Academic Recommendation using Citation Analysis with the**advisor**

Erik Saule
joint work with Onur Küçüktunç, Kamer Kaya, Ümit V. Çatalyürek

esaule@bmi.osu.edu

Department of Biomedical Informatics
The Ohio State University

CSTA 2013

Table of Contents

Once upon a time : a survey paper

The Jimmy John's scheduling problem

scheduling partitioning mapping	$\times$	pipeline workflow data flow task graph	$\times$	linear chain sequences (tree) (serial parallel)

Once upon a time : a survey paper

The Jimmy John's scheduling problem

scheduling partitioning mapping	$\times$	pipeline workflow data flow task graph	$\times$	linear chain sequences (tree) (serial parallel)

But also...

"Scheduling problems in parallel query optimization"
"Bringing skeletons out of the closet: A pragmatic manifesto for skeletal parallel programming"

Once upon a time : a survey paper

The Jimmy John's scheduling problem

scheduling partitioning mapping	$\times$	pipeline workflow data flow task graph	$\times$	linear chain sequences (tree) (serial parallel)

But also...

"Scheduling problems in parallel query optimization"
"Bringing skeletons out of the closet: A pragmatic manifesto for skeletal parallel programming"

After 6 months, unknown papers where still uncovered

Once upon a time : a survey paper

The Jimmy John's scheduling problem

scheduling partitioning mapping	$\times$	pipeline workflow data flow task graph	$\times$	linear chain sequences (tree) (serial parallel)

But also...

"Scheduling problems in parallel query optimization"
"Bringing skeletons out of the closet: A pragmatic manifesto for skeletal parallel programming"

After 6 months, unknown papers where still uncovered

Develop software to make the search easier!

Design Goals

Personalized

The user should be able to make a query that describes precisely what she is looking for.

Conceptual

The system should free of linguistic problems. Ambiguity and synonymy should be taken into accounts.

Exploratory

Different perspective should be available. The system should enhance the user's search.

Easy to use

The user should not need to know anything about data mining or algorithms.

The Academic Web Service Ecosystem

DBLP

List of CS papers with clean reference and disambiguated names.

The Academic Web Service Ecosystem

DBLP

List of CS papers with clean reference and disambiguated names.

Citeseer, {Ref,Ack,Collab}Seer

Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

The Academic Web Service Ecosystem

DBLP

List of CS papers with clean reference and disambiguated names.

Citeseer, {Ref,Ack,Collab}Seer

Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

CiteUlike

Social paper tagging application. Find paper from researchers with similar interest.

The Academic Web Service Ecosystem

DBLP

List of CS papers with clean reference and disambiguated names.

Citeseer, {Ref,Ack,Collab}Seer

Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

CiteUlike

Social paper tagging application. Find paper from researchers with similar interest.

ArnetMiner

Academic network analysis.

The Academic Web Service Ecosystem

DBLP

List of CS papers with clean reference and disambiguated names.

Mendeley

Application for managing references. Database of reference.

Citeseer, {Ref,Ack,Collab}Seer

Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

CiteUlike

Social paper tagging application. Find paper from researchers with similar interest.

ArnetMiner

Academic network analysis.

The Academic Web Service Ecosystem

DBLP

List of CS papers with clean reference and disambiguated names.

Mendeley

Application for managing references. Database of reference.

Citeseer, {Ref,Ack,Collab}Seer

Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

Google Scholar

Keyword-based search engine (with citation informations).

CiteUlike

Social paper tagging application. Find paper from researchers with similar interest.

ArnetMiner

Academic network analysis.

The Academic Web Service Ecosystem

DBLP

List of CS papers with clean reference and disambiguated names.

Mendeley

Application for managing references. Database of reference.

Citeseer, {Ref,Ack,Collab}Seer

Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

Google Scholar

Keyword-based search engine (with citation informations).

CiteUlike

Social paper tagging application. Find paper from researchers with similar interest.

Microsoft Academic Search

Keyword-based search engine and Academic network analysis.

ArnetMiner

Academic network analysis.

The Academic Web Service Ecosystem

DBLP

List of CS papers with clean reference and disambiguated names.

Mendeley

Application for managing references. Database of reference.

Citeseer, {Ref,Ack,Collab}Seer

Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

Google Scholar

Keyword-based search engine (with citation informations).

CiteUlike

Social paper tagging application. Find paper from researchers with similar interest.

Microsoft Academic Search

Keyword-based search engine and Academic network analysis.

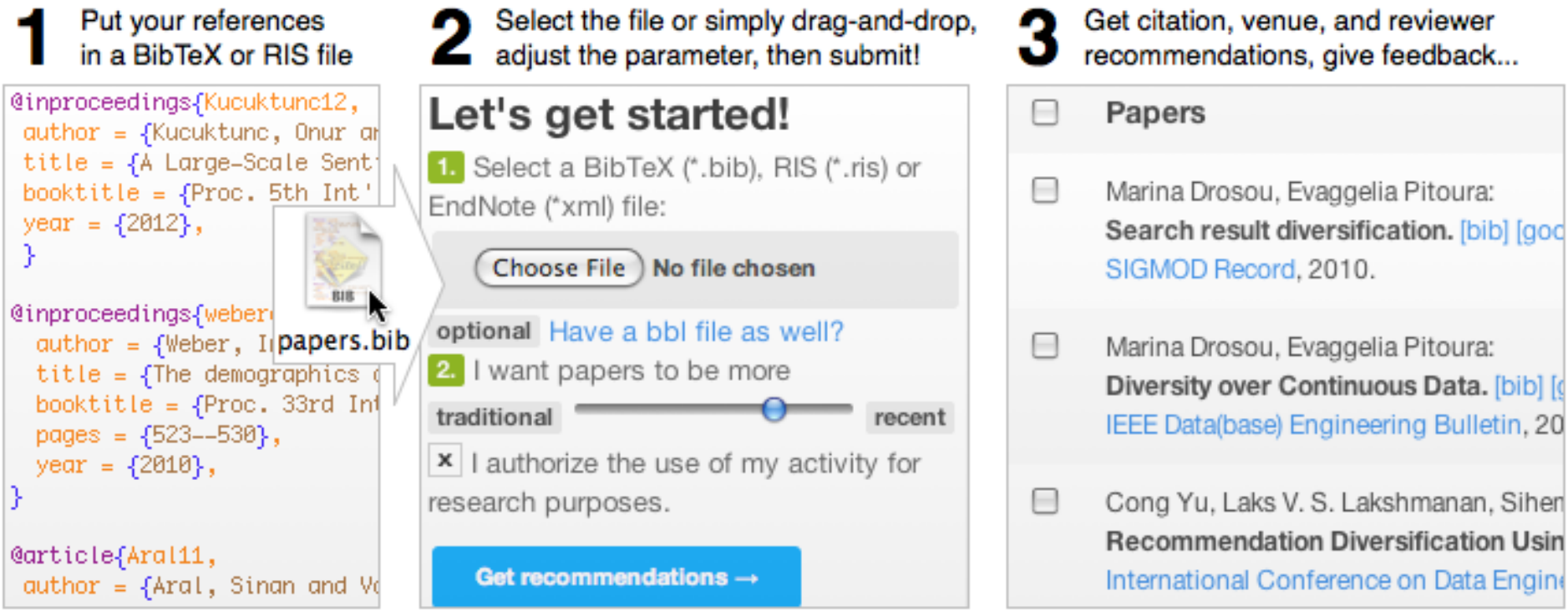
ArnetMiner

Academic network analysis.

IEEE, ACM, Elsevier, JSTOR, ...

Publishers or digital libraries with complete text and references. Some suggestions.

A Use Case

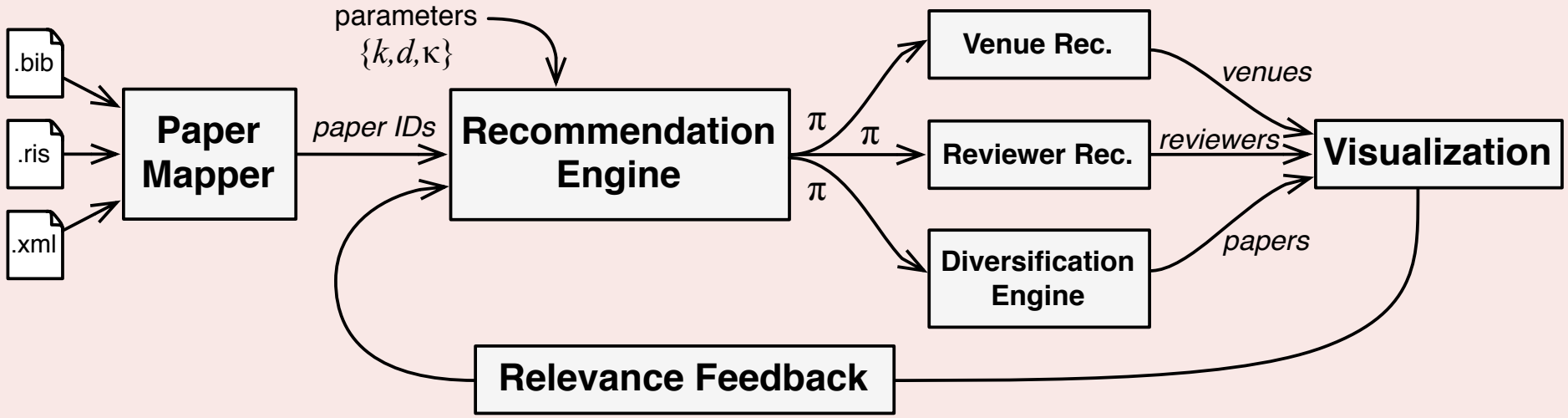

1 Put your references in a BibTeX or RIS file
@inproceedings{Kucuktunc12,
author = {Kucuktunc, Onur a
title = {A Large-Scale Sent
booktitle = {Proc. 5th Int'
year = {2012},
}
@inproceedings{weber
author = {Weber, I
title = {The demographics
booktitle = {Proc. 33rd Int
pages = {523--530},
year = {2010},
}
@article{Aral11,
author = {Aral, Sinan and V
papers.bib
2 Select the file or simply drag-and-drop, adjust the parameter, then submit!
Let's get started!
1. Select a BibTeX (*.bib), RIS (*.ris) or EndNote (*xml) file:
Choose File
No file chosen
optional
Have a bbl file as well?
2. I want papers to be more
traditional
recent
I authorize the use of my activity for research purposes.
Get recommendations →
3 Get citation, venue, and reviewer recommendations, give feedback...
Papers
Marina Drosou, Evaggelia Pitoura:
Search result diversification. [bib] [goo
SIGMOD Record, 2010.
Marina Drosou, Evaggelia Pitoura:
Diversity over Continuous Data. [bib] [
IEEE Data(base) Engineering Bulletin, 20
Cong Yu, Laks V. S. Lakshmanan, Sihem
Recommendation Diversification Usin
International Conference on Data Engine

System Overview

Architecture

A web-server as a front end. A cluster in the back-end. New instances are dynamically created as the load varies.

