Средства Визуального анализа Информационного Наполнения Порталов, входящих в облако Linked Open Data

З.В. Апанович[1], П.С. Винокуров[1], Т.А. Кислицина[2]

[1]Институт систем информатики СО РАН

[2]НГУ

630090, Новосибирск, проспект Лаврентьева, 6, Россия

apanovich@iis.nsk.su

Что было:

До этого мы уже работали с визуализацией онтологий и информационного наполнения научных порталов (археология, компьютерная лингвистика). Имели несколько удачных результатов вылавливания ошибок проектирования онтологии и ошибок ручного ввода информационного наполнения и при помощи визуализации, в основе которой лежали методы совместного изображения специфических отношений.

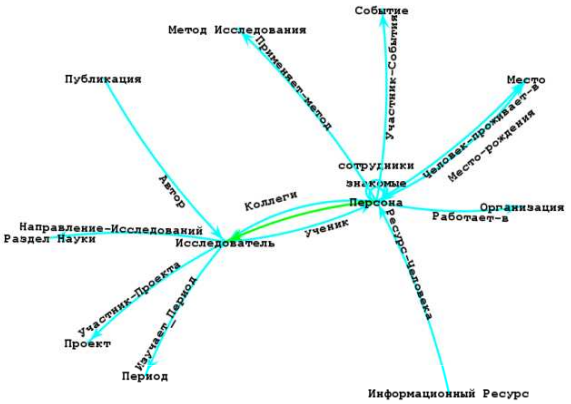
Для поиска ошибок:

1) Совместное изображение отношений наследования и ассоциативных отношений при визуализации онтологий
2) Совместное изображение отношений партономии и ассоциативных отношений для информационного наполнения.

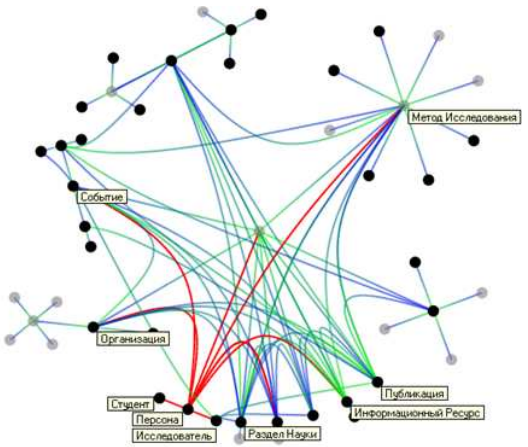
Для наукометрического анализа:

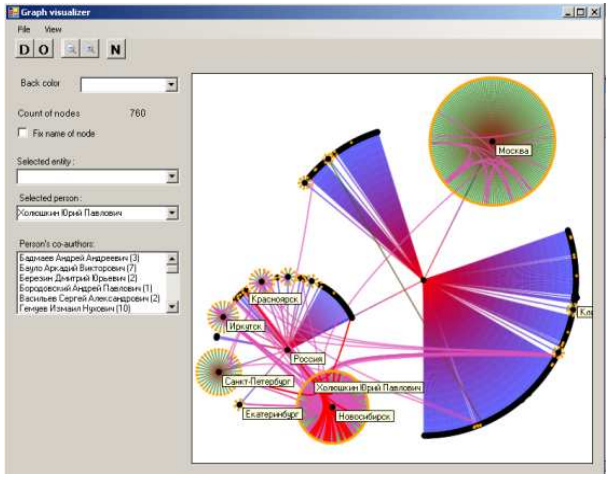
1) Генерация новых отношений (соавторства) и их совместное изображение с различными иерархическими отношениями (отношение партономии) при помощи иерархических жгутов ребер
2) Кластеризация сетей соавторства.

Что было:


Событие
Метод Исследования
Публикация
Место
Участник-Событие
Принимает-метод
Человек-проживает-в
Место-рождения
сотрудники
знакомые
Автор
Коллеги
Персона
Организация
Работает-в
Направление-Исследований
Раздел Науки
Исследователь
ученик
Ресурс-Человека
Участник-Проекта
Изучает_Период
Проект
Период
Информационный Ресурс


Железный век
Бронзовый век
Палеолит
Энеолит
Мезолит
Каменный век
Неолит
Средний палеолит
Верхний палеолит
Античность
Средневековая археология
Нижний палеолит


Метод Исследования
Событие
Организация
Студент
Персона
Исследователь
Раздел Науки
Публикация
Информационный Ресурс

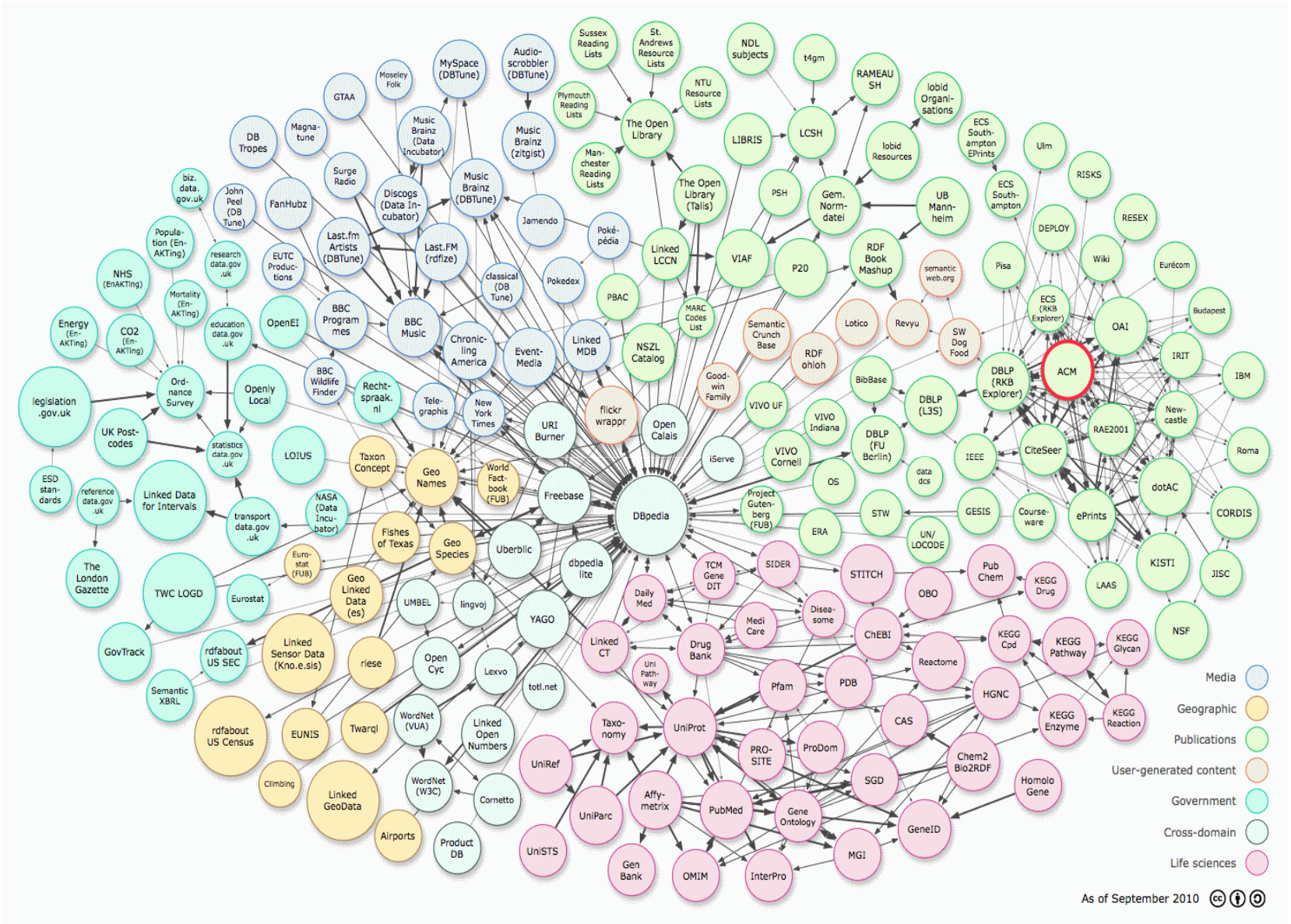

Graph visualizer
File
View
D
O
N
Back color
Count of nodes
760
Fix name of node
Selected entity :
Selected person :
Холошкин Юрий Павлович
Person's co-authors:
Бадмаев Андрей Андреевич (3)
Бакуло Аркадий Викторович (7)
Березин Дмитрий Юрьевич (2)
Бородовский Андрей Павлович (1)
Васильев Сергей Александрович (2)
Генчев Измаил Нуксович (10)
Москва
Красноярск
Иркутск
Россия
Санкт-Петербург
Холошкин Юрий Павлович
Екатеринбург
Новосибирск

