

Applying Science Models for Re-Ranking in IR

Introducing bibliometrically enhanced metadata to IR

By Andreas Kruff, Anh Huy Matthias Tran

Agenda

1. Introduction
2. Creating a Baseline
3. Data Preparation
4. Creating Graphs
 - a. *Co-citation Graph*
 - b. *Lotka-inspired Graph*
 - c. *Citation Graph*
 - d. *Core Journal Graph*
5. Experiments & Evaluation
6. Conclusion

1. Introduction

Project Motivation

Dissertation: 'Re-Ranking auf Basis von Brafordizing für die verteilte Suche in Digitalen Bibliotheken' (Mayr, 2009)

Introduction of new metrics for bibliometrically enhanced Information Retrieval (BIR) in the context of Re-Ranking.

Application Case

- ❖ Recent data sets: **TREC-COVID**
- ❖ In combination with *Graph Construction & Network Analysis*

Project Motivation

Papers: 'Science models as value-added services for scholarly information systems' (*Mutschke, 2011*)

Introduction of scholarly Information Retrieval (IR) as a further developed models for improving retrieval quality, involving features such as Bradford law of Information and co-authorship networks.

Application Case

- ❖ Recent data sets: **TREC-COVID**
- ❖ In combination with *Graph Construction & Network Analysis*

Ranking	Document	Score	Journal	coreness	
1	Doc 10	15.4646	bioRxiv	0.35	↓↑
2	Doc 15	14.3549	Emerg Infect Dis	0.24	↓↑
3	Doc 101	14.3542	Journal of virology	0.12	↓↑
[...]	[...]	[...]	[...]	[...]	↓↑
998	Doc 17	1.636	J Biomed Sci	0.01	↓↑
999	Doc 4	0.002	Emerg Infect Dis	0.12	↓↑
1000	Doc 90	0.000	bioRxiv	0.35	

Further Motivations

As mentioned in the lectures concerning topics such as



Network Analysis

Centrality, Betweenness



Power Law's

Lotka's Law, Bradford law and Zipf's law



Re-Ranking

Based on Bibliometrics and authorships

Further Motivations

As mentioned in the lectures concerning topics such as

Stratagems (*as defined by Marcia Bates*)

◆ Citation Search

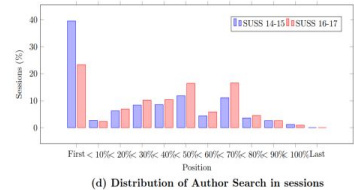
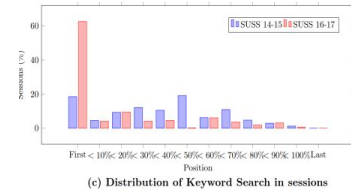
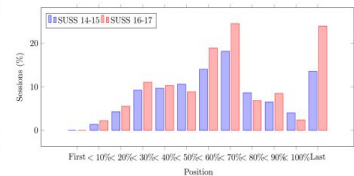
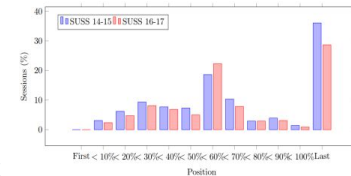
Usage of provided citation connections

◆ Author Search

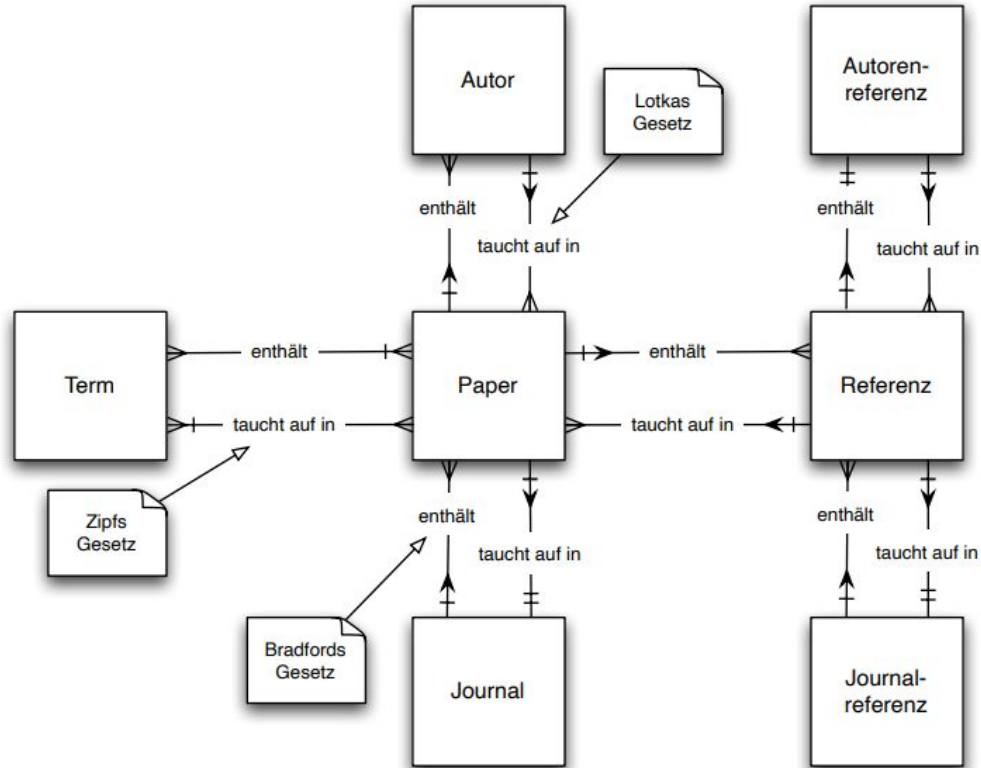
Usage of provided author connections

◆ Journal Run

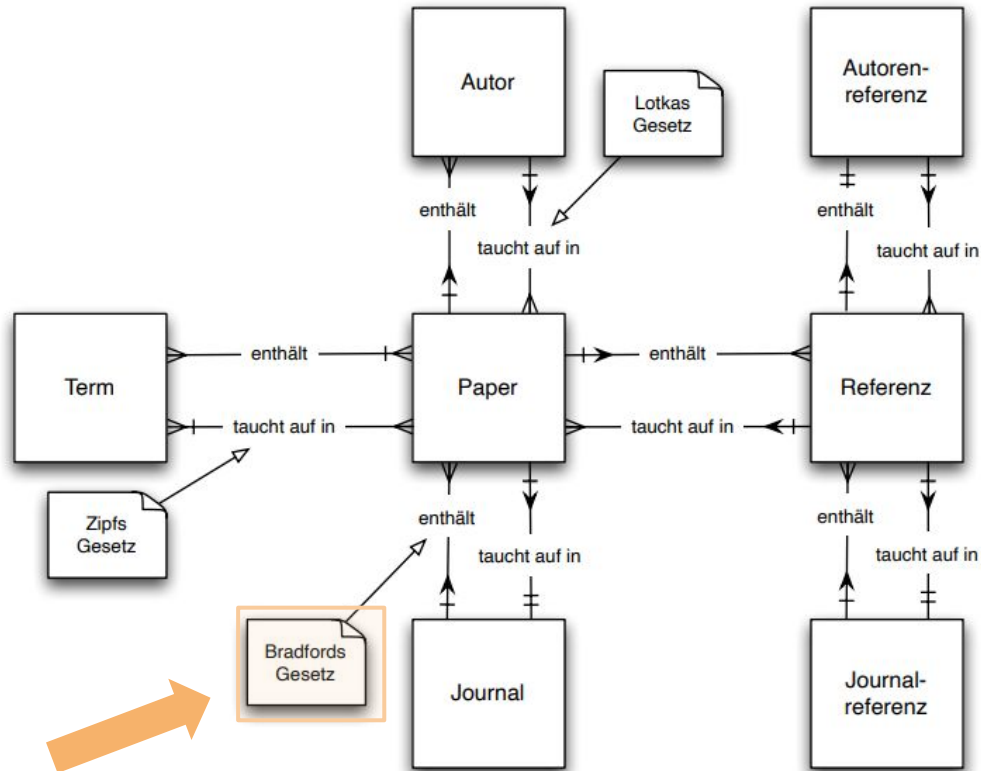
Usage of given journal connections



General Approach



General Approach



General Approach

Bradford's Law

- ❖ Identifying **Core Journals**
- ❖ Boosting Papers by **occurrences** of journal

