations from each top-level PACS code into other Physical Review papers. Spatially, each journal volume is represented by a constructed node in the center of the area it covers on the map.

To overlay the Nobel laureate winners whose Nobel winning papers were published in Physical Review, the winning papers were crosschecked with the Physical Review database. If the Nobel winning papers were found, then a Nobel icon was placed where the paper belongs on the base map determined by looking at its year, PACS code, and journal. If the Nobel's paper was predicted by Thomson ISI, then a larger Nobel icon was used, see Fig. 1.

With the final reference system and data overlays, the growth and decline of the number of papers in different PACS codes and journals can be seen. The number and density of the overlaid 2005 citation links shows that physics draws extensively on old and new knowledge and from different areas of physics.

4. Interpretation

The first and most obvious pattern illustrated by this map is the huge growth in physics papers in Physical Review over time. The relative amounts of publications in the various PACS codes are also evident. Somewhat more subtly, the relative frequencies of papers from different journals in each PACS over time can be seen, because within each PACS the papers are organized and aggregated by the volume and colored by the journal.

Citation patterns appear to span the whole history of the Physical Review and cross all disciplinary boundaries within physics – at least for 2005 paper citations. Very few volumes are not cited by papers published in 2005. Also, many lines converge at almost all volumes indicating that many different papers cite a paper in these volumes.

Inter-PACS cross-citation can be seen. Specifically, PACS 8 -- the 'interdisciplinary category' -- as expected cites heavily in other PACS codes. This can be seen by looking at the bright line arching downwards, as it is near