Данная работа преследовала две цели:

- 1) При помощи визуализации познакомиться поближе с данными из облака Open Linked Data.
- 2) Опробовать старые и новые методы визуализации на общеизвестных данных в стандартных форматах (RDF/ OWL)достаточно большого объема

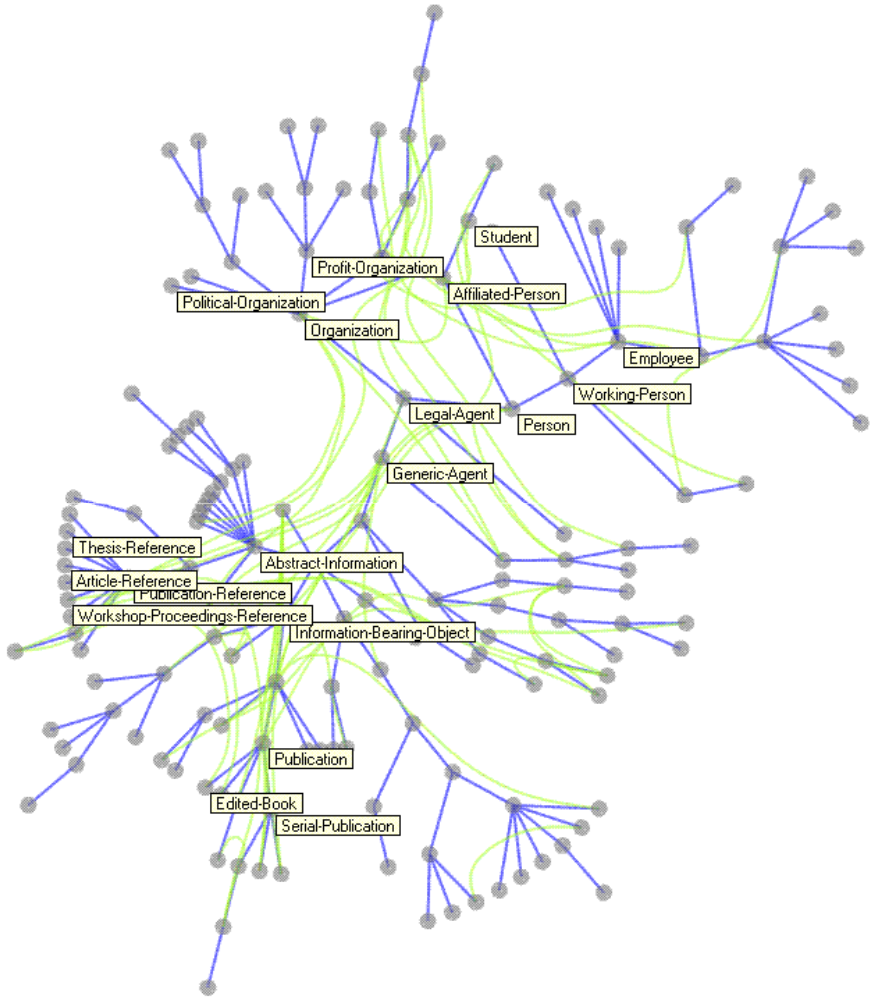
  - визуализация сетей соавторства

  - визуализация сетей цитирования


Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2010

Citeseer, ACM, DBLP…

- Данные предоставляются в формате RDF и имеют весьма внушительные объемы.
- Например, RDF-данные, предоставленные порталом Citeseer содержат 8 146 852 троек RDF,
- данные портала ACM насчитывают 12,402,336 троек RDF,
- портал DBLP предоставил 28 384 790 троек RDF. Пользователь может либо скачивать файлы в формате RDF, либо генерировать данные при помощи запросов sparql.

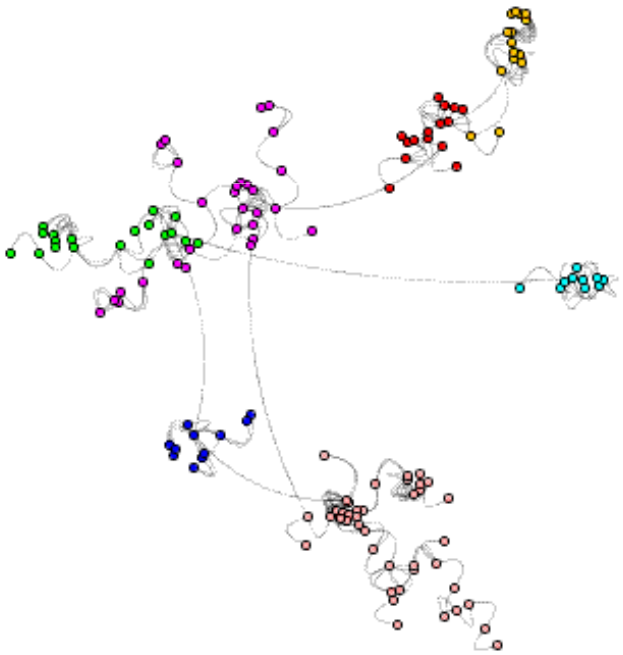


AKT Reference ontology =
Support ontology + portal ontology + Extensions ontology + RDF compatibility ontology

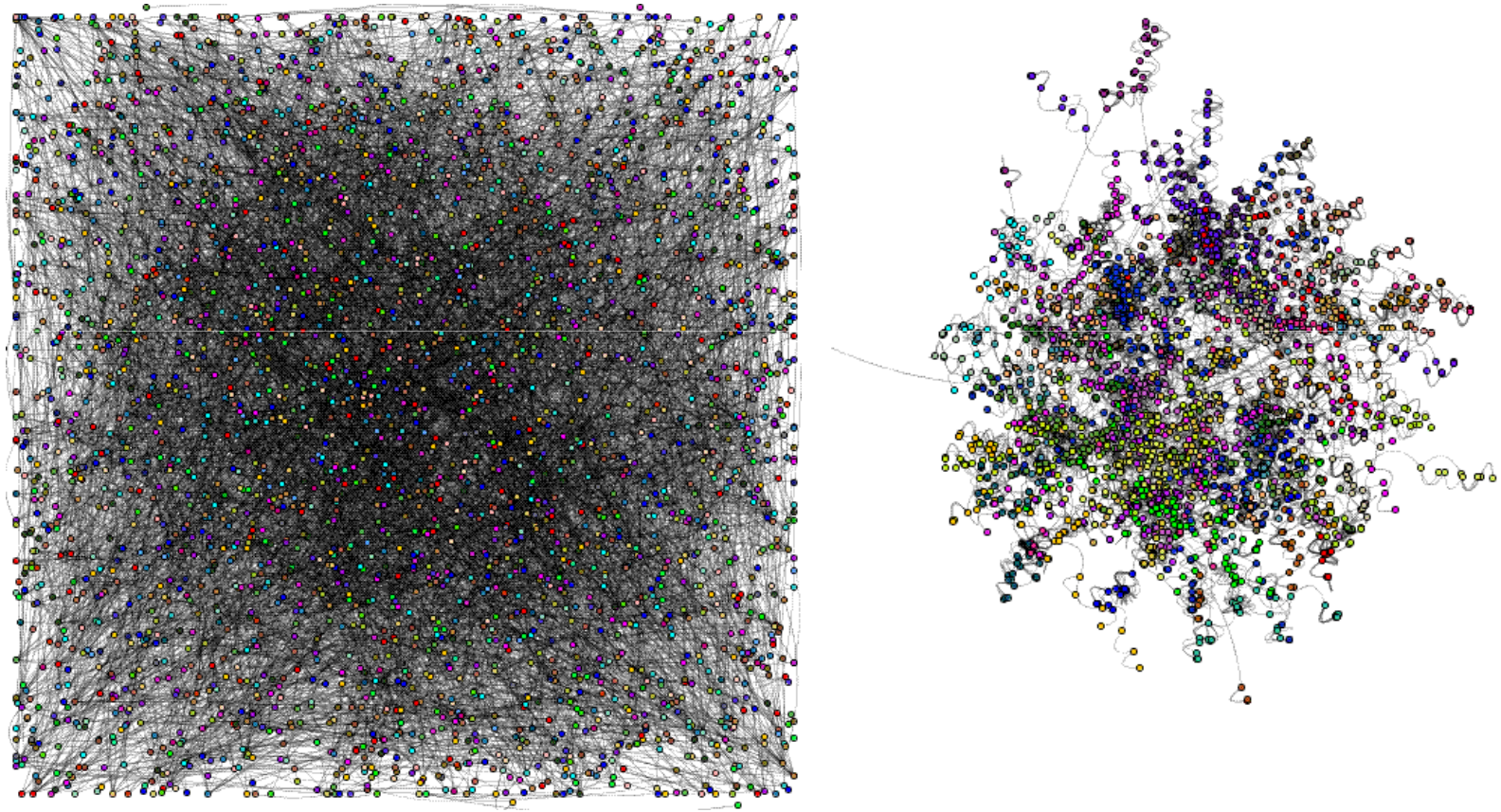
http://www.aktors.org/ontology/portal

Генерация сети соавторства

- CONSTRUCT{?y :co_author ?z}
- WHERE{
- ?x akt:has-author ?y ;
- akt:has-author ?z ;
- a ?type .
- FILTER(?y != ?z &&(?type = akt :Publication-Reference) }
- LIMIT N.