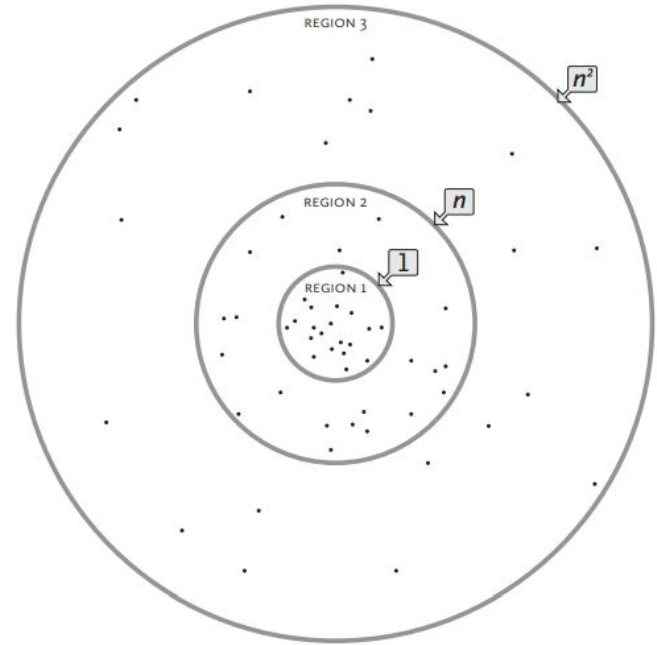
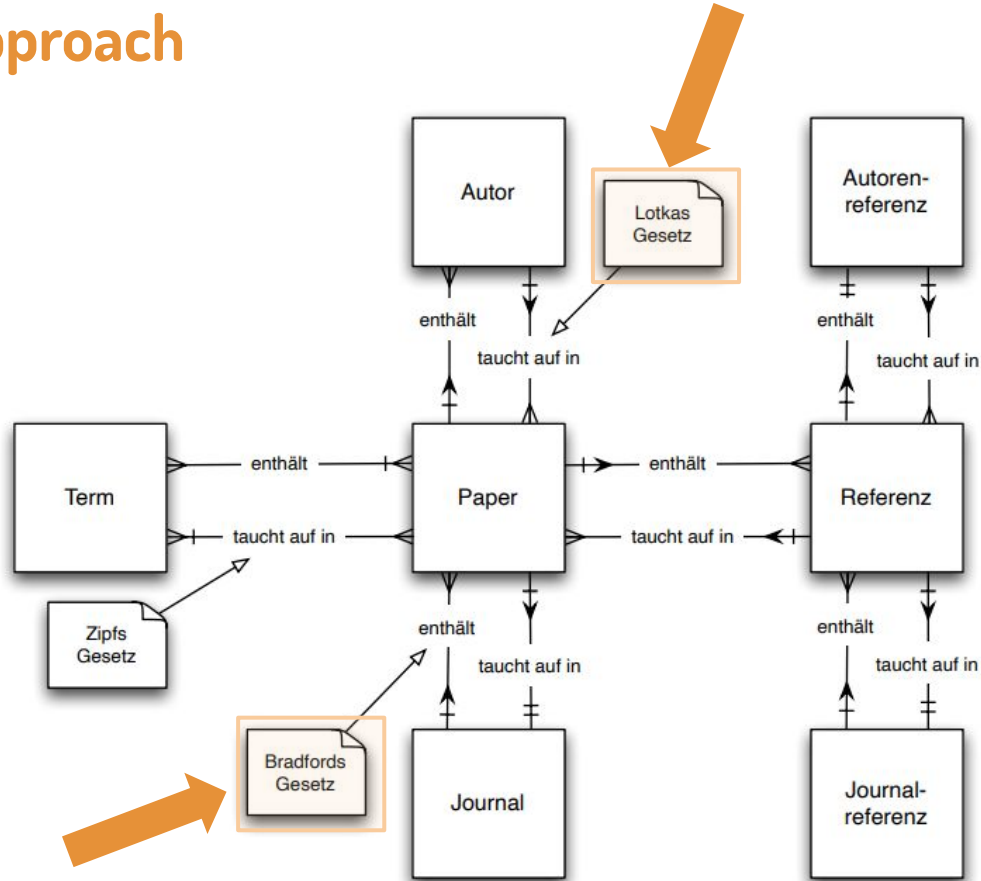


FIG. 1. The Bradford regions. Each search region contains one-third of the articles on the subject. Each ring is five times the area of the next smaller one.

General Approach



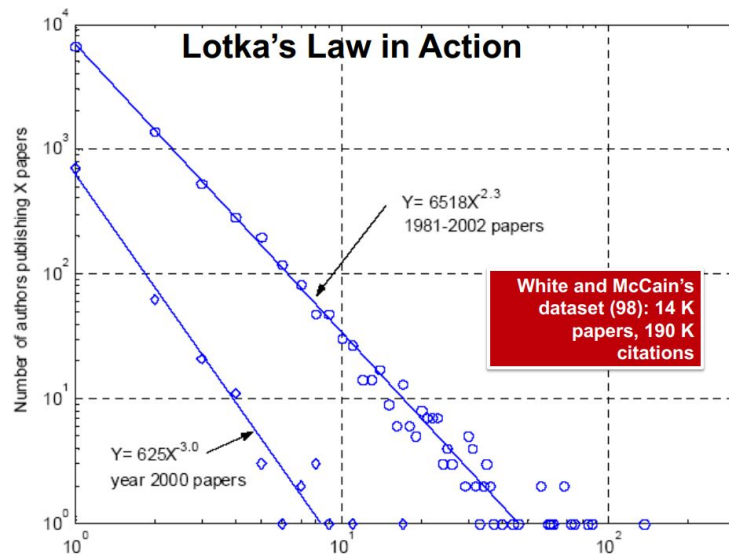
General Approach

Bradford's Law

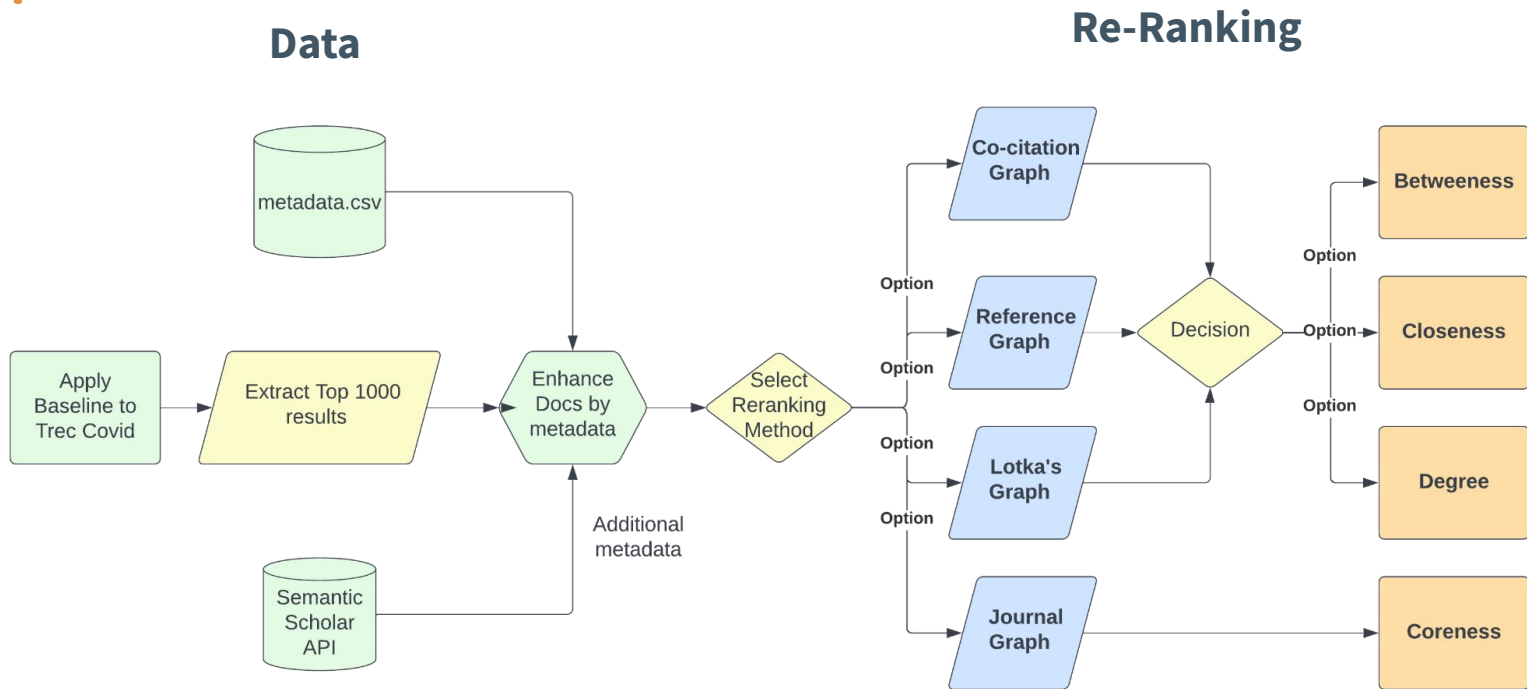
- ▷ Identifying Core Journals
- ▷ Boosting Papers by occurrence of journal

