Time aware visualization of document collections

Aretha Barbosa Alencar

barbosaa@indiana.edu

Abstract. Scientific articles are the major mechanism for researchers to report their scientific results. Analyzing this scientific production is possible to detect interesting topical events along time, like the emergence of a new topic. In this project, we intend to prove that content-based document visualizations may be adapted to generate visualizations that convey topical events along time. In these content-based document maps, each document is identified as a visual marker in the two-dimensional space, and (ideally) documents close in the map share similar content, whereas distant documents are uncorrelated in terms of their content. The maps allow visually identifying groups of highly related documents (addressing similar topics) and frontiers between groups. For this we investigate: (i) how to adapt a content based projection technique to construct gradual maps of a time-stamped collection, while preserving user context at successive time-stamps, i.e., the maps should not change drastically from one time stamp to the next; (ii) how to identify topics on groups of similar documents based on the sequence of document maps generated and analyze the similarities of the extracted topics along time; (iii) how to meaningfully integrate the exhibition of topics in the maps, while visually emphasizing important topical changes along time.

Content-based Document Maps

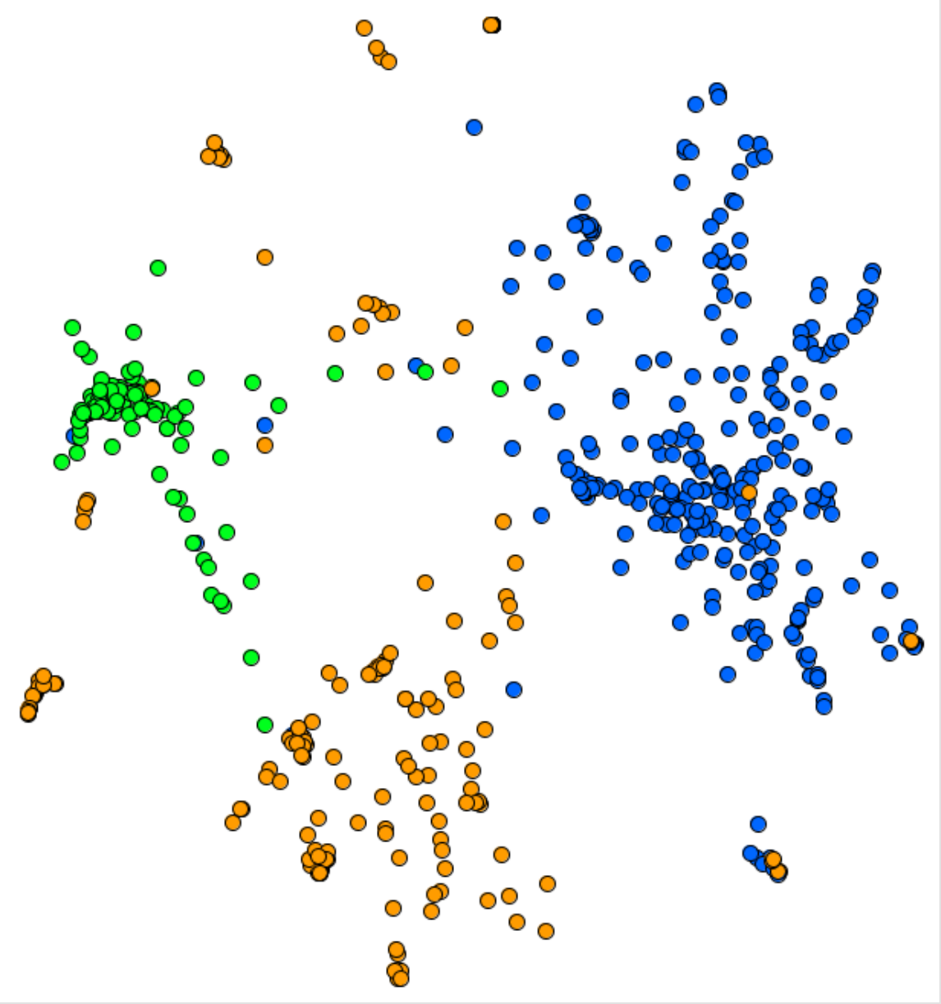
Content-based document maps are created using **multidimensional projection techniques** to generate graphical representations of documents collections based on their content similarity