Следует сказать, что при таком способе генерации сетей соавторства их связность и плотность напрямую связаны с объемом. Например, для портала DBLP при установке лимита на количество ребер в сети соавторства, равном десяти тысячам, наибольшая связная компонента этой сети имеет всего 140 вершин и 191 ребро, 7 научных сообществ, показаны разными цветами



При возрастании лимита на объем сети до 50000 ребер, наибольшая связная компонента имеет уже 3001 вершину и 4983 ребра.

Модулярность является свойством сети и оценивает качество разбиения сети на сообщества. [Newman M. E. J., Girvan M. Finding and evaluating community structure in networks// *Physical Review E*, 69.—26113.— 2004.]

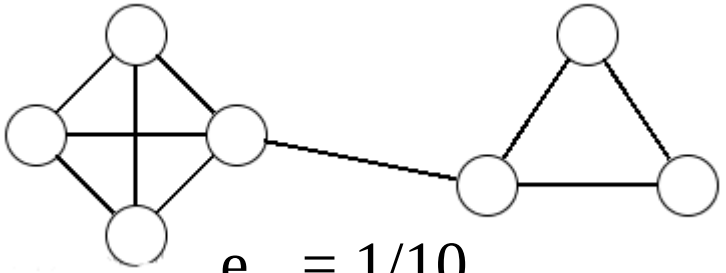
Предположим, что вершины графа сотрудничества разбиты на сообщества, c_i.

e_{ij} - доля всех ребер, соединяющих сообщество c_i и сообщество c_j

$a_i = \Sigma_j e_{ij}$ - доля всех ребер, связанных с вершинами сообщества c_i.

Модулярность выражается через a_i *и* e_{ij} следующим образом:

$$Q = \sum_i (e_{ii} - a_i)$$



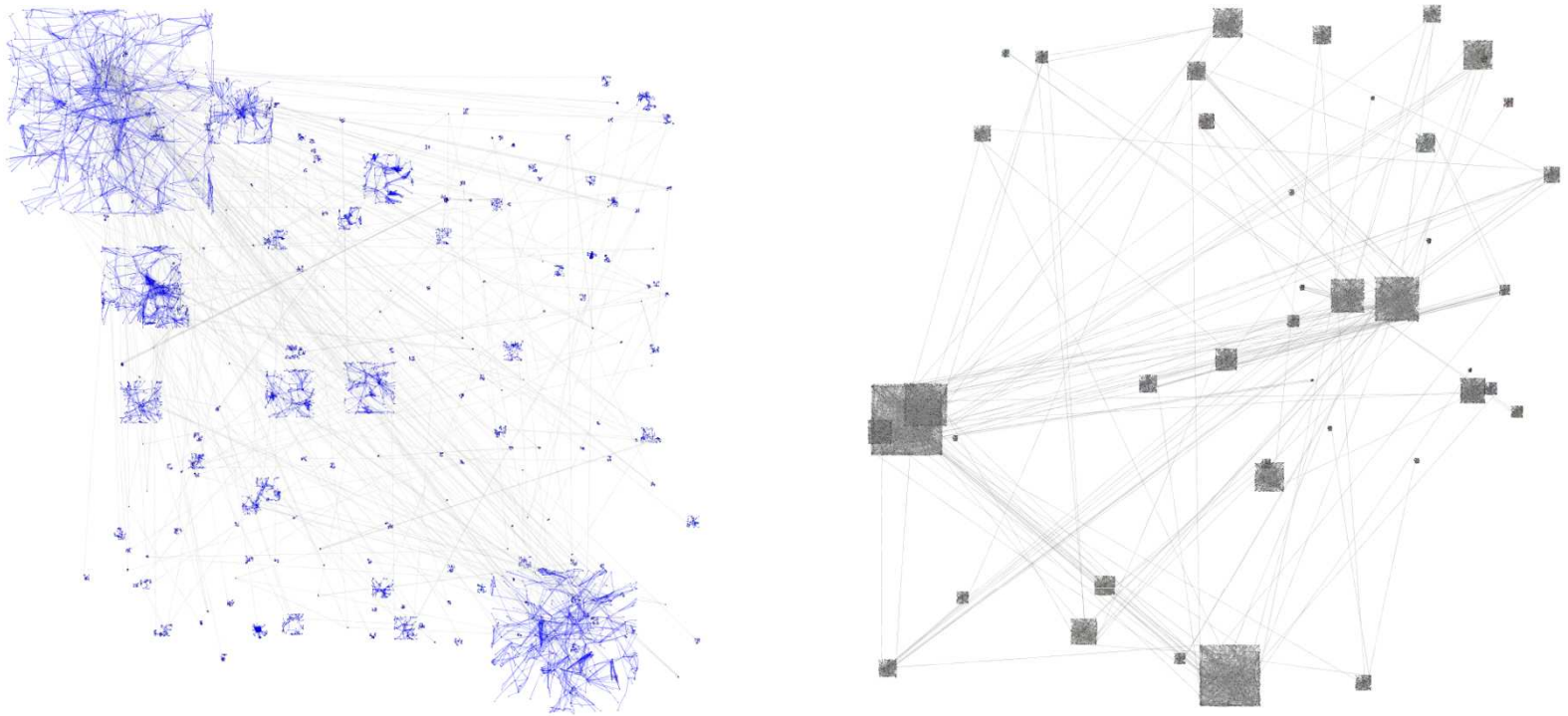
C_1 C_2

$e_{12} = 1/10$,
$e_{11} = 6/10$,
$e_{22} = 3/10$,
$a_1 = 7/10$,
$a_2 = 4/10$
$Q = 41/100$

Экспериментально обнаружено, что значения модулярности, превышающие 0.3, являются указателем на факт наличия сообществ в сети.

Новая реализация:

1) Начальное состояние- каждая вершина является отдельным кластером, объединяются два кластера, дающие наибольшее улучшение модулярности
2) Итеративное улучшение
3) Многоуровневость



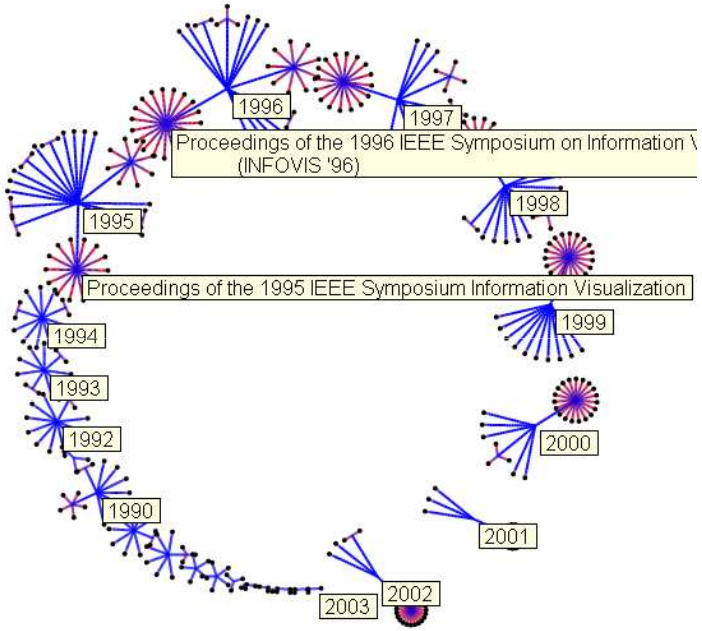
(а) (б)