Lotkas Law

- ▷ Creating co-citation and reference graph
- ▷ Calculate and compare different graph measures
- ▷ Boost by average and maximum measures for paper authors

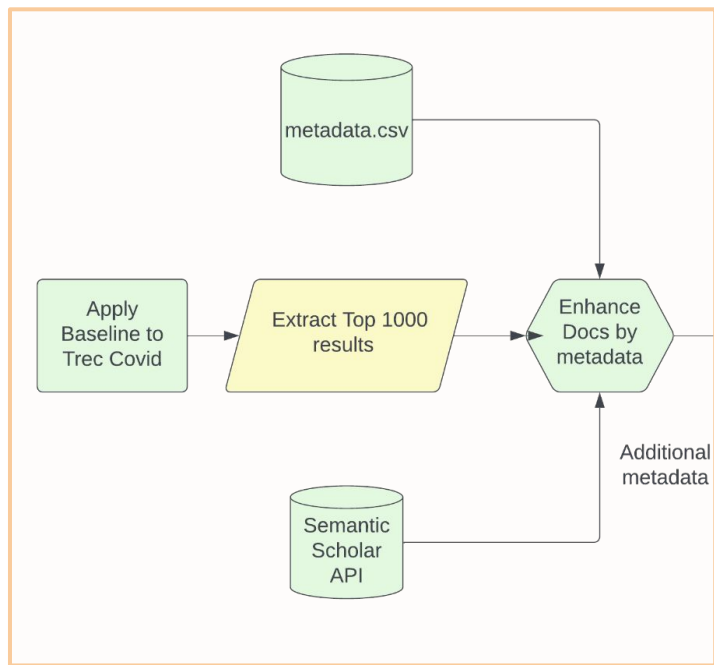


Pipeline

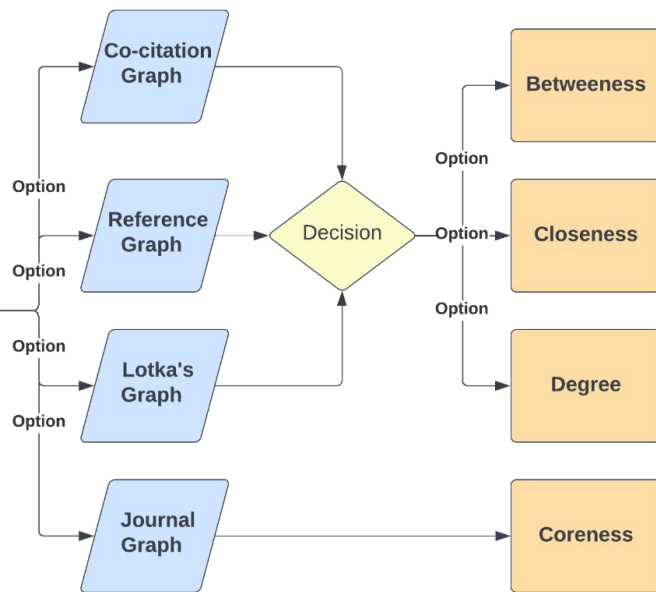


Pipeline

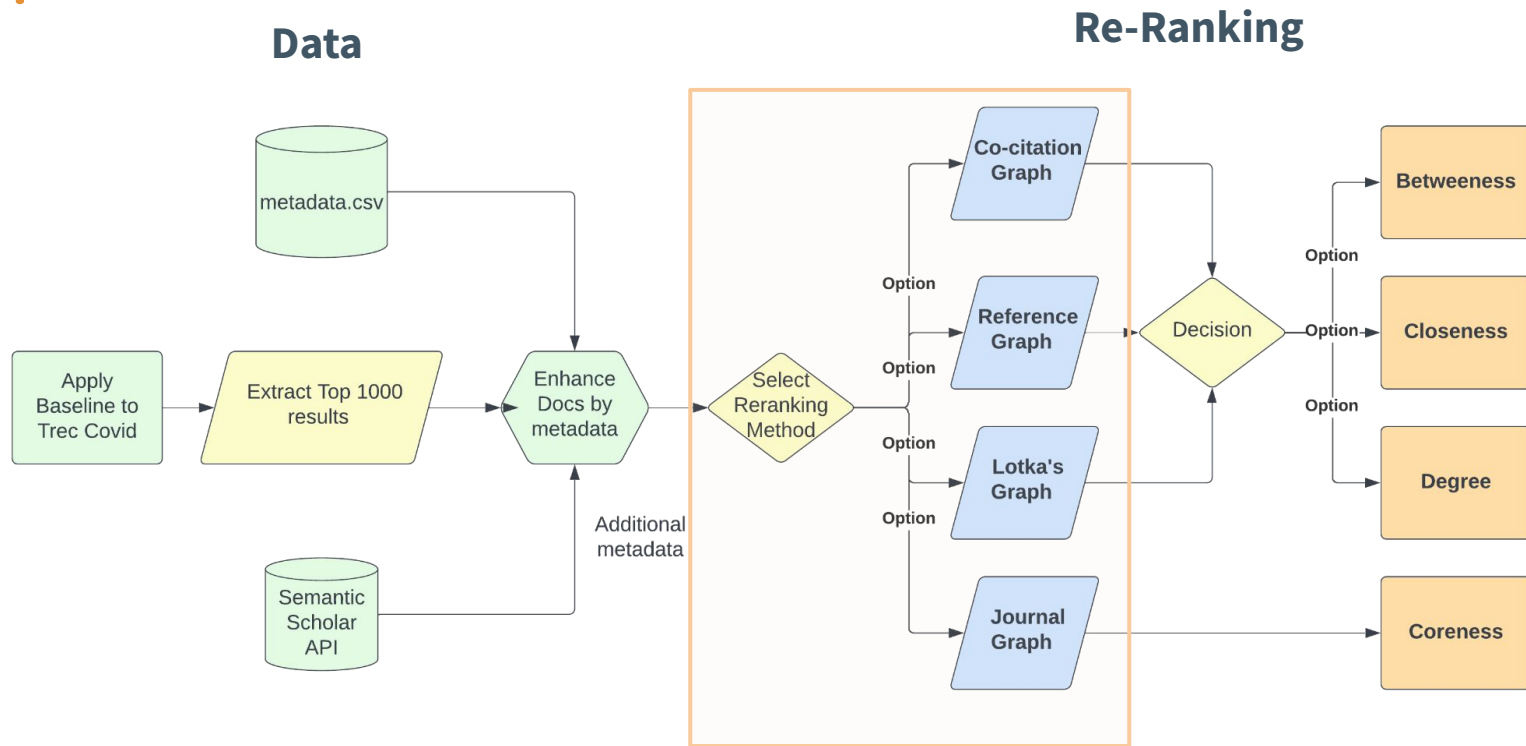
Data



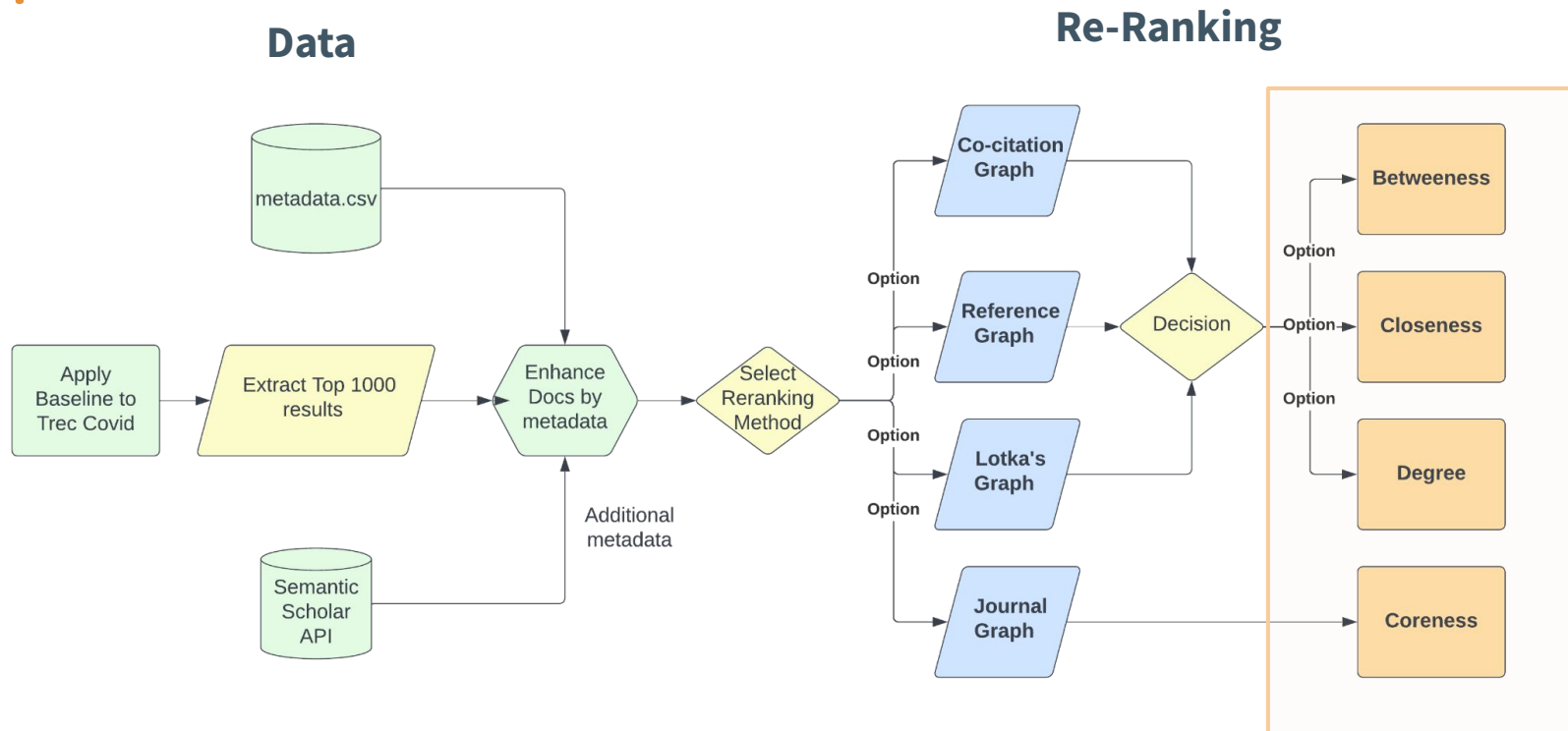
Re-Ranking



Pipeline



Pipeline



2. Creating a Baseline

Creating a Baseline

Cord-19

Data Set (*Version 2020-07-16*)

Weighting Model

BM25

Applied fields

title & abstract

Selection of Data

- *Top 1000* documents per query
- Removing elements without DOI
- Removing duplicates

Baseline: Performance

Cord-19 Data Set

(Version 2020-07-16),

Query 1: 'coronavirus origin'

