Wykrywanie społeczności w grafie współautorstwa DBLP

Grupa - Piątek 17^{50}-19^{20}

Paweł Banach
Michał Biel

Input set

- Downloaded from https://dblp.uni-trier.de/
- List of computer science articles, papers from conferences, proceedings, thesis published in journals
- Data collected since 1995

Input set

- One *big* xml file - around 2,5 GB
- More than 56 100 000 records
- Every type of paper has own tag
- Example

```
<article mdate="2017-05-17" key="journals/displays/LindbergNM06">
    <author>Tomas Lindberg</author>
    <author>Risto Nasanen</author>
    <author>Kiti Muller</author>
    <title>How age affects the speed of perception of computer icons.</title>
    <pages>170-177</pages>
    <year>2006</year>
    <volume>27</volume>
    <journal>Displays</journal>
    <number>4-5</number>
    <ee>https://doi.org/10.1016/j.displa.2006.06.002</ee>
    <url>db/journals/displays/displays27.html#LindbergNM06</url>
</article>
```

Input set

- We had to clean data, because it contained not valuable records for detecting society, for example
  - websites:

```
<www mdate="2018-06-05" key="homepages/220/4232">
<author>Luis Puche Rondon</author>
<title>Home Page</title>
</www>
```

  - phd/master thesis

```
</phdthesis><phdthesis mdate="2016-05-04" key="phd/de/Zivic2008">
<author>Natasa Zivic</author>
<title>Joint channel coding and cryptography.</title>
<year>2008</year>
<school>University of Siegen</school>
<pages>1-117</pages>
<isbn>978-3-8322-7180-0</isbn>
<ee>http://d-nb.info/99003707X</ee>
</phdthesis><phdthesis mdate="2016-05-04" key="phd/de/Kohlhase2008">
```

Input set

- We omitted dtd validation during parsing to speed up creation of input graph
- Around 25 times faster (~ 6 hours vs ~15 minutes)
- Dtd file was used to encoded special characters
- We validated data with comparing our result graph with an one small file made from original one

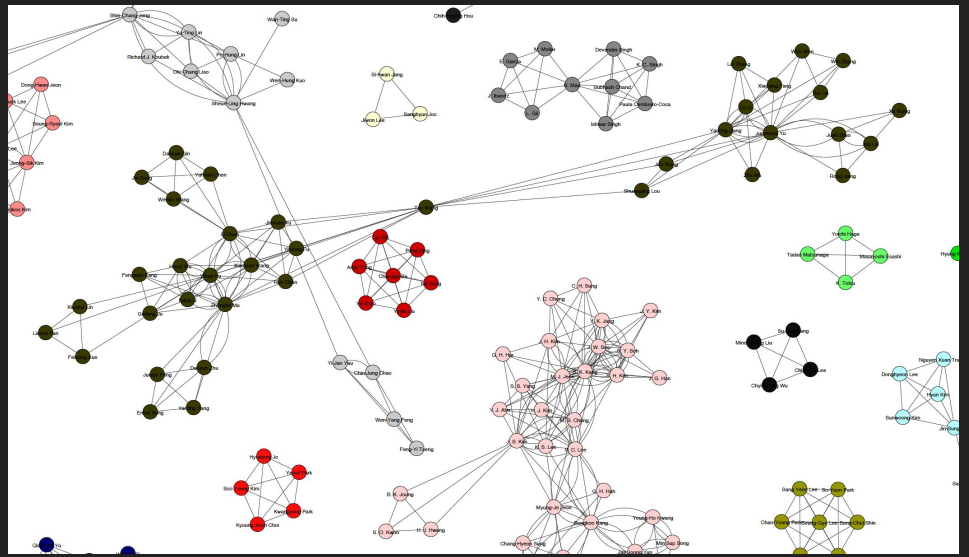
Input set - first insight

Articles with more than 200 authors:

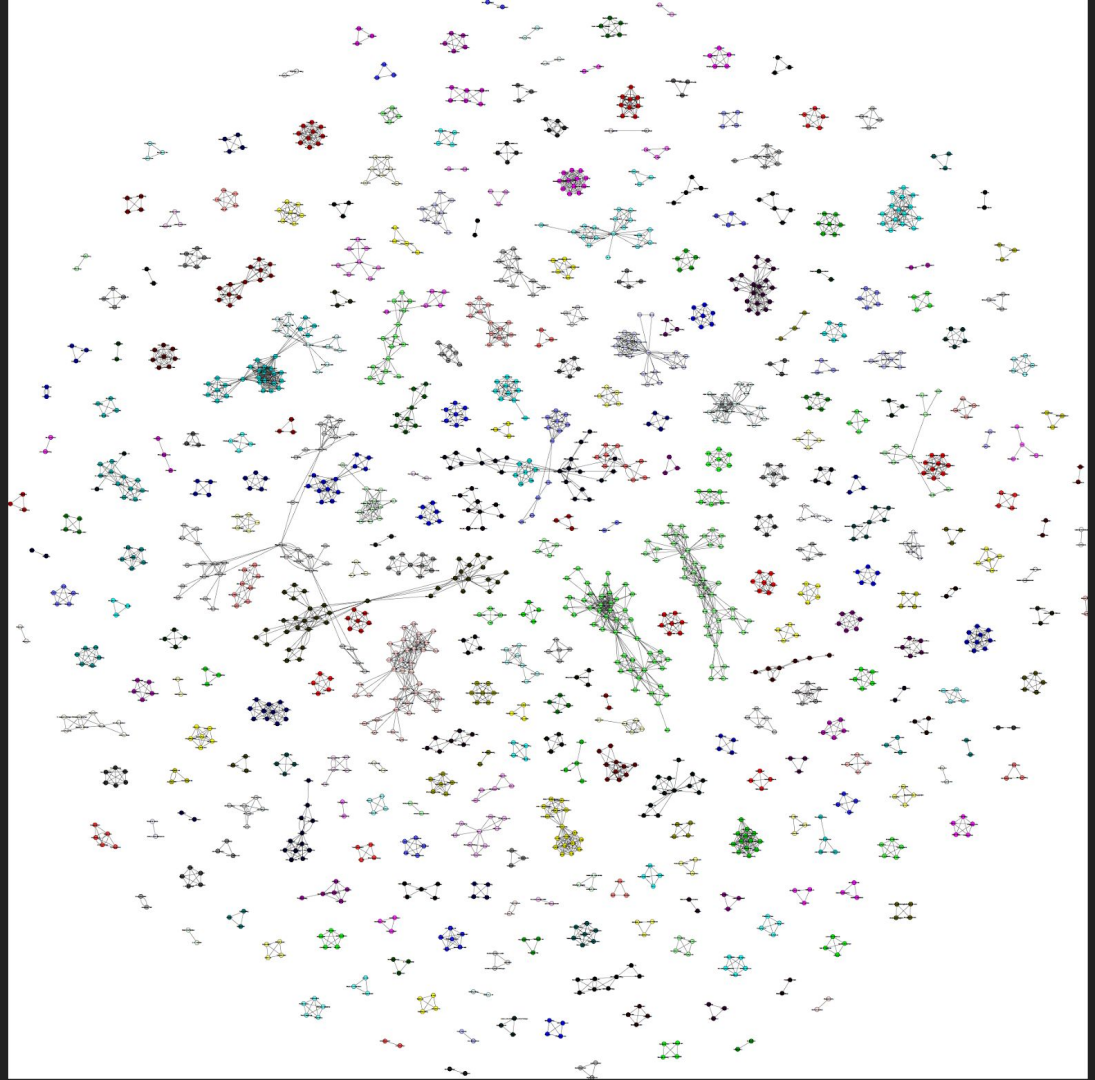
- A promoter-level mammalian expression atlas (263) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4529748/ (RIKEN Omics Science Center)
- The IceProd Framework: Distributed Data Processing for the IceCube Neutrino Observatory (287) https://arxiv.org/abs/1311.5904

First attempt

- Small file to choose graph library, algorithm for community detection, evaluate and show results



Fragment of graph for journal “Displays”