Functional



Outline

Using the Citation Graph

Hypothesis

If two papers are related or treat the same subject, then they will be close to each other in the citation graph (and reciprocal)

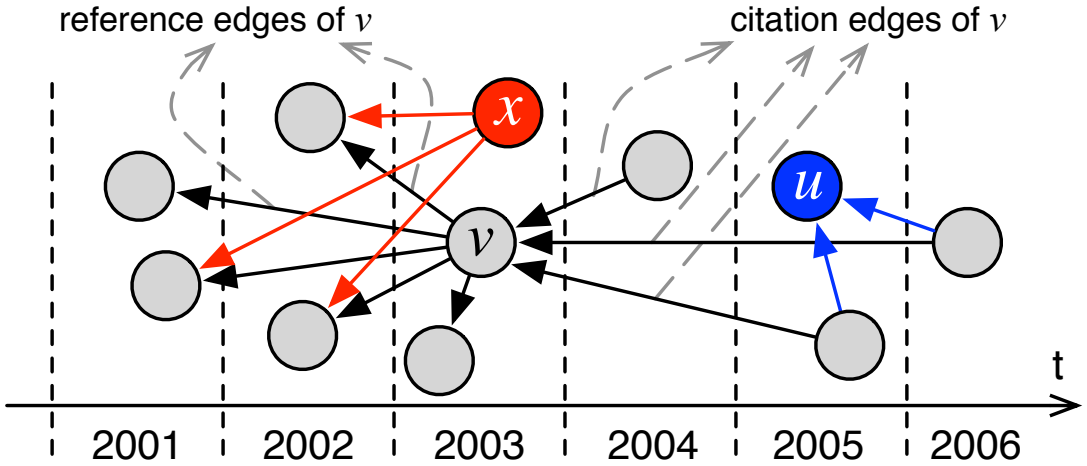
Benefits

- No linguistic => no synonymy, no ambiguity
- Automatically crowd source by researchers

Drawbacks

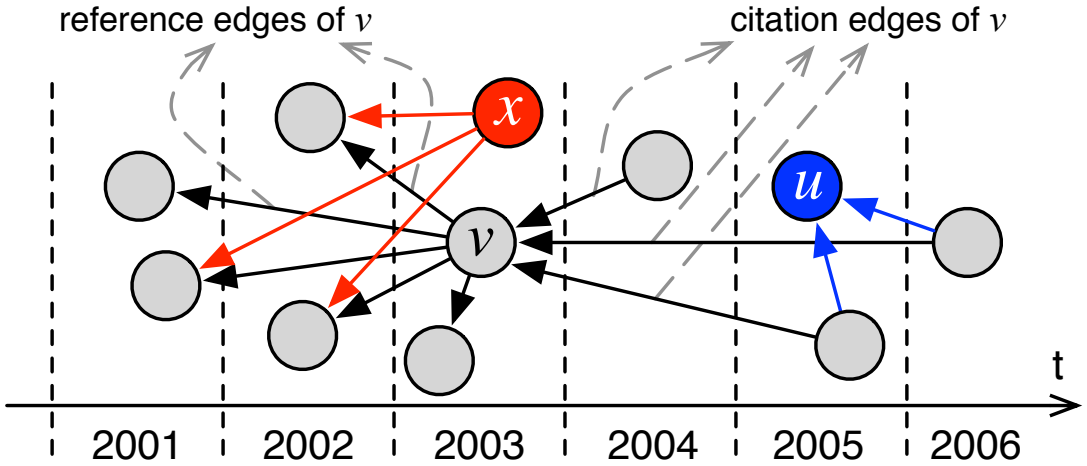
- Difficult to gather the data (But thanks Citeseer)
- Relies on researcher already having made similar connections

Local Approaches



- **Bibliographic coupling [Kessler63]**: papers having similar references are related
- **Cocitation [Small73]**: papers which are cited by the same papers are related
- **CCIDF [Lawrence99]**: cocitations weighted with inverse frequencies

Local Approaches



- **Bibliographic coupling [Kessler63]**: papers having similar references are related
- **Cocitation [Small73]**: papers which are cited by the same papers are related
- **CCIDF [Lawrence99]**: cocitations weighted with inverse frequencies

Problem: Considers only level-2 papers based on level-1 information.

Global Approaches

Graph distance-based

- Katz: number of paths between two papers [Strohman07]

Random walk with restarts (RWR) based

- ArticleRank [Li09] (PageRank [Brin98] extension)
- PaperRank [Gori06] (Personalized PageRank [Haveliwala02] extension)

RWR treats the citations and references in the same way

Global Approaches

Graph distance-based

- Katz: number of paths between two papers [Strohman07]

Random walk with restarts (RWR) based

- ArticleRank [Li09] (PageRank [Brin98] extension)
- PaperRank [Gori06] (Personalized PageRank [Haveliwala02] extension)

RWR treats the citations and references in the same way

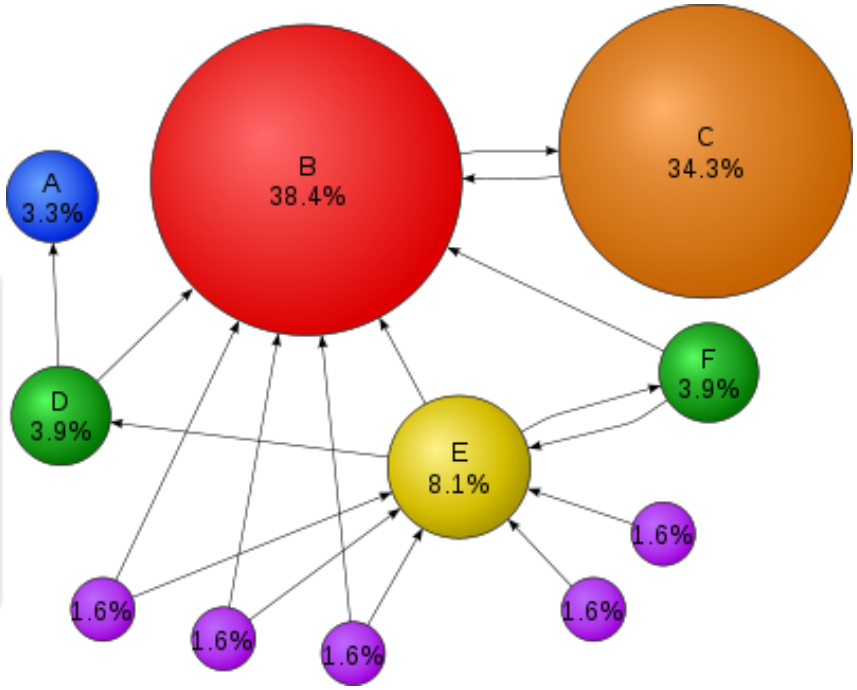
This is not exploratory!

PageRank

Let $G = (V, E)$ be the citation graph

PageRank [Brin98]

$\pi_i(u) = d\frac{1}{|V|} + (1-d)\sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$
with $d \in (0:1)$ is the damping factor.
It converges to a stable distribution.



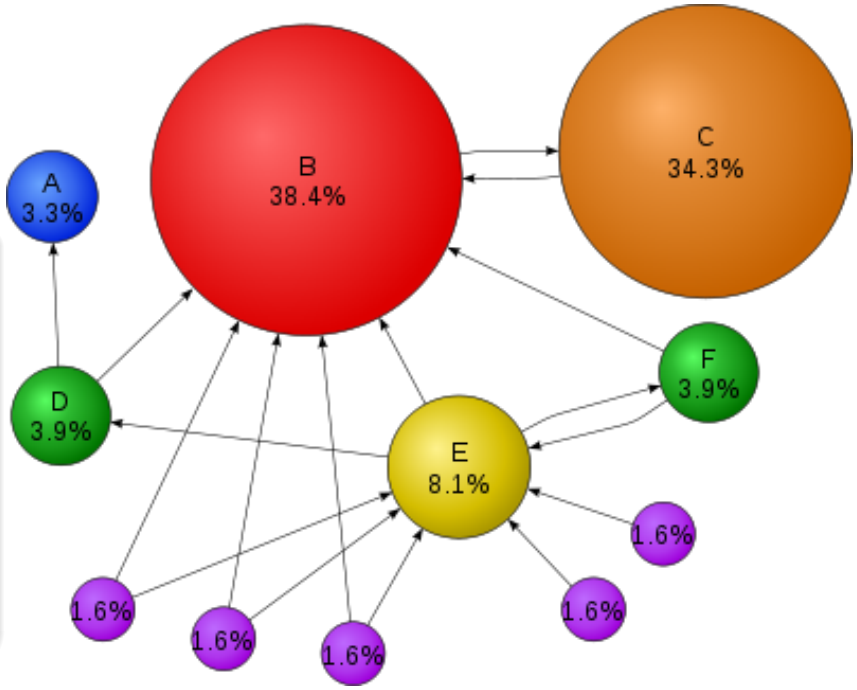
source: wikipedia

PageRank

Let $G = (V, E)$ be the citation graph

PageRank [Brin98]

$\pi_i(u) = d\frac{1}{|V|} + (1-d)\sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$
with $d \in (0:1)$ is the damping factor.
It converges to a stable distribution.
But it is not personalized.



source: wikipedia

PageRank

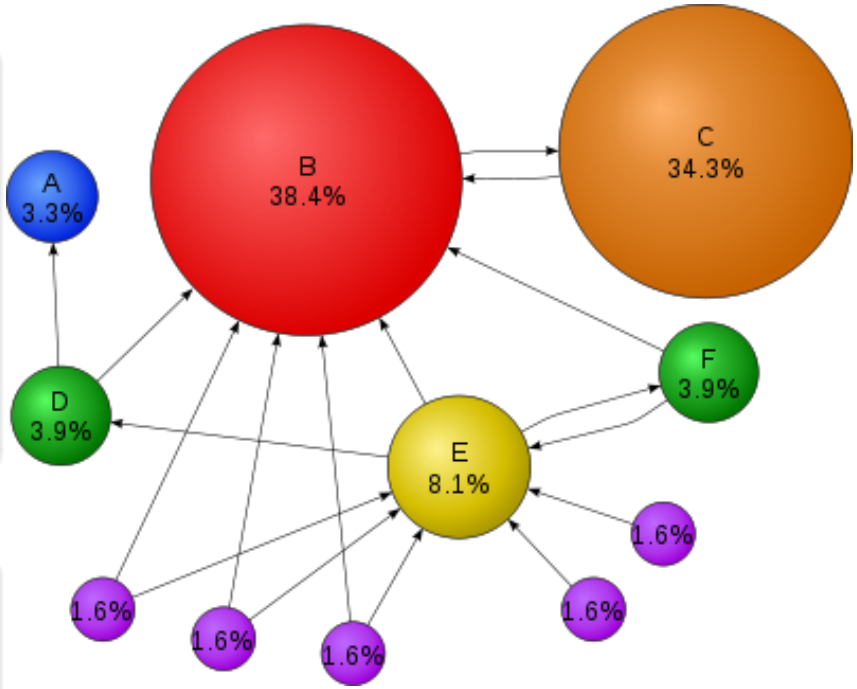
Let $G = (V, E)$ be the citation graph

PageRank [Brin98]

$\pi_i(u) = d\frac{1}{|V|} + (1-d)\sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$
with $d \in (0:1)$ is the damping factor.
It converges to a stable distribution.
But it is not personalized.

Personalized PageRank [Jeh03]

$\pi_i(u) = dp^*(u) + (1-d)\sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$
with $\sum p^*(u) = 1$.



source: wikipedia

PageRank

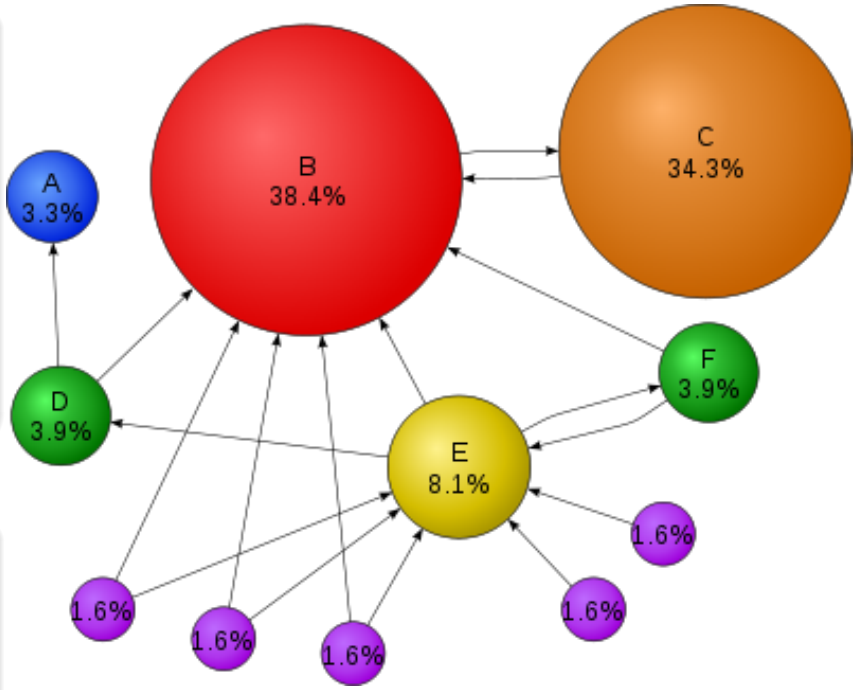
Let $G = (V, E)$ be the citation graph

PageRank [Brin98]

$\pi_i(u) = d\frac{1}{|V|} + (1-d)\sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$
with $d \in (0:1)$ is the damping factor.
It converges to a stable distribution.
But it is not personalized.

Personalized PageRank [Jeh03]

$\pi_i(u) = dp^*(u) + (1-d)\sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$
with $\sum p^*(u) = 1$.
But it is not exploratory.



source: wikipedia

Direction Awareness

Time exploration

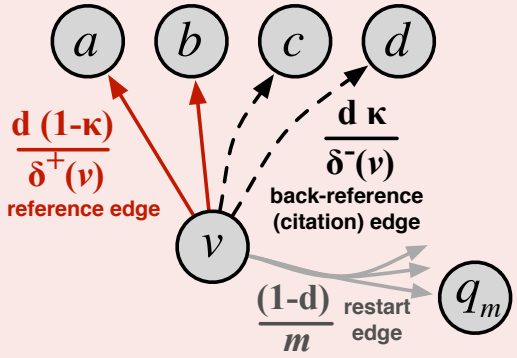
What if we are interested in searching papers per years. Recent papers? Traditional papers?