Run Name	map	P@20	Recall@20	MRR	ndcg_cut_20
Baseline	0.1028	0.7	0.02	0.5	0.5219

Weighting Model

BM25

Applied fields

title & abstract

3. Data Preparation

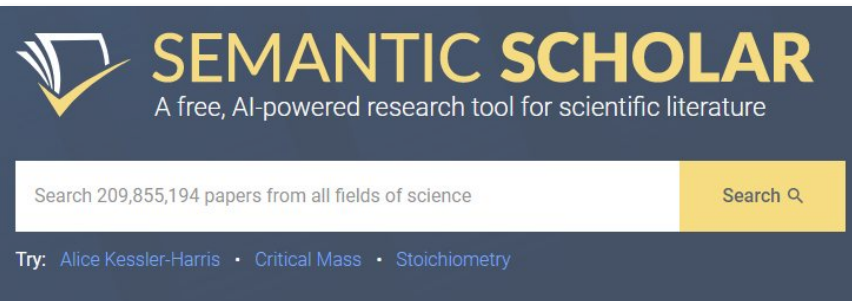
Data Enrichment

Enrichening metadata with *Semanticscholar API*

Metadata used

Authors

- ▶ *authors.name*
- ▶ *authors.affiliations*
- ▶ *authorId*



Papers

- ▶ *fieldsOfStudy*
- ▶ *s2FieldsOfStudy*
- ▶ *Citations.authors*
- ▶ *paperId*
- ▶ *Journal*

4. Creating Graphs

Creating Graphs

Creating various graphs and their bibliometric measures

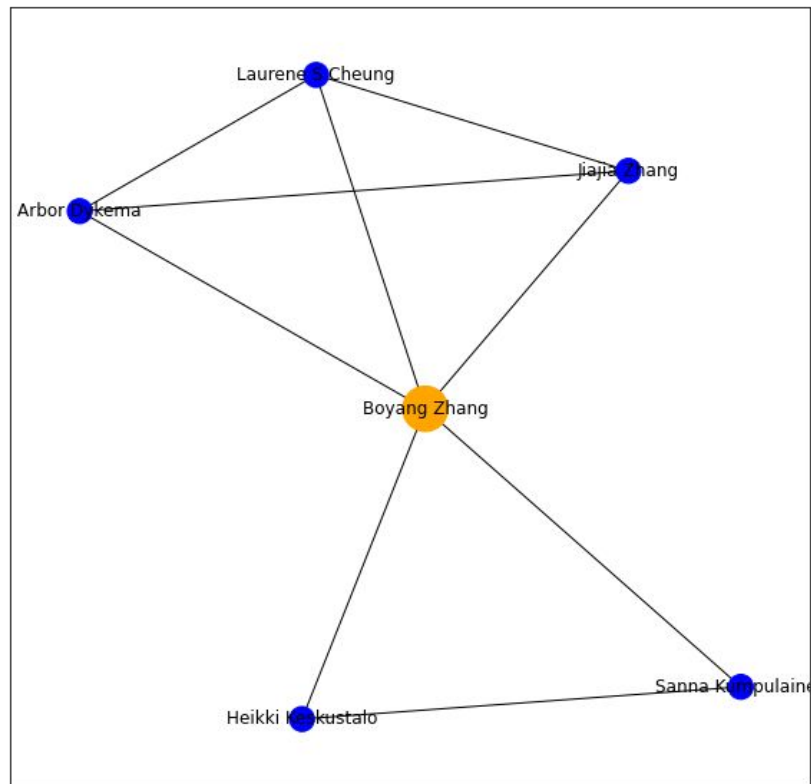
1. **Co-citation** Graph
2. **Lotka-inspired** Graph
3. **Citation** Graph
 - a. Between authors
 - b. Between papers
4. **Journal** Graph

