Bull. Nov. Comp. Center, Comp. Science, 32 (2011), 19–35

Visualization of citation networks for large science portals*

Z.V. Apanovich, T.A. Kislicina

Abstract. We examine a number of methods for probing and understanding a large-scale structure of networks that evolve over time. In particular, we focus on citation networks, networks of references between documents such as research papers. We describe three different methods of visualization: the first is based on a hierarchical edge bundles algorithm, the second implements dynamic layered drawing, and the third utilizes a geometry-based edge bundling. Using the datasets of Linked Open Data portals, we generate citation networks and demonstrate how each of these methods can reveal significant features of the considered networks.

Key words*: hierarchical edge bundles, science portal, content, information visualization, layered drawing, ontology, citation networks, Open Linked Data*

1. Introduction

Due to the fast progress of Semantic Web and its new branch, Linked Open Data, large amounts of structured information on various scientific fields are becoming available. The main part of the content of scientific digital libraries and specialized portals constitute research publications, the most reliable source of information in any research area. The most active and influential researchers, organizations in which they work, and geographic locations of the research units – all this information is currently available in the rdf / xml format. This information evolves over time and rapidly grows in volume. To optimize the science management, new tools for investigation and analysis of these data are needed. A generally accepted way to facilitate the understanding of large and complex data sets is graph visualization. The topic of our paper is several visualization methods for citation networks. Previously, we considered the methods of visualization of information on scientific cooperation, represented by co-authorship networks derived from small information portals [1–3]. Our current work is a further development of this research. The data sets under consideration have a significantly greater volume, and newly developed algorithms are presented to analyze and visualize these data.

A citation network is a network in which the vertices represent documents and the edges between them represent references between documents. Ci-

*Partially supported by the RFBR grants N 09-07-00400, 11-07-00388- and RAS project 14/12.

tation networks are directed: citations go from one document to another. Citation networks evolve over time as new documents are created. The citation network analysis started with the paper by Garfield et al. [14] and has been studied by many authors [11, 16]. Force-directed methods of visualization used to be the main tool of investigation for these networks.

In this paper, we present three different methods of visualization for the citation networks: the first is based on a hierarchical edge bundles algorithm, the second implements dynamic layered drawing, and the third utilizes the geometry-based edge bundling. For testing, we make extensive use of citation networks designed from the data sets of Open Linked Data portals. The paper is organized as follows. Section 2 discusses extracting citation networks from the content of Linked Open Data portals. Section 3 demonstrates some problems of the citation networks visualization by the hierarchical edge bundles method. Section 4 describes some results of visualization of citation networks by a layered dynamic method. Section 5 demonstrates the citation networks visualization with a geometry-based edge bundling method. Finally, section 6 presents conclusion and perspectives for further research.

2. Open Linked Data and citation networks generation

The datasets of Linked Open Data (LOD) portals such as DBLP, Citeseer, CORDIS, NSF, EPSRC, ACM, IEEE, [4-9], etc., have been used as test data. These datasets are described in RDF format and have a very impressive size. For example, the data provided by the Citeseer portal consist of 8,146,852 triples, ACM portal data comprise 12,402,336 triples, and DBLP portal has granted 28,384,790 triples. A user can either download the files in the RDF format, or generate data using a sparql query. All datasets of these portals are described according to a single ontology AKT Reference Ontology [6], which is the union of several ontologies (Support Ontology, Portal Ontology, Extensions Ontology and RDF Compatibility Ontology).

The main one is Portal Ontology, which describes such concepts as organizations, persons, projects, publications, geographic data, etc. AKT Ontology has a rather deep hierarchical structure (Figure 1). For example, to describe the publications, there exist two root classes “Information-Bearing-Object” and “Abstract-Information”. The subclasses of “Information-Bearing-Object” are the classes “Recorded-Audio”, “Recorded-Video”, “Publication”, “Edited-Book”, “Composite-Publication”, “Serial-Publication”, “Periodical-Publication” and “Book”. All individuals of the class “Information-Bearing-Object” have a relationship “has-publication-reference”, pointing to an object of the class “Publication-Reference”, which is a subclass of the class “Abstract-Information”. In turn, the class “Publication-Reference” has, as its subclasses, the classes “Web-Reference”, “Book-Reference”, “Edited-Book-Reference”, “Conference-Proceedings-Reference”, “Workshop-

Proceedings-Reference", "Book-Section-Reference", "Article-Reference", "Proceedings-Paper-Reference", "Thesis-Reference" and "Technical-Report-Reference". The individuals of the class "Publication-Reference" have such relationships as "has-date", "has-title", "has-place-of-publication", "cites-publication-reference", etc. There exists a class "Organization", which is a subclass of the class "Legal-Agent", and the class "Legal-Agent" is a subclass of the class "Generic-Agent". The class "Person" is a subclass of the class "Generic-Agent".