Let M be a set of known relevant papers.

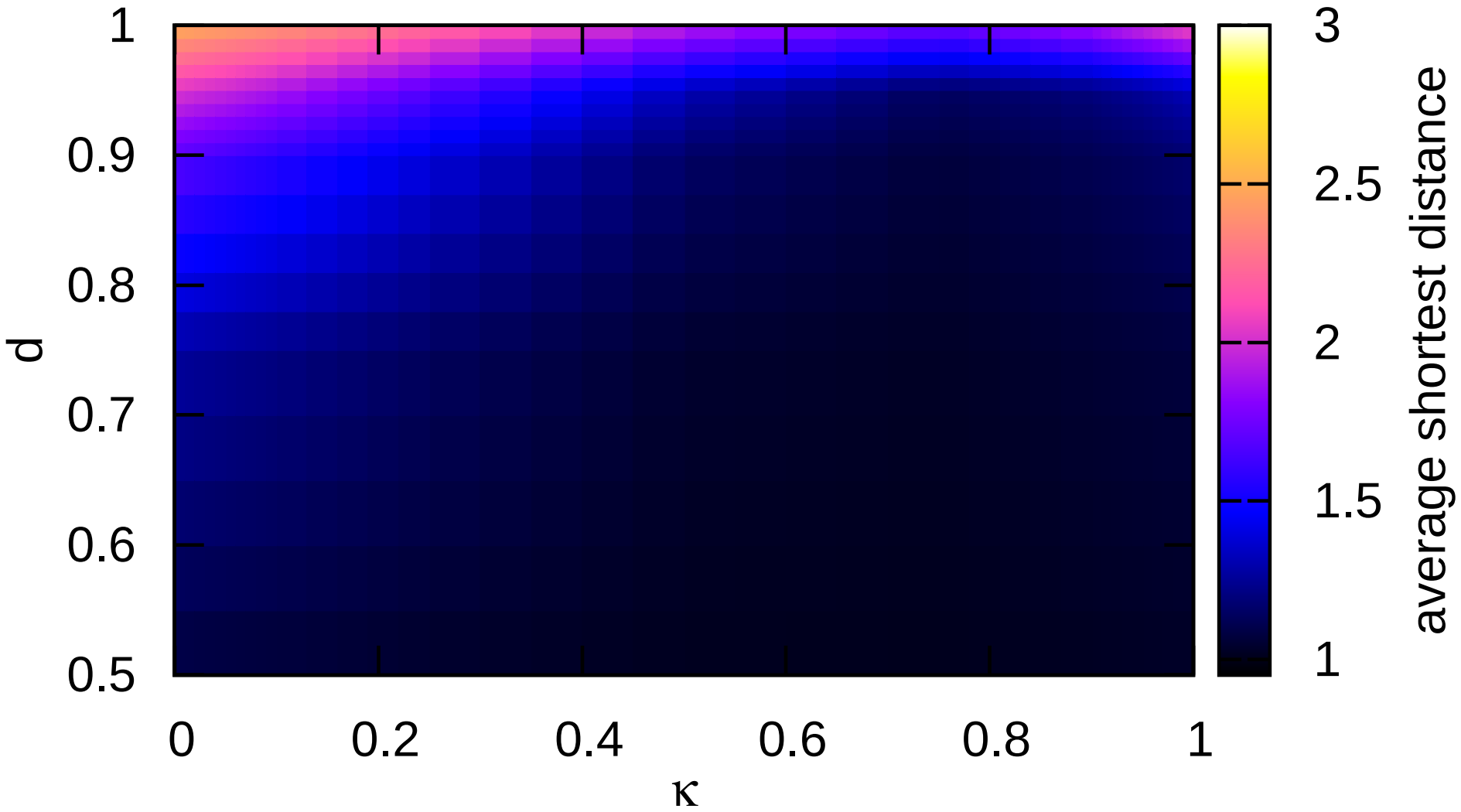
Direction Aware Random Walk with Restart

$$\pi_i(u) = dp^*(u) + (1-d)(\kappa \sum_{v \in N^+(u)} \frac{\pi_{i-1}(v)}{\delta^-(v)} + (1-\kappa) \sum_{v \in N^-(u)} \frac{\pi_{i-1}(v)}{\delta^+(v)})$$

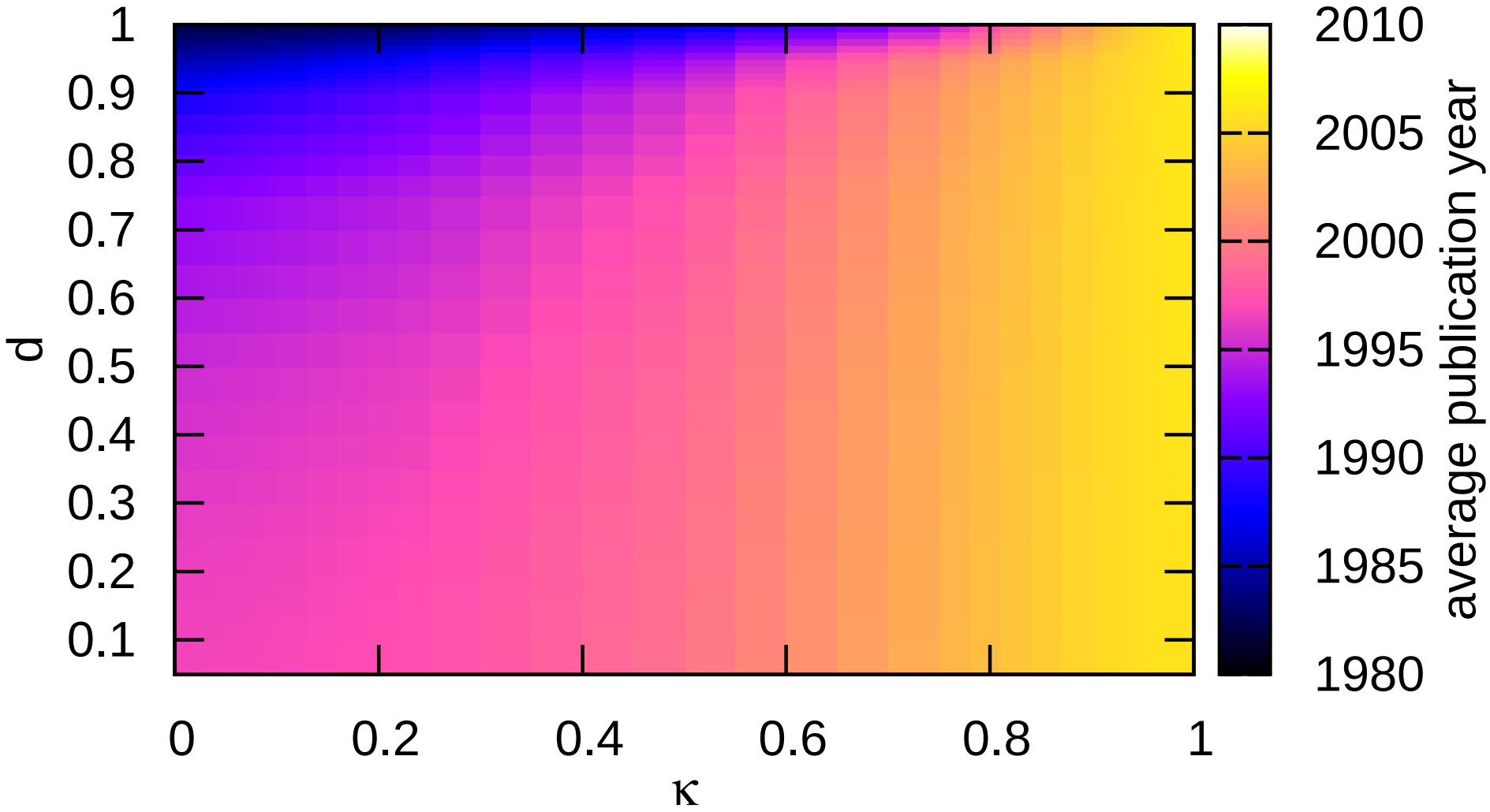
- $d \in (0:1)$ is the damping factor.
- $\kappa \in (0:1)$.
- $p^*(u) = \frac{1}{|M|}$, if $u \in M$, $p^*(u) = 0$, otherwise



Exploring in Depth



Exploring in Time



Hidden Reference Discovery

The recovery test

Let's hide some references from a paper and see if an algorithm can find them

Results of the experiments with mean average precision (MAP@50) and 95% confidence intervals.

	hide random			**hide recent**			**hide earlier**		
	mean	interval		mean	interval		mean	interval	
DARWR	48.00	46.80	49.20	**42.22**	**40.95**	**43.50**	**60.64**	**59.48**	**61.80**
P.R.	**56.56**	**55.31**	**57.80**	38.75	37.50	40.00	58.93	57.76	60.10
Katz$_\beta$	46.33	45.16	47.50	34.56	33.42	35.70	44.19	42.97	45.40
Cocit	44.60	43.39	45.80	14.22	13.25	15.20	55.97	54.64	57.30
Cocoup	17.28	16.36	18.20	17.56	16.61	18.50	2.93	2.57	3.30
CCIDF	18.05	17.11	19.00	18.97	17.94	20.00	3.55	3.10	4.00

Outline

A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting DaRWR

$$\pi_i(u) = dp^*(u) + (1-d)\left(\kappa \sum_{v \in N^+(u)} \frac{\pi_{i-1}(v)}{\delta^-(v)} + (1-\kappa) \sum_{v \in N^-(u)} \frac{\pi_{i-1}(v)}{\delta^+(v)}\right)$$

$$\pi_i(u) = dp^*(u) + \sum_{v \in N^+(u)} \frac{(1-d)\kappa}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{(1-d)(1-\kappa)}{\delta^+(v)} \pi_{i-1}(v)$$

$$\pi_i = dp^* + \mathbf{A}\pi_{i-1}$$

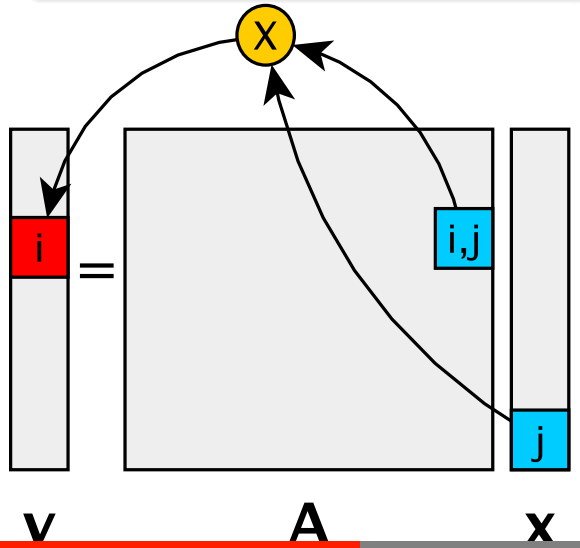
A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting DaRWR

$$\pi_i(u) = dp^*(u) + (1-d)\left(\kappa \sum_{v \in N^+(u)} \frac{\pi_{i-1}(v)}{\delta^-(v)} + (1-\kappa) \sum_{v \in N^-(u)} \frac{\pi_{i-1}(v)}{\delta^+(v)}\right)$$

$$\pi_i(u) = dp^*(u) + \sum_{v \in N^+(u)} \frac{(1-d)\kappa}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{(1-d)(1-\kappa)}{\delta^+(v)} \pi_{i-1}(v)$$

$$\pi_i = dp^* + \mathbf{A}\pi_{i-1}$$



CRS Full

- Traverse **A** column per column.
- Skip columns where $\pi_{i-1}(v) = 0$.
- Per edge: 2 non-zeros (2 indices, 2 values)

A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting DaRWR

$$\pi_i(u) = dp^*(u) + \sum_{v \in N^+(u)} \frac{(1-d)\kappa}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{(1-d)(1-\kappa)}{\delta^+(v)} \pi_{i-1}(v)$$