Co-citation Graph

Co-authors with
Author → **Author**
 → Undirected Relationship

ID
 → By authorId

Relevant fields
 → authorId

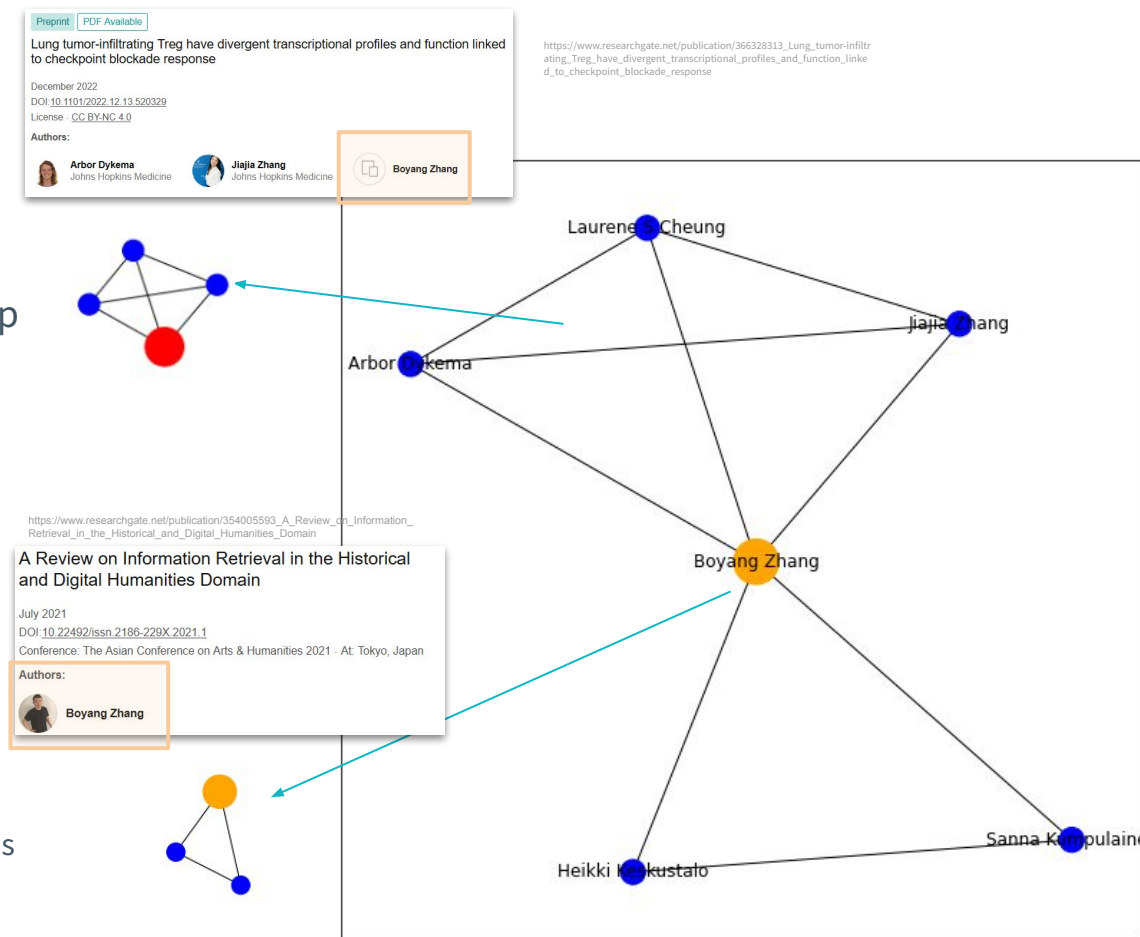


Co-citation Graph

- Co-authors with
- Author** → **Author**
- Undirected Relationship
- ID
- By authorId
- Relevant fields
- authorId