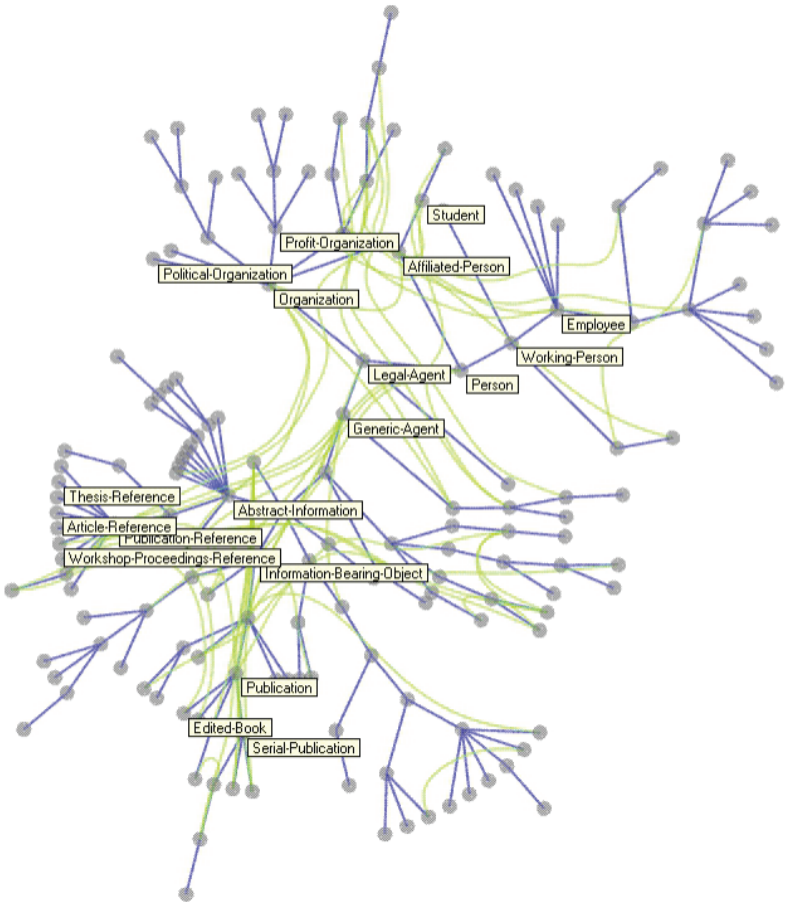


Figure 1. AKT Reference ontology

There are several problems of using the LOD datasets. Although all bibliographic datasets of the LOD cloud use AKT Ontology as a common vocabulary, the contents of these sets are very heterogeneous and are based on very narrow subsets of this vocabulary. To describe real objects, the classes of the highest level of hierarchy are normally used. For example, the classes "Publication-Reference" and "Article-Reference" are used for description of

publications while such classes as “Proceedings-Paper-Reference” are not used at all. This feature makes it difficult to generate the hierarchical structure needed for applying the hierarchical edge bundles method. Also, the data sets are not complete and many attributes remain to be filled in. Besides, the citation relationship (*akt: cites-publication-reference*) existing in AKT Reference Ontology is described explicitly only for several datasets, such as Citeseer and ACM [7, 9]. However, the common mechanism of access simplifies working with these data. It is easy enough to generate a simple citation network for any storage of the LOD cloud if the publications described in these datasets have the relationship “cites-publication-reference”. An example of a user interface and SPARQL 1.0 query intended for citation networks generation from the ACM dataset is shown in Figure 2.

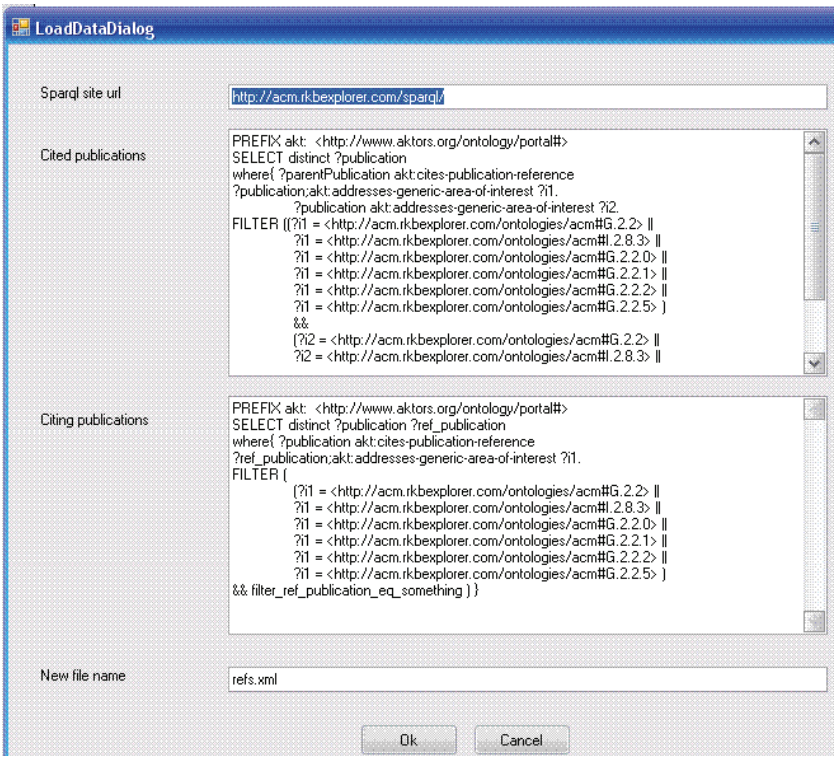


Figure 2. An example of a user interface and SPARQL 1.0 query intended for citation networks generation

To select the desired volume of data, the query modifier LIMIT N was used. We could relatively easy extract citation networks of 20 000–30 000 vertices.

