Fabrikant, S.I. and Salvini, M. M. (2011) Charting the ICA World of Cartography 1999-2009. *Proceedings,* 25th Cartographic Conference, International Cartographic Association, Jul. 3-8, Paris, France (ISBN: 978-1-907075-05-6).

CHARTING THE ICA WORLD OF CARTOGRAPHY 1999–2009

Sara Irina Fabrikant and Marco Michele Salvini
Department of Geography, University of Zurich
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
{sara.fabrikant, marco.salvini}@geo.uzh.ch

ABSTRACT

In this paper we are proposing a cognitively adequate and perceptually salient spatialization workflow to efficiently analyze and effectively visualize the intellectual development of cartographic knowledge since the turn of the 21st century, as documented by published submissions to the International Cartographic Association's bi-annual Cartographic Conference 1999-2009. We explore the salient cartographic research themes and threads in the last ten years by means of a novel network spatialization method, including semantic and cartographic generalization. As the conference location itself has also moved across the globe during this period, we investigate if and how geography might have had an influenced on the evolving research themes and threads. Our preliminary results suggest that conference themes have indeed changed over the years in terms of submission magnitude, and location might have played a role in the relative distribution of themes. We additionally validate the results of our proposed approach with a second spatialization procedure, as spatialization outcomes indeed need to be cross-checked with various methods to assure stable, data-driven, and not method driven patterns.

INTRODUCTION

The mapping of scientific production has been a long-lasting pursuit (Chen, 2003). This has been documented over time with so-called maps of science, for example, like those compiled by information scientists in a recently published Atlas of Science (Börner, 2010). Science maps aim at visually encoding the structure and evolution of scholarly knowledge, including its scientific products (i.e., papers, books, grants, patents, etc.), and contributing actors (i.e., individual researchers, scientific communities, research institutions, etc.). Similar to maps of geographic data representing the Earth's surface, science maps allow a reader to visually explore and navigate the evolving landscape of scientific production. The mapmakers of science are typically scientometricians, information visualization researchers, and graphic designers (Börner, 2010). Geographers, or formally trained mapping experts including cartographers, seem to be mostly absent in this endeavor (Skupin and Fabrikant, 2007). Therein lies a fruitful new research avenue and application area for the ICA community.

With this contribution we outline an empirically validated, cognitively adequate, and perceptually salient science mapping approach to uncover the latent structure of scientific cartographic production since the turn of the 21st century, as documented by published submissions to the International Cartographic Association's bi-annual Cartographic Conference 1999-2009. We explore the salient cartographic research themes and threads in the last ten years by means of a novel network spatialization approach. In the next sections we describe and discuss the spatialization approach in more detail.

METHODS

While our study is inspired by previous work based on self-organizing maps derived from the 2001 and 2003 ICA conference proceedings (Skupin and de Jongh 2005), we pursue conceptually and methodologically a different research avenue. Following the empirically validated spatialization framework proposed by Fabrikant and Skupin (2005), including semantic and cartographic generalization, we conceptualize the ICA research landscape as a semantic network. In this network, the nodes represent paper submissions to the ICA conference series during the studied time period. The links between the papers represent the latent semantic associations, and thus explicitly reveal topical relatedness.

DATA

We extracted titles and abstracts of paper submissions including additional metadata from the ICA Proceedings 1999-2009, available in digital format at ESRI's GIS Bibliography web site[1]. The extracted plain-text data for the six conferences was subsequently error-checked, and cleaned (i.e., deletion of problematic characters, inconsistencies, deletion of empty records, etc.). Pre-processing resulted in a tab-delimited text file including 2693 submissions (rows), structured in columns in the following way: sequential record identifier (ID), conference date, conference location, authors, title, and abstract.

SEMANTIC GENERALIZATION

The first step includes the semantic analysis of topical content captured both in the title and the abstract of a submission. We purposefully chose the paper titles and abstracts (instead of full paper

[1] on the Web at: http://training.esri.com/campus/library/index.cfm

text) for our approach, to overcome potential language biases that might be either due to submissions varying significantly in length, and/or in semantic or syntactic quality, considering the international (and not necessarily natively English speaking) ICA authorship. We employ the probabilistic topic models method for the semantic analysis (Steyvers, 2007). Topic Models (TM) belong to the class of (automated) latent semantic analysis methods (Landauer et al., 2007). LSA involves typically three main steps: extraction of semantic information from word-document co-occurrence matrices, dimensionality reduction, and representation of words and documents as points in a lower-dimensional Euclidean semantic space. While TM also include the first two steps of LSA, they differ in the third step, as TM are based on the idea that documents are a mixtures of topics, where a topic is a probability distribution over words in the text corpus (Steyvers, 2007). Another advantage of TM is that word order (sequences of words) is also considered with TM, in contrast to latent semantic indexing (LSI), for example, as one of the more known LSA methods. We applied TM to the ICA database, using the *Text Visualization Toolbox* (TVT), a very powerful MATLAB toolbox aimed at automatically analyzing very large text corpora (Rebich Hespanha and Hespanha, 2010). TVT includes the MATLAB Topic Modeling Toolbox (1.3.2) with its Latent Dirichlet Allocation implementation in MATLAB, developed by Mark Steyvers and Tom Griffiths (2007).

