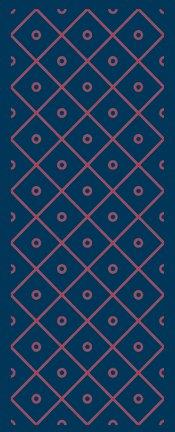
Charles University

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University



Master thesis

Social network analysis in academic environment

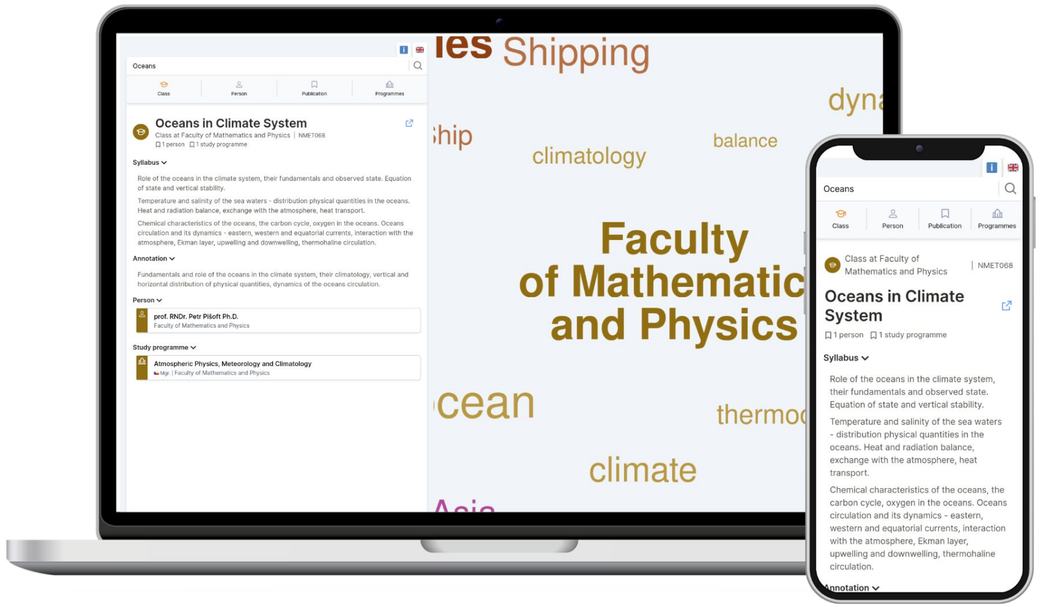
Bc. Jindřich Bär, 2024

FACULTY OF MATHEMATICS AND PHYSICS
Charles University

Charles Explorer

- Open-source academic search engine developed for Charles University
- Search in **people**, **courses**, **study programmes** and **publications**
- Goal: improve UX over existing solutions + exploratory tool

- Available at https://explorer.cuni.cz
- Source available in GitLab.



Problems

- **Existing data exports** (SIS, WHOIS, Verso) **are flawed**
  - Data input inconsistencies, schema irregularity
  - Missing identifiers for guest authors(!), only names
- **In-application ranking**
  - The search result ranking is easily skewed
  - Producing unexpected results for certain queries
- **Visualization tool** provides unsatisfactory results

  => **Solution(?)**: Mining the **social networks** in the data

Identity inference

- Reconstructing the missing identifiers for external authors
- **Baseline**: Merging all authors of the **same name**
  - Test data evaluation:

Method	Value
Macro-averaged F_1 score	0.874959
Micro-averaged F_1 score	0.900994

- **Proposed approach**:
  - Social-network based hierarchical merging
    - Performant implementation of distance matrix calculation
    - Hierarchical clustering with parametrized cutoff cophenetic distance