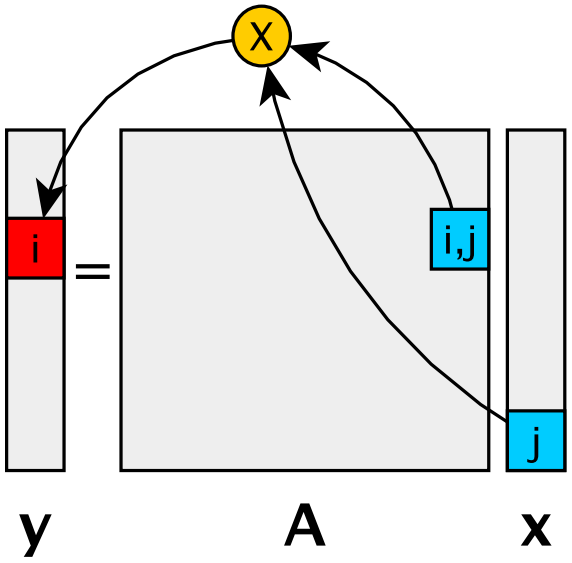
$$\pi_i = dp^* + \mathbf{B}^- \frac{(1-d)\kappa}{\delta^-} \pi_{i-1} + \mathbf{B}^+ \frac{(1-d)(1-\kappa)}{\delta^+} \pi_{i-1}$$

A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting DARWR

$$\pi_i(u) = dp^*(u) + \sum_{v \in N^+(u)} \frac{(1-d)\kappa}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{(1-d)(1-\kappa)}{\delta^+(v)} \pi_{i-1}(v)$$

$$\pi_i = dp^* + \mathbf{B}^- \frac{(1-d)\kappa}{\delta^-} \pi_{i-1} + \mathbf{B}^+ \frac{(1-d)(1-\kappa)}{\delta^+} \pi_{i-1}$$



CRS Half

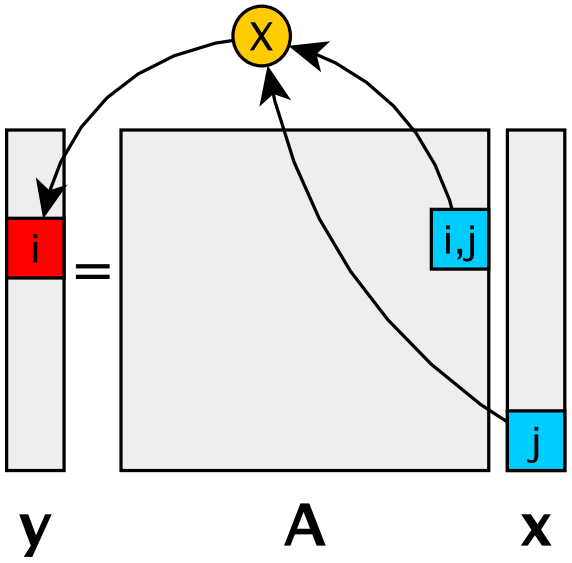
- pre-compute: $\frac{(1-d)\kappa}{\delta^-}\pi_{i-1}$ and $\frac{(1-d)(1-\kappa)}{\delta^+}\pi_{i-1}$
- $\mathbf{B}^-$ and $\mathbf{B}^+$ are 0/1 and symmetric
- Traverse the matrix twice ($\mathbf{B}^-$ and $\mathbf{B}^+$)
- Skip columns where $\pi_{i-1}(v) = 0$.
- Per edge: 1 non-zeros (1 index, 0 values)

A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting DaRWR

$$\pi_i(u) = dp^*(u) + \sum_{v \in N^+(u)} \frac{(1-d)\kappa}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{(1-d)(1-\kappa)}{\delta^+(v)} \pi_{i-1}(v)$$

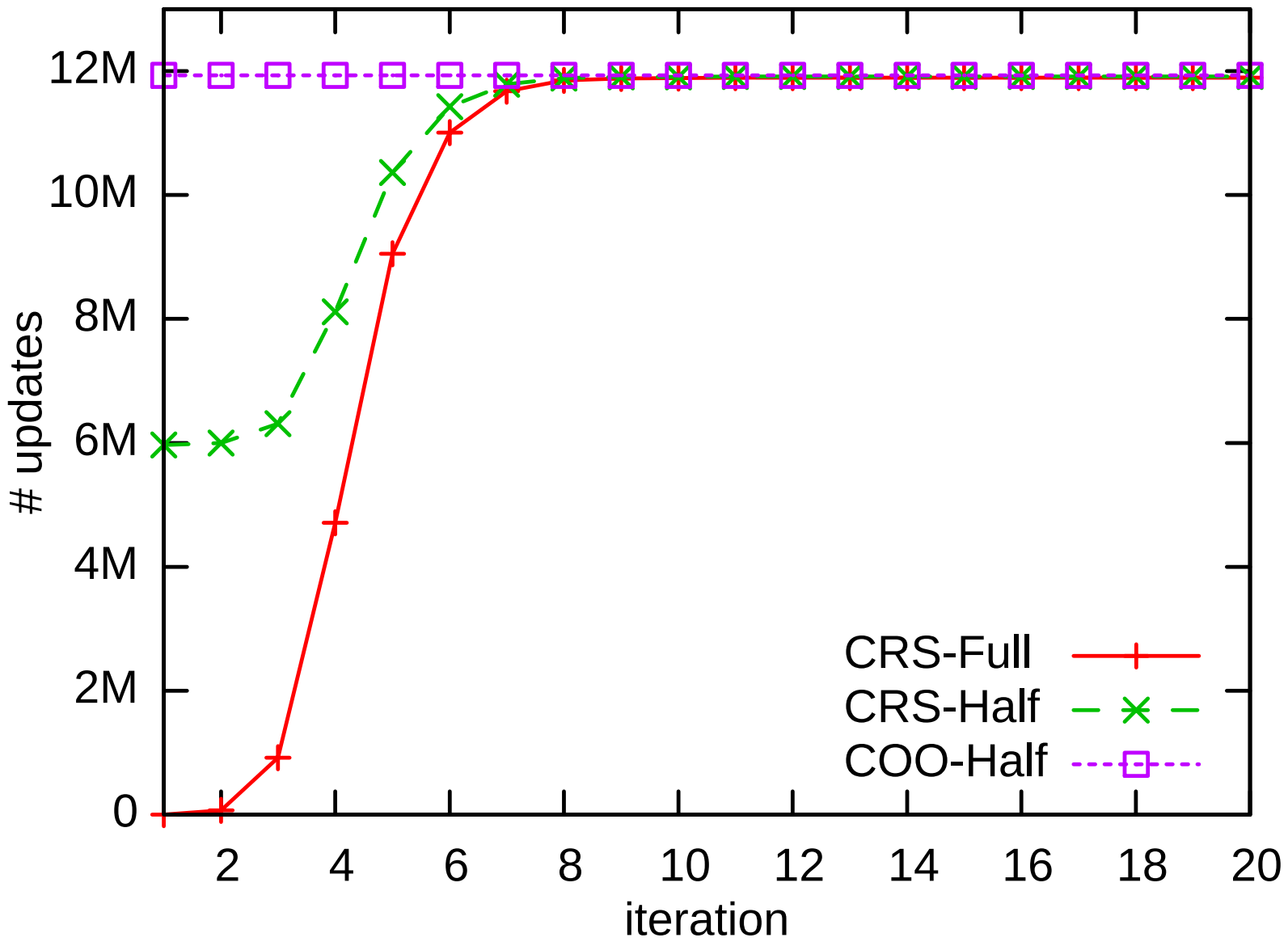
$$\pi_i = dp^* + \mathbf{B}^- \frac{(1-d)\kappa}{\delta^-} \pi_{i-1} + \mathbf{B}^+ \frac{(1-d)(1-\kappa)}{\delta^+} \pi_{i-1}$$



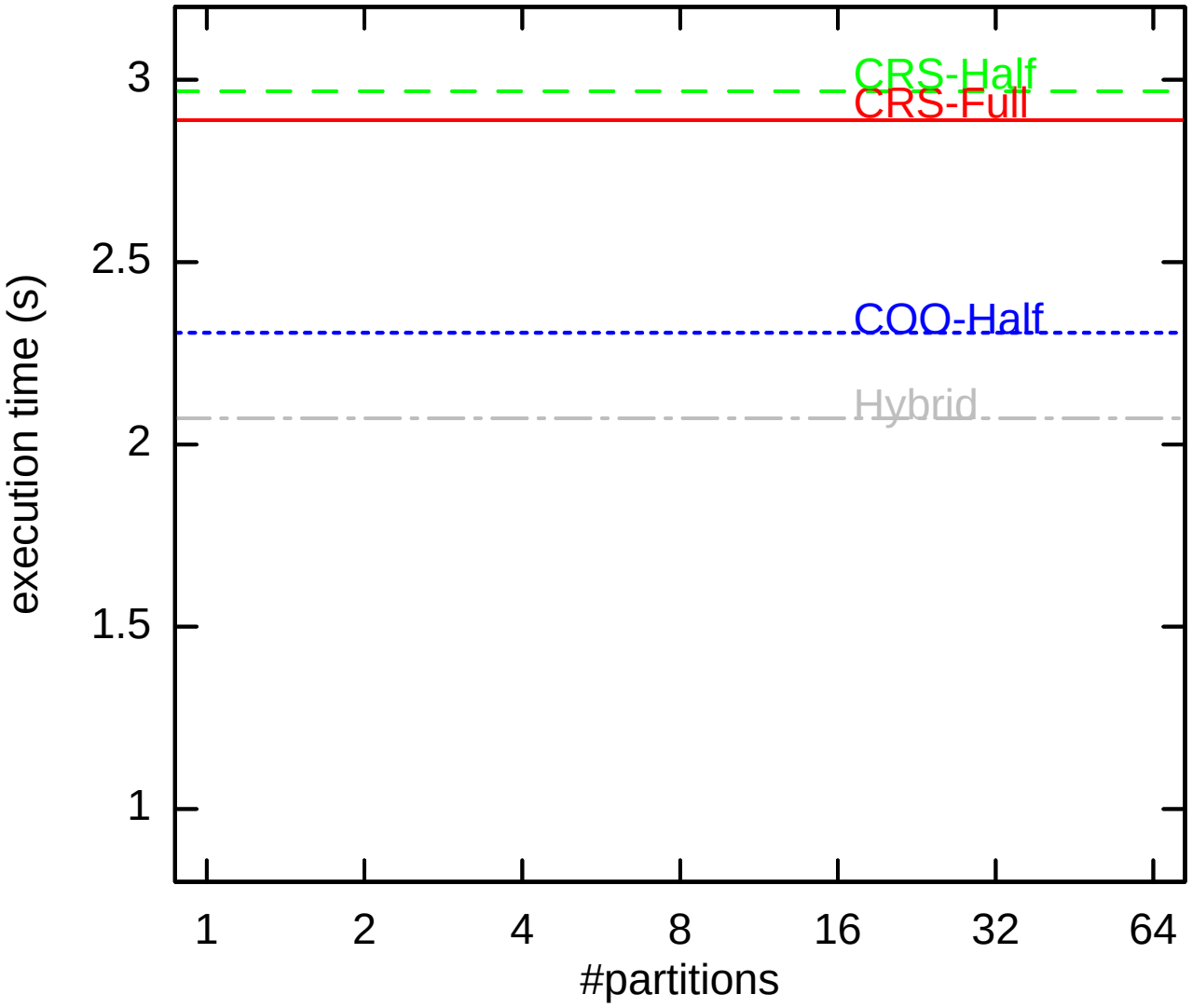
COO Half

- pre-compute: $\frac{(1-d)\kappa}{\delta^-}\pi_{i-1}$ and $\frac{(1-d)(1-\kappa)}{\delta^+}\pi_{i-1}$
- $\mathbf{B}^-$ and $\mathbf{B}^+$ are 0/1 and symmetric
- Traverse the matrix once ($\mathbf{B}^-$ and $\mathbf{B}^+$)
- Arbitrary order. Don't skip anything.
- Per edge: 1 non-zeros (2 indices, 0 values)

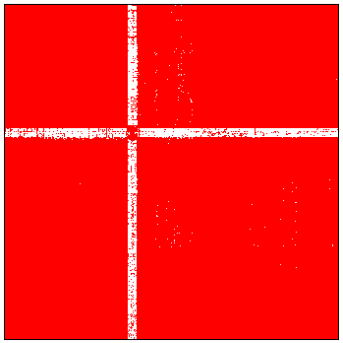
Number of updates



Runtimes


CRS-Half
CRS-Full
COO-Half
Hybrid
execution time (s)
3
2.5
2
1.5
1
1
2
4
8
16
32
64
#partitions

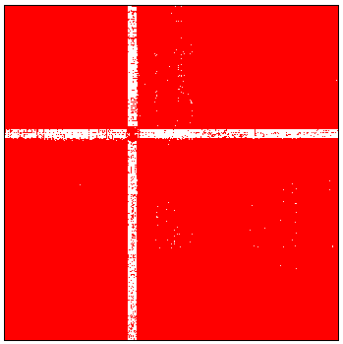
Ordering

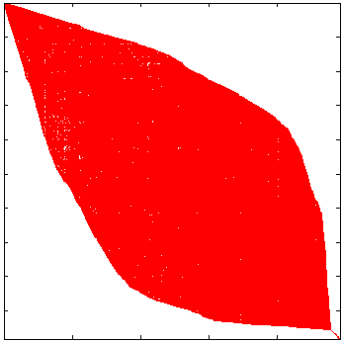


Locality

SpMV is sensitive to non-zero locality.