We identified twenty salient topics with TM that meaningfully capture the ICA text corpus. One result of the TM analysis is a rectangular two-mode paper-topic table, where each document (row) includes a vector with 20 probabilities (columns), describing the documents' topic probability distribution. For further geometric generalization (see below) we transformed this two-mode matrix to a one-mode paper-to-paper similarity matrix with UCINET 6[2]. This new one-mode matrix indicates how similar each paper is to every other paper in the corpus by means of the cross-product of the topic association vectors. We further processed the square matrix in Pajek[3], a social network analysis package (Batagelj and Mrvar, 1998), to remove loops (i.e., document self-similarity), and one half of the matrix (as similarity between document A and B is equal to B and A). For the following geometric generalization the original similarity weights ranged zero to one were scaled to the value range zero to thousand, to avoid potential computational problems.

[2] on the Web at: http://www.analytictech.com/ucinet/

[3] on the Web at: http://pajek.imfm.si/doku.php

GEOMETRIC GENERALIZATION

The resulting half-matrix was input to the Network Workbench[4] (NWB Team 2006), a large-scale network analysis, modeling, and visualization toolkit. As mentioned earlier, we conceptualize a paper's location in a topological network space based on the distance-similarity metaphor (Fabrikant et al. 2004). The latent structure of the network is investigated with the NWB tool both on the nodes and their respective linkages. We first compute the degree (i.e., how many connections) and the strength of each node (i.e., weighted sum of the connections). We then apply the Blondel community detection algorithm (Blondel, 2008) to the network, to identify clusters of papers with high semantic similarities with the aim to uncover latent themes in the ICA publication landscape. As each paper is almost connected to every other paper in the network (i.e., 2962 links), we apply the Pathfinder network scaling algorithm to identify only the structural most salient linkages. The resulting minimum spanning tree (MST) is the input for a forced directed placement algorithm (DrL), that computes the position of the document in a two-dimensional space (Griffith et al. 2007). The obtained layout embodies the distance-similarity metaphor: papers that are more similar in paper title and paper abstract content are placed closer to one another on the network than papers that are less similar (Fabrikant et al. 2004).

RESULTS

The automatically computed configuration in Figure 1, depicted with NWB, already is able to convey some structure in the ICA proceedings landscape. We can identify more central themes (i.e., clusters of papers) in the center of the network, and major trajectories between semantic document clusters. The Blondel community algorithm identifies seven clusters, distinguished by node color in Figure 1 below. For example, one can detect a central cluster with mostly lime green colored (papers) nodes, that has six links to other, differently colored paper clusters. The stronger the semantic association between nodes, the closer the distance on the network, and the thicker and darker the edges in Figure 1. As stronger linkages are typically found within the cluster, line width and color value changes are barely visible at this viewing scale in Figure 1. Another drawback with the NWB tool (and similar network visualization toolkits) is the lack of further advanced cartographic processing (i.e., label placement, etc.).