Evaluation

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

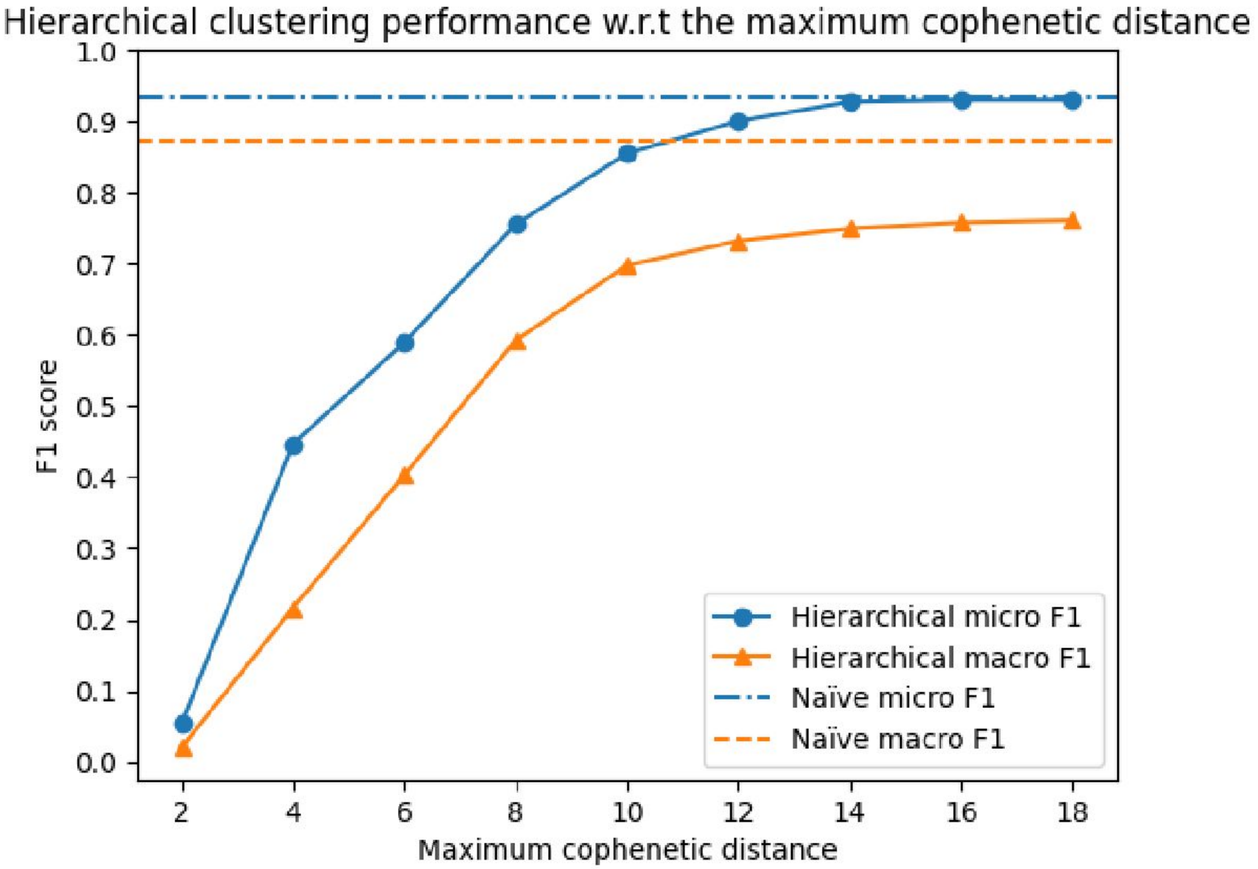


Figure 2.11: The F1 scores for the hierarchical clustering algorithm for different threshold values.[14]

- Proposed solution dominated by the naïve merging
- Likely caused by the specific data distribution in the input dataset

Result reranking

- Mitigating the effect of “**SEO optimization**”
- Global author / publication relevance:
  - Is not in the CUNI data
  - Federated queries to larger citation databases are **infeasible**
  - *Can it be mined from semi-local neighborhoods in the social network?*
- Searching for **correlation** between:
  - Node degree, **centrality measures** (Betweenness, Katz), node cut size
  - Search result ranking from **Scopus** (via nDCG), or
  - **Citation** and **reference count** for publications

Benchmark design

- Evaluating the proposed algorithms on a **precomputed query set**
  - The results must **fairly cover** the original dataset, w.r.t. faculty distribution
  - Wordnet seeding + “simulated annealing” search for the optimal KL
- Comparing:
  - Proposed ranking based on local social network structure
    - vs
  - Result rankings from Scopus, or
  - Rankings based on citation / reference count

Evaluation

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Feature	Coefficient
charles_explorer_ranking	-0.23068
centrality_1	0.08317
centrality_2	0.01018
degree	-0.08756
katz_centrality	-0.00491
node_cut	-0.02733

Figure 3.10: Coefficients of the linear regression model for the linear combination of the original relevance scores and the social network metrics.