Multidimensional Projection Techniques:

Data is projected form a *m*-dimensional space into a *d*-dimensional space with *d* << *m*, while preserving, on the projected space, the maximum of information possible about similarity relationships among the data-instances in the original high dimensional space.

Example

Document collection: 574 articles of three research areas: Cased-based Reasoning; Inductive Logic Programming; and Information Retrieval



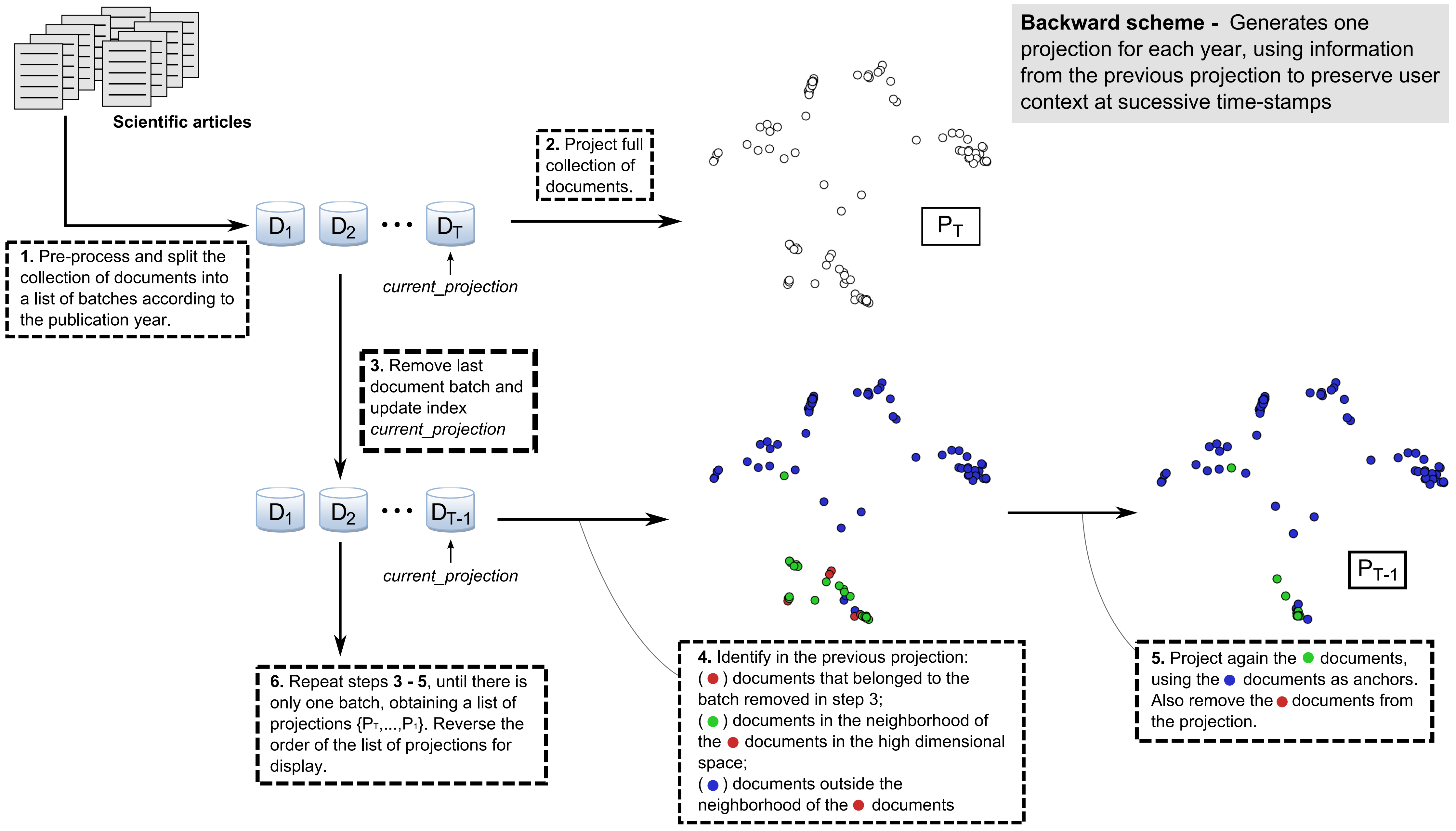
- Each circle represents a document in the two-dimensional space, and (ideally) documents close in the map share similar content, whereas distant documents are uncorrelated in terms of their content
- Color indicates the research area of the article:

🟩 Inductive Logic Programming

🟧 Information Retrieval

🟦 Case-based Reasoning

Dynamic Content-based Document Map



Backward scheme - Generates one projection for each year, using information from the previous projection to preserve user context at sucessive time-stamps

1. Pre-process and split the collection of documents into a list of batches according to the publication year.

2. Project full collection of documents.

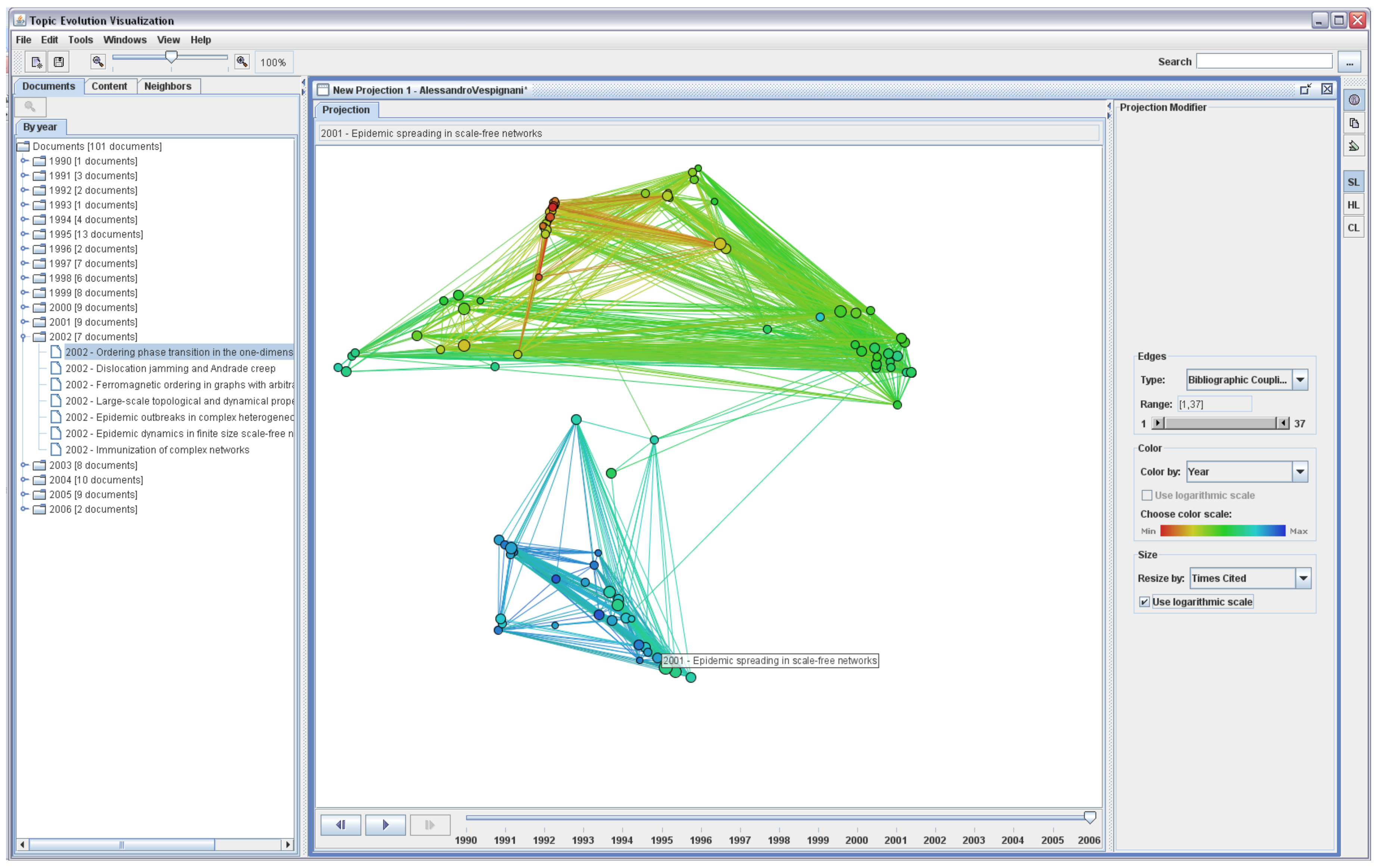
3. Remove last document batch and update index *current_projection*