However: Potential conflict

- Same name, different persons
- In different research fields



Lotka-inspired Graph

written by

Paper → **Author**

→ Directed Relationship

ID

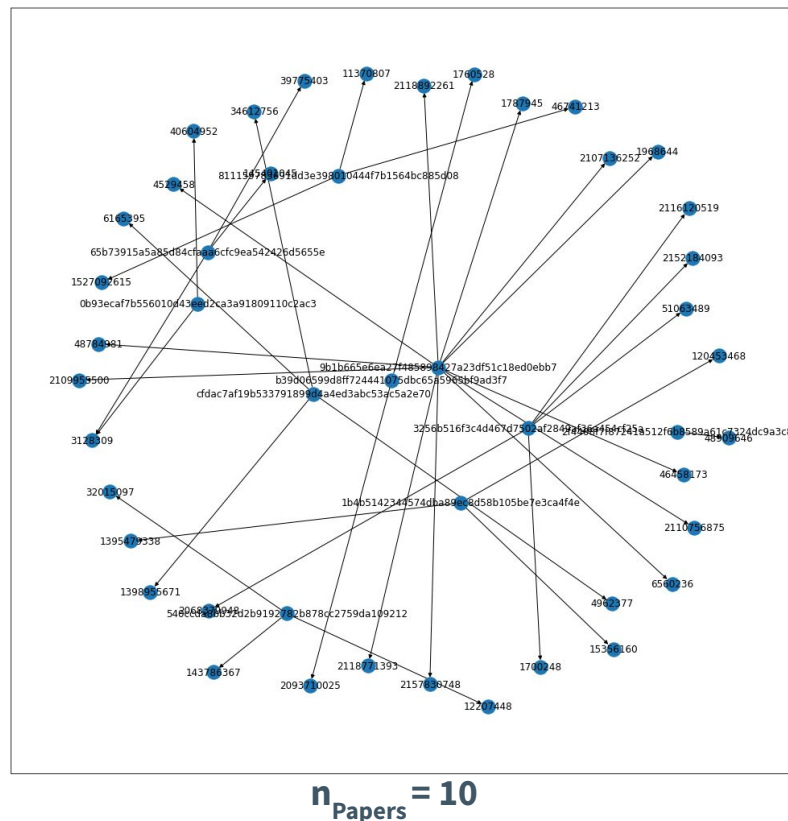
→ By authorId

Relevant fields

→ authorId

→ paperId

→ Highlights **author prominent** in many papers



Enriched Lotka-inspired Graph

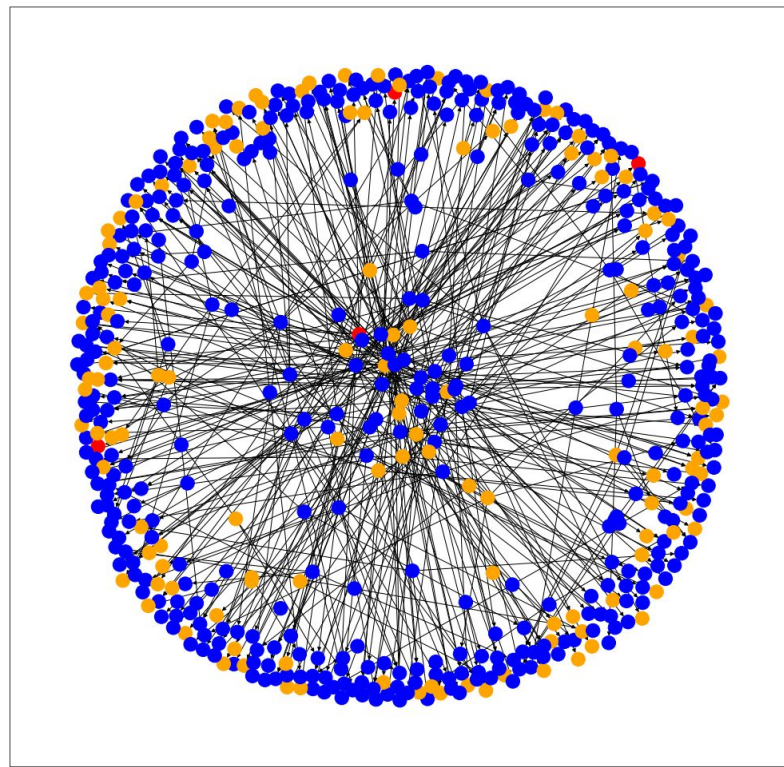
Paper ^{written by} **Author**

→ Directed Relationship

Secondary Relationships

→ Related authors through citations in the papers

→ Highlights **authors prominent** in many papers



$n_{\text{Papers}} = 2$

Enriched Lotka-inspired Graph

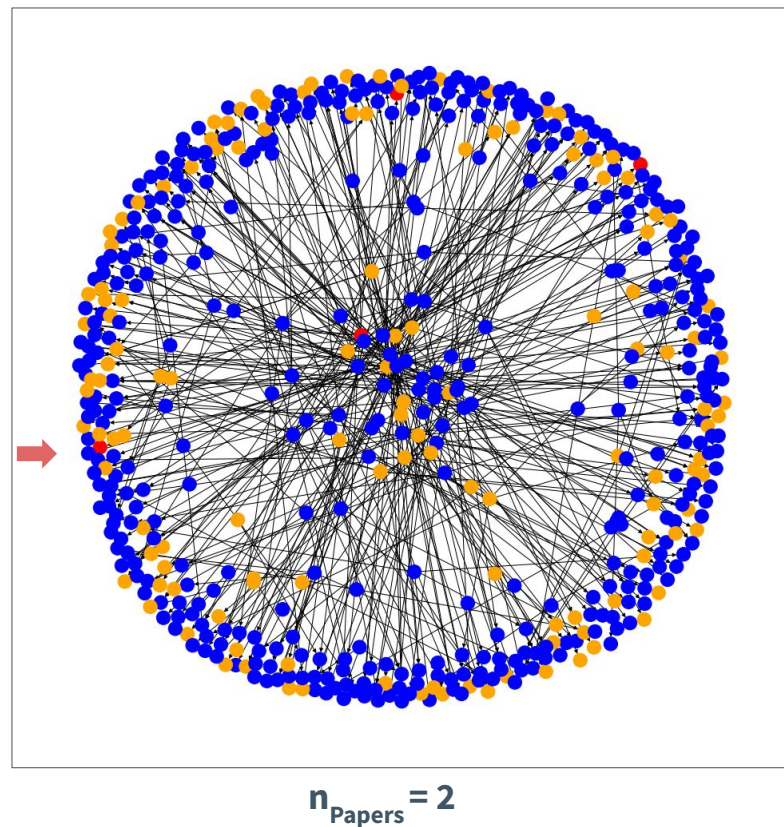
Paper ^{written by} **Author**

→ Directed Relationship

Secondary Relationships

→ Related authors through citations in the papers

→ Highlights **authors prominent** in many papers



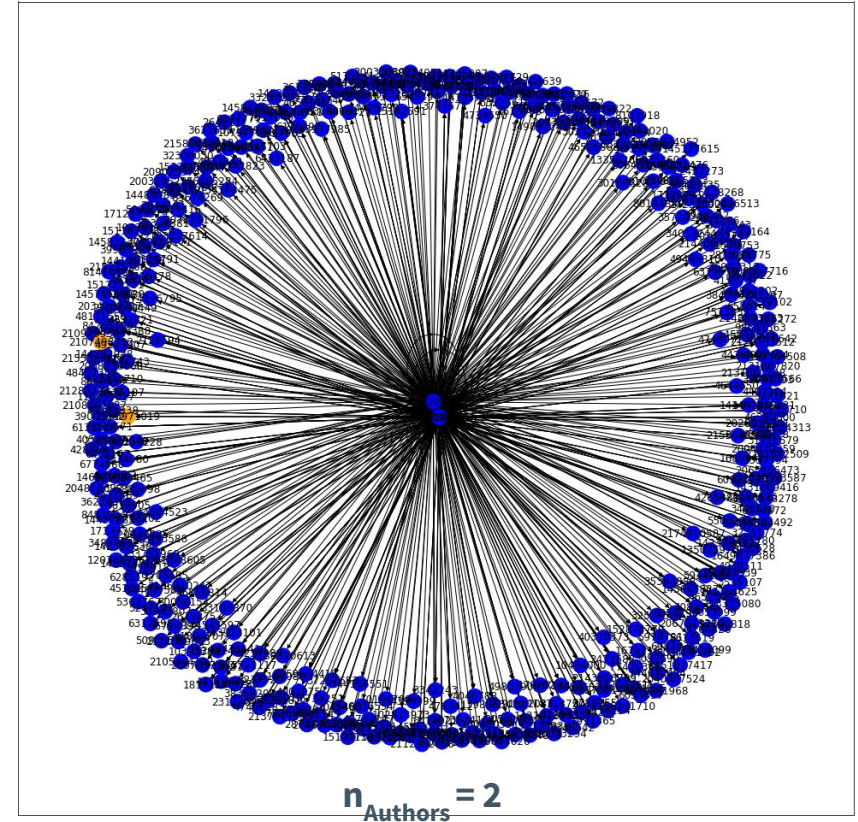
Citation Graph (between Authors)

Based on

Author ^{cites} Author
 → Directed Relationship

Describes

- The source of **citations**
- **Authors** that often get cited



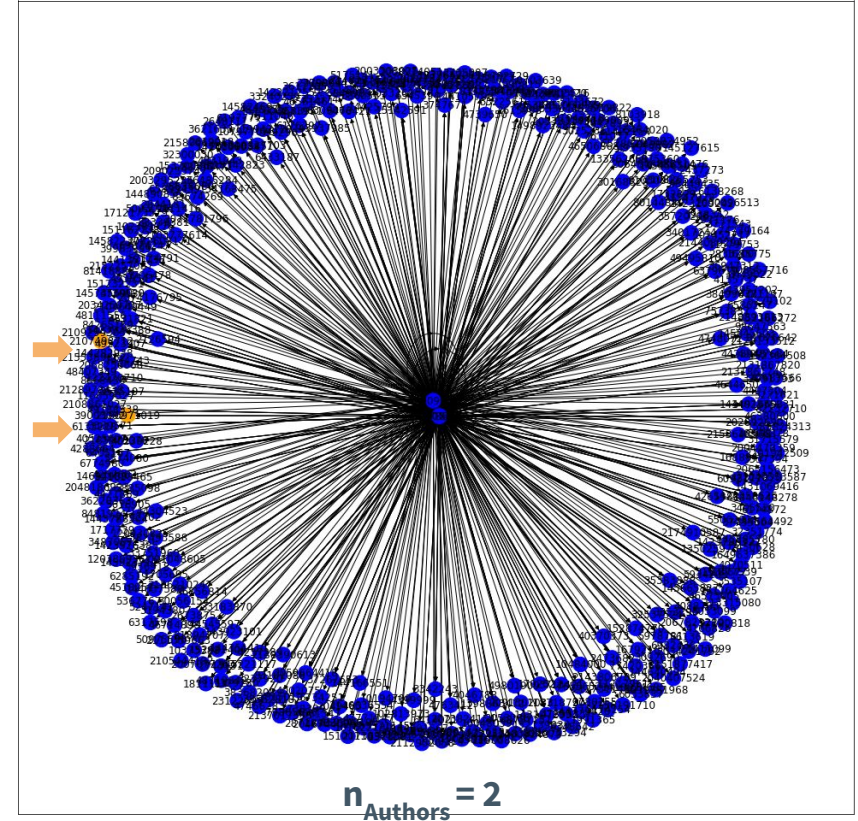
Citation Graph (between Authors)

Based on

■ Author ^{cites} → ■ Author
→ Directed Relationship

Describes

- The source of **citations**
- ■ **Authors** that often get cited



Citation Graph (between Papers)

Based on

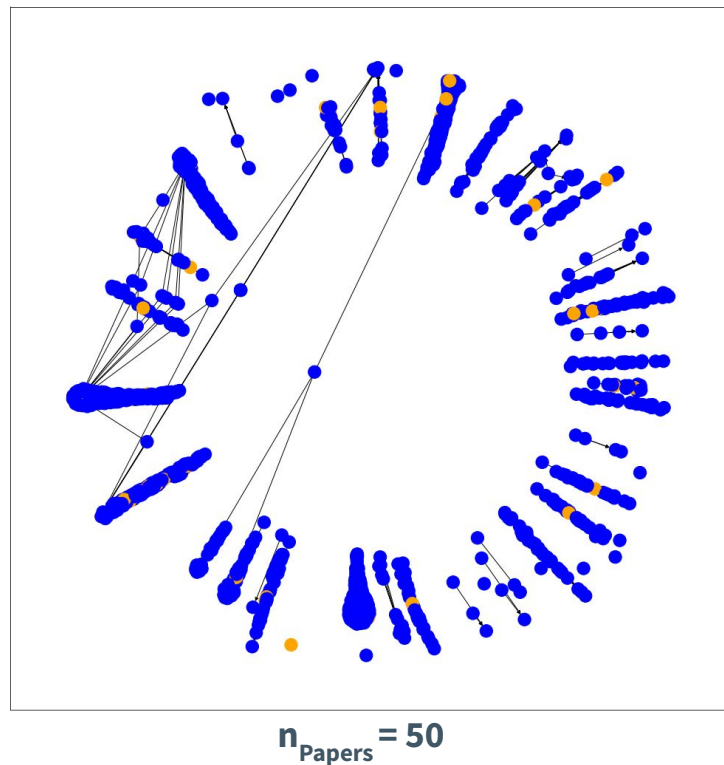
references

■ Paper → ■ Paper

→ Directed Relationship

Describes

- Direction of **references**
- ■ **Papers** that often get **references**
- Distinctive reciprocal citation clusters



Citation Graph (between Papers)