Ordering





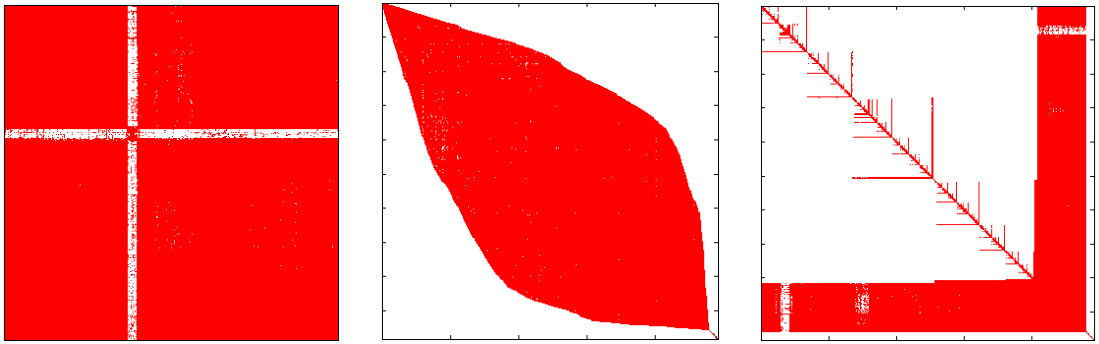
Locality

SpMV is sensitive to non-zero locality.

Reverse Cuthill-McKee [Cuthill, McKee, 69]

Order with respect to a Breadth First Search ordering. (Do 10 times, pick best)

Ordering



Locality

SpMV is sensitive to non-zero locality.

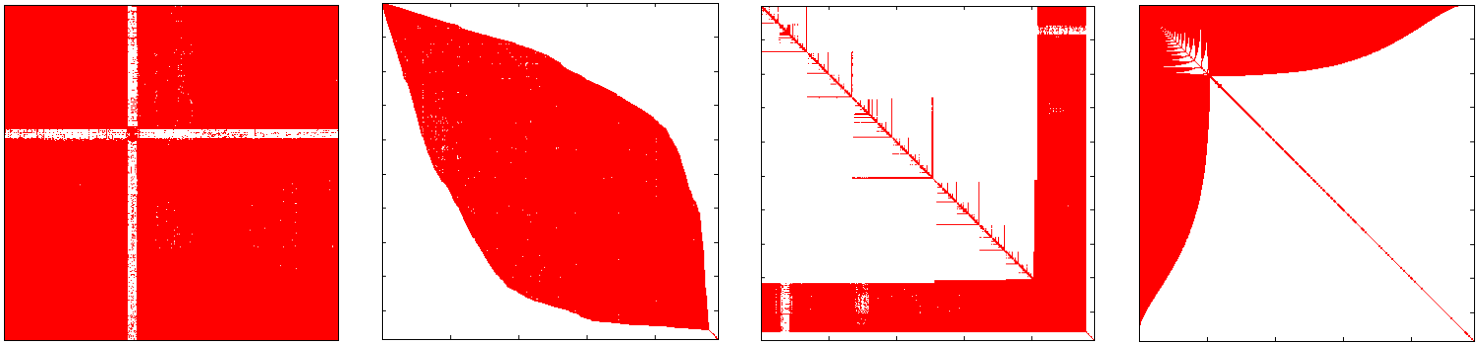
Reverse Cuthill-McKee [Cuthill, McKee, 69]

Order with respect to a Breadth First Search ordering. (Do 10 times, pick best)

Approximate Minimum Degree [Amestoy et al.,96]

Greedily, add the vertex whose degree is minimum.

Ordering



Locality

SpMV is sensitive to non-zero locality.

Reverse Cuthill-McKee [Cuthill, McKee, 69]

Order with respect to a Breadth First Search ordering. (Do 10 times, pick best)

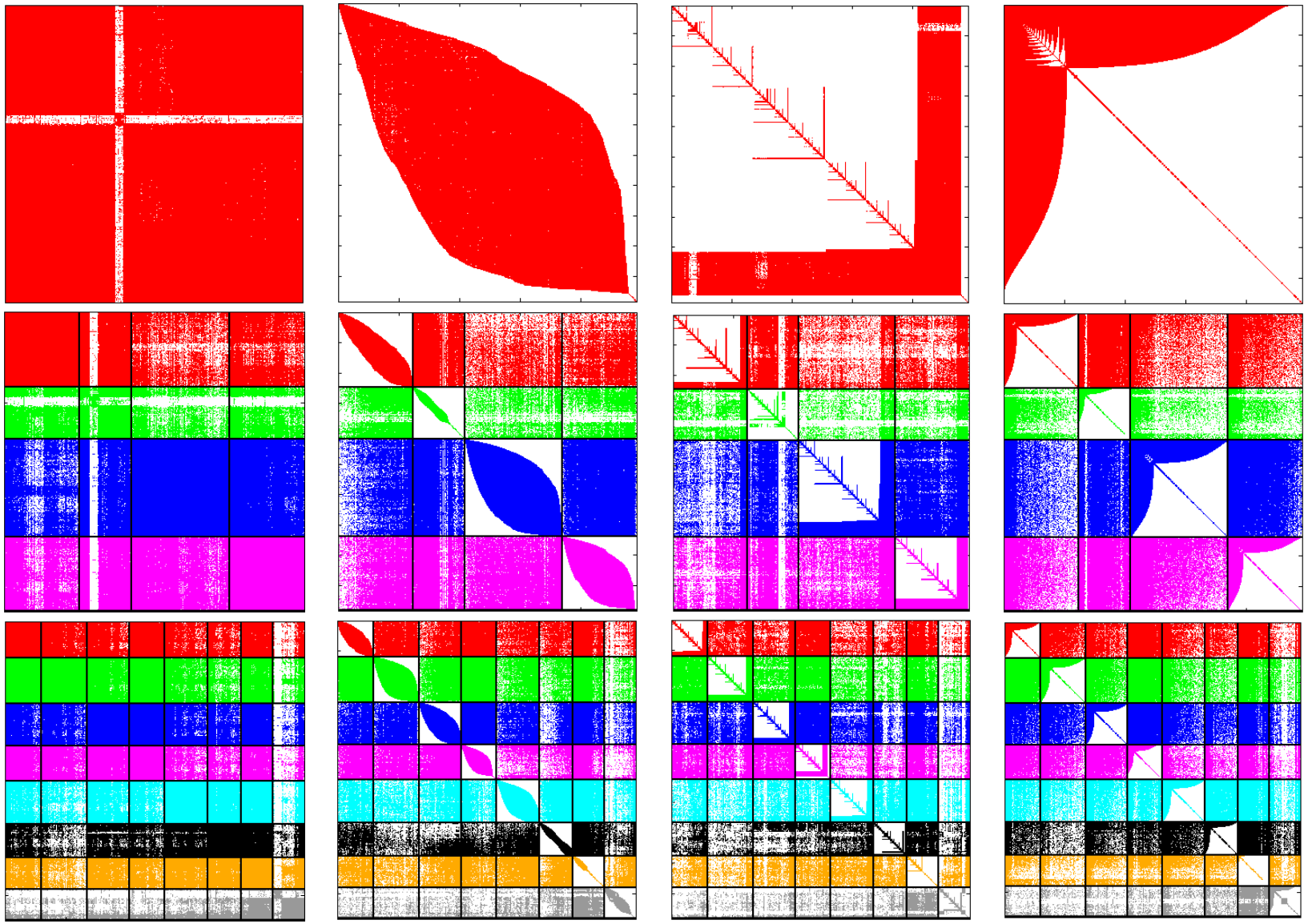
Approximate Minimum Degree [Amestoy et al.,96]

Greedily, add the vertex whose degree is minimum.

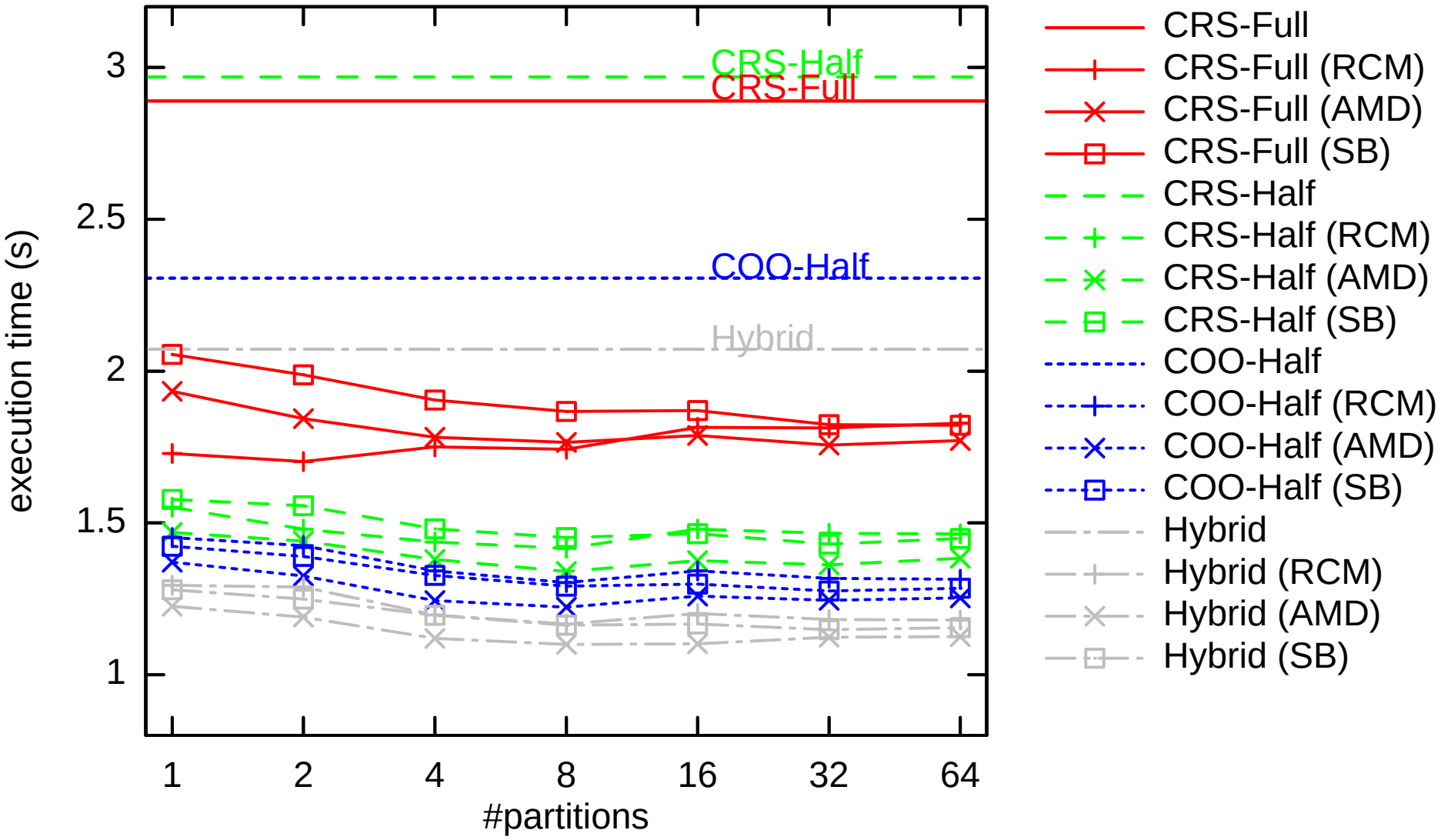
Slashburn [Kang, Faloutsos,11]

Order by connected components. Remove the highest degree vertex. Repeat.