3. Visualization of citation networks using the hierarchical edges bundles

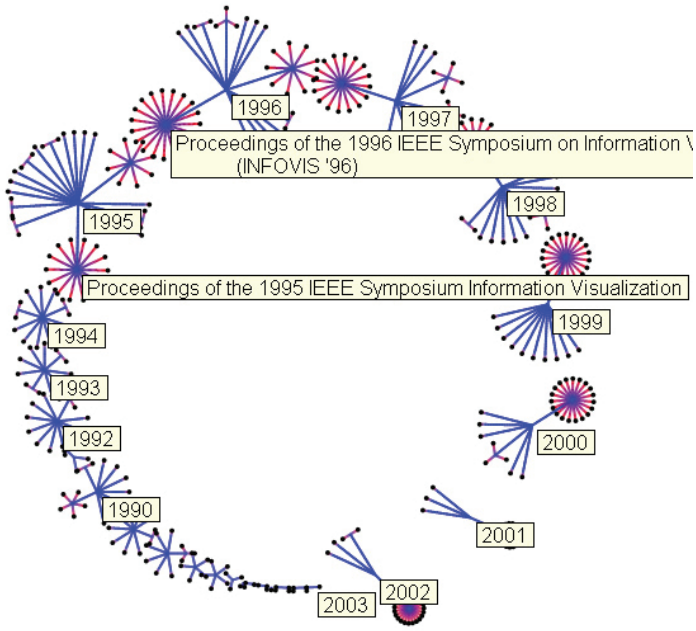
We have started our experiments by applying our implementation of the hierarchical edge bundles method [15, 2] for the citation networks visualization. This method allows a drawing of a citation network to be combined with drawings of other elements of the portal content. It is implemented as follows. Some predefined hierarchical structure is drawn as a tree whose leaves are research papers. Then each link of the citation network is modeled as a single B-spline [19] using the control points along the shortest path in the tree layout from one leaf point to another. A test set of 561 publications on information visualization for 10 years is shown in Figure 3. A three-level hierarchy consisting of years, conferences, and publications is depicted with a balloon tree method (Figure 3(a)), and the citation links are drawn with the hierarchical edge bundles method. When looking at this drawing, users can easily identify the years with the largest number of publications (the years 1995 and 1996). We have slightly improved the drawing comprehensibility by depicting the citation index of papers by the radius of nodes. Since we do not want the area of drawing to grow up due to the node size enlargement, the node overlaps are permitted. However, these overlaps should not spoil the drawing readability. So we allowed users to calculate the nodes visibility as a function of the citation index, as it is shown in Figure 3(b), where the number of visible nodes and the number of the node overlaps have been reduced. Further on, users can also change the width of reference links and their opacity as a function of the citation index of incident nodes.

Some possible functions for calculation of these parameters are:

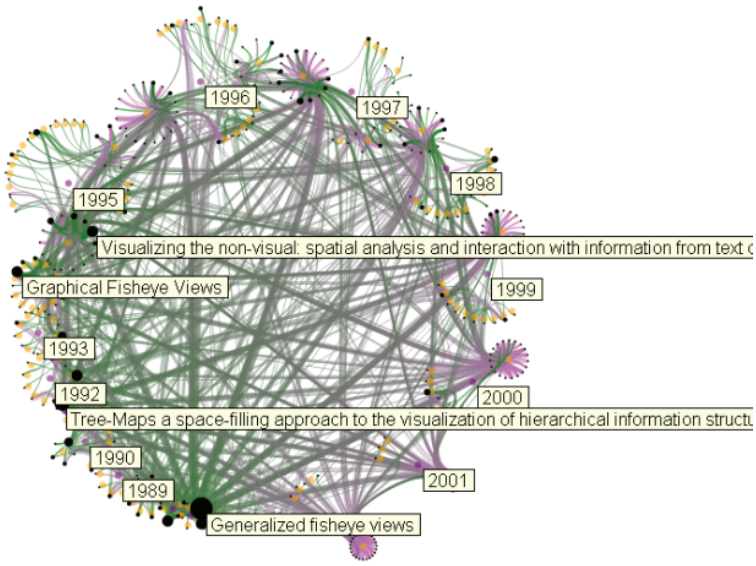
$$y = (o_{\max} - o_{\min}) \frac{I - I_{\min}}{I_{\max} - I_{\min}} + o_{\min}, \tag{1}$$

$$y = (o_{\max} - o_{\min}) \cdot \left(1 - \sqrt{\frac{I_{\max} - I}{I_{\max} - I_{\min}}}\right), \tag{2}$$

where I is the citation index, I_{max} and I_{min} are the largest and smallest citation indices in the citation network under consideration, o_{max} and o_{min} are the upper and lower bounds of values for y.



(a)



(b)

Figure 3. Hierarchical structure and citation network. (a) A three-level hierarchy consisting of years, conferences, and publications. (b) Hierarchical edge bundles drawing of a citation network

Formula (1) helps us to identify the group of the most cited publications, since the node sizes are proportional to their citation indices. Formula (2) helps us to find the most cited publication since it assigns a much larger radius to the node with the highest citation index.

After the most cited papers are identified, a user can choose such a node with a mouse pointer and examine its name, the list of its authors and all the papers citing it as shown in Figure 4.