[4] on the Web at: http://nwb.cns.iu.edu/

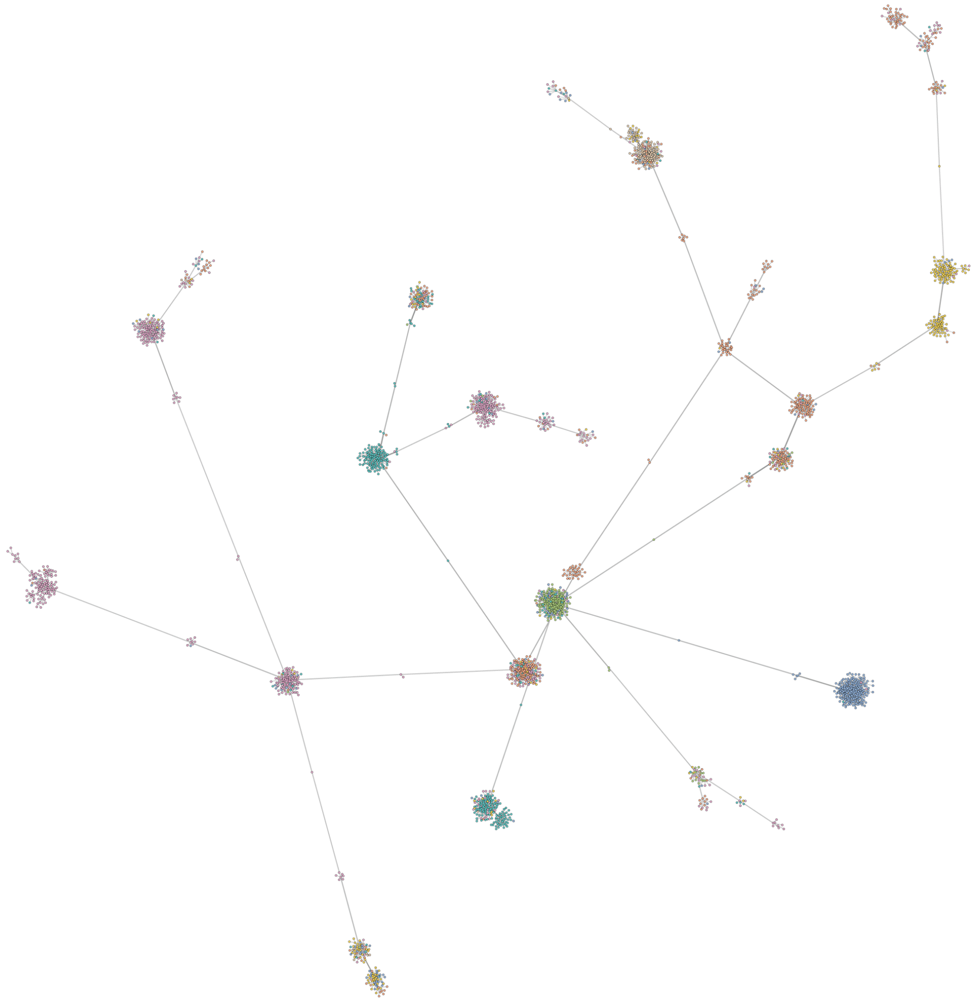


Figure 1. The semantic ICA 1999-2009 network spatialization based on 2693 proceedings contributions.

In a next step, we apply empirically validated (cartographic) design principles to the network configuration aimed at improving the legibility and perceptual salience of the uncovered latent semantic structure (Fabrikant et al. 2004). Basic spatial analysis functions available in off-the-shelf Geographic Information Systems (GIS) help to further analyze and visualize the latent network structure. By means of a python script we transfer the (x-y) node locations and the (from-to) link table from NWB to ArcGIS for spatial analysis. Specifically, we first identify semantically central nodes (i.e., those with the highest node-centrality value) in the MST network structure. We then aggregate the less central nodes within a node cluster to its closest center node, using node-distance analyses in ESRI's ArcGIS. While it is possible to further graphically process the network in a GIS, we preferred to post-process the visualization in Adobe Illustrator for various reasons relating to thematic mapping issues, map assembly, legend design, etc. We employed the statistical mapping plug-in for Adobe Illustrator developed by the Department of Cartography at the ETH Zurich (http://www.ika.ethz.ch/plugins/) to generate the graduated pie chart diagrams seen in Figure 2.

While the edge widths are continuously scaled based on computed semantic connectivity in Figure 2, the legend includes only the symbolization for the highest and lowest values (i.e., strongest and weakest relations). We redundantly apply color value to the edges to additionally communicate relatedness. The result is shown in Figure 2 below.