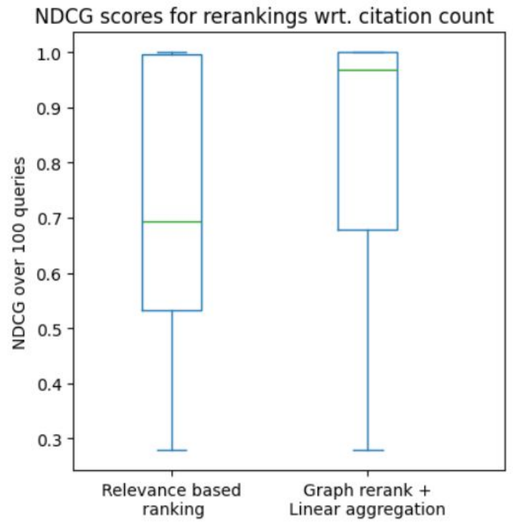


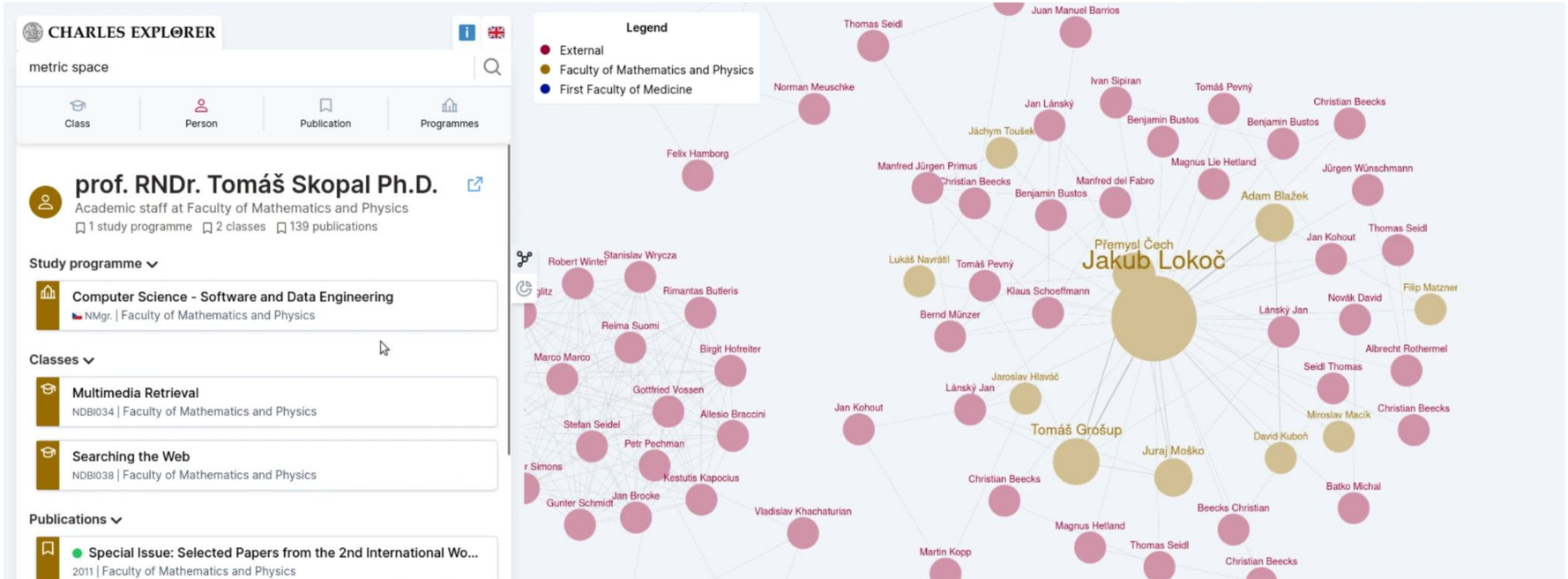
Figure 3.13: nDCG scores of 100 queries with different citation count prediction methods.

- Default Scopus ranking seems to be heavily full-text-relevance based
  - This means also potentially vulnerable to ASEO(!)
- Search for linear aggregation of local SocNet features yields promising results
  - nDCG weighs top results more, just like users would