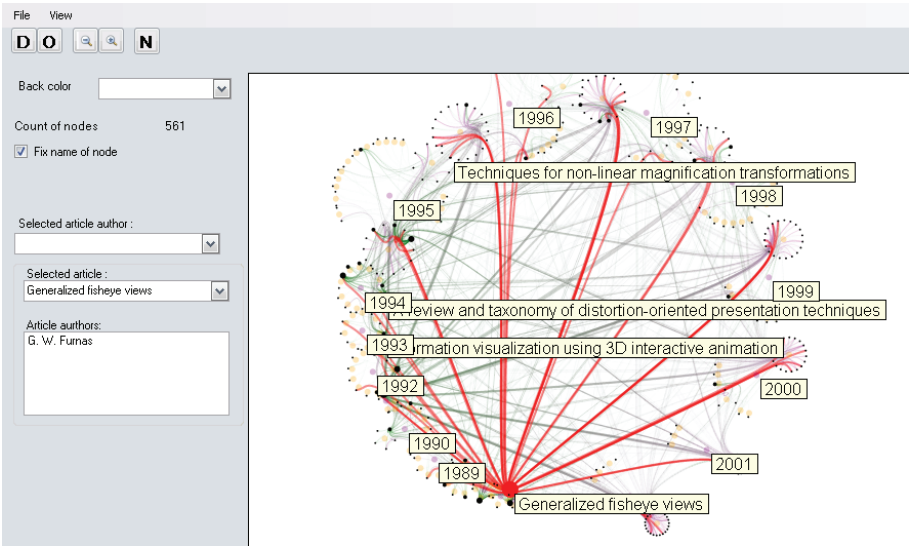
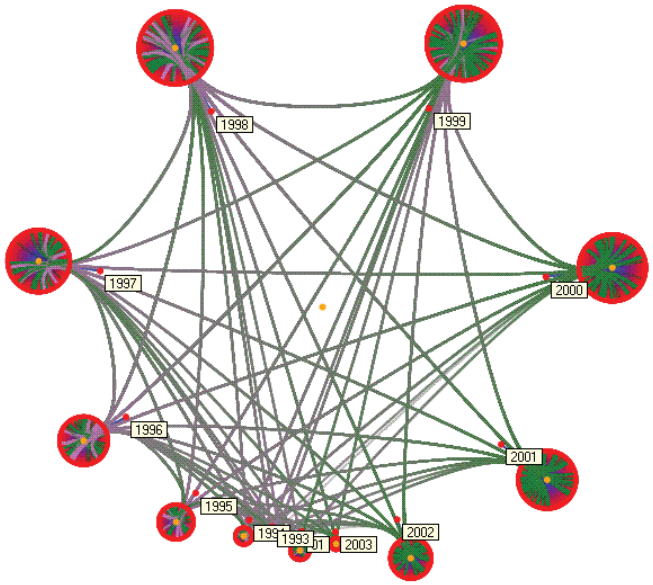
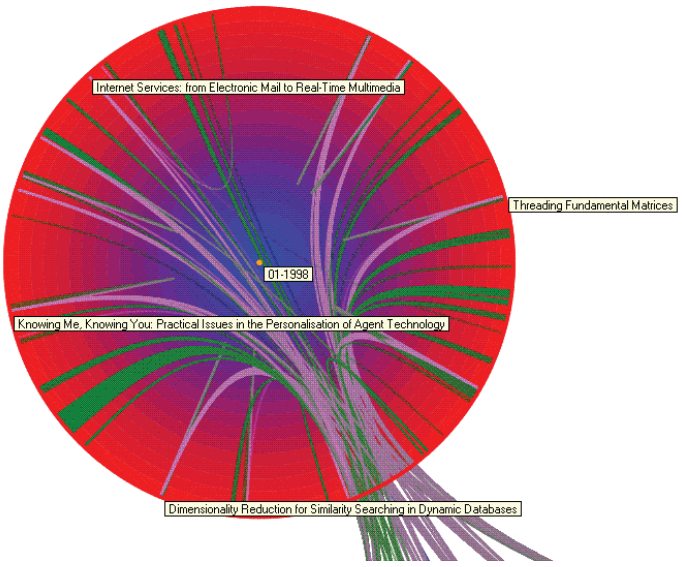


Figure 4. The most cited paper and links citing it

When the size of a citation network increases, the hierarchical edge bundles method becomes more difficult to use. For example, a drawing of a citation network of 20 000 vertices, retrieved from the Citeseer database, is shown in Figure 5. We have only managed to create a two-level hierarchy for the Citeseer dataset: the year of publication – the month of publication, which results in a drawing, quite sparse in the center (Figure 5(a)) and very dense at the periphery (Figure 5(b)). The time interval of the dataset of these publications covers the period from 1993 to 2003. The drawing permits us to compare the number of publications by year: the largest number of publications of the test set falls on the years 1998 and 1989 while publications of the year 2003 are not numerous. It does seem possible to extract more detailed information from this drawing. The central part of the drawing is a complete graph stating that there exist citing links from any year to any posterior year in this network. The publications of every year are so numerous that it is very difficult to select a vertex by the mouse pointer for further investigation.



(a)



(b)

Figure 5. A citation network of 20 000 vertices retrieved from Citeseer portal. (a) A global view. (b) One-month publications of the 1998 year

In addition, the circular drawing of the chronologically ordered data does not always look natural. Therefore, we have implemented a dynamic layered method of visualization to emphasize the directed nature of links in the citation networks.

4. Dynamic layered drawing of the citation networks

A citation network is a directed graph, so it is desirable that all edges have the same direction. The direction of the edges corresponds to the chronological order of publications. Also, the citation networks are assumed to be acyclic, even if it is not always the case. For example, if a research paper sometimes cites another one, forthcoming but not yet published, the resulting network will have a closed loop. However, such loops are rare and short.

The construction of a layered graph drawing [17] proceeds in a sequence of standard steps:

1. **Layer assignment**. The vertices of the directed acyclic graph are assigned to layers, such that each edge goes from left to right. In the current implementation, each layer corresponds to a publishing year, i.e. the papers published in the same year are assigned to the same layer. We are going to parameterize the length of time intervals in the nearest future. Edges that span multiple layers are replaced by paths of dummy vertices so that, after this step, each edge in the expanded graph connects two vertices on adjacent layers of the drawing.
2. **Crossing minimization.** The vertices within each layer are permuted in an attempt to reduce the number of crossings among the edges connecting it to the previous layer. Since finding the minimum number of crossings is NP-complete, we place each vertex in a position determined by the average of the positions of its neighbors at the previous level and then permuting adjacent pairs as long as this improves the number of crossings.
3. **Coordinate assignment.** A coordinate within its layer, consistent with the permutation calculated in the previous step is assigned to each vertex. The dummy vertices are removed from the graph and the vertices and edges are drawn.