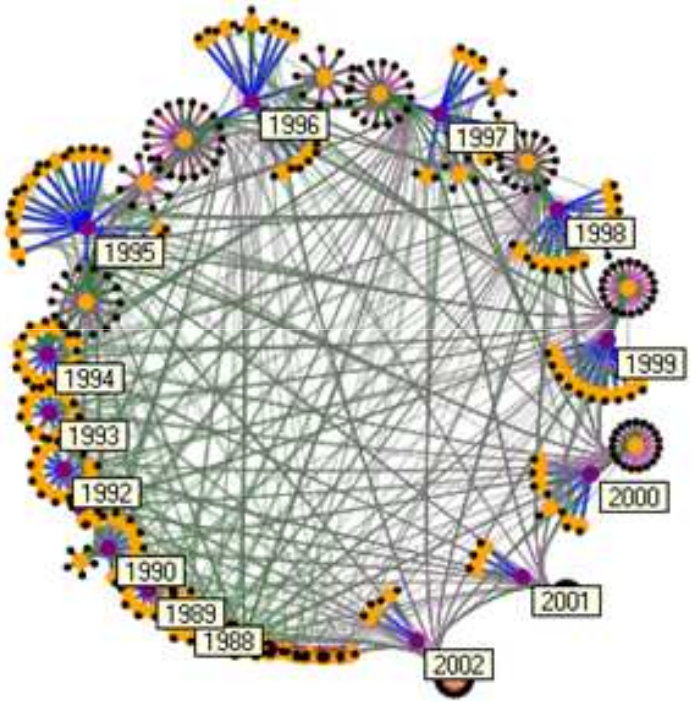
(а) разбиение на сообщества прежним алгоритмом кластеризации (количество вершин 5625, количество ребер 10103, модулярность 0.922, 197 сообществ. (б) разбиение на сообщества той же самой сети многоуровневым алгоритмом (48 сообществ, Модулярность 0.948) .

Методы визуализации сетей цитирования

- 1. Построение списков цитируемой литературы для каждой публикации требует гораздо больших технических усилий, поэтому в открытом доступе эта информация предоставляется только небольшим количеством порталов. Среди порталов облака LOD мы обнаружили эти данные для CiteSeer и ACM.
- 2 . Для генерации информативных сетей цитирования нужны дополнительные усилия. В случае портала CiteSeer нами применялась многоуровневая схема генерации сетей цитирования, а в случае портала ACM дополнительно использовалась собственная онтология этого портала, позволяющая выбирать публикации, относящиеся к определенному разделу науки.

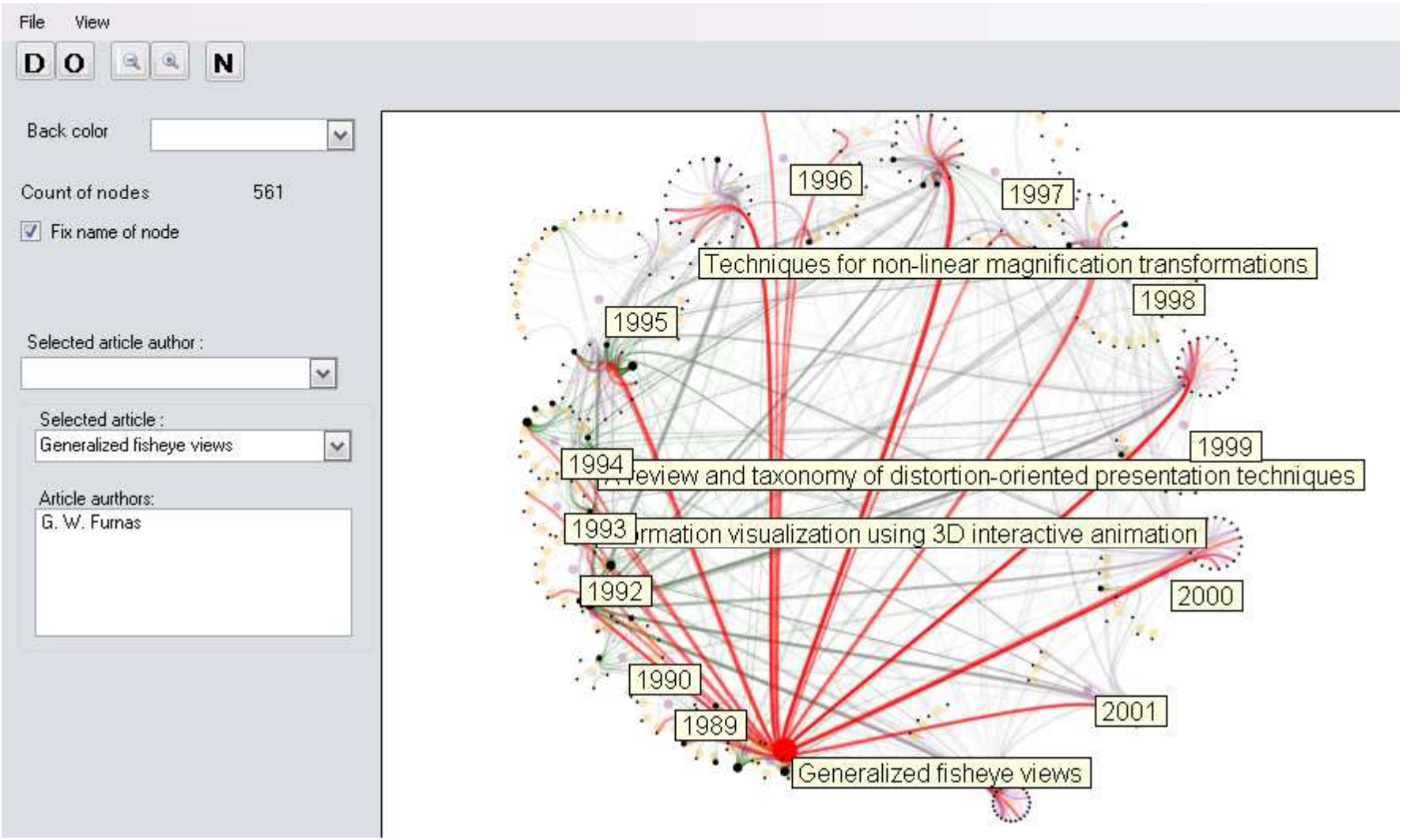
Визуализация сетей цитирования при помощи иерархических жгутов ребер





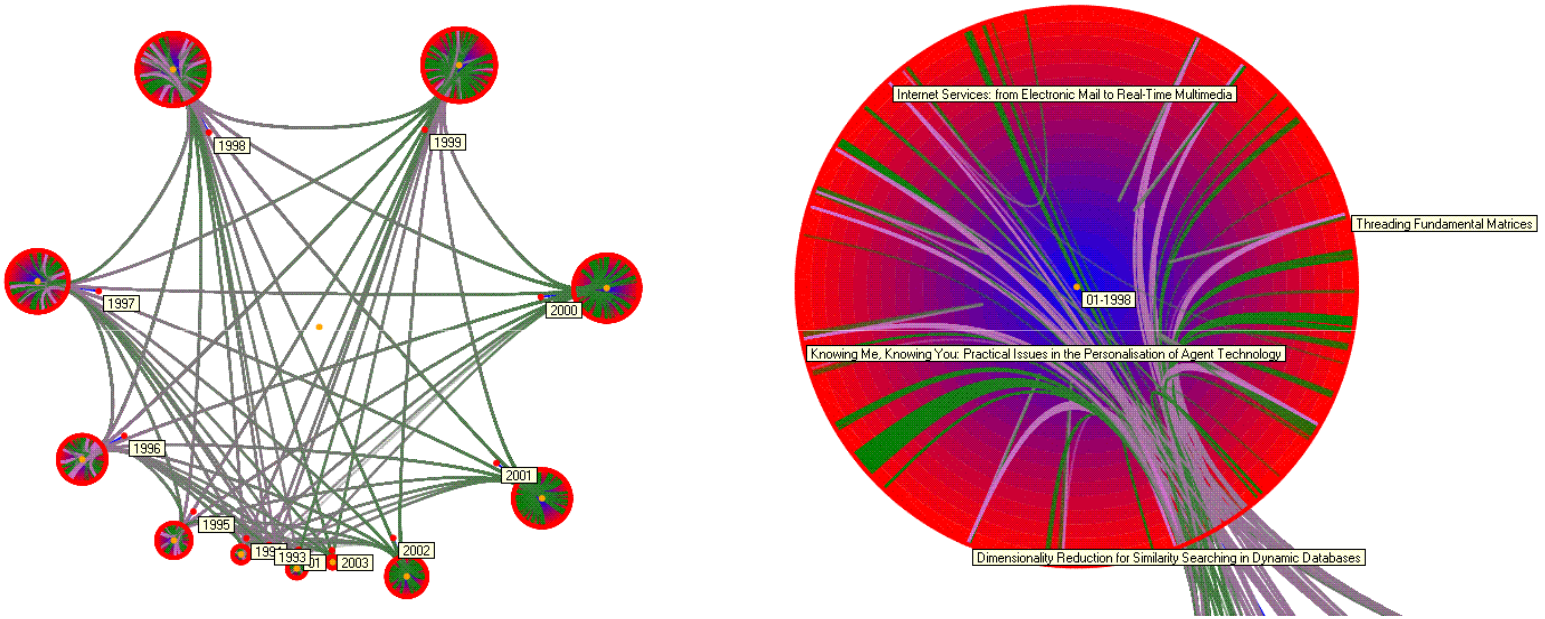
Проблемы:

1) Отсутствие иерархии, на которую можно натянуть жгуты
2) Неестественное изображение для ориентированного графа
3) Визуальная перегруженность



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

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

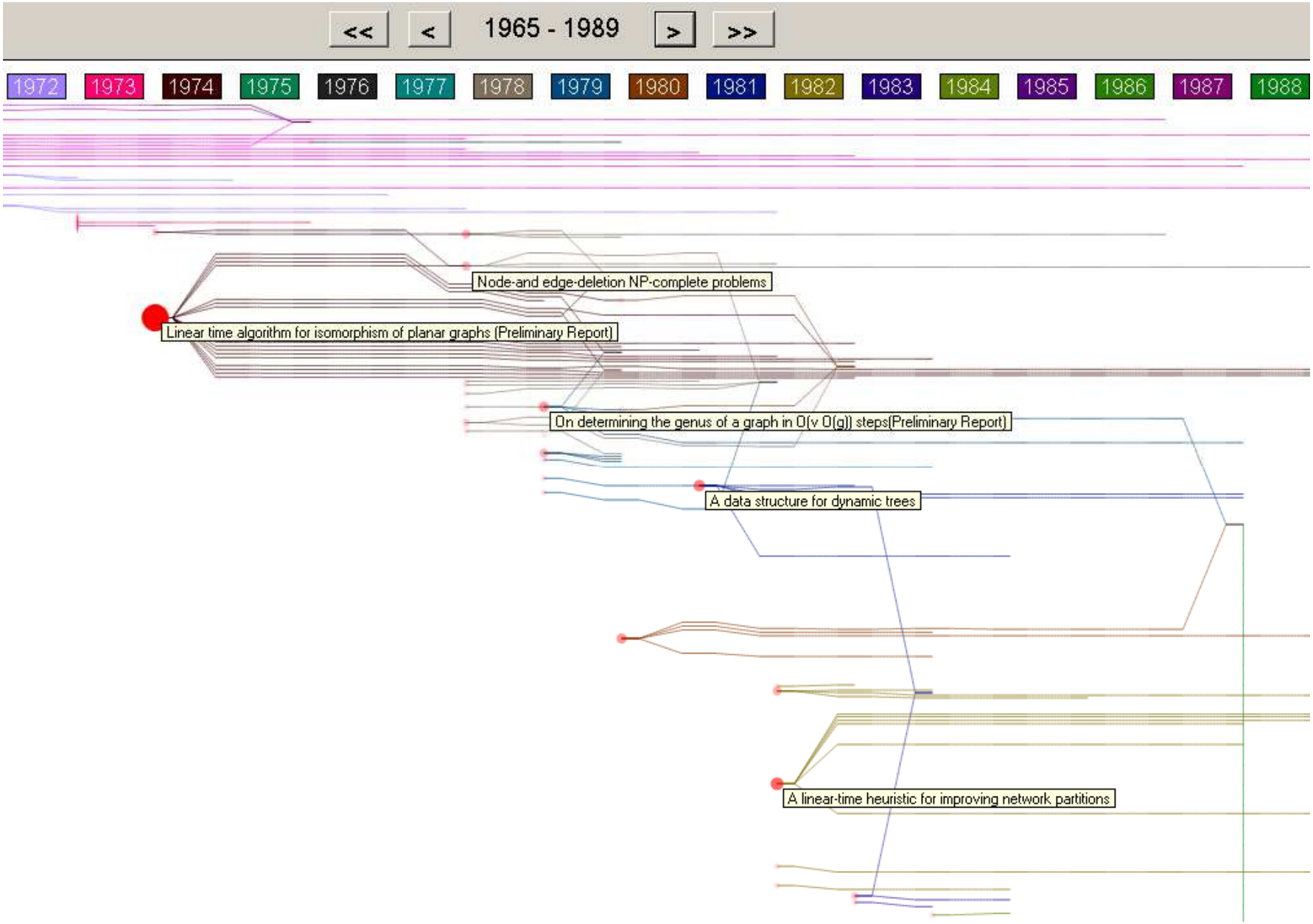


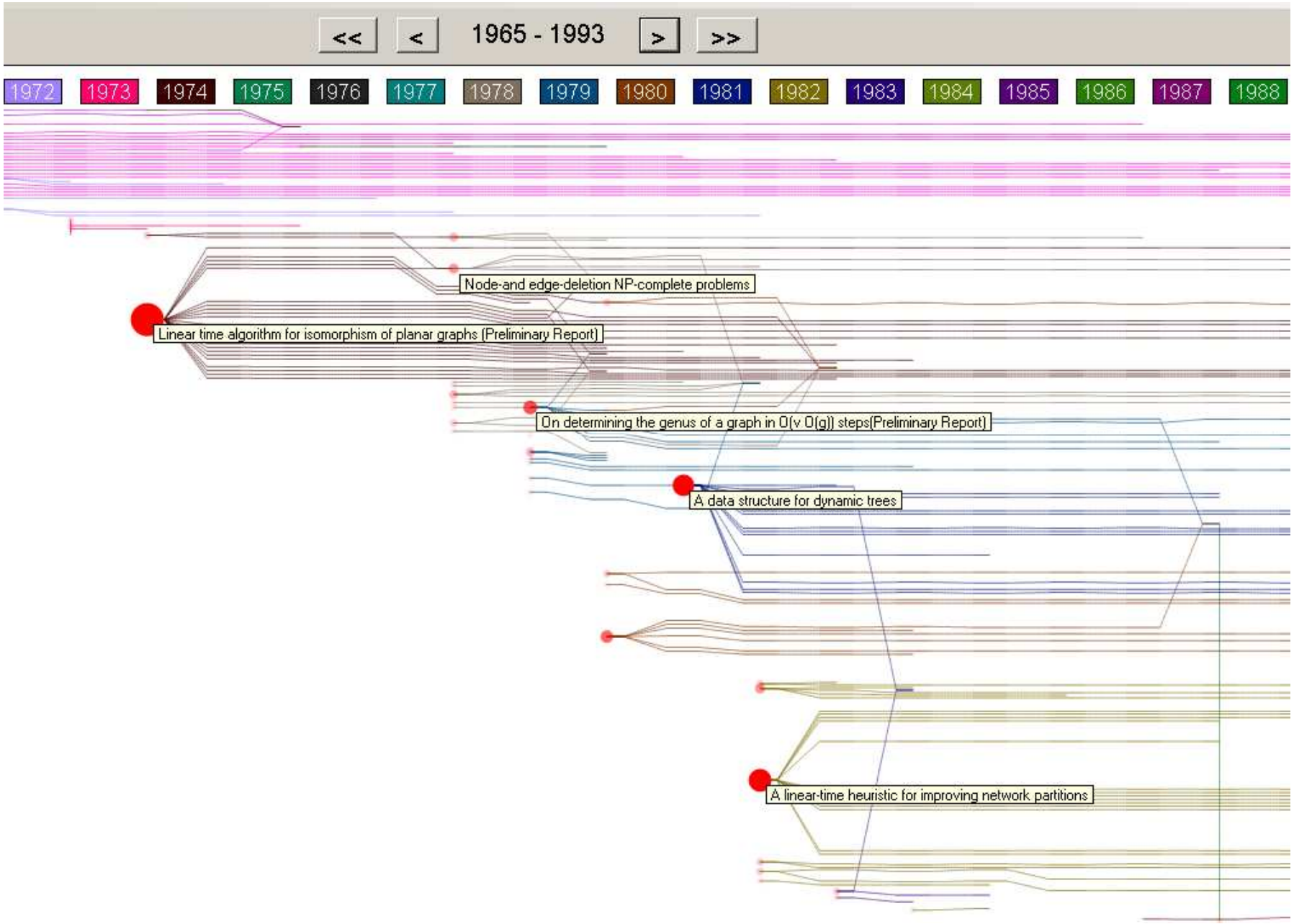
– Изображение сети цитирования, извлеченной из RDF-данных портала Citeseer и содержащей 20 000 вершин. (а) общий план изображения, (б) публикации за один месяц 1998 года