FACULTY OF MATHEMATICS AND PHYSICS
Charles University

Visualization

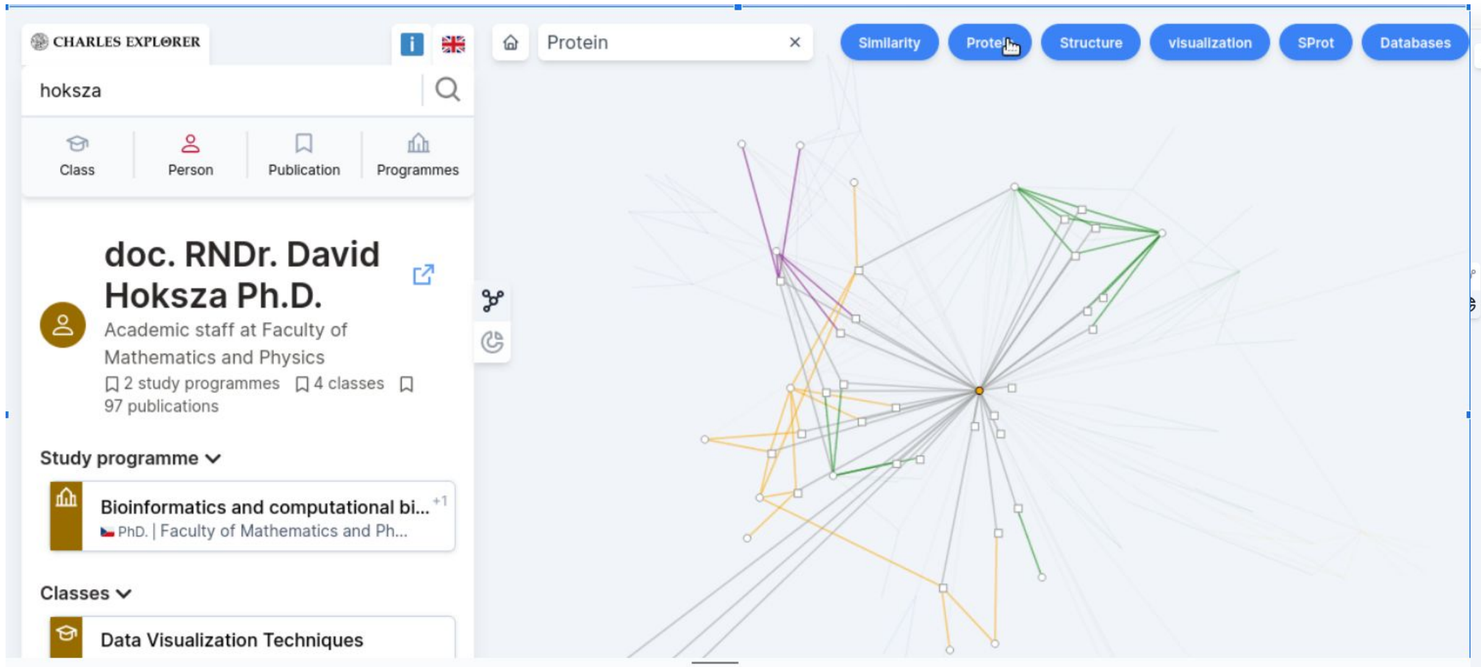
- Exploratory tool for the university structure
- More information passed “per view” + preattentive processing
- First implementation = causing user confusion and performance issues

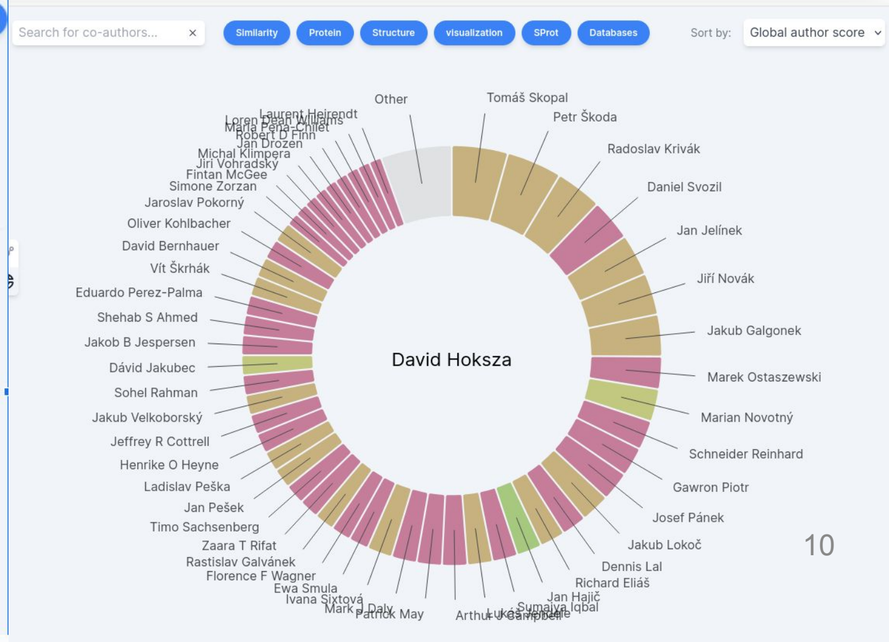


FACULTY OF MATHEMATICS AND PHYSICS
Charles University

Visualization, refined

- Removing unintuitive graph preprocessing (monopartite projection)
- Adding different modes (network / pie chart) - less load on each visualization
- Adding interactive search => from static visualization to true exploratory tool