4. Identify in the previous projection:
(●) documents that belonged to the batch removed in step 3;
(●) documents in the neighborhood of the ● documents in the high dimensional space;
(●) documents outside the neighborhood of the ● documents

5. Project again the ● documents, using the ● documents as anchors. Also remove the ● documents from the projection.

6. Repeat steps **3 - 5**, until there is only one batch, obtaining a list of projections {P_T,...,P_1}. Reverse the order of the list of projections for display.

Topic Evolution Visualization Tool



Documents [101 documents]
1990 [1 documents]
1991 [3 documents]
1992 [2 documents]
1993 [1 documents]
1994 [4 documents]
1995 [13 documents]
1996 [2 documents]
1997 [7 documents]
1998 [6 documents]
1999 [8 documents]
2000 [9 documents]
2001 [9 documents]
2002 [7 documents]
2002 - Ordering phase transition in the one-dimens
2002 - Dislocation jamming and Andrade creep
2002 - Ferromagnetic ordering in graphs with arbitra
2002 - Large-scale topological and dynamical prope
2002 - Epidemic outbreaks in complex heterogenec
2002 - Epidemic dynamics in finite size scale-free n
2002 - Immunization of complex networks
2003 [8 documents]
2004 [10 documents]
2005 [9 documents]
2006 [2 documents]

2001 - Epidemic spreading in scale-free networks

Edges
Type: Bibliographic Coupli...
Range: [1,37]
1 37

Color
Color by: Year
□ Use logarithmic scale
Choose color scale:
Min Max

Size
Resize by: Times Cited
☑ Use logarithmic scale

Currently, the tool encompasses techniques to create and interact with dynamic content-based document map.

A wizard guides users in creating new maps for document collections in one of the input formats:

- Endnote Export Format
- ISI
- BibTeX

Once obtained the projections in their correct order {P_1, ..., P_T}, an **animation** that interpolates between successive projection is build to **enhance the perception of changes** along the projections.

The animations uses two graphical events:

- the emergence of new documents added to the projection by using a transparency effect that decreases progressively

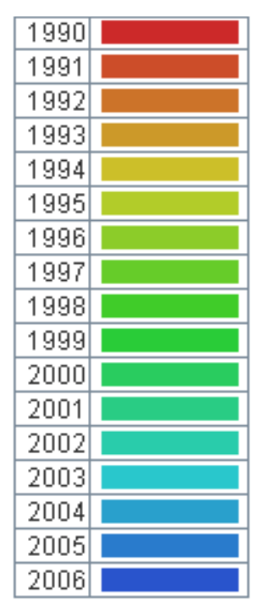
- the movement in straight line from documents that had their position altered from the initial to the final position

Preliminary Results

Alessandro Vespignani is an Italian physicist and Professor of Informatics and Cognitive Science at Indiana University, Bloomington campus.