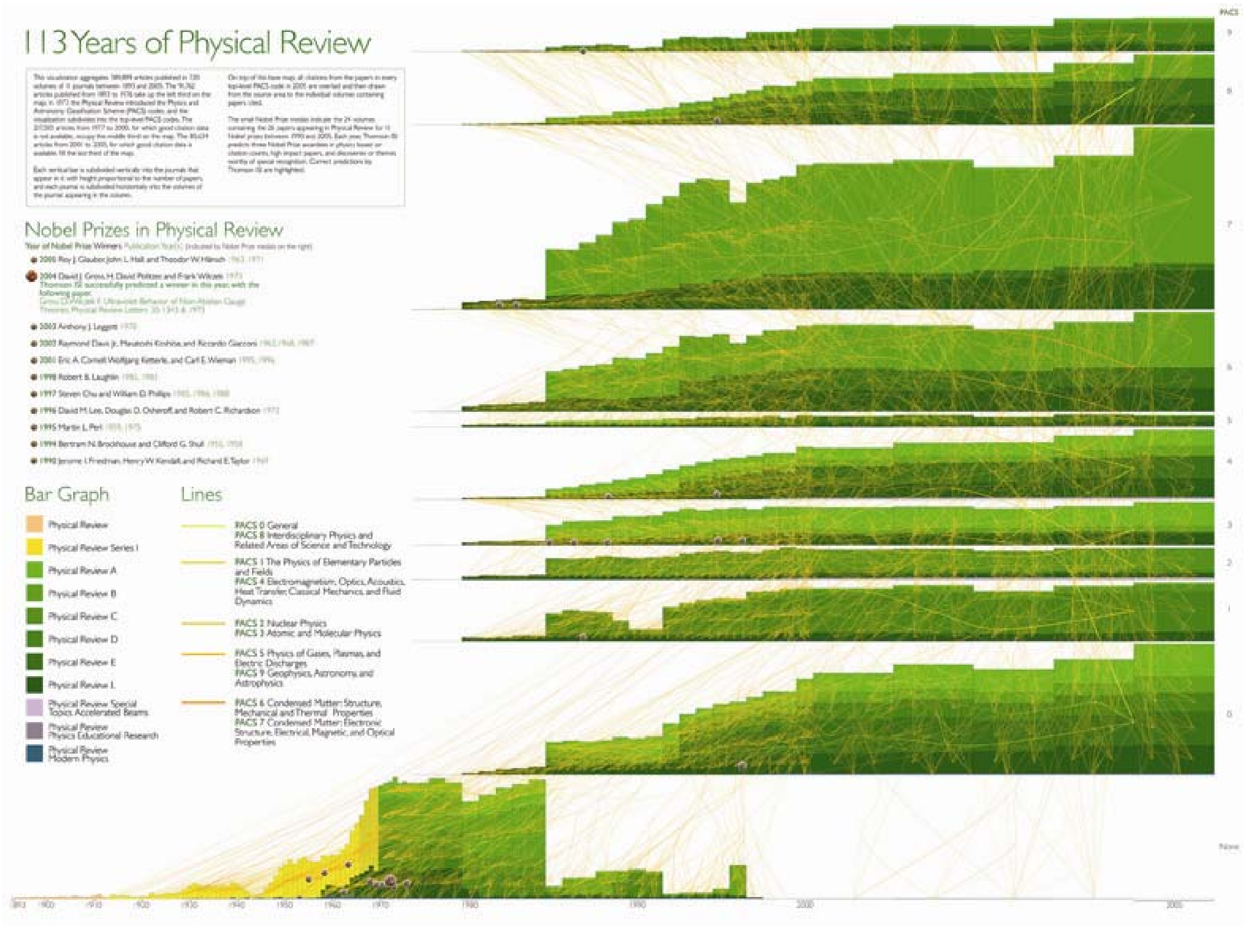

Figure 2 113 Years of Physical Review

the top of the map. Somewhat more interestingly, a large segment of citations from the 'general' PACS code 0 go to PACS 3 representing research in 'atomic and molecular physics'.

The map also highlights that most recent Nobel Prizes have been awarded for work from a period when physics was a relatively small field. Today, the Nobel Prize committee must increasingly choose between works spanning a field many times the size it was just twenty five years ago.

5. Related Work

Movement maps or flow maps help communicate the movement of tangible objects, e.g., people, bank notes, goods, in geographical space but also of intangible objects, e.g., energy, ideas, or reputation, in digital spaces [6]. In *113 Years of Physical Review* they are used to show the flow of knowledge via citation linkages.

Some of the first flow maps were drawn by hand by Harness, Belpaire, and Minard. A close-up of one of Minard's maps showing the approximate amount of cotton imports by Europe in 1864 [7] is shown in Fig. 3.

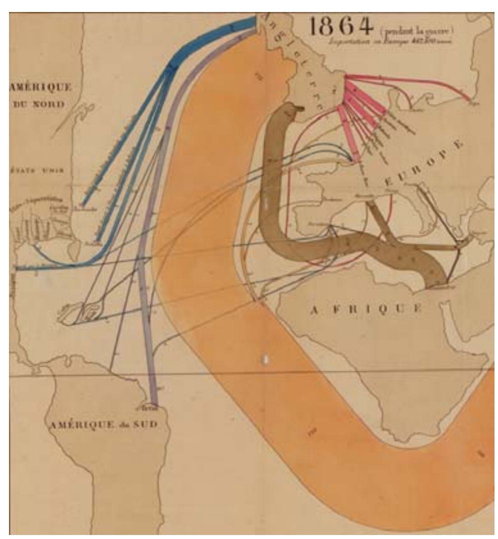

Figure 3 Minard's 1864 flow map of cotton brut imports

Flow maps are generated from tables that represent the amount of flow occurring between pairs of places, e.g., migration tables produced by the US Bureau of the Census. Typically, the largest amount of movement is between spatially close places modulated by the size of the places. The amount of traffic follows a power law, i.e., there are few pairs of places that have enormous traffic and many pairs with rather little traffic.

There are two types of flow maps: continuous and discrete. *Continuous flow maps* use vector fields or streamlines to show continuous flow patterns. As an example, the maps by Tobler given in Figures 4-7 show estimated state to state net migration in US depicted as a vector field with scalar potential, contour lines, and estimated trajectories [8].

Figure 4 Gaining and losing states based on the marginals of a 48 by 48 migration table.