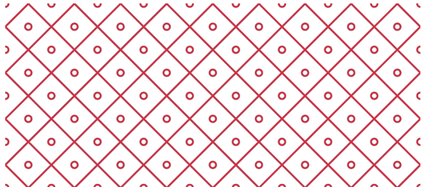
Questions

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Bc. Jindřich Bär

https://github.com/barjin

https://jindrich.bar



Deployment details

- 2 threads on Intel Xeon (~2015 - 2018)
- 8 GiB RAM (+ 8 GiB swap)
- 100 GiB HDD space

- PostgreSQL, Memgraph, Apache Solr, Node.JS, Python microservices
  - ~1.1 million DB records, ~3 million (implicit & explicit) edges

- Over different categories:
  - *p50* (median time) *1.19s* for response (round-trip)
    - *p95: 3.38s* (round-trip)

Evaluation results

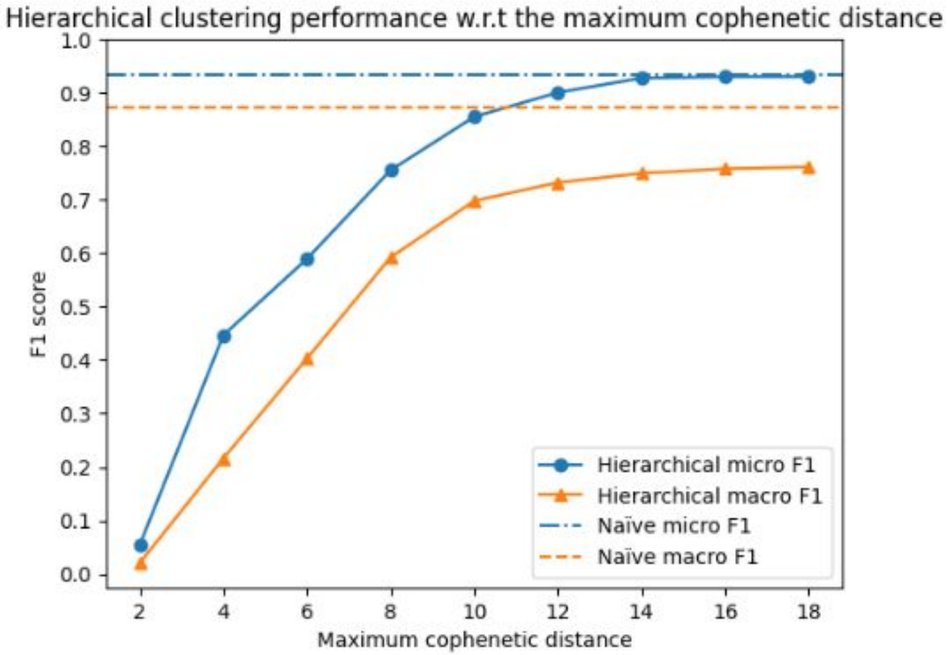


Figure 2.11: The F1 scores for the hierarchical clustering algorithm for different threshold values.[14]

After aggregation over all the queries, this gives us the following unfavorable statistics:

Mean	0.208727
Standard deviation	0.211699
Minimum	0.010101
25%	0.074786
50%	0.137028
75%	0.265263
Maximum	1.000000

Figure 3.7: Aggregated statistics of the F_1 score for the search results of Charles Explorer.

	dcg	idcg	ndcg
mean	14.919819	19.167405	0.761607
std	16.810894	17.665142	0.180979
min	0.094340	0.094340	0.405669
25%	5.250473	7.704989	0.627563
50%	9.527864	14.840570	0.736246
75%	18.064385	24.112511	0.934206
max	104.693354	104.693354	1.000000

Figure 3.9: Aggregated statistics of the nDCG score for the original search results of Charles Explorer (*query count* = 149).

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Bc. Jindřich Bär

https://github.com/barjin

https://jindrich.bar