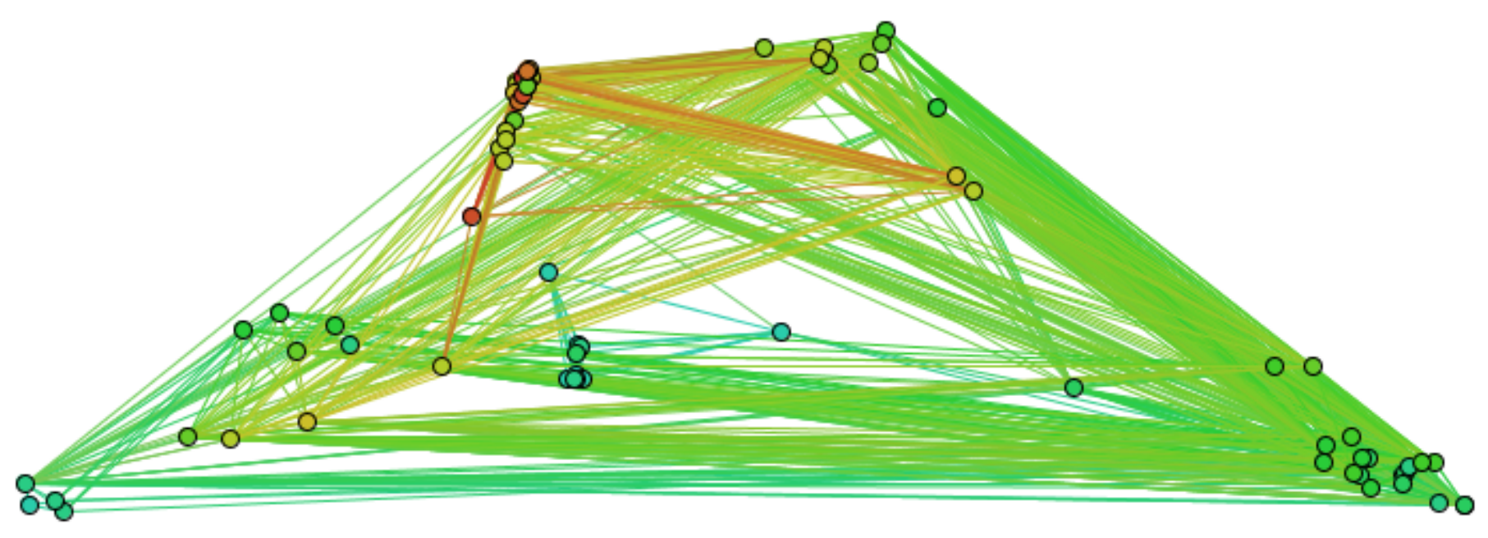
Document collection:
101 publications authored or co-authored by Alessandro Vespignani from 1990 to 2006.

Edges - represent bibliographic coupling (number of references two documents share). Every time two documents share a least one reference, an edge is added connecting them.