Based on

references

!Paper → !Paper

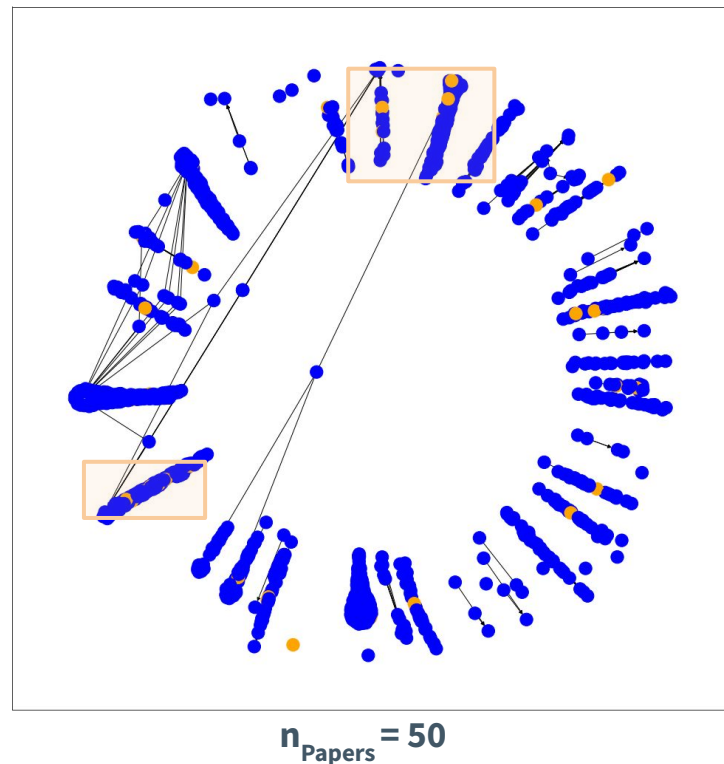
→ Directed Relationship

Describes

→ Direction of **references**

→ !Papers that often get **references**

→ Distinctive reciprocal citation clusters



Journal Graph

Based on

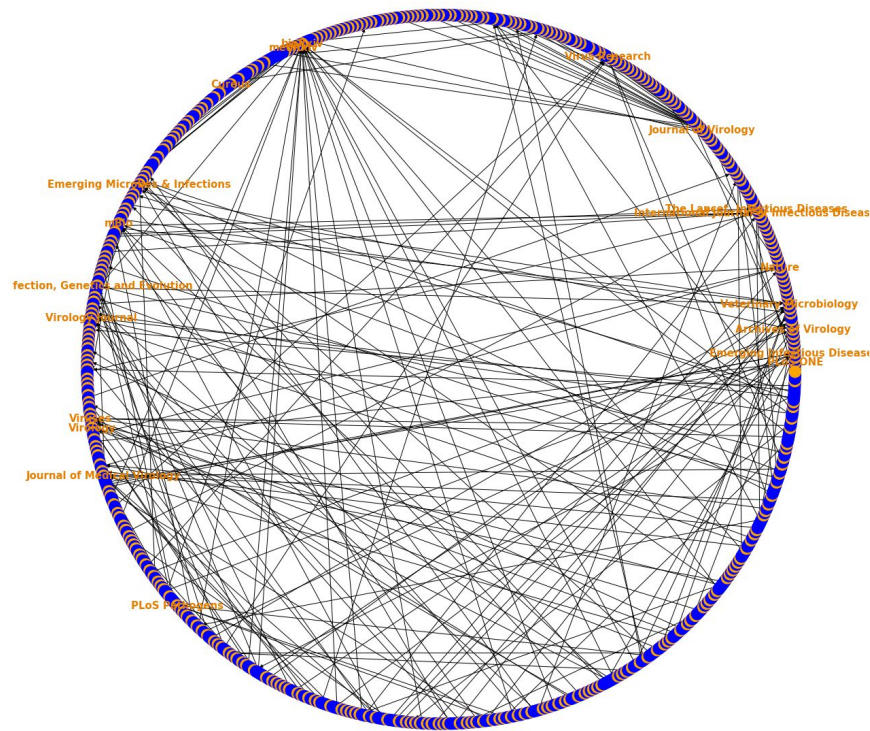
is part of

Paper → **Journal**

→ Directed Relationship

Describes

- **Prominent Core Journals** that feature most papers
- Visualizes the influence of a core journal in the subject



$n_{\text{Papers}} = 596$

5. Experiments & Evaluation

Observing the effects of boosting in re-ranking

Centrality Measures

Degree Centrality

→ D_{fss}

Betweenness Centrality

→ F_{das}

Closeness Centrality

Experiments

Boost by

→ maximum {centrality measure} from **author/papers/journals**

Boost by

→ average {centrality measure} from **all authors/papers/journals**

Boost by **connection** to **most popular author**

→ For *high/low* distance

...in the re-ranking process

Execute PyTerrier Runs on

→ Experimental graphs with bibliometrical metadata

1. Author co-citation
2. Author popularity
3. Citation between papers
4. Journal Coreness

Compare IR metrics

→ [map], [P@10], [P@20], [P@100], [Recall@20], [Recall@100] [RecipRank], [ndcg_cut_20]

Results: Author Co-Citation

Based on → centrality

	name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
	Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.5	0.519000
	reranker_degree_mean	0.103915	0.9	0.70	0.48	0.020029	0.068670	0.5	0.524213
	reranker_degree_max	0.103710	0.8	0.60	0.50	0.017167	0.071531	0.5	0.487725
	reranker_closeness_mean	0.103760	0.8	0.65	0.51	0.018598	0.072961	0.5	0.499142
	reranker_closeness_max	0.105412	0.8	0.65	0.52	0.018598	0.074392	1.0	0.579050
	reranker_betweenness_mean	0.103230	0.9	0.70	0.48	0.020029	0.068670	0.5	0.525544
	reranker_betweenness_max	0.102602	0.8	0.70	0.49	0.020029	0.070100	1.0	0.563768

→ Slight increases in [*map, recip_rank & recall@100 & ndcg*]

Results: Author Popularity

Based on

→ Distance to most popular author

name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.50	0.519000
reranker_most_popular_user_high_dist_mean	0.092684	0.7	0.70	0.46	0.020029	0.065808	1.00	0.561533
reranker_most_popular_user_high_dist_max	0.092689	0.7	0.70	0.46	0.020029	0.065808	1.00	0.561533
reranker_most_popular_user_short_dist_mean	0.092371	0.4	0.60	0.40	0.017167	0.057225	0.50	0.329489
reranker_most_popular_user_short_dist_max	0.091112	0.3	0.55	0.40	0.015737	0.057225	0.25	0.294177

name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.5	0.519000
reranker_most_popular_user_high_dist_mean	0.103087	0.8	0.70	0.49	0.020029	0.070100	1.0	0.603958
reranker_most_popular_user_high_dist_max	0.103077	0.8	0.70	0.49	0.020029	0.070100	1.0	0.604214
reranker_most_popular_user_short_dist_mean	0.101879	0.6	0.65	0.48	0.018598	0.068670	0.5	0.477122
reranker_most_popular_user_short_dist_max	0.101573	0.6	0.65	0.47	0.018598	0.067239	0.5	0.475326

name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.5	0.519000
reranker_most_popular_user_high_dist_mean	0.103023	0.8	0.70	0.49	0.020029	0.070100	1.0	0.603958
reranker_most_popular_user_high_dist_max	0.102999	0.8	0.70	0.49	0.020029	0.070100	1.0	0.604214
reranker_most_popular_user_short_dist_mean	0.103076	0.7	0.60	0.49	0.017167	0.070100	0.5	0.462294
reranker_most_popular_user_short_dist_max	0.103077	0.7	0.65	0.50	0.018598	0.071531	0.5	0.476451

No log + cutoff 10

Log10, cutoff 10

Log10, cutoff 5

- Boosting by short distance to most popular author worsens the results
- Boosting by long distance slightly increases [*map*, *recip_rank*, *ndcg*]

Results: Author Popularity

Based on

→ Distance to most popular author

name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.50	0.519000
reranker_most_popular_user_high_dist_mean	0.092684	0.7	0.70	0.46	0.020029	0.065808	1.00	0.561533
reranker_most_popular_user_high_dist_max	0.092689	0.7	0.70	0.46	0.020029	0.065808	1.00	0.561533
reranker_most_popular_user_short_dist_mean	0.092371	0.4	0.60	0.40	0.017167	0.057225	0.50	0.329489
reranker_most_popular_user_short_dist_max	0.091112	0.3	0.55	0.40	0.015737	0.057225	0.25	0.294177
reranker_most_popular_user_short_dist_mean	0.101879	0.6	0.65	0.48	0.018598	0.068670	0.5	0.477122
reranker_most_popular_user_short_dist_max	0.101573	0.6	0.65	0.47	0.018598	0.067239	0.5	0.475326

No log+ cutoff 10

Log10, cutoff 10

name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.5	0.519000
reranker_most_popular_user_high_dist_mean	0.103023	0.8	0.70	0.49	0.020029	0.070100	1.0	0.603958
reranker_most_popular_user_high_dist_max	0.102999	0.8	0.70	0.49	0.020029	0.070100	1.0	0.604214
reranker_most_popular_user_short_dist_mean	0.103076	0.7	0.60	0.49	0.017167	0.070100	0.5	0.462294
reranker_most_popular_user_short_dist_max	0.103077	0.7	0.65	0.50	0.018598	0.071531	0.5	0.476451