Partitioning



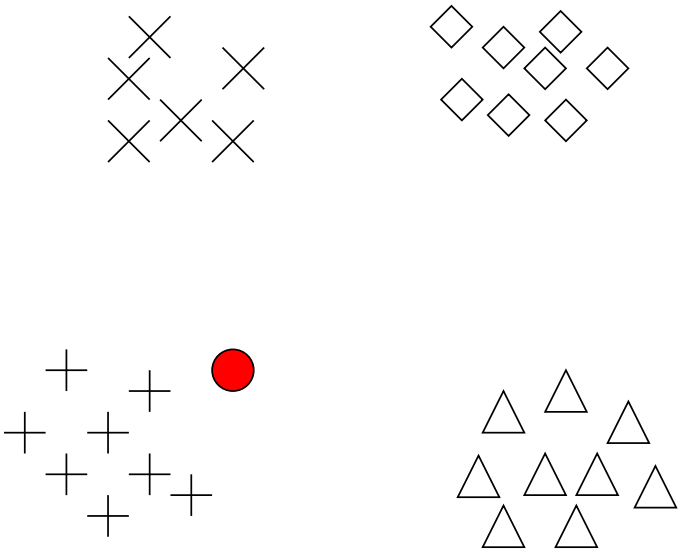
Runtimes


execution time (s)
3
2.5
2
1.5
1
1
2
4
8
16
32
64
#partitions
CRS-Half
CRS-Full
COO-Half
Hybrid
CRS-Full
CRS-Full (RCM)
CRS-Full (AMD)
CRS-Full (SB)
CRS-Half
CRS-Half (RCM)
CRS-Half (AMD)
CRS-Half (SB)
COO-Half
COO-Half (RCM)
COO-Half (AMD)
COO-Half (SB)
Hybrid
Hybrid (RCM)
Hybrid (AMD)
Hybrid (SB)

Outline

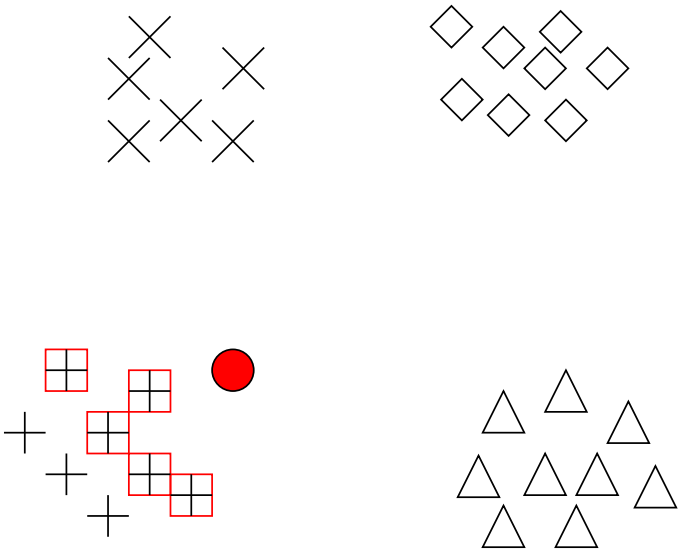
Principle

The goal of diversity is to avoid clustered answers.



Principle

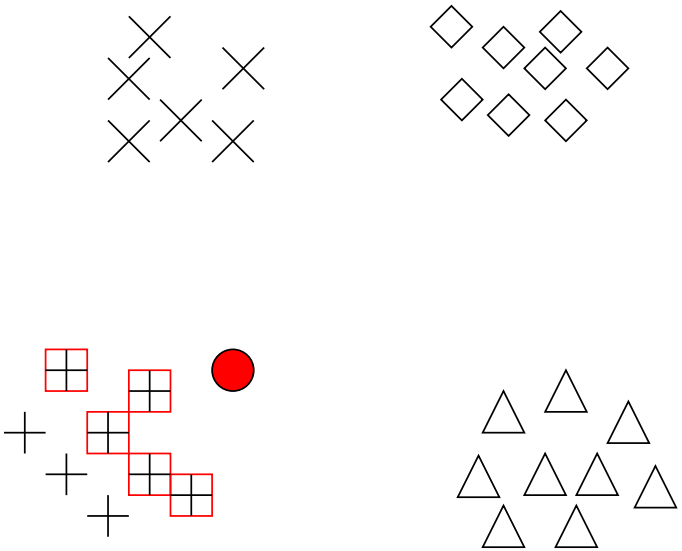
The goal of diversity is to avoid clustered answers.



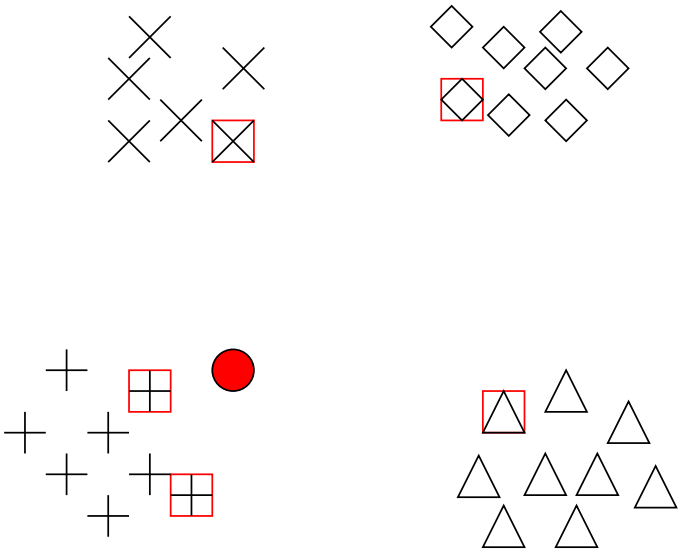
Relevant

Principle

The goal of diversity is to avoid clustered answers.