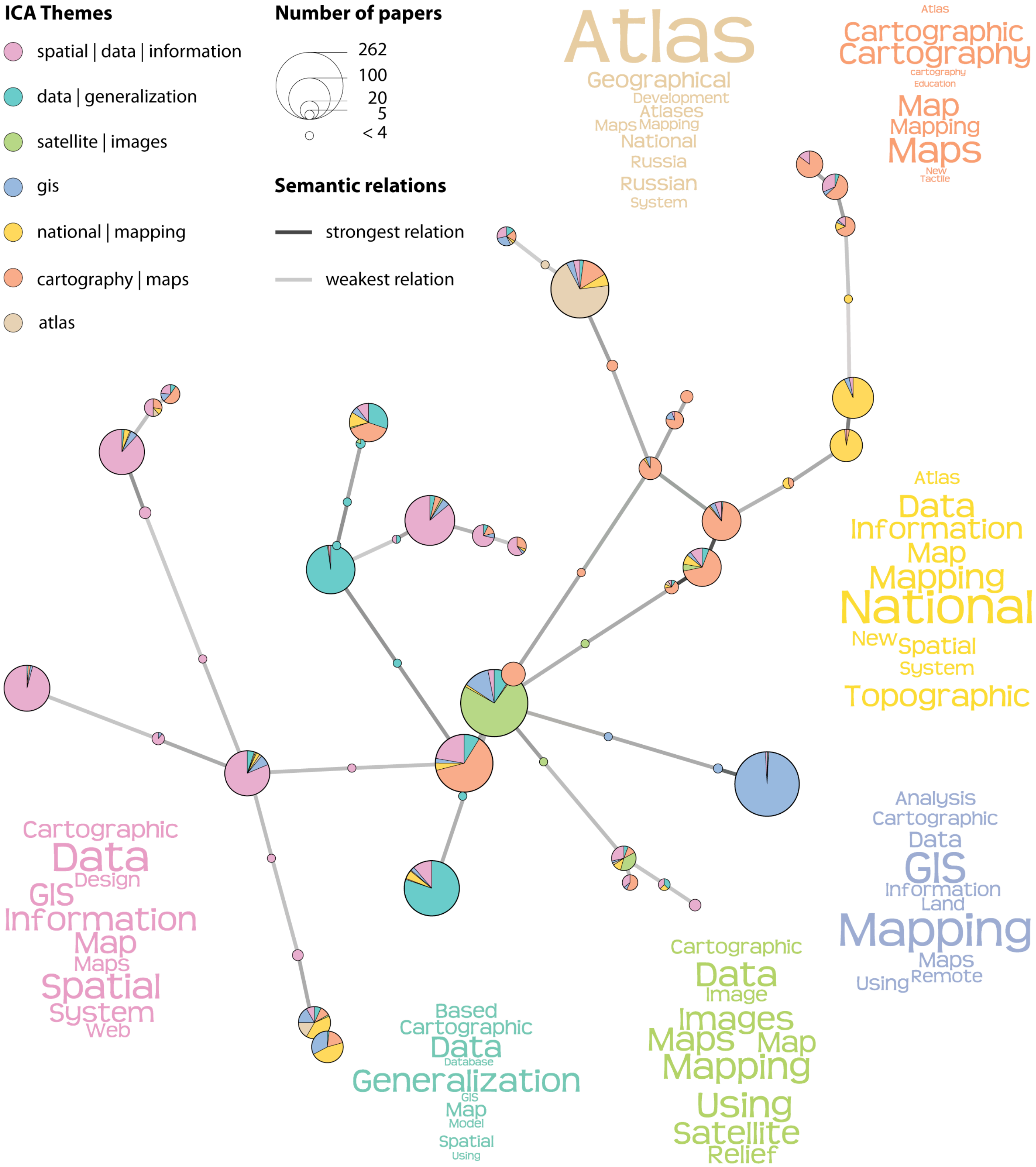


Figure 2. Cartographically enhanced ICA 1999-2009 network spatialization.

The nodes in Figure 2 are scaled based on the number of papers that have been aggregated due to their close proximity in the network space, as explained above. As the initial node clusters (see

Figure 1) are semantically not completely homogeneous, i.e., papers with varying latent themes might be directly associated with each other in a cluster, we depict the distribution of themes within a cluster in Figure 2 with pie chart segments. The colored segments of the nodes represent the proportion of papers belonging to a specific semantic cluster, as identified by the Blondel algorithm. For example, in the center of the network configuration in Figure 2, we can again recognize the largest (most central) node group from Figure 1, dominated by papers assigned to the lime green topic, but also including papers from the blue and orange themes amongst others. To automatically label found clusters, we employ the word cloud approach, using Wordle (http://www.wordle.net/) on paper titles belonging to a semantic cluster. More frequently appearing words in the cluster appear larger in the word cloud. The lime green cluster dominated by the words *cartographic data, images, maps, remote sensing, satellite* and *relief* might suggest an existing ICA theme that has been coined *geographic information,* as published in the ICA research agenda (Virrantaus et al. 2009). Some clusters (i.e., blue and tan) are more concentrated, and more compact than others (i.e., pink and orange). The pink cluster with labels *data, gis, design, spatial, system,* etc. includes quite a few themes from the ICA research agenda (i.e., geographic information, cartographic design, map production, etc.), and this might explain its expansion and connectivity to many other themes in the network. The dark green cluster with *data, generalization, model,* etc. maps well onto the geospatial analysis and modeling theme identified in the ICA research agenda. In fact, one might even suggest a semantic division in the network, perhaps with data, analysis, and modeling related themes on the left-hand side of the network, compared to more applied, map production, and map display-centric themes on the right hand side of the central (dark green and lime green), horizontal axis.