Figure 5 shows the drawing of a citation network generated by the layered method of placement. The years of publishing for papers in the citation network are shown as rectangles of different colors at the top of the image. All papers published in the same year are placed in a vertical column corresponding to this year. The edges of the network correspond to citations. The color of each edge is identical to the color of the label of the year of the citing publication. The more citation links a publication has, the more

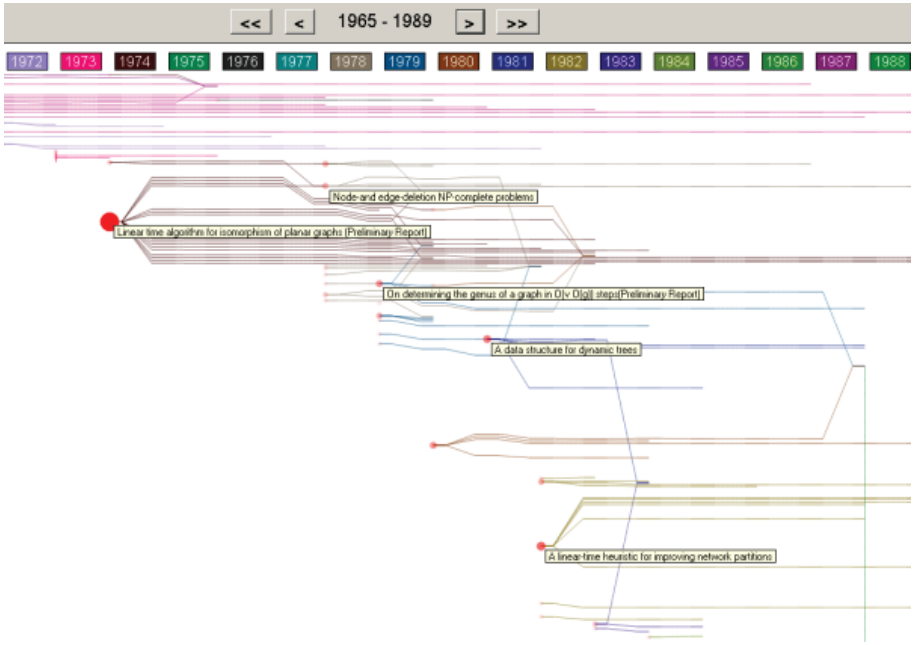
input edges the corresponding vertex has and the greater its radius is. As a result, the citation links of publications form highly visible bundles. Four buttons at the top of the screen are used to track the dynamics of the citation network year by year. The buttons "<" and ">" are designed to move through the drawing and observe the evolution of the citation network over time. Technically, this feature is implemented by filtering the vertices and edges of the citation network. Pressing the ">>" button displays the entire citation network, and the "<<" button is used to clean the drawing.

The evolution over time of a citation network of papers devoted to the graph theory is shown in Figure 5. The four fragments of this figure show different intervals of time between 1965 and 2005. In the period from 1965 to 1989 (Figure 6(a)), the test set of publications is dominated by the paper "Linear-time algorithm for isomorphism of planar graphs". The corresponding vertex has the largest radius and a large brown tail of input edges. In 1993 (Figure 6 (b)), the number of references to the papers "A data structure for dynamic trees" and "A linear-time heuristic for improving network partition" increases. In 1995 (Figure 6 (c)), these two papers have the same level of the citation index as the paper "Linear-time algorithm for isomorphism of planar graphs". Finally, in 2005 (Figure 6 (d)), the paper "A linear-time heuristic for improving network partition" becomes the most cited. So this dynamic visualization really improves the citation networks readability.

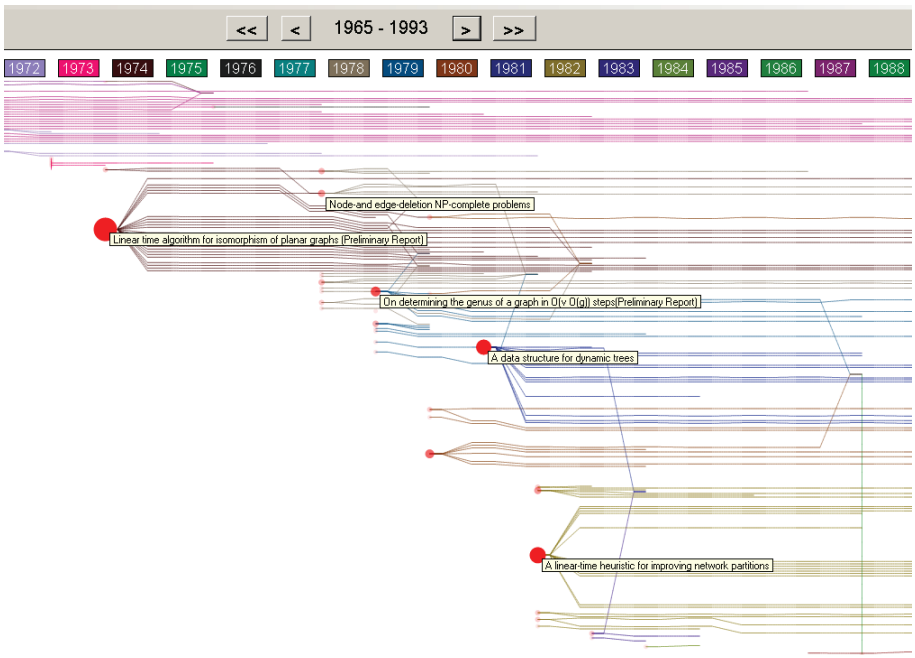
It is also possible to observe a growing interest in the paper "Node-and-edge-deletion NP-complete problems", which refers to the previously dominating paper "Linear-time algorithm for isomorphism of planar graphs", i.e. a chain of highly cited related publications arises.