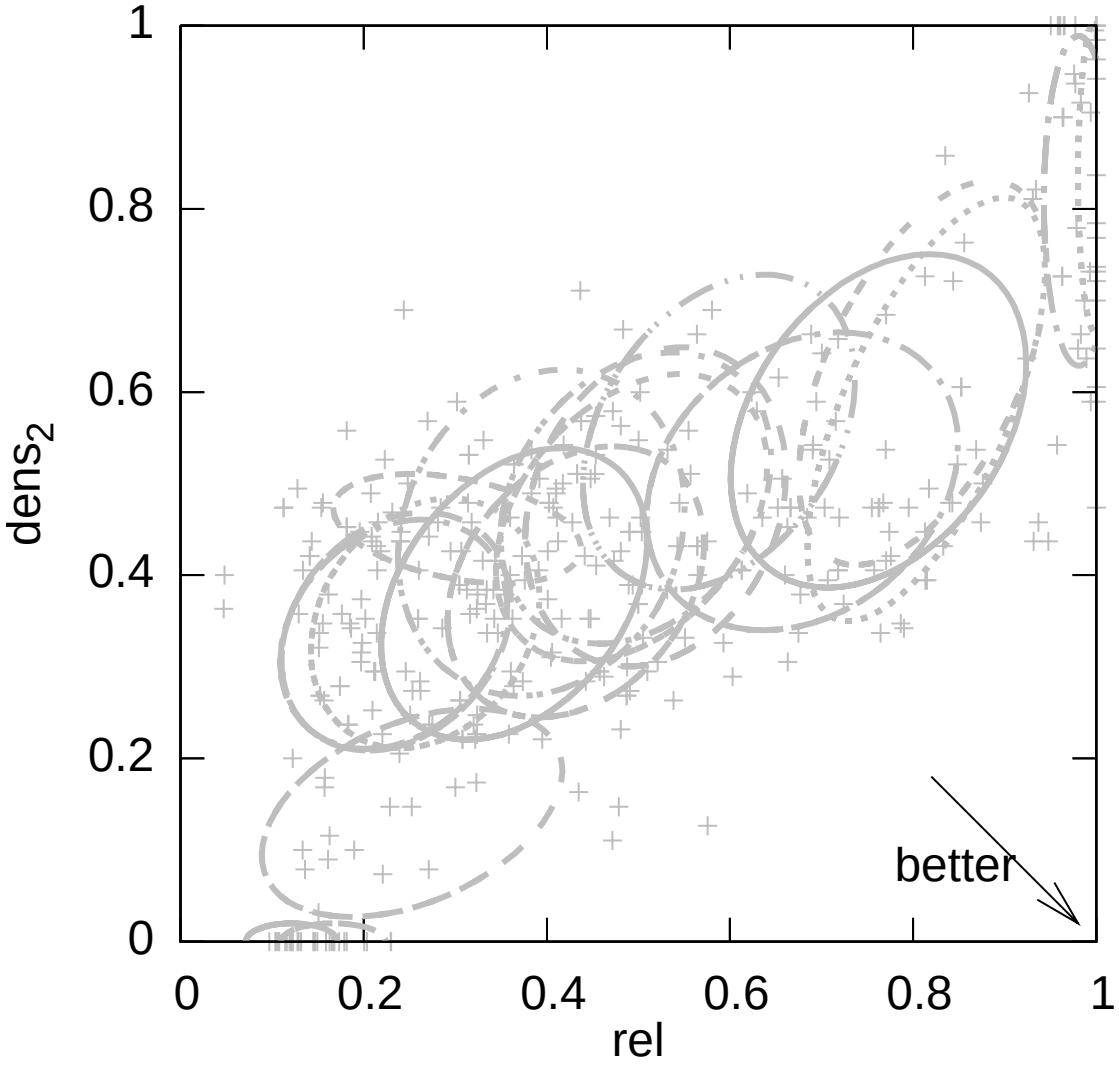


Relevant



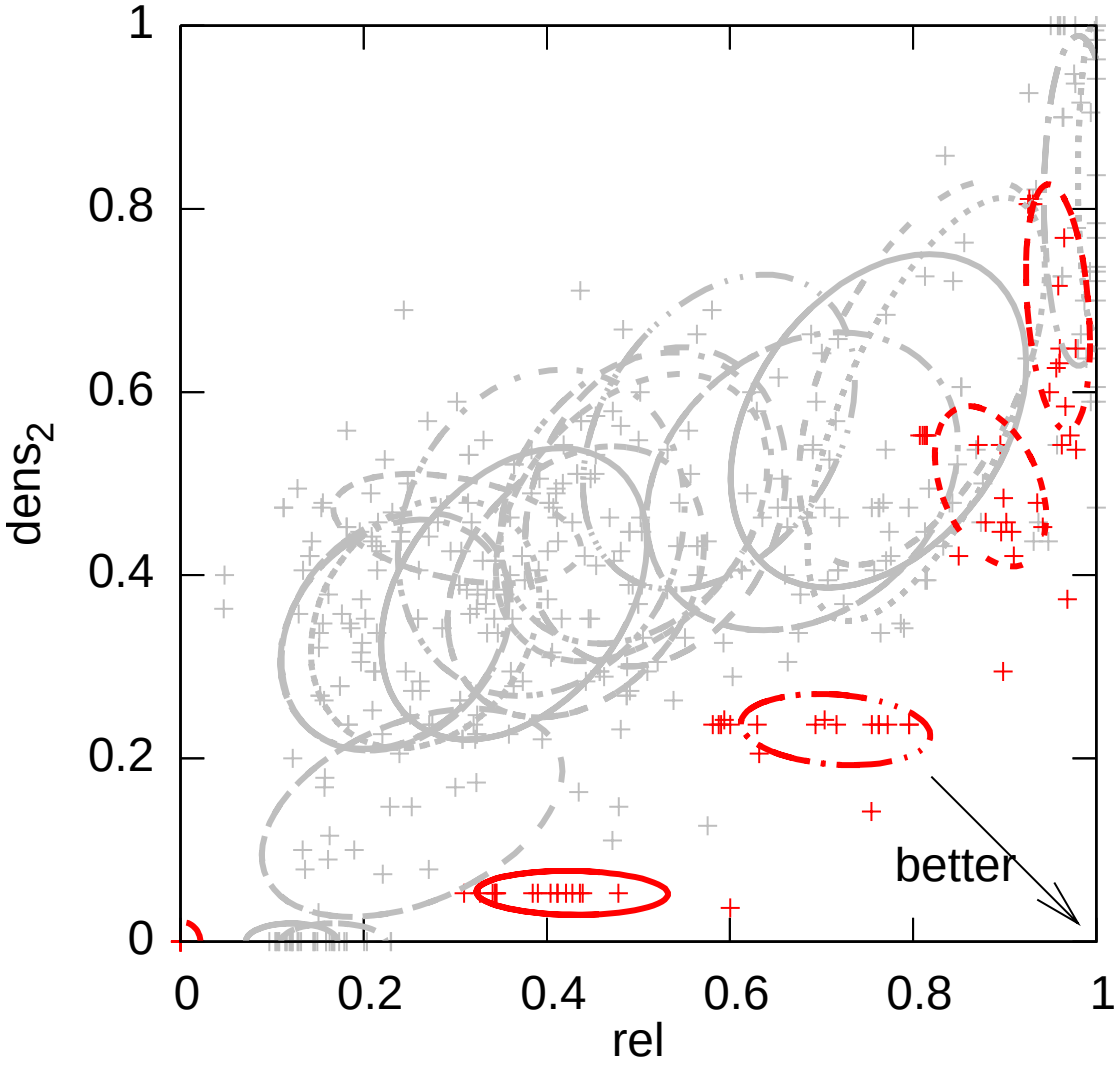
Relevant Diverse

A Modelization problem



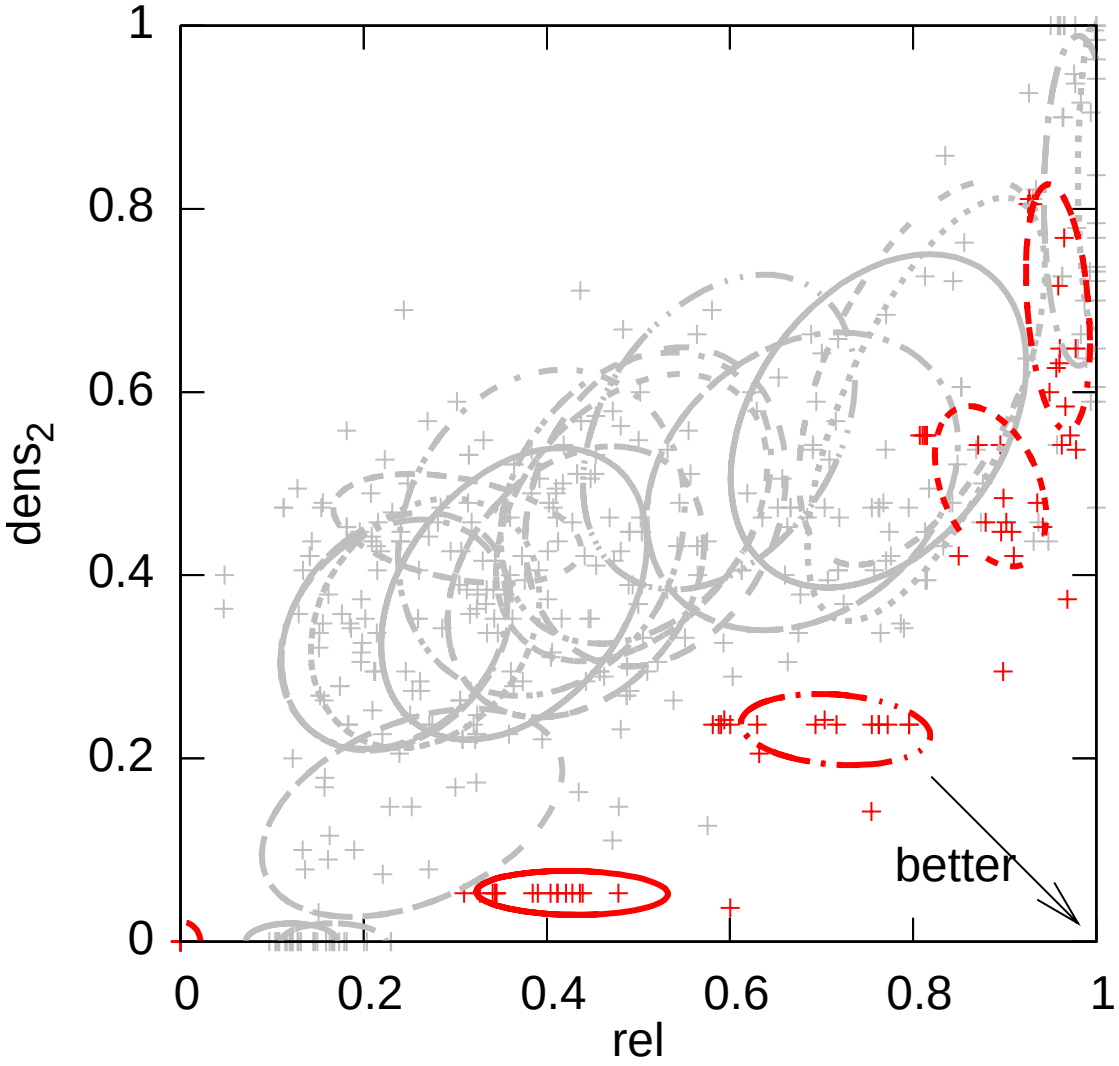
Here is a distribution of known algorithms

A Modelization problem



Would such an algorithm be of interest?

A Modelization problem

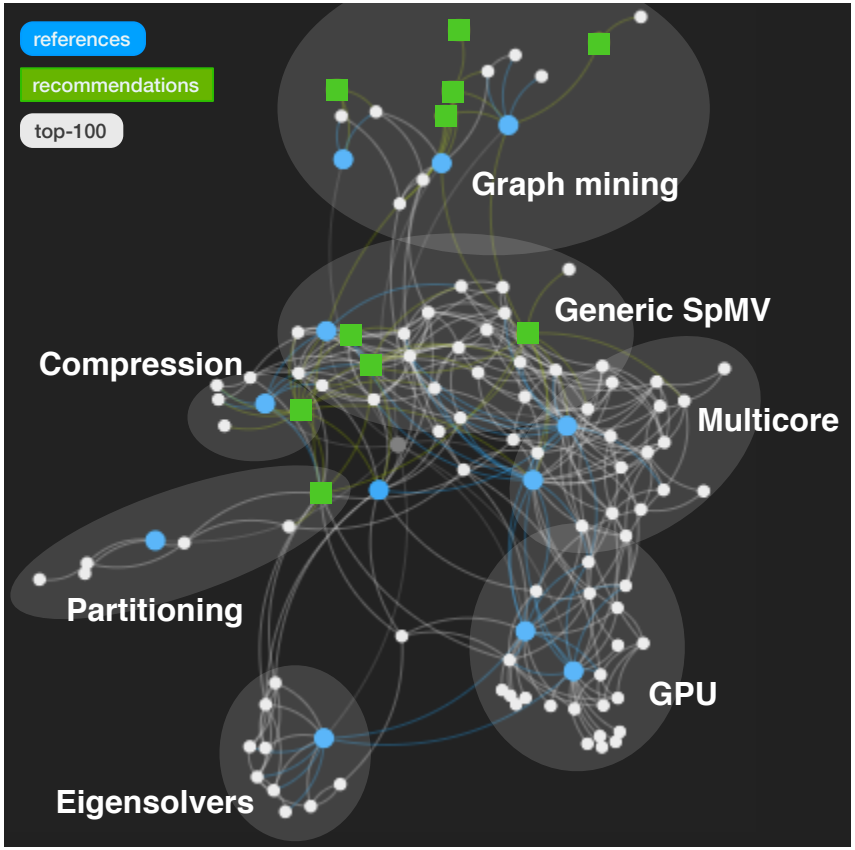


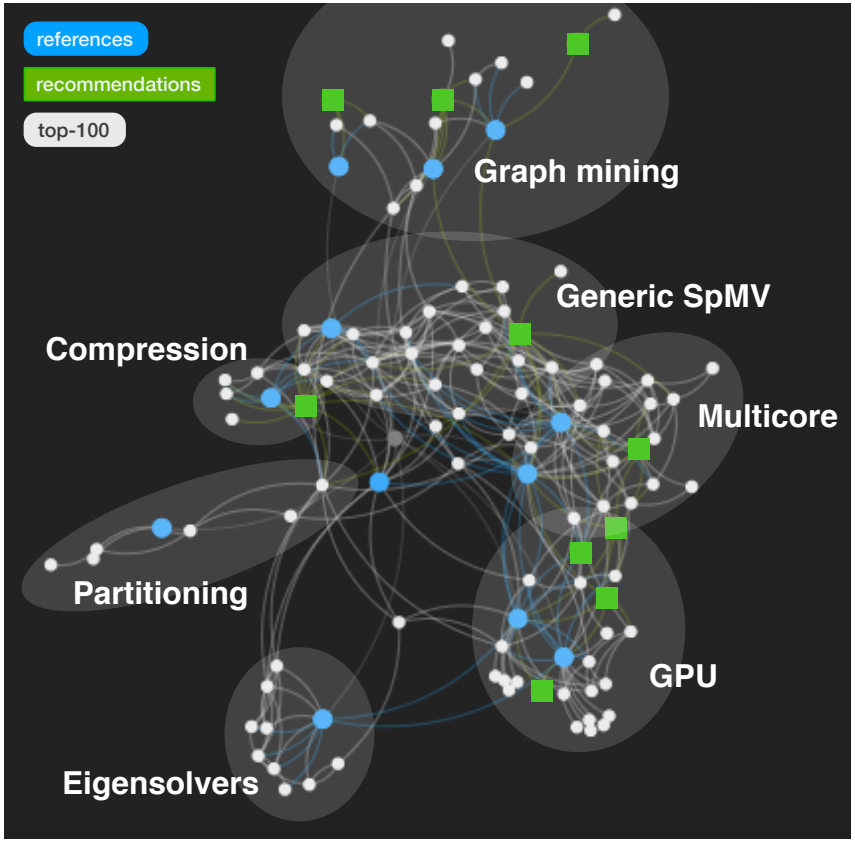
Would such an algorithm be of interest?
That algorithm is random!

What to do?

See later talk!

Results


references
recommendations
top-100
Graph mining
Generic SpMV
Compression
Multicore
Partitioning
GPU
Eigensolvers


references
recommendations
top-100
Graph mining
Generic SpMV
Compression
Multicore
Partitioning
GPU
Eigensolvers

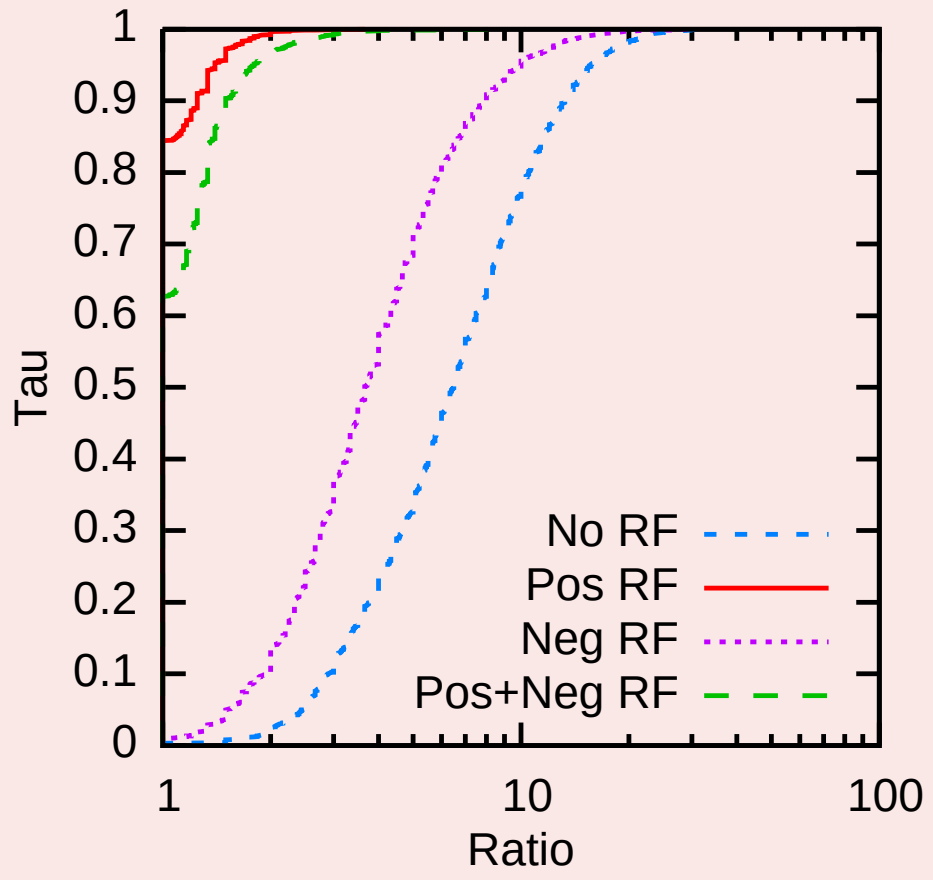
Outline

Relevance Feedback

Papers can be tagged are relevant or irrelevant.

- **Positive feedback (+RF):** Relevant results are added to Q
- **Negative feedback (-RF):** Irrelevant results are removed from the graph

How long does it take to find the first level-3 paper?