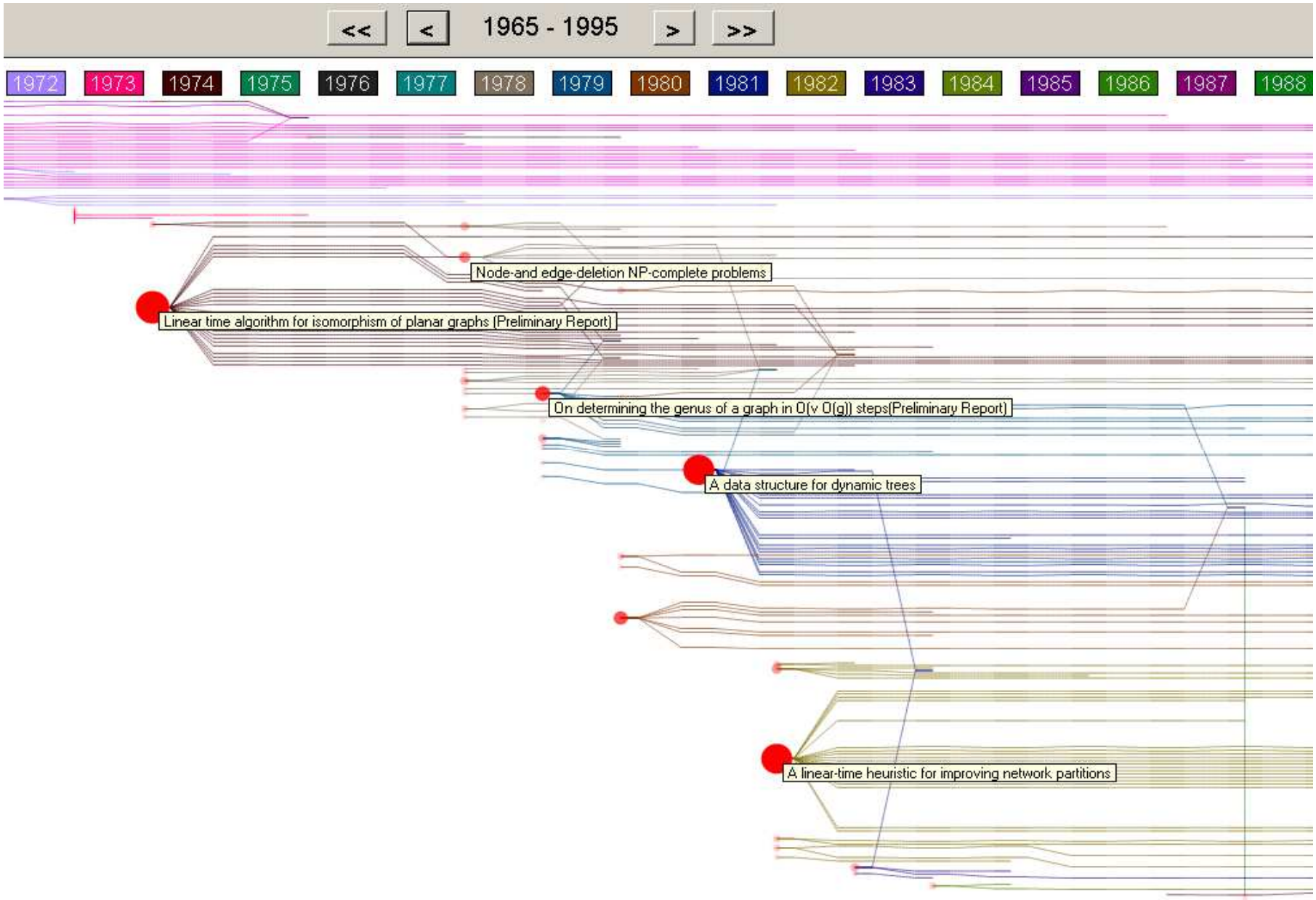
Остаются проблемы

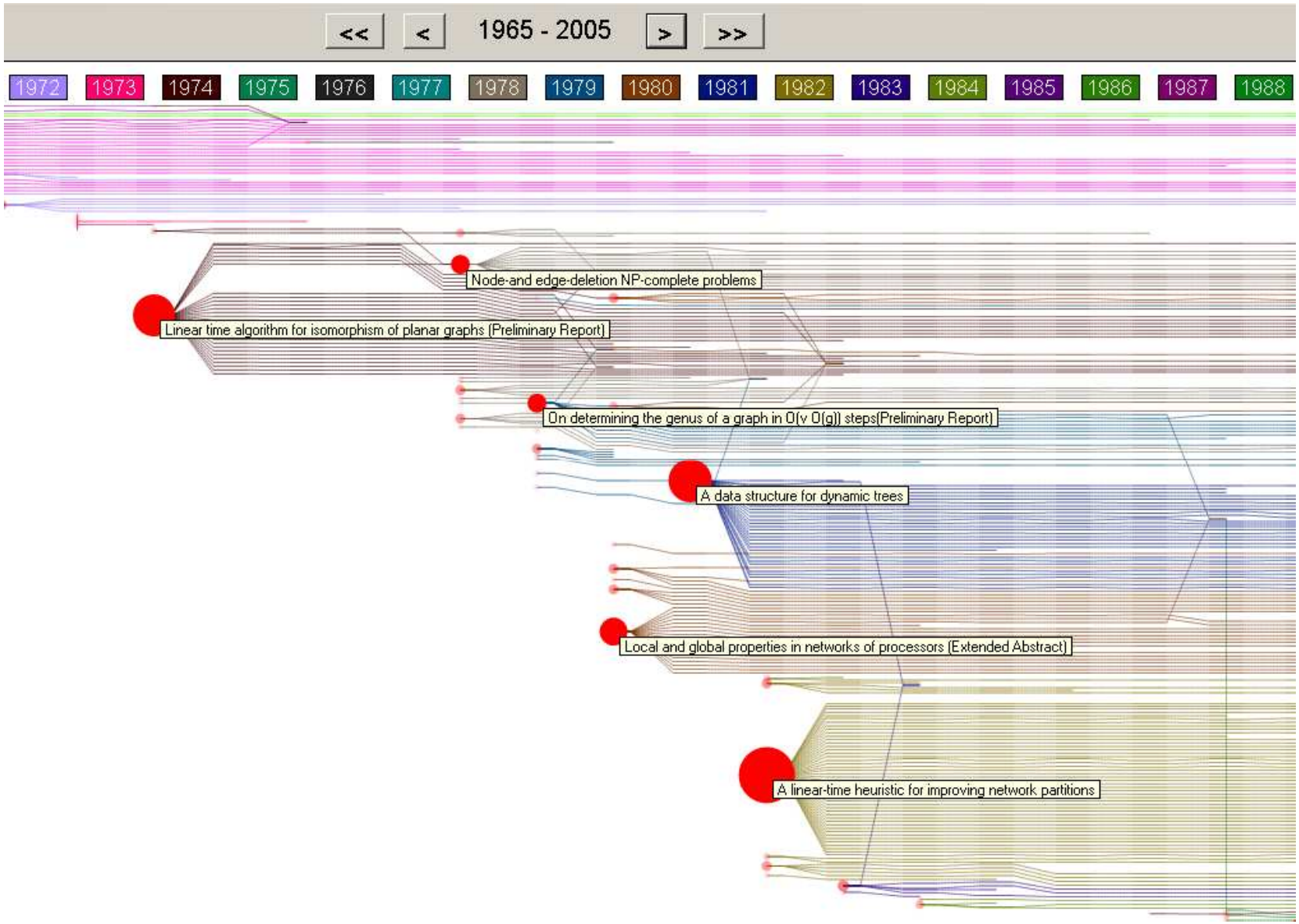
- 1) Визуальная перегруженность
- 2) Неестественное изображение хронологических данных

- Поэтому для визуализации сетей цитирования был реализован динамический алгоритм поуровневой визуализации

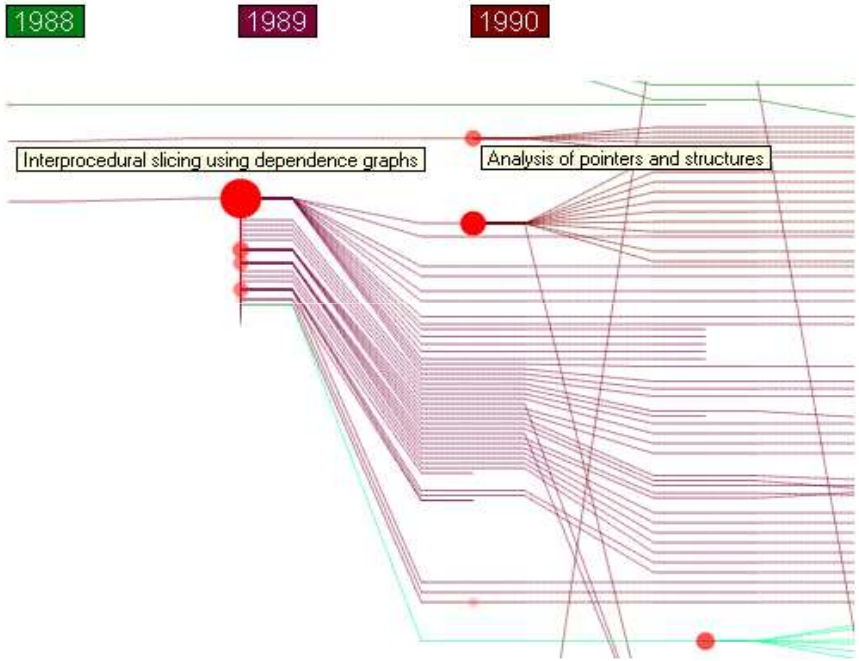

<<
<
1965 - 1989
>
>>
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
Node-and edge-deletion NP-complete problems
Linear time algorithm for isomorphism of planar graphs (Preliminary Report)
On determining the genus of a graph in O(v O(g)) steps(Preliminary Report)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