One of our research questions at the outset was whether space and time might have influenced the latent semantic structure of the ICA conference proceedings landscape. The graph in Figure 3 below sheds some light on this research question.

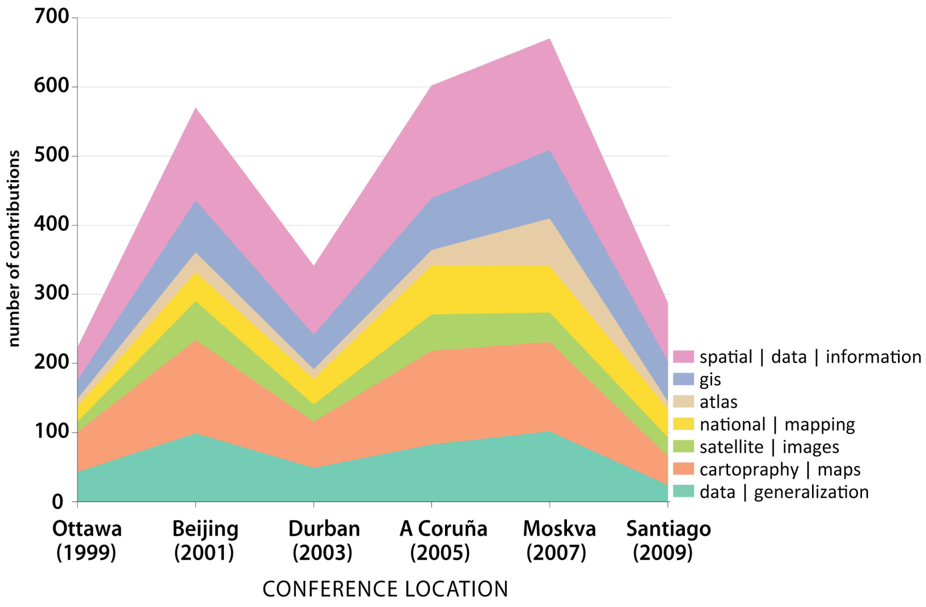


Figure 3. ICA themes in absolute submission numbers over time and across space.

Figure 3 seems to suggest an overall steadily growing trend in proceedings submissions with peaks in Beijing, China (2001), A Coruña, Spain (2005), and Moscow, Russia (2007). This might mean several things, such as geographic uniqueness and thus attractiveness of a conference location for attendees, higher submission acceptance rates by conference organizers to raise attendance numbers, etc. For example, in Beijing (2001) paper presentations were shorter than at previous conferences (i.e., 10 minutes), as to allow for more participation. While proceedings submissions seem to be at an all-time high in Moscow (2007) for the studied time period, it is interesting to note that the Moscow proceedings contain quite a few abstracts, without full papers, of people who did not attend the conference. The relatively low numbers for Durban, South Africa (2003) and Santiago de Chile (2009) might have had other reasons. Perhaps travel distance might have had an effect? Travel safety issues raised prior to the conference due to political reasons might have impacted the Durban conference. The move of the traditional conference meeting time from (North) summer (during summer holiday season) to (North) winter (during the school year) could have been another reason for relatively low submission rates at the Santiago de Chile meeting. Clearly, ICA's General Assembly (every four years) did not make a difference in submission rates, since Ottawa had lowest submission numbers, followed by Durban (third lowest) and then Moscow with the highest submission rate.

To analyze the thematical distribution over space and time it might be more useful to look at relative proportions, as shown in Figure 4 below.