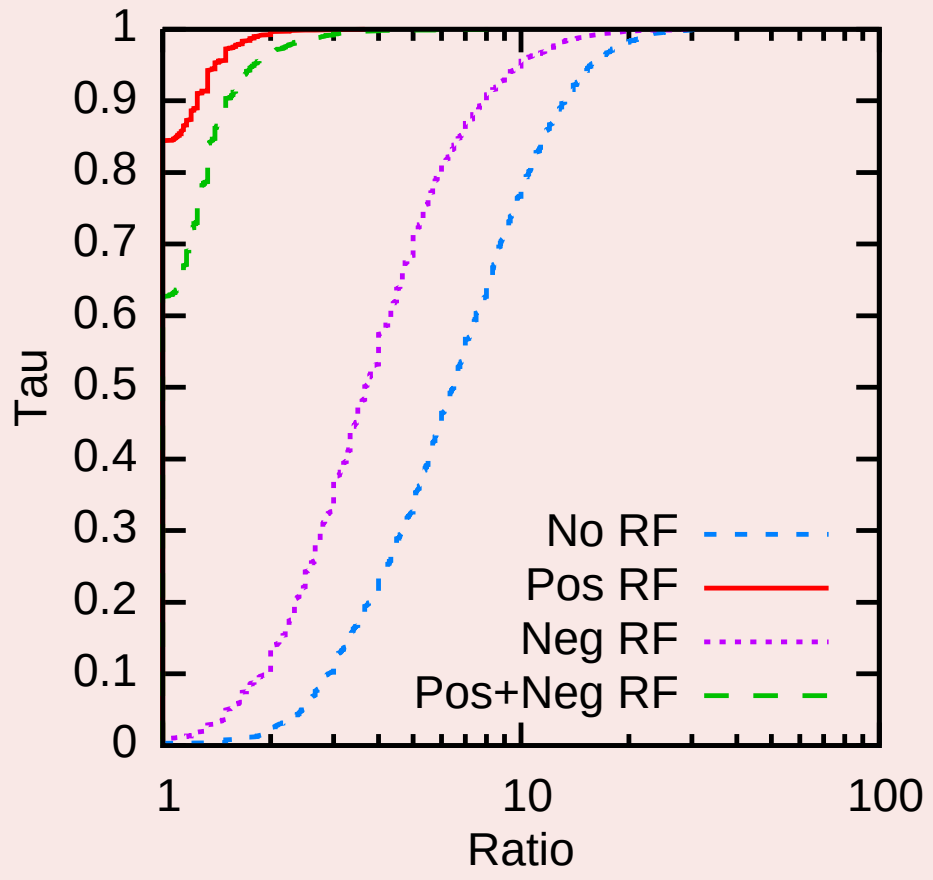


Relevance Feedback

Papers can be tagged are relevant or irrelevant.

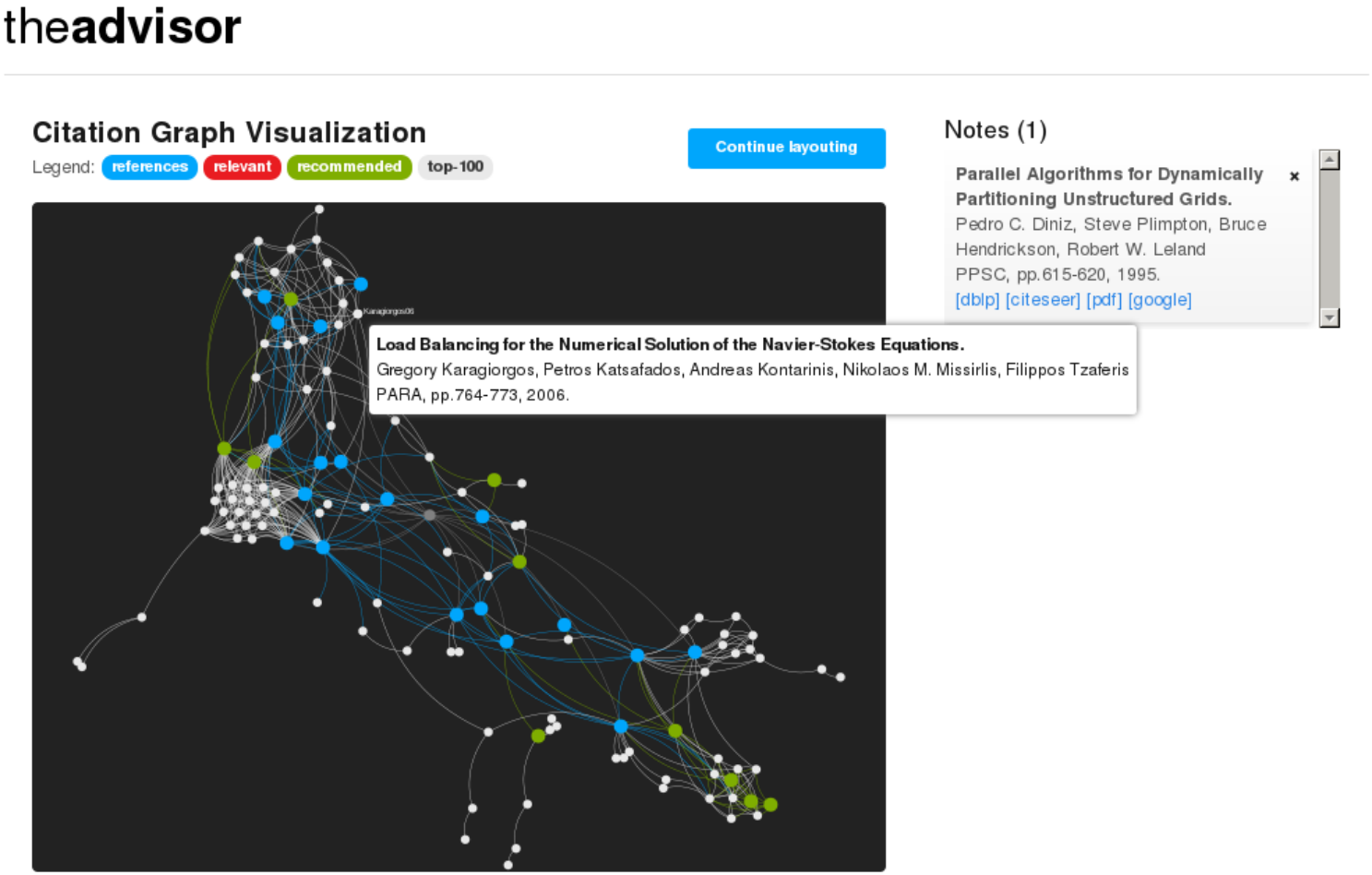
- **Positive feedback (+RF):** Relevant results are added to Q
- **Negative feedback (-RF):** Irrelevant results are removed from the graph

How long does it take to find the first level-3 paper?

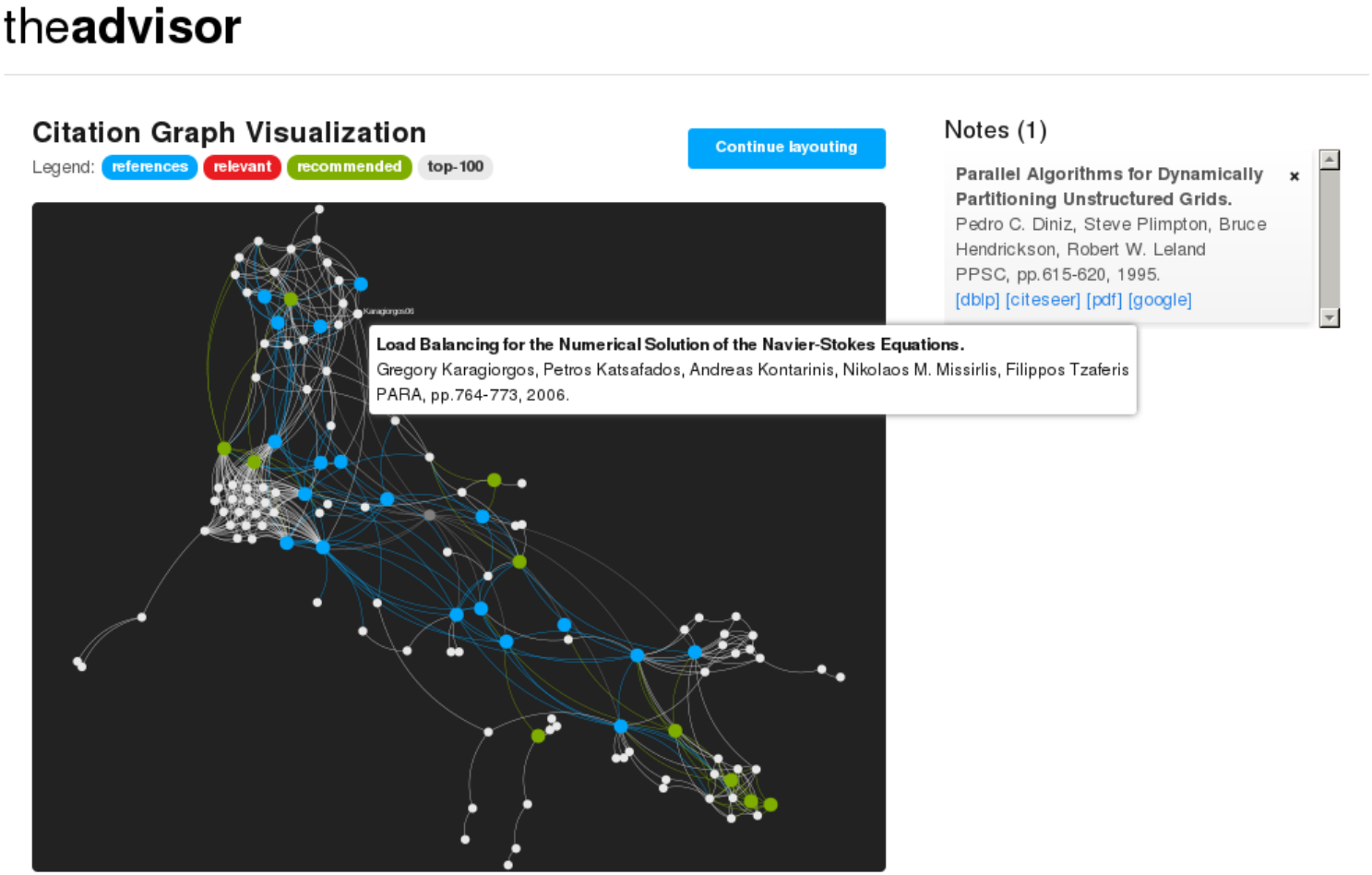


More exploration!

Visualization



Visualization



More exploration!

Application Programming Interface

Web service

the**advisor** can be accessed programmatically. Emit HTTP requests and obtain JSON encoded replies.

Application Programming Interface

Web service

the**advisor** can be accessed programmatically. Emit HTTP requests and obtain JSON encoded replies.

Potential Applications

- Interfacing with article editors (e.g., TexShop)
- Recommendation in bibliography manager (e.g., Mendeley)
- Suggesting reviewers to program committees (e.g., EasyChair)
- Suggesting sessions of interest at conferences (e.g., iConference)

Application Programming Interface

Web service

the**advisor** can be accessed programmatically. Emit HTTP requests and obtain JSON encoded replies.

Potential Applications

- Interfacing with article editors (e.g., TexShop)
- Recommendation in bibliography manager (e.g., Mendeley)
- Suggesting reviewers to program committees (e.g., EasyChair)
- Suggesting sessions of interest at conferences (e.g., iConference)

Easier to use!

Outline

Design Goals - Are they matched?

Personalized

The query is expressed in a very precise way.

Conceptual

Using the citation graph allows to avoid all linguistic. Though, it may not be enough to find all relevant papers.

Exploratory

Direction Awareness (to choose time), Diversification (to see more topics), Visualization (for manual crawling)

Easy to use

Efficient (recommendation in less than 2 seconds), web-based.

Design Goals - Are they matched?

Personalized

The query is expressed in a very precise way.

Conceptual

Using the citation graph allows to avoid all linguistic. Though, it may not be enough to find all relevant papers.

Exploratory

Direction Awareness (to choose time), Diversification (to see more topics), Visualization (for manual crawling)

Easy to use

Efficient (recommendation in less than 2 seconds), web-based.

Is it good enough? Tell us!

Future works

Clustering

Let's assume for an instant that we have accurate disambiguated tags for every document. We could restrict analysis to some fields. Improve diversification.

Betweenness Centrality

DaRWR provides recommendation *around* the query set. What about recommending what is *between* it?

Contextual information

Distinguishing types of papers and citations. Survey, Method, Application...

Thank you

More information

contact : esaule@bmi.osu.edu

visit: `http://theadvisor.osu.edu`

(or `http://bmi.osu.edu/hpc/`

or `http://bmi.osu.edu/~esaule`)

Research at HPC lab is supported by

NSF
Department of Energy United States of America
National Institutes of Health
National Cancer Institute
Sandia National Laboratories
INL Idaho National Laboratory
DAGSI The Dayton Area Graduate Studies Institute
NVIDIA