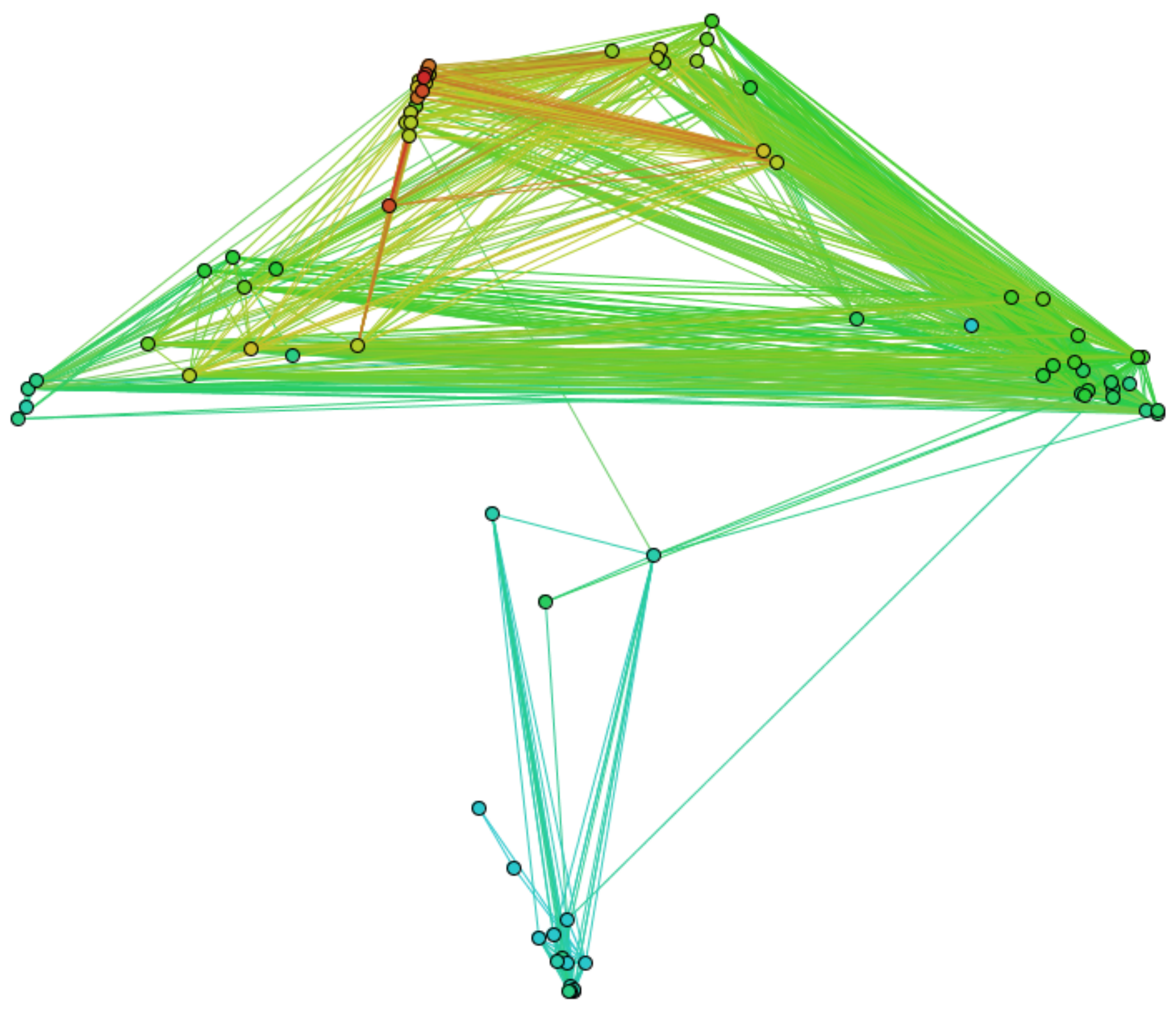
Color of circles - represents the publication year of the corresponding article

Color of edges - intermediate color between the color of the circles that an edge connects.