Besides, this visualization method helps us to detect errors and inaccuracies in bibliographic data. Figure 6(a) shows a fragment of a citation network generated for the ACM dataset on the time interval from 1988 to 1990. A brown link connects the node of the paper "Analysis of pointers and structures" published in 1990 and the paper "Interprocedural slicing using dependence graphs" published in 1988. Since the color of the link corresponds to 1990, it should mean that the arc is oriented backward and a paper published in 1988 cites a paper published in 1990. By checking the ACM dataset (Figure 6(b)), we have discovered that the paper "Interprocedural slicing using dependence graphs" has several dates of publication and the corresponding node is placed in the layer of the earliest date of publication.

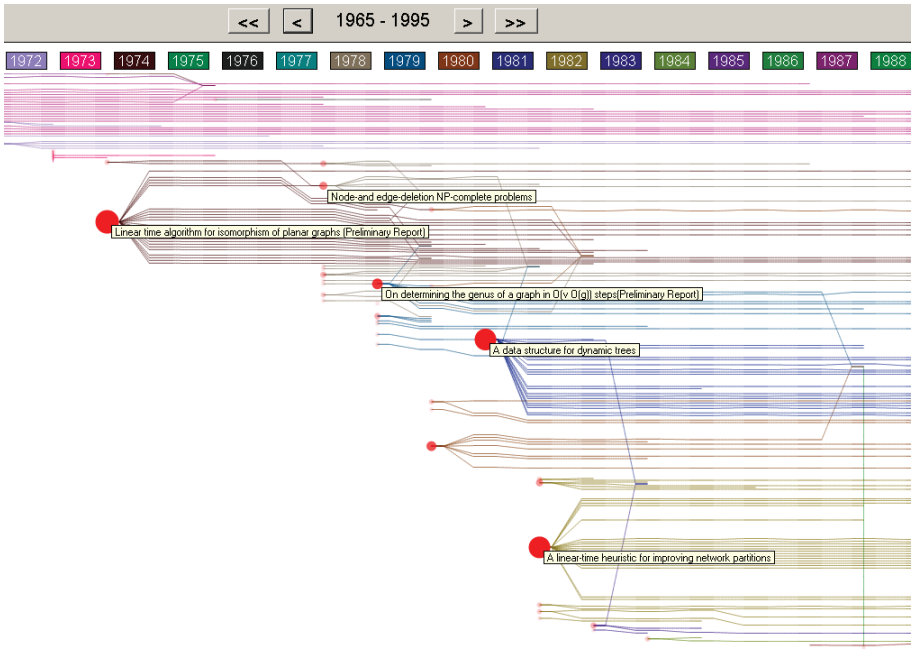
The main problem with the conventional layered method is that drawings become overloaded very quickly and filtering removes irrelevant papers but distorts reality: irrelevant publications are the major contributors in determining the significance of other publications. The hierarchical edge bundles method is not applicable in the absence of any external hierarchical structure. Therefore, we have implemented an algorithm which can reduce the


<< < 1965 - 1989 > >>
1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988
Node-and edge-deletion NP-complete problems
Linear time algorithm for isomorphism of planar graphs (Preliminary Report)
On determining the genus of a graph in O(v O(g)) steps(Preliminary Report)
A data structure for dynamic trees
A linear-time heuristic for improving network partitions

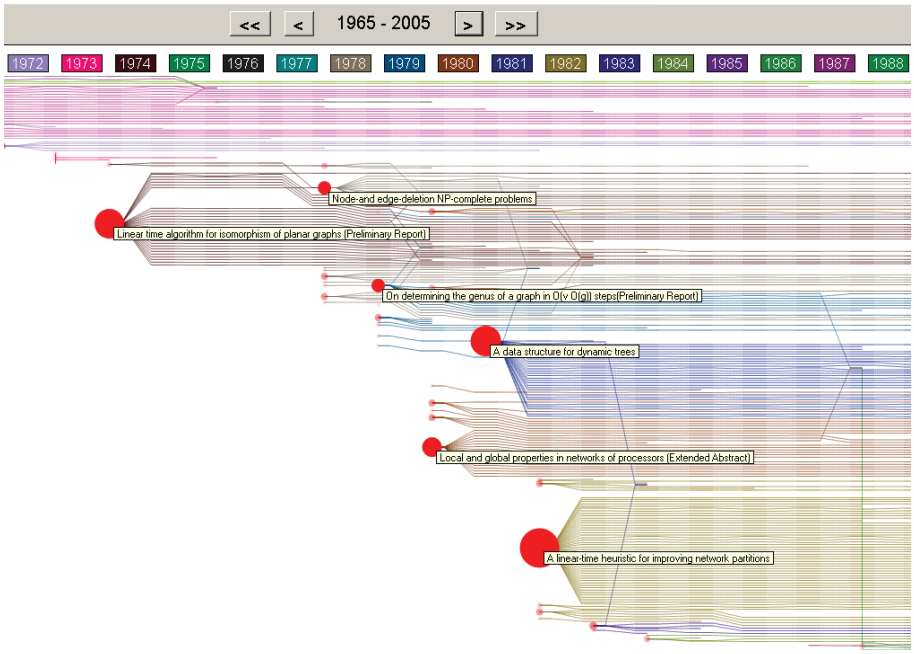
(a)