Libraries/algorithm

- We chose igraph as main library to manage graph, because it had a wide range of community detection algorithms and it had an option to plot results
- Community_multilevel was a main algorithm for community detection, because it run quite quick on this graph, its modularity was acceptable and it supported weighted edges

Detecting community using multilevel - algorithm

Community structure based on the multilevel algorithm of Blondel et al.

```
g = igraph.Graph()
g.add_vertices(graph_vertices)  # add a list of unique vertices to the graph
g.add_edges(graph_edges)  # add the edges to the graph.
p = g.community_multilevel()
print('Communities:')
```

Communities

```
[57185] Yvan Pelletier, Robert Caissie
[57186] Henry Meeter, Henry Kurz
[57187] Osama Abouelkhair, Amr Elsaadny
[57188] A. V. Duplinskiy, M. G. Alexanina, E. O. Kiktenko, George V. Tarasov,
        Pavel Postupalski, V. L. Kurochkin, R. R. Yunusov, P. A. Tregubov, L.
        Yegoshin, L. Semenov, Alec Miloslavski, A. S. Sokolov, A. I. Kotov, J.
        Boyle, M. Scharf, Brian Galvin, Konstantin Kishinski, Dmitriy I.
        Kharitonov, R. P. Ermakov, Y. V. Kurochkin, V. E. Ustimchik, V. Gusev,
        N. O. Pozhar, A. S. Trushechkin, Aleksey Simanchuk, A. I. Alexanin,
        Herbert Ristock, Nikolay A. Anisimov, I. A. Kabanov, M. N. Anufriev,
        Aleksey Kouvalenko, S. Dolgobrodov, D. A. Kronberg, S. E. Diyakov,
        Catherine Cooper-Weidner, Alan McCord, V. Antonov, C. C. W. Lim,
        Evgeniy A. Golenkov, A. K. Fedorov, A. V. Brodsky, S. S. Vorobey,
        Gregory Pogosyants, A. A. Kanapin, A. I. Lvovsky, S. Shkrabov, P. V.
        Babyak, H. Horch, A. V. Miller, Nikolay Korolev, Aleksey Kavalenko
[57189] Gianpaolo Spadini, Dipankar Pramanik
[57190] David G. Goldstein, Major Hal Clark
[57191] Ramona Lumpkin, Marjorie Armstrong-Stassen, Margaret Landstrom
[57192] Mehrdad Poorhosseini, Ali Reza Hejazi
[57193] Jaime Miranda Junior, Bruno Bessa, Caroline De Medeiros, Simone
        Santos, Thiago Mendes
[57194] Jianfang Shi, Jing Yang 0013, Qingshuang Zhu, Fei Wang 0022, Hongbiao
        Tang
[57195] Fernando Laudares Camargos, Benoit des Ligneris
[57196] Wun-Ting Hsu, Wen-Shu Lai
[57197] Richard A. Regueiro, Beichuan Yan
[57198] Chen-feng Ren, Yi-chen Ma, Xin-qiang Qin, Zhi-peng Zhang
[57199] Marian Piekarski, Adam Okninski, Marceli Uruski
```

Communities printed in console

Analysis

- Details about graph
- Most popular journals inside the five biggest communities
- Journals categories in communities
- Authors universities inside communities

Graph details - part I:

Number of communities:	57015
Degree distribution:	N = 2111917, mean +- sd: 18.3728 +- 55.9872
Average clustering coefficient:	0.7265728393390077
Vertices:	2111917
Edges:	19400929
Modularity:	0.738460033820342

Graph details - part II:

Average degree distribution:	18.37281389372946
Clique number:	287
Density:	8.699595009331 73e-06
Max degree:	4203
Person with max degree:	Wen Gao
Eigenvector centrality:	5.1713464105268 05e-06

The most popular journals inside the five biggest communities

- We found the biggest communities
- We count number of articles of each journal inside community
- The result presents the six most popular journals in the five biggest communities