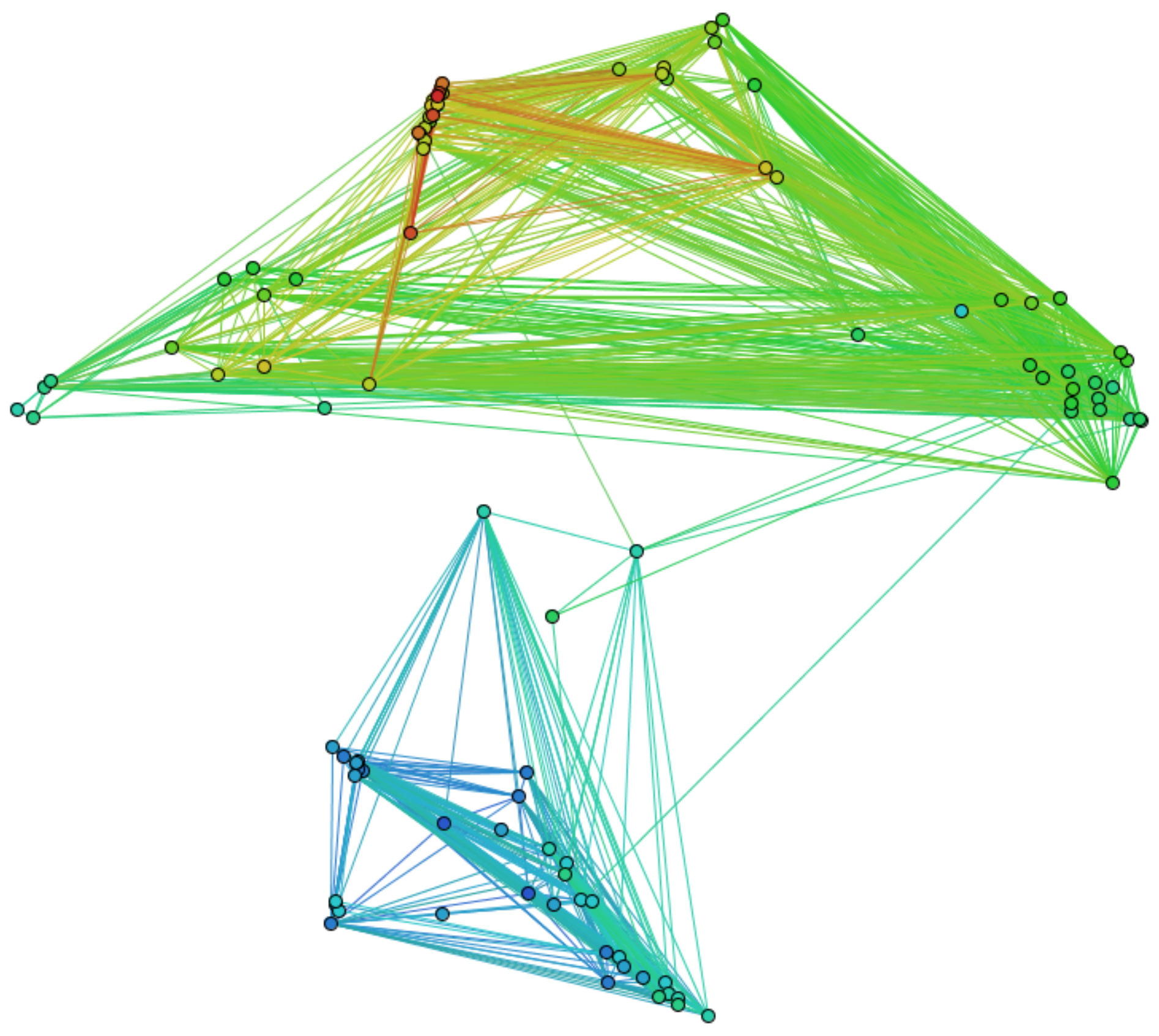
2002

One large group related to Alessandro Vespignani work in *Physics* at Europe



2003

A second group related to Alessandro Vespignani´s work in *Complex Networks* research area appears.



2006

More documents added to the second group. This is explainable by Alessandro Vespignani´s research being more focused in the *Complex Networks* area in the latest years at Indiana University.

cyberinfrastructure for NETWORK SCIENCE CENTER
USP
CAPES
FAPESP