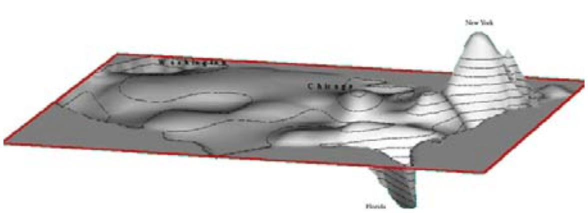

Figure 5 The pressure to move in US based on a continuous spatial gravity model.

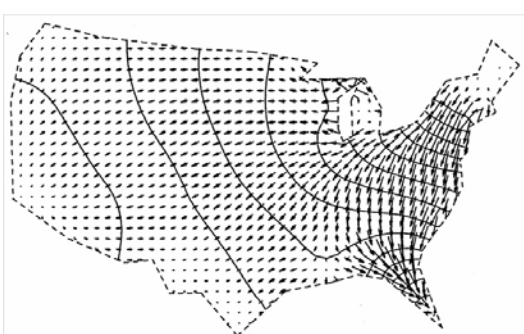

Figure 6 Migration potentials and gradients with the potentials shown as contours.

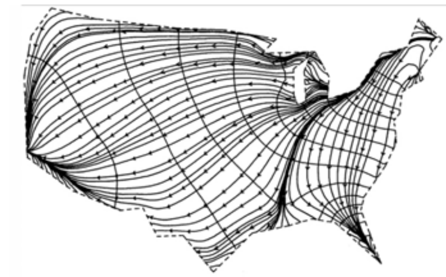

Figure 7 Migration potentials and streaklines with the gradient vectors connected to form streaklines.

Discrete flow maps use bands or arrows whose width is proportional to the volume moved. One of the first computer generated flow maps showed a 'Cartographatron display of 9,931,000 desire line traces of personal trips in Chicago' and was published in 1959 [9].

In 1987, Tobler designed the Flow Mapper program. An updated version is freely available via the *Center for Spatially Integrated Social Science* [10]. The program was used to generate the flow map of the 1995 – 2000 net migration shown in Fig. 8.

Figure 8 Net migration, 1995 – 2000

It has also been used to map journal inter-citations using data by Combs et al [11], see Fig. 9. In the latter, both citation directions are shown using half arrows.

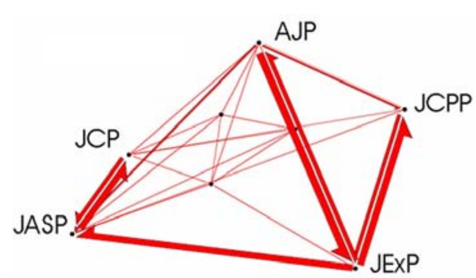

Figure 9 Journal inter-citations

The *113 Years of Physical Review* map uses a Flow Map Layout Algorithm developed and made freely available by Phan and colleagues at Stanford University.

The algorithm uses hierarchical clustering to generate a flow map tree for a particular root given a given a set of nodes, positions, and flow data between the nodes [5]. During layout, the flow map tree is used to merge edges that run in the same direction, thus minimizing edge crossings and improving readability.

A flow map generated with this algorithm of migration from California to all other US states for 1995-2000 is shown in Fig. 10.

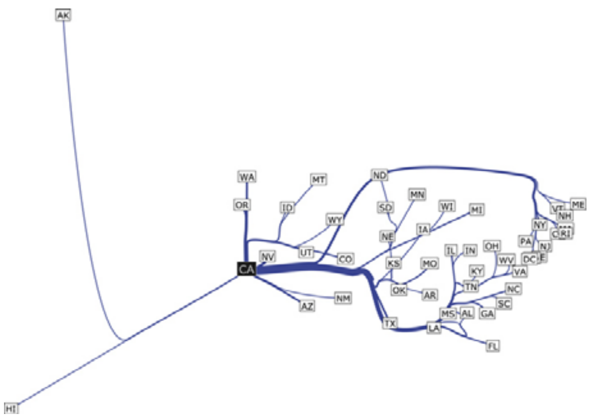

Figure 10 California migrations flow map

Output from the algorithm is easily combined to show multiple flows in the same space. Shown here are the top-10 states that migrate to California and to New York.

