

MENTOR AND MENTEE RELATIONS BASED ON AUTHORSHIP GRAPHS

Reza Karimi

Elsevier



Outline

- Introduction:
 - Elsevier and Scopus Data
- Mentorship Model
- Training and Validation
- (Big) Graph Visualization (in Spark)



ELSEVIER

Who are we?

- Elsevier: the #1 provider of scientific, technical, and medical information, and a technology company originally established in 1880 in Netherlands. Now part of RELX group.
- Privileged to publish about 25% of cited (what matters) scientific publications
- Publishing 400K articles per year in about 2500 journals such as Lancet, Cell, Trends, Current Opinions, Artificial Intelligence (generally accessed via [ScienceDirect](#) platform (about 900m full text download per year)
- Curating Scientific publication data in products such as Scopus, SciVal, Pure



Full Text Access in ScienceDirect

Search results: 463 results found for (Apache Spark).

 Save search alert |  RSS

Refine filters

Year

- 2017 (2)
- 2016 (118)
- 2015 (74)
- 2014 (32)
- 2013 (13)

[View more >>](#)

Publication title

- Procedia Computer Science (43)
- Clinical Microbiology and Infection (30)
- Future Generation Computer Systems (17)
- Diagnostic Microbiology and Infectious Disease (10)
- Journal of Systems and Software (10)

[View more >>](#)

Topic

- patient (80)
- result (24)
- mrsa (18)
- por (18)
- spark (17)

[View more >>](#)

Content type

- Journal (463)

[Apply filters](#)

 Download PDFs

 Export

 Relevance

 All access types

- Learning distributed discrete Bayesian Network Classifiers under MapReduce with Apache Spark Original Research Article

Knowledge-Based Systems, In Press, Corrected Proof, Available online 22 June 2016

Jacinto Arias, Jose A. Gamez, Jose M. Puerto

[Abstract](#) | [Research highlights](#) |  PDF (786 K)

- Scaling machine learning for target prediction in drug discovery using Apache Spark Original Research Article

Future Generation Computer Systems, In Press, Corrected Proof, Available online 24 May 2016

Dries Harnie, Mathijs Saey, Alexander E. Vapirev, Jörg Kurt Wegner, Andrey Gedich, Marvin Steijaert, Hugo Ceulemans, Roel Wuyts, Wolfgang De Meuter

[Abstract](#) | [Research highlights](#) |  PDF (1057 K)

- A Framework for Fast and Efficient Cyber Security Network Intrusion Detection Using Apache Spark Original Research Article

Procedia Computer Science, Volume 93, 2016, Pages 824-831

Govind P. Gupta, Manish Kulariya

[Abstract](#) |  PDF (170 K)

- Apache Spark a Big Data Analytics Platform for Smart Grid Original Research Article

Procedia Technology, Volume 21, 2015, Pages 171-178

Shyam R., Bharathi Ganesh H.B., Sachin Kumar S., Prabaharan Poornachandran, Soman K.P.

[Abstract](#) |  PDF (506 K)

- A Hybrid Distributed Collaborative Filtering Recommender Engine Using Apache Spark Original Research Article

Procedia Computer Science, Volume 83, 2016, Pages 1000-1006

Sasmita Panigrahi, Rakesh Ku. Lenka, Ananya Stilipragyan

[Abstract](#) |  PDF (452 K)

- kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data Original Research Article

Knowledge-Based Systems, In Press, Corrected Proof, Available online 14 June 2016

Jesus Maillo, Sergio Ramirez, Isaac Triguero, Francisco Herrera

[Abstract](#) |  PDF (1127 K)



Feedback 

What is Scopus?

Scopus is the largest abstract and citation database of peer-reviewed literature, and features smart tools that allow you to track, analyze and visualize scholarly research.

The screenshot shows the Scopus search interface. At the top, there's a navigation bar with links for Scopus, SciVal, Library catalogue, Register, Login, and Help. A banner on the right says "Brought to you by Scopus Team". Below the navigation is a dark teal header with tabs for Search, Alerts, Lists, and My Scopus. The main search area has a "Document search" tab selected. It includes a search bar with placeholder text "Search for..." and "Eg. 'heart attack' AND stress", a dropdown for "Article Title, Abstract, Keywords", and a search button. Below the search bar are filters for "Date Range (inclusive)" (set to "Published All years to Present"), "Document Type" (set to ALL), and "Subject Areas" (checkboxes for Life Sciences, Health Sciences, Physical Sciences, and Social Sciences & Humanities). To the right, there's a sidebar with links: "Learn more about how to Improve Scopus", "Stay up-to-date on Scopus. Follow @Scopus on Twitter", "Watch tutorials and learn how to make Scopus work for you", "Get citation alerts pushed straight to your inbox", and "Get started with Scopus APIs".