<<
<
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


<<
<
1965 - 1995
>
>>
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


<<
<
1965 - 2005
>
>>
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
Local and global properties in networks of processors (Extended Abstract)
A linear-time heuristic for improving network partitions

Данные портала Citeseer

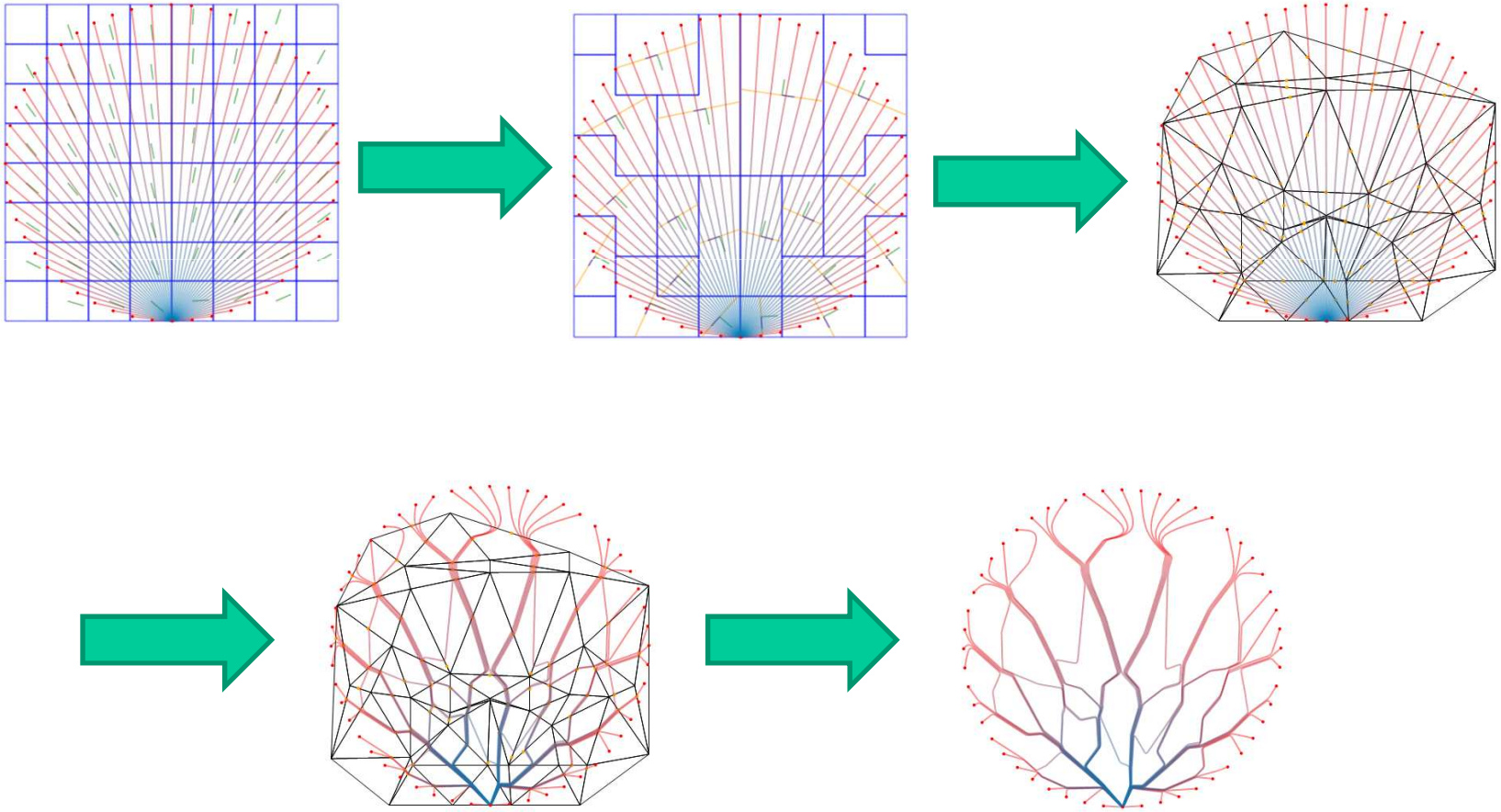


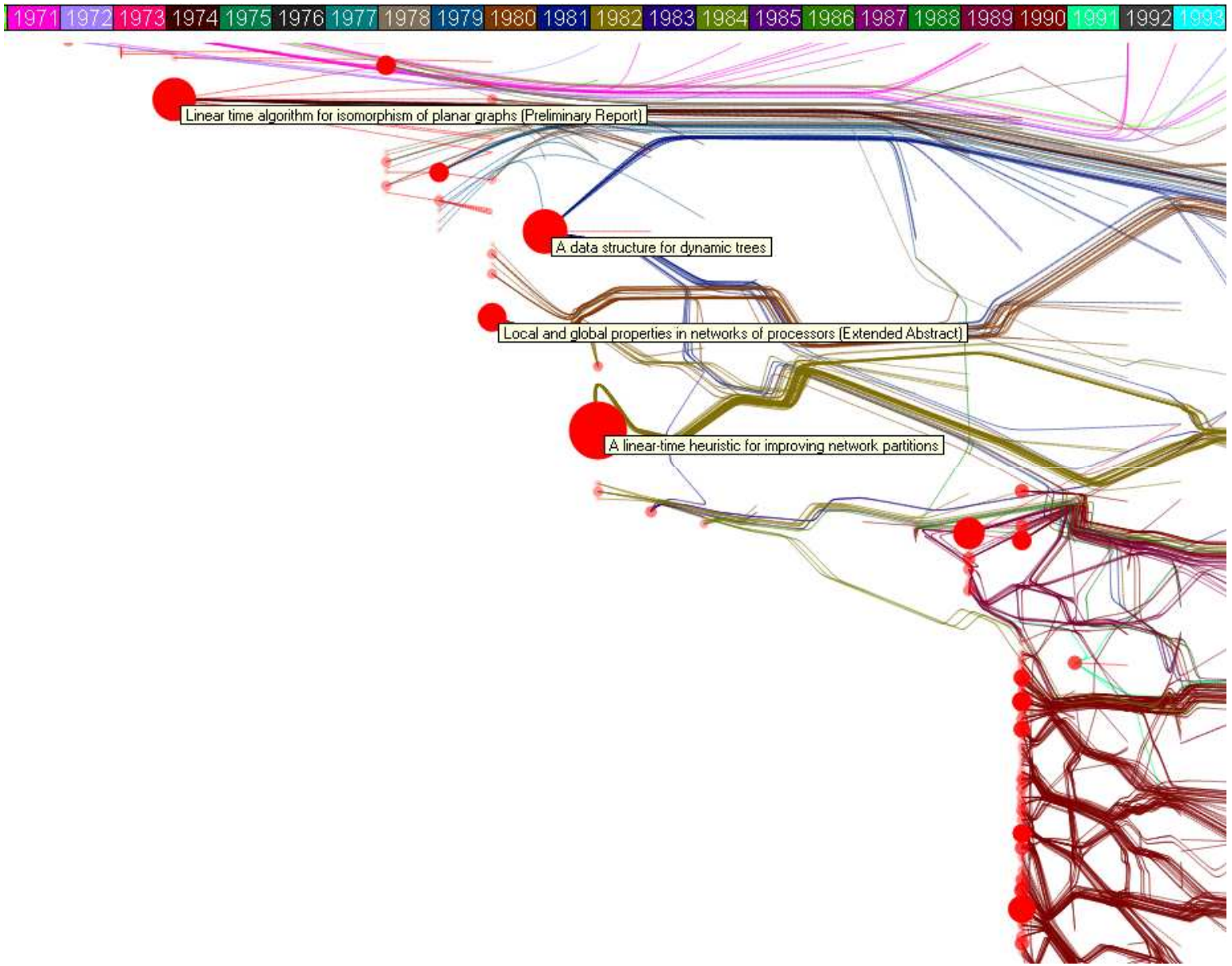
Analysis of pointers and structures	akt:has-date	1990-01-01
Interprocedural slicing using dependence graphs	akt:cites-publication-reference	Analysis of pointers and structures
Interprocedural slicing using dependence graphs	akt:has-date	1988-01-01
Interprocedural slicing using dependence graphs	akt:has-date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