The most popular journals inside the five biggest communities

Community	1	2	3	4	5
Size	210328	106019	90386	76031	72997
Journal 1	'EAI Endorsed Trans. Ubiquitous Environments', 111242	'EAI Endorsed Trans. Ubiquitous Environments', 58847	'EAI Endorsed Trans. Ubiquitous Environments', 35551	'EAI Endorsed Trans. Ubiquitous Environments', 44725	'EAI Endorsed Trans. Ubiquitous Environments', 31096
Journal 2	'SIGMOD Record', 64227	'SIGMOD Record', 35998	'SIGMOD Record', 21220]	'SIGMOD Record', 23346	'SIGMOD Record', 19641
Journal 3	'World Wide Web', 40208	'World Wide Web', 20199	'NeuroImage', 17324	'World Wide Web', 20177	'CoRR', 15904
Journal 4	'CoRR', 19181	'IEICE Transactions', 13982	'World Wide Web', 16589	'CoRR', 14063	'World Wide Web', 13724
Journal 5	'Int. J. Hum.-Comput. Stud.', 13347	'Int. J. Hum.-Comput. Stud.', 7508	'CoRR', 7153	Int. J. Hum.-Comput. Stud.', 11987	'Discrete Mathematics', 5182
Journal 6	'Sensors', 10240	'Systems and Computers in Japan', 4199	'IEEE Trans. Med. Imaging', 5223	'IEEE Trans. Communications', 6088	'Discrete Applied Mathematics', 4758

Journals categories in communities

- We used the most popular journals in the biggest 200 communities to check theirs categories
- We got categories from https://www.scimagojr.com/journalrank.php, this set contain a lot of journals from ours list, but we didn't find any categories in some communities
- Unfortunately, many communities had journals like 'EAI Endorsed Trans. Ubiquitous Environments', 'World Wide Web', that wasn't interesting and the main category was Software, but we found some interesting results

Journals categories in communities

- The category is present with counter
  - 10 [('Computer Science Applications', 11), ('Molecular Biology', 9), ('Computational Mathematics', 7), ('Computational Theory and Mathematics', 7), ('Software', 7), ('Artificial Intelligence', 7), ('Genetics', 5), ('Biochemistry', 5), ('Information Systems', 4), ('Applied Mathematics', 4)]
  - 13. [('Information Systems', 11), ('Computer Science Applications', 7), ('Computer Networks and Communications', 6), ('Library and Information Sciences', 6), ('Management of Technology and Innovation', 6), ('Software', 5), ('Management Science and Operations Research', 5), ('Strategy and Management', 5), ('Economics and Econometrics', 5), ('Information Systems and Management', 4)]
  - 110. [('Agronomy and Crop Science', 1), ('Animal Science and Zoology', 1), ('Computer Science Applications', 1), ('Forestry', 1), ('Horticulture', 1)
  - 113 [('Language and Linguistics', 3), ('Linguistics and Language', 3), ('Software', 2), ('Communication', 1), ('Hardware and Architecture', 1), ('Computer Networks and Communications', 1), ('Artificial Intelligence', 1), ('Arts and Humanities (miscellaneous)', 1), ('Human-Computer Interaction', 1), ('Psychology (miscellaneous)', 1)]
  - 114 [('Ecological Modeling', 1), ('Environmental Engineering', 1), ('Software', 1), ('Computational Theory and Mathematics', 1), ('Management of Technology and Innovation', 1), ('Management Science and Operations Research', 1), ('Modeling and Simulation', 1), ('Strategy and Management', 1), ('Numerical Analysis', 1), ('Statistics, Probability and Uncertainty', 1)]
  - 123. [('Analytical Chemistry', 1), ('Atomic and Molecular Physics, and Optics', 1), ('Electrical and Electronic Engineering', 1), ('Instrumentation', 1), ('Medicine (miscellaneous)', 1)]

Authors universities in communities

- Check universities where authors works
- Problem with getting real data, solution was to parse google scholar page and get data by div id
- No standard form, text with academic title or different order
- Problem with Chinese names, their spelling diffres
- Many authors don’t have universities on google scholar

Authors universities in communities

- Test, simple communities works fine (around 30), there was a few authors from Stanford University, two from Harvard, other from different universities and more than a half didn't present the university
- There was problem with Google in real research. The biggest communities contained more than 100 unique authors and crawling scholar webpage resulted in recaptcha requests, even when we put a long timeouts.

Links

GitHub: https://github.com/PawelBanach/data-mining