< 1965 - 1993 > >>
1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988
Node-and edge-deletion NP-complete problems
Linear time algorithm for isomorphism of planar graphs (Preliminary Report)
On determining the genus of a graph in O(v O(g)) steps(Preliminary Report)
A data structure for dynamic trees
A linear-time heuristic for improving network partitions

(b)

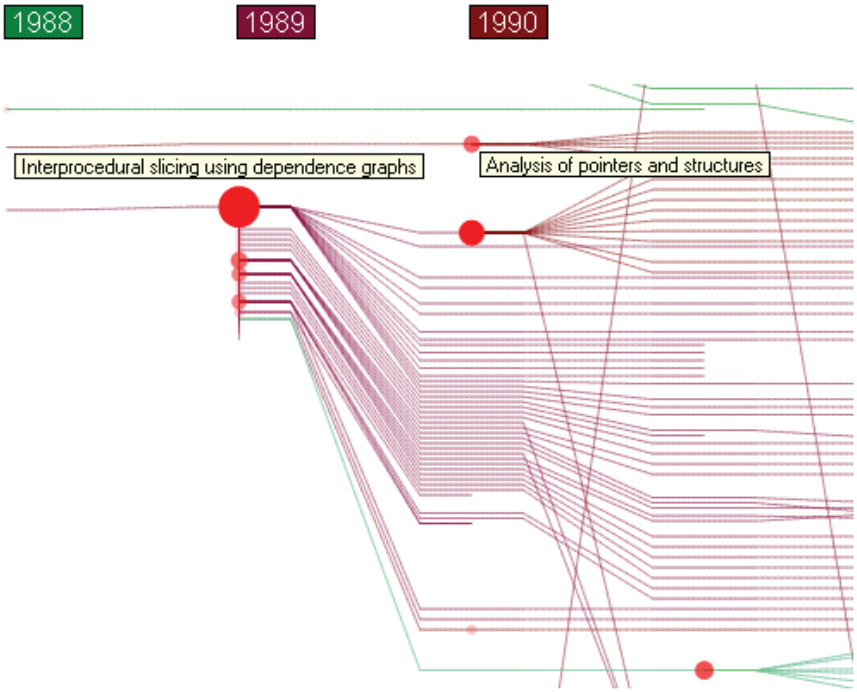


(c)



(d)

Figure 6. The evolution of a citation network over time



(a)

Analysis of pointers and structures	akt:has-date	1990-01-01
Interprocedural slicing using dependence graphs	akt:cites-publication-reference	Analysis of pointers and structures
Interprocedural slicing using dependence graphs	akt:has-date	1988-01-01
Interprocedural slicing using dependence graphs	akt:has-date	1988-07-01
Interprocedural slicing using dependence graphs	akt:has-date	1990-01-01

(b)

Figure 7. Datasets inaccuracies. (a) Backward link representing a paper published in 1988 citing a paper published in 1990. (b) Multiple dates of publishing for a paper

drawing density by forming bundles of edges based on their own geometry, and not introduced from outside.

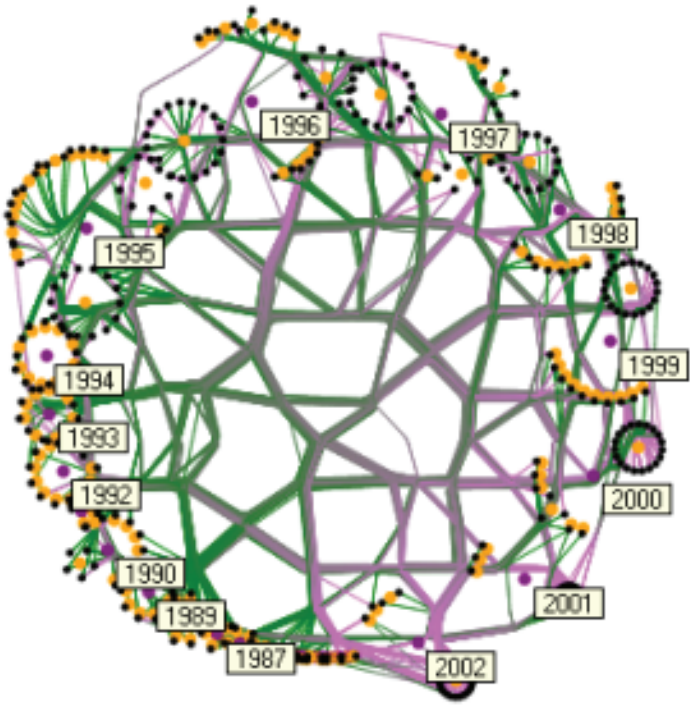
5. Geometry-based edge bundling method

The main idea of the geometry-based edge bundling method [12] is to reduce the visual clutter of an image by bending the edges through a special control grid without changing the original locations of graph vertices. This method proceeds as follows:

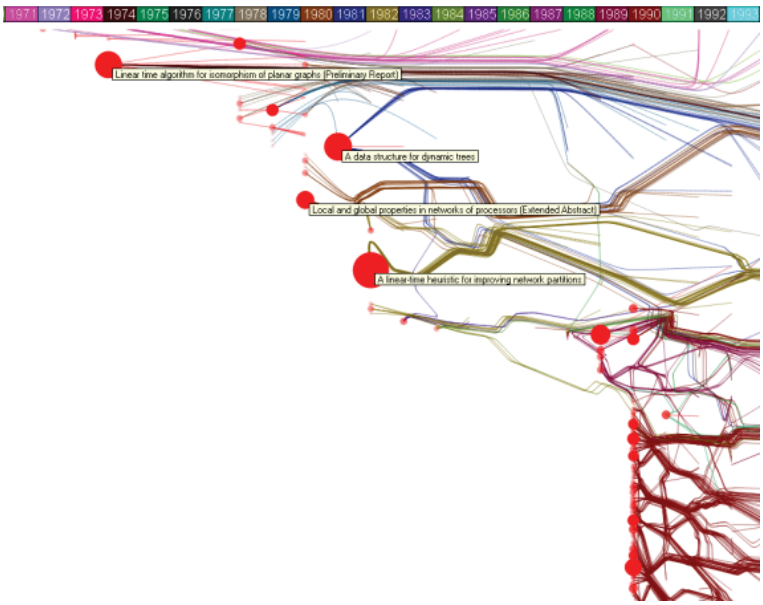
1. Generate a rectangular NxN grid and put it over a graph drawing constructed in any way.
2. For each grid cell, calculate the main direction of edges crossing the cell.
3. Merge the adjacent cells having directions that differ by no more than the threshold value into zones.
4. Calculate the basic direction and normal vector to the main direction of each zone.
5. Calculate the points of intersection of the normal segments with zones' boundaries.
6. Use the resulting points to construct a triangulation.
7. For each edge of the constructed triangulation, find the point of intersection with the edges of the original graph drawing. Calculate the centers of these points.
8. Use the resulting points as control points of b-splines.

Figure 8(a) demonstrates an application of the geometry-based edge bundling strategy to the drawing obtained by the circular drawing method from Figure 2(b). Figure 8(b) shows an application of the geometry-based edge bundling strategy to the drawing obtained by the layered visualization method from Figure 5(d).

No doubt, this methodology reduces the drawing congestion. At this stage, however, this method poses more questions than gives answers. How to choose the best direction for a rectangular grid? How does the direction of the edge bundles depend on the size of the grid? How to choose the best edge direction within each zone in the function of the underlying visualization method? Nevertheless, we intend to develop this method to the point where it can be used to examine trends in the research field.



(a)



(b)

Figure 8. The application of the geometry-based edge bundling strategy to the drawing obtained by the layered visualization method. (a) The application of the geometry-based edge bundling strategy to the circular drawing. (b) The application of the geometry-based edge bundling strategy to the layered drawing

Conclusion

In this paper, we have demonstrated three visualization methods for citation networks generated for datasets of Linked Open Data portals. These drawings are quite helpful for understanding the datasets of large volumes. They also enable users to observe the evolution of datasets over time. In the nearest future, we are going to apply the previously developed clustering methods to the citation networks analysis and to compare the results obtained by the two groups of methods.

References

[1] Apanovich Z.V., Vinokurov P.S. Ontology and content of knowledge portals analysis with information visual // Complex systems control and modeling problems. – Proc. of the XI Internat. Conf. – Samara, 2009. – P. 556–562.

[2] Apanovich Z.V., Kislicyna T.A. Extending the subsystem of content visualization of informational portal by visual analytics tools // Complex systems control and modeling problems. – Proc. of the XII Internat. Conf. – Samara, 2010. – . 518–525.

[3] Apanovich Z.V., Vinokurov P.S. Ontology based portals and visual analysis of scientific communities // First Russia and Pacific Conf. on Computer Technology and Applications, 6–9 September, 2010. – Vladivostok, 2010. – P. 7–11.

[4] Bizer, C., Heath, T. and Berners-Lee, T. Linked Data – The Story So Far // Internat. J. Semantic Web Inf. Syst. . – 2009. – Vol 5 (3). – P. 1–22.

[5] Linked Open Data datasets. – http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets.

[6] AKT ontology description. – http://www.aktors.org/ontology.

[7] CiteSeer dataset. – http://citeseer.rkbexplorer.com/.

[8] DBLP dataset. – http://dblp.rkbexplorer.com/.

[9] ACM dataset. – http://acm.rkbexplorer.com/.

[10] Cui W., Zhou H., Qu H., Wong P.C., Li X. Geometry-based edge clustering for graph visualization // IEEE Trans. on Visualization and Computer Graphics. – Proceedings Visualization. – Information Visualization 2008. – 2008. – Vol. 14, N 6.

[11] Chen h., Song I-Y., Weizhong Zhu W. Trends in conceptual modeling: Citation analysis of the ER conference papers (1979–2005) // Proc. of the 11th Internat. Conf. on the International Society for Scientometrics and Informatrics. – Madrid: CSIC, 2007. – P. 189–200.

[12] Cui W., Zhou H., Qu H., Wong P.C., Li X. Geometry-based edge clustering for graph visualization // IEEE Trans. on Visualization and Computer Graphics. – Proceedings Visualization. – Information Visualization 2008. – 2008. – Vol. 14, N 6.

[13] Fruchterman T.M.J., Reingold E.M. Graph drawing by force-directed placement // Software – Practice and Experience. – 1991. – Vol. 21, N 11. – P. 1129–1164

[14] Garfield E, Sher I.H, Torpie R.J. The Use of Citation Data in Writing the History of Science. – Philadelphia: The Institute for Scientific Information, 1964. – http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf

[15] Holten D. Hierarchical edge bundles: visualization of adjacency relations in hierarchical data // IEEE Trans. Visualization and Computer Graphics. – 2006. – Vol. 12, N 5. – P. 741–748.

[16] Small H. Visualizing science by citation mapping // J. of the American Society for Information Science. – 1999. – Vol. 50(9). – P. 799–813.

[17] Sugiyama K., Tagawa S., Toda M. Methods for visual understanding of hierarchical system structures // IEEE Trans. Systems, Man, and Cybernetics. – 1981. – P. 109-125.

[18] Small H. Visualizing science by citation mapping // J. of the American Society for Information Science. – 1999. – Vol. 50(9). – P. 799–813.

[19] http://ru.wikipedia.org/wiki/B-spline