The spatial pattern reveals that New York tends to attract people from the East Coast, while California residents come from diverse geographic regions in the United States.

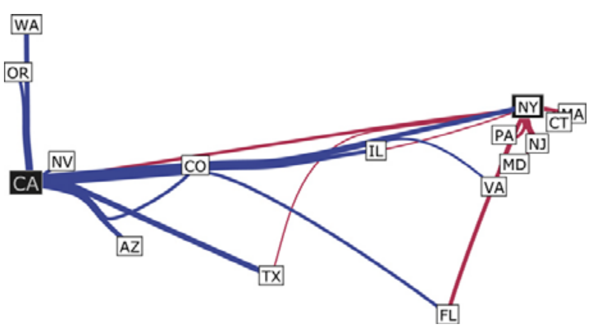

Figure 11 Immigration to California and New York

6. Conclusions

113 Years of Physical Review makes clear the difficulties of depicting large numbers of documents and citation-based information flows. A two-dimensional reference system is used and a flow map layout applied to reduce the quantity of ink and numbers of edge crossings and improve readability.

The map provides several basic insights about publication and citation patterns in Physical Review. These insights underscore how interdisciplinary physics remains, even as it becomes impossible for any one person to be an expert in more than a small area of it.

7. Further Work

Flow maps are a significant improvement over drawing edges directly between related points with data on this scale. However, they do not make it easy to note anything beyond the broadest patterns. An improved flow map layout algorithm might improve the situation significantly, as might additional annotation the data flow map segments aggregate.

One particular improvement would be a multi-sourced flow map sensitive to the presence of other sources when laying out edges, reducing the ‘tangle’ factor. An advanced version might even allow edges from separate sources to merge before reaching their destinations.

8. Acknowledgements

Soma Sanyal retrieved and processed the Nobel Laureate predictions from ISI, W. Bradford Paley and Daniel Zeller helped conceptualize the layout and design, Jan Witkowski, CSHL provided inspiration and feedback, Waldo A. Tobler and Doantam Phan provided sample flow maps and advise, Gavin La Rowe was involved in parsing the data, and the Physical Review data set was provided by the American Physical Society. This material is based upon work supported by the National Science Foundation under grants SBE-0738111, IIS-0715303, and IIS-0534909. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] American Physical Society. 2008. The Physics and Astronomy Classification Scheme (PACS). http://publish.aps.org/PACS/ (accessed 4/15/2008).

[2] Redner, S. 2005. Citations Statistics from 110 Years of Physical Review. *Physics Today*, vol. 58, 49.

[3] Kessler, M. M. 1963 Bibliographic coupling between scientific papers. *American Documentation*, vol. 14, no. 1: 10-25.

[4] Kohonen, T., 1995. Self-Organizing Maps. *Springer Series in Information Sciences*, vol. 30.

[5] Phan, D, L. Xiao, R. Yeh, P. Hanrahan, and T. Winograd. 2005. Flow Map Layout. *Proceedings of the 2005 IEEE Symposium on Information Visualization*, vol. 29: 219-224.

[6] Börner, Katy, Shashikant Penumarthy, Mark Meiss, and Weimao Ke. 2006. Mapping the Diffusion of Information Among Major U.S. Research Institutions. In *Scientometrics, Dedicated issue on the 10th International Conference of the International Society for*

Scientometrics and Informetrics, vol. 68, no. 3, 415-426. Stockholm, Sweden.

[7] Robinson, Arthur H. 1955. The 1837 Maps of Henry Drury Harness. *The Geographical Journal*, vol. 121, no. 4, 440-450.

[8] Tobler, W. A. 1981. Model of Geographical Movement. *Geographical Analysis*, vol. 13, no. 1, 1-20.

[9] Final Report. *Chicago Area Transportation Study*, 1959; fig. 23, p. 46.

[10] Tobler, W. A. 1987. An Experiment in Migration Mapping by Computer. *The American Cartographer*, vol. 14, no. 2, 155-163.

[11] Coombs, C., J. Dawes, A. Tversky. 1970 *Mathematical Psychology*. Engelwood Cliffs, NY: Prentice Hall, 73-75.