Остается проблема визуальной перегруженности:

- Фильтрация ребер нарушает соответствие реальности
- Для построения жгутов не хватает иерархии
- Решение: построение жгутов на основе собственной геометрии ребер

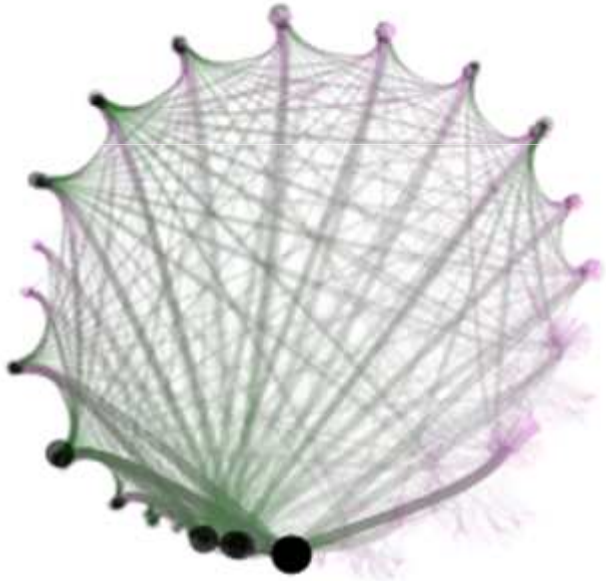
Жгуты на основе геометрии ребер



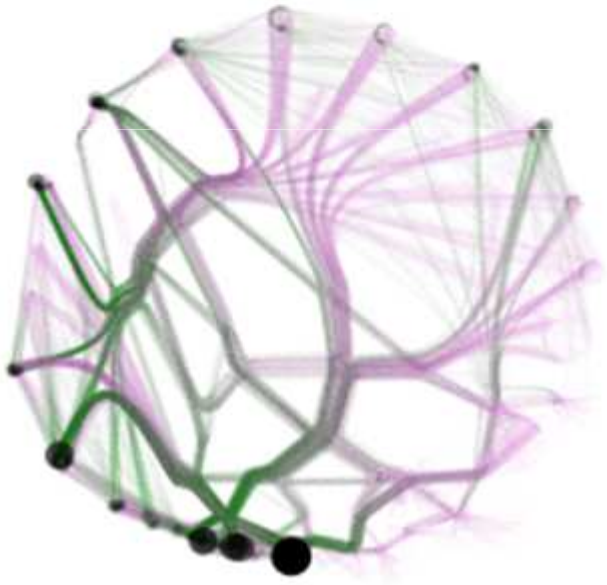

1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993
Linear time algorithm for isomorphism of planar graphs (Preliminary Report)
A data structure for dynamic trees
Local and global properties in networks of processors (Extended Abstract)
A linear-time heuristic for improving network partitions

Иерархические и геометрические жгуты

- 3000 публикаций



ИРР

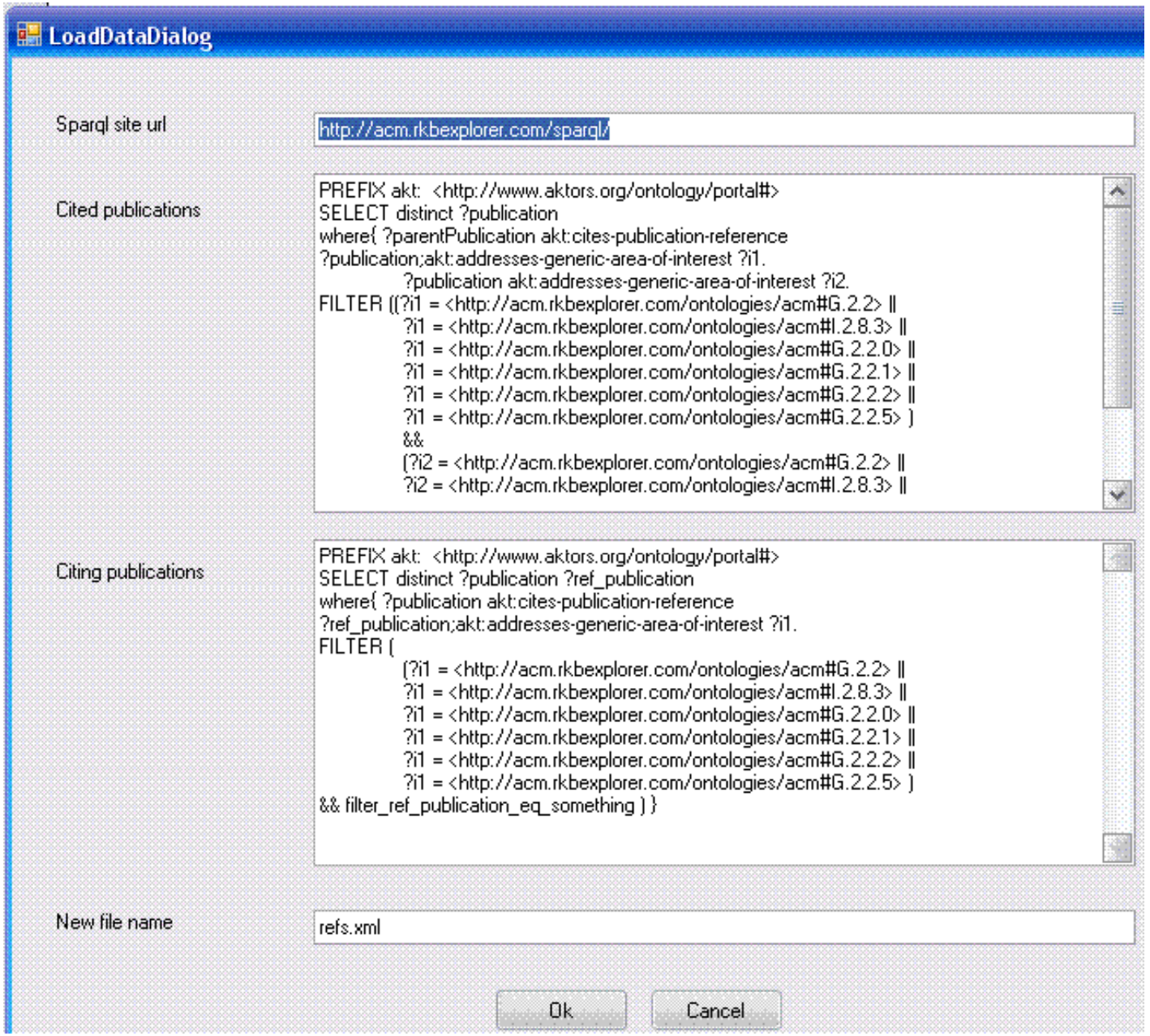


ГРР

Дальнейшие планы

- Дальнейшее исследование метода построения жгутов на основе геометрии от различных параметров.
- Потестировать эти методы на российских источниках больших данных для определения наиболее полезных направлений развития.

- **СПАСИБО ЗА ВНИМАНИЕ!**


LoadDataDialog
Sparql site url
http://acm.rkbexplorer.com/sparql/
Cited publications
PREFIX akt: <http://www.aktors.org/ontology/portal#>
SELECT distinct ?publication
where{ ?parentPublication akt:cites-publication-reference
?publication;akt:addresses-generic-area-of-interest ?i1.
?publication akt:addresses-generic-area-of-interest ?i2.
FILTER ((?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#I.2.8.3> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2.0> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2.1> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2.2> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2.5>)
&&
(?i2 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2> ||
?i2 = <http://acm.rkbexplorer.com/ontologies/acm#I.2.8.3> ||
Citing publications
PREFIX akt: <http://www.aktors.org/ontology/portal#>
SELECT distinct ?publication ?ref_publication
where{ ?publication akt:cites-publication-reference
?ref_publication;akt:addresses-generic-area-of-interest ?i1.
FILTER (
(?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#I.2.8.3> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2.0> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2.1> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2.2> ||
?i1 = <http://acm.rkbexplorer.com/ontologies/acm#G.2.2.5>)
&& filter_ref_publication_eq_something) }
New file name
refs.xml
Ok
Cancel