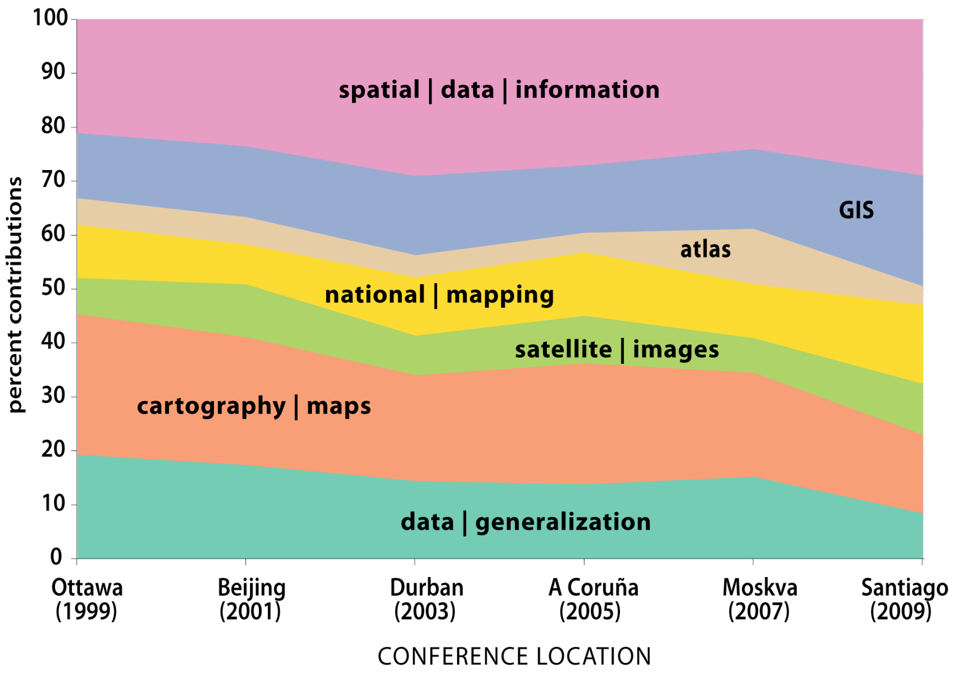


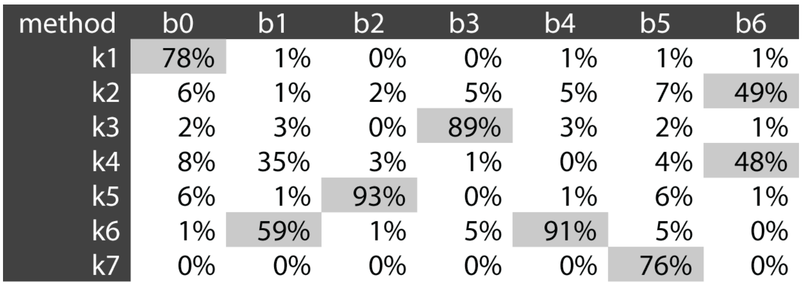
Figure 4. Relative distribution of ICA themes over time and across space.

Inspecting the proportion of submissions by themes of one might detect a growing trend for the*,* *spatial, data, information* cluster (top area in Figure 3, in pink), also containing the keywords *visualization, system, web, virtual, navigation, services, interactive, mobile*, but not shown in the Figure. The traditional ICA themes *cartography, maps* in orange, including the words *production, education, history, and atlas* (not shown) and *data, generalization* in dark green, including the words *automated, database, model, qualit*y (not shown) seem to be in decline over the past years. The thread labeled *atlas* in tan color, including the words *geographical, names, Russian* (not shown) stays relatively small throughout, but surges at the Moscow conference (one explanation for the relative high word frequency for Russia and Russian), to almost disappear again after that. Another perhaps localized favorite in Santiago (2009) is the *gis* stream (in blue) also containing the keywords *planning, monitoring, environmental, case, application,* etc.

EVALUATION

As mentioned in the methods section above, a series of transformation steps are involved in spatialization, and respective data-driven and methodological decisions need to be made at the outset. This is not unlike the traditional cartographic process. However, in cartography, one can evaluate the appropriateness of the map product by assessing its fidelity to representing reality. As spatializations are based on metaphors, thus are already representations of reality, evaluation must involve systematic validation of the mapping procedure. In other words, we must ask how stable the uncovered latent structure is, for example, when changing spatialization methods, modifying method parameters, etc. While many possibilities for validating spatializations exist, very few researchers have actually taken on the extra work to empirically validate their outputs (Boyak et al. 2005). This is in fact also true for cartography, as rarely cartographers evaluate their maps systematically by empirical means (Fabrikant and Lobben, 2009).

Here we present one of the various ways to validate the ICA publication spatialization. Firstly, we subjected the two-mode paper-topic matrix (computed with TM) including document vectors with probabilities for each of the 20 topics to a self-organizing map (SOM) routine, available in the SOM_PAK 3.1 software. We then applied k-means clustering, using the same numbers of clusters as identified with the Blondel community detection algorithm for the network spatialization, mentioned earlier. We can now quantitatively assess the two clustering results by means of a cross-tabulation, shown in Table 1 below.



method	b0	b1	b2	b3	b4	b5	b6
k1	78%	1%	0%	0%	1%	1%	1%
k2	6%	1%	2%	5%	5%	7%	49%
k3	2%	3%	0%	89%	3%	2%	1%
k4	8%	35%	3%	1%	0%	4%	48%
k5	6%	1%	93%	0%	1%	6%	1%
k6	1%	59%	1%	5%	91%	5%	0%
k7	0%	0%	0%	0%	0%	76%	0%

Table 1. Proportions of paper co-occurrences in Blondel (b_{0-7}) and k-means (k_{1-7}) clusters.