Log10, cutoff 5

- Boosting by short distance to most popular author worsens the results
- Boosting by long distance slightly increases [*map*, *recip_rank*, *ndcg*]

Results: Author Popularity

Based on

→ Distance to most popular author

	name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
	Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.50	0.519000
	reranker_most_popular_user_high_dist_mean	0.092684	0.7	0.70	0.46	0.020029	0.065808	1.00	0.561533
	reranker_most_popular_user_high_dist_max	0.092689	0.7	0.70	0.46	0.020029	0.065808	1.00	0.561533
	reranker_most_popular_user_short_dist_mean	0.092371	0.4	0.60	0.40	0.017167	0.057225	0.50	0.329489

	name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
	Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.5	0.519000
	reranker_most_popular_user_high_dist_mean	0.103087	0.8	0.70	0.49	0.020029	0.070100	1.0	0.603958
	reranker_most_popular_user_high_dist_max	0.103077	0.8	0.70	0.49	0.020029	0.070100	1.0	0.604214
	reranker_most_popular_user_short_dist_mean	0.101879	0.6	0.65	0.48	0.018598	0.068670	0.5	0.477122
	reranker_most_popular_user_short_dist_max	0.101573	0.6	0.65	0.47	0.018598	0.067239	0.5	0.475326

	Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.5	0.519000
	reranker_most_popular_user_high_dist_mean	0.103023	0.8	0.70	0.49	0.020029	0.070100	1.0	0.603958
	reranker_most_popular_user_high_dist_max	0.102999	0.8	0.70	0.49	0.020029	0.070100	1.0	0.604214
	reranker_most_popular_user_short_dist_mean	0.103076	0.7	0.60	0.49	0.017167	0.070100	0.5	0.462294
	reranker_most_popular_user_short_dist_max	0.103077	0.7	0.65	0.50	0.018598	0.071531	0.5	0.476451

Log10 + cutoff 10

Log10 , cutoff 10

Log10, cutoff 5

- Boosting by short distance to most popular author worsens the results
- Boosting by long distance slightly increases [*map*, *recip_rank*, *ndcg*]

Results: Author Popularity

Based on

→ Distance to most popular author

	name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
	Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.50	0.519000
	reranker_most_popular_user_high_dist_mean	0.092684	0.7	0.70	0.46	0.020029	0.065808	1.00	0.561533
	reranker_most_popular_user_high_dist_max	0.092689	0.7	0.70	0.46	0.020029	0.065808	1.00	0.561533
	reranker_most_popular_user_short_dist_mean	0.092371	0.4	0.60	0.40	0.017167	0.057225	0.50	0.329489
	reranker_most_popular_user_short_dist_max	0.091112	0.3	0.55	0.40	0.015737	0.057225	0.25	0.294177

	name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
	Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.5	0.519000
	reranker_most_popular_user_high_dist_mean	0.103087	0.8	0.70	0.49	0.020029	0.070100	1.0	0.603958

	name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
	Baseline	0.102866	0.9	0.70	0.49	0.020029	0.070100	0.5	0.519000
	reranker_most_popular_user_high_dist_mean	0.103023	0.8	0.70	0.49	0.020029	0.070100	1.0	0.603958
	reranker_most_popular_user_high_dist_max	0.102999	0.8	0.70	0.49	0.020029	0.070100	1.0	0.604214
	reranker_most_popular_user_short_dist_mean	0.103076	0.7	0.60	0.49	0.017167	0.070100	0.5	0.462294
	reranker_most_popular_user_short_dist_max	0.103077	0.7	0.65	0.50	0.018598	0.071531	0.5	0.476451

Log10 + cutoff 10

Log10, cutoff 10

Log10, cutoff 5

- Boosting by short distance to most popular author worsens the results
- Boosting by long distance slightly increases [*map*, *recip_rank*, *ndcg*]

Results: Lotka-Inspired Graph

Based on

→ activity of an author

name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
Baseline	0.102866	0.9	0.7	0.49	0.020029	0.07010	0.5	0.519000
reranker_lotka_degree_mean	0.103161	0.9	0.7	0.49	0.020029	0.07010	0.5	0.520775
reranker_lotka_degree_max	0.103370	0.9	0.7	0.48	0.020029	0.06867	0.5	0.523919
reranker_lotka_closeness_mean	0.103345	0.9	0.7	0.49	0.020029	0.07010	0.5	0.521625
reranker_lotka_closeness_max	0.103685	0.9	0.7	0.48	0.020029	0.06867	0.5	0.525544

→ Slight increases in $[map \& ndcg_cut@20]$

Results: Citation Graph (between papers)

Based on

→ centrality between papers

name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
Baseline	0.102866	0.9	0.7	0.49	0.020029	0.070100	0.5	0.519000
reranker_citation_paper_degree	0.104598	0.8	0.7	0.51	0.020029	0.072961	0.5	0.504761
reranker_citation_paper_closeness	0.105153	0.8	0.7	0.52	0.020029	0.074392	1.0	0.514877

→ Slight increases in [*map*, *recip_rank* & *recall@100*]

Results: Journal Graph

Based on

→ Coreness of the Journals

name	map	P_10	P_20	P_100	recall_20	recall_100	recip_rank	ndcg_cut_20
Baseline	0.102866	0.9	0.7	0.49	0.020029	0.0701	0.5	0.519000
reranker_graph_coreness	0.106561	0.9	0.8	0.49	0.022890	0.0701	1.0	0.643262

→ Increases in *[map, recall, recip_rank and ndcg_cut@20]*

Results

Run Name	:affix	:weight	map	P@20	Recall@20	MRR	ndcg_cut_20
Baseline	default	-	0.1028	0.7	0.02	0.5	0.5219
Co-Citation	avg. centrality	0.3	0.xxxx	0.xxxx	0.xxxx	0.xxxx	0.xxxx
Co-Citation	max. centrality	0.3	0.xxxx	0.xxxx	0.xxxx	0.xxxx	0.xxxx
Citation	default	0.3	0.xxxx	0.xxxx	0.xxxx	0.xxxx	0.xxxx
Citation	dithered	0.3	0.xxxx	0.xxxx	0.xxxx	0.xxxx	0.xxxx
Journal Coreness	default	1.3	0.xxxx	0.xxxx	0.xxxx	0.xxxx	0.xxxx

6. Conclusion

Conclusion

Observations concerning the boosting of certain metrics in the re-ranking

- **Author co-citation:** Leads to ► slight increases in **map**, **recall@100** & **ndcg_cut@20**, but leads to worse results in **P@20** and **Recall@20**
- **Popularity:** Boosting by **shortest path** to the **most popular node**
▼ worsens the results for top results but ▲ improves results within the Top 100

Conclusion

Observations concerning the boosting of certain metrics in the re-ranking

- **Lotka:** Boosting papers according to the **productivity** of an author leads to
▶ small gains in **map** and **ndcg_cut_20**, but ▼ losses in **recall@100**
- **Citations between papers:** Results in ▲ gains in **map**, **p@100** and **recall@100**
- **Coreness:** Improves ranking only ▶ slightly with no visible losses

Lesson's learned

NetworkX's bad scalability for extensive graph analysis

- External programs such as GraphVis, Cytoscape, etc. might be more suitable

Semanticscholar API

- lead to insufficient metadata for **Field of Science (FOS)** and **Affiliations** for Graph Analysis

Future Work

Limited scope

- Incorporate graphs of other topic queries
- Apply and compare with a more robust baseline

Limited Interactivity

- NetworkX graphs only allow static views of graphs

Contributions

Andreas Kruff:

Research, Preprocessing, Implementation of Graphs & Metrics, Visualizations of Graphs

Anh Huy Tran:

Research, Implementation of Metrics, Experiments, Analysis and Evaluation of Experiments

- [1] Mayr, P. (2009, März). *Re-Ranking auf Basis von Bradfordizing für die verteilte Suche in Digitalen Bibliotheken*. <https://www.researchgate.net>. Abgerufen am 28. November 2022, von https://www.researchgate.net/publication/260282769_Re-Ranking_auf_Basis_von_Bradfordizing_fur_die_verteilte_Suche_in_Digitalen_Bibliotheken
- [2] Sahraoui, A. K. & Mayr, P. (2018, März). *Users are not influenced by high impact and core journals while searching*. <https://www.researchgate.net>. Abgerufen am 28. November 2022, von https://www.researchgate.net/publication/324562131_Users_are_not_influenced_by_high_impact_and_core_journals_while_searching

THANKS FOR LISTENING!

Any questions?