The rows in Table 1 represent the seven k-means clusters (k_{1-7}) and the columns contain the seven Blondel clusters (b_{0-6}), respectively. The cell values are percentages of papers that are represented in each cluster. The higher the percentage of clustered papers in a cell, the better the match between the two methods. The cells shaded in grey represent the highest clustering matches. For example, 78% of the papers in Blondel cluster zero are also in k-means cluster one. Interesting to note that Blondel cluster *b6* is almost equally split into k-means clusters two and four. This quantitative assessment can be complemented with a more qualitative (semantic) approach. Figure 5 shows the output of the SOM spatialization with individual documents (i.e., points) aggregated to seven k-means clusters, and visualized by means of a Voronoi tessellation. One can now qualitatively compare the word clouds for the network spatialization approach with Blondel clustering (shown in Figure 2) with the SOM & k-means word clouds depicted in Figure 5. To facilitate comparison, we employ the same color scheme as in Figures 1 & 2, using the Blondel classification as a baseline. As Table 1 suggests, the Blondel cluster six is evenly split between k-means clusters two and four. We therefore apply two shades of pink to the respective cluster zones (upper left corner) in Figure 5. Similarly, a texture fill with tan and orange hatches is employed for k-means cluster six, to highlight its topical mix with Blondel clusters one and four. The labels of the zones in Figure 5 are again the most frequent (i.e., salient) words for each cluster generated with Wordle word clouds.

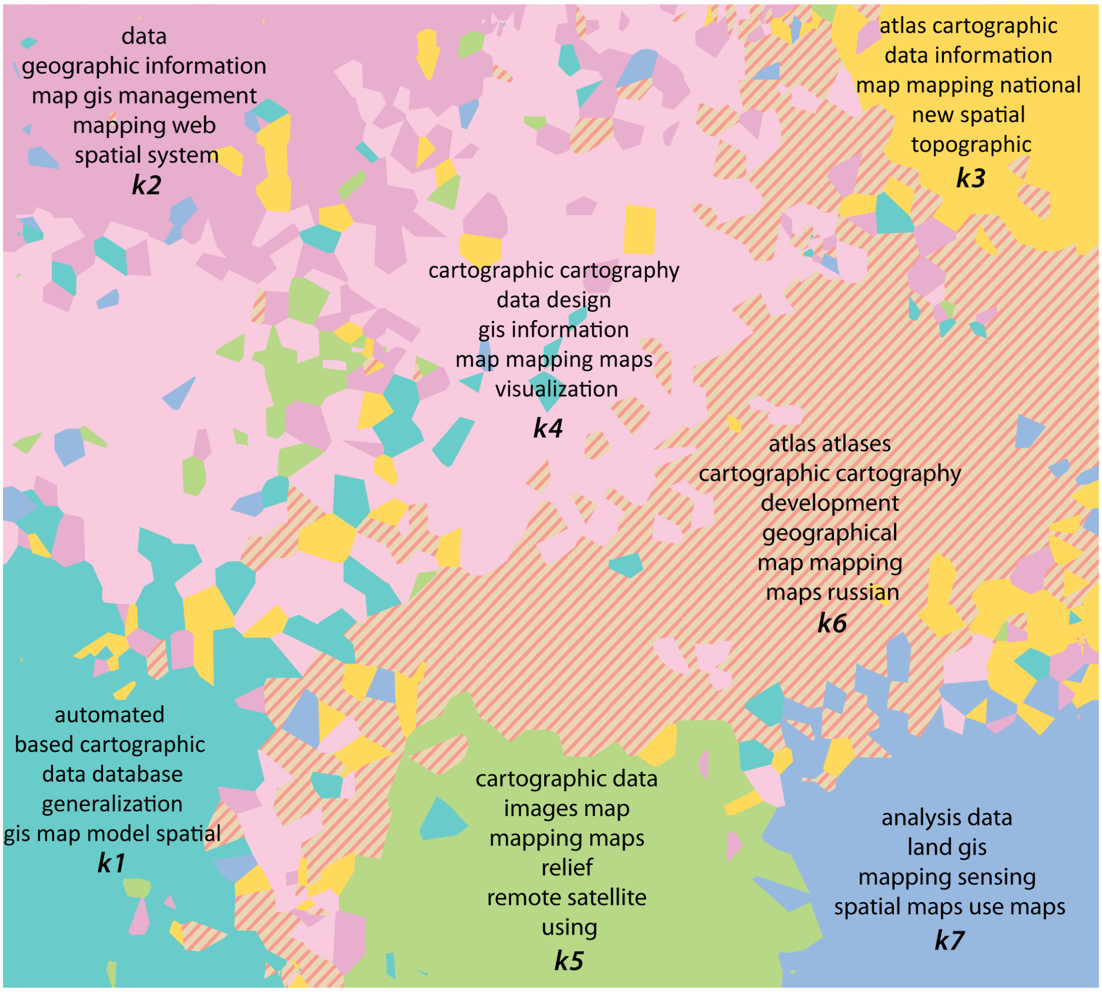


Figure 5. The ICA 1999-2009 region spatialization (SOM & k-means clustering).

The pink *k4* and hatched *k6* clusters are not only the largest zones in the Figure 5, but also contain most articles. Hence, polygon size in the SOM communicates thematic dominance adequately. The pink cluster labeled *spatial, data, information* in Figure 3 also contains the largest amount of papers, and this cluster is equally dispersedly distributed on the network. In contrast to the SOM in Figure 5, however, the lime green cluster labeled *satellite, images* is the second largest in the network (Figure 3). The Wordle clouds in Figure 2 and 5 suggest high semantic similarity of clusters identified with two conceptually and methodologically different methods. In other words, the uncovered latent themes seem meaningful and stable, irrespective of the applied spatialization approach.

CONCLUSIONS

We proposed a cognitively adequate and perceptually salient network spatialization approach to efficiently analyze and effectively visualize the intellectual development of cartographic knowledge since the turn of the 21st century, as documented by published submissions to the International Cartographic Association's bi-annual Cartographic Conference 1999-2009. The proposed approach allows us to reveal the latent cartographic research themes and threads buried in almost 3000 proceedings contributions during the last ten years. Conference location seems to have influenced thematic distribution. The uncovered latent structure has been validated with two different spatialization methods, suggesting a meaningful and stable result. With this knowledge domain mapping approach we invite the cartographic community to visually explore the evolving landscape of their own discipline, and hope to have provided well-designed charts to help uncover the latent structure of the continuously evolving body of cartographic knowledge.

ACKNOWLEDGMENTS

We would like to thank Martin Karcher for help with pre-processing the text data. We are indebted to Stacy Rebich Hespanha and Joao Hespanha who not only introduced us to TM, but also shared their excellent TVT toolkit with us, and always provided invaluable respective feedback when needed.

REFERENCES

Batagelj, V. and Mrvar, A. (1998). Pajek – Program for Large Network Analysis. *Connections,* vol. 21, no. 2: 47-57.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of community hierarches in large networks (http://arxiv.org/abs/0803.0476, accessed Feb. 2011).

Börner, K. (2010). *Atlas of Science: Visualizing What We Know*, MIT Press, Cambridge, MA.

Boyack, K. W., Klavans, R., Börner, K. (2005). Mapping the backbone of science. *Scientometrics,* vol. 64, no. 3: 351—374.

Chen, C. (2003). *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*, Springer, London, U.K.

Fabrikant, S. I. and Lobben, A., (eds.). (2009). *Cognitive Issues in Geographic Information Visualization.* Cartographica, vol. 44, no. 3.

Fabrikant, S.I., Montello, D.R., Ruocco, M., Middleton, R.S. (2004). The Distance-Similarity Metaphor in Network-Display Spatializations. Cartography and Geographic Information Science, vol. 31, no. 4, 237-252.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. T. (2007). Topics in Semantic Representation. *Psychological Review,* vol. 114, no. 2: 211-244.

NWB Team. (2006). Network Workbench Tool. Indiana University and Northeastern University. Available at: http://nwb.slis.indiana.edu (accessed Feb. 2011).

Rebich Hespanha, S. and Hespanha, J. (2010). Text Visualization Toolbox — A MATLAB toolbox to visualize a large corpus of documents. [http://www.ece.ucsb.edu/~hespanha (accessed Feb. 2011)].

Skupin, A. and Fabrikant, S. I. (2007). Spatialization. In: Wilson, J., and Fotheringham, S., (eds.), *Handbook of Geographic Information Science*, Blackwell Publishers, Oxford, UK: 61-79.

Steyvers, M. (2007). Probabilistic Topic Models. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W., (eds.), *Handbook of Latent Semantic Analysis* Lawrence Erlbaum Associate Publishers, Mahwah, N. J.: 427-448.

Virrantaus, K., Fairbairn, D., Kraak, M-J. (2009). ICA Research Agenda on Cartography and GIScience, *The Cartographic Journal,* vol. 46, no. 2: 63-75.