Scopus includes content from more than 5,000 publishers and 105 different countries

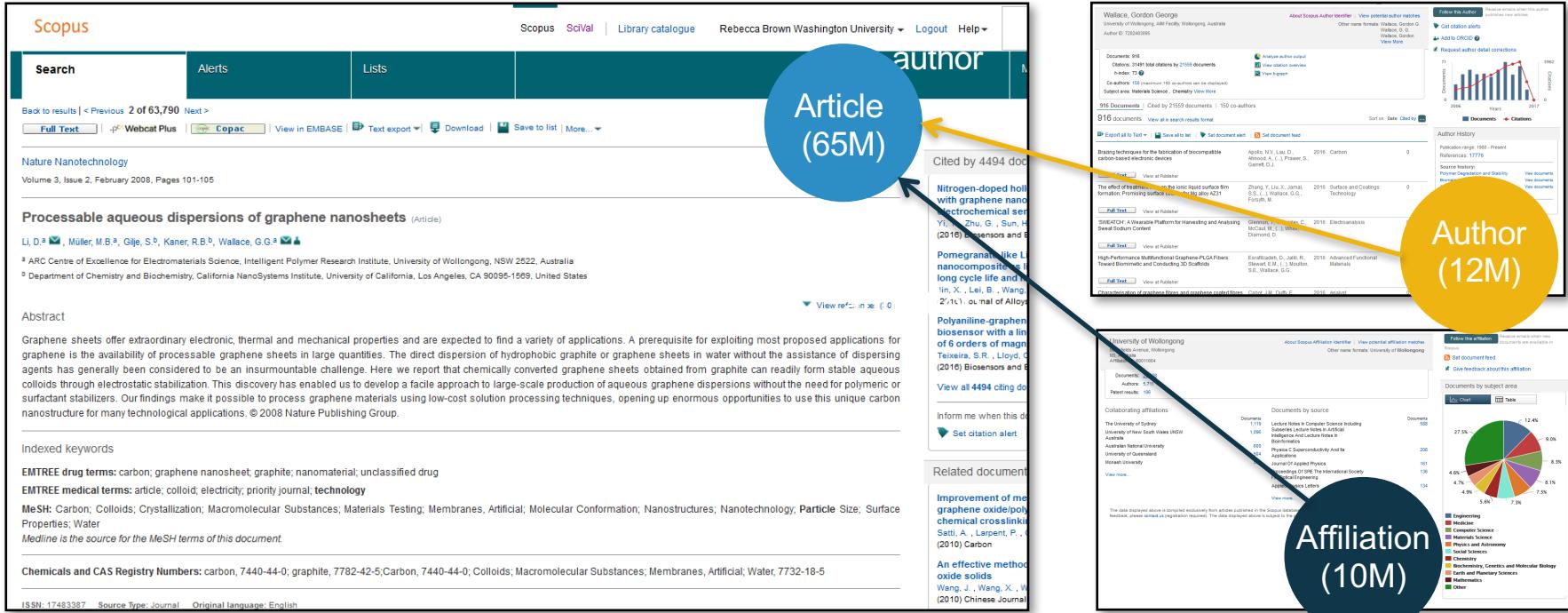
65M records from 22K serials, 90K conferences and 120K books

- Updated daily
- Records back to 1823
- “Articles in Press” from > 3,750 titles
- 40 different languages covered
- 3,715 active Gold Open Access journals indexed

| | JOURNALS | CONFERENCES | BOOKS | PATENTS* |
|-----------------------------------|--|---|---|---|
| Physical Sciences 7,443 | 21,568 peer-reviewed journals 361 trade journals <ul style="list-style-type: none">• Full metadata, abstracts and cited references (ref's post-1995 only)• Funding data from acknowledgements• Citations back to 1970 | 90K conference events 7.3M conference papers | 531 book series 30K Volumes / 1.2M items 119,882 stand-alone books 974K items | 27M patents From 5 major patent offices <ul style="list-style-type: none">- WIPO- EPO- USPTO- JPO- UK IPO |
| Health Sciences 6,795 | | Mainly Engineering and Computer Sciences | | |
| Social Sciences 8,086 | | | Focus on Social Sciences and A&H | |
| Life Sciences 4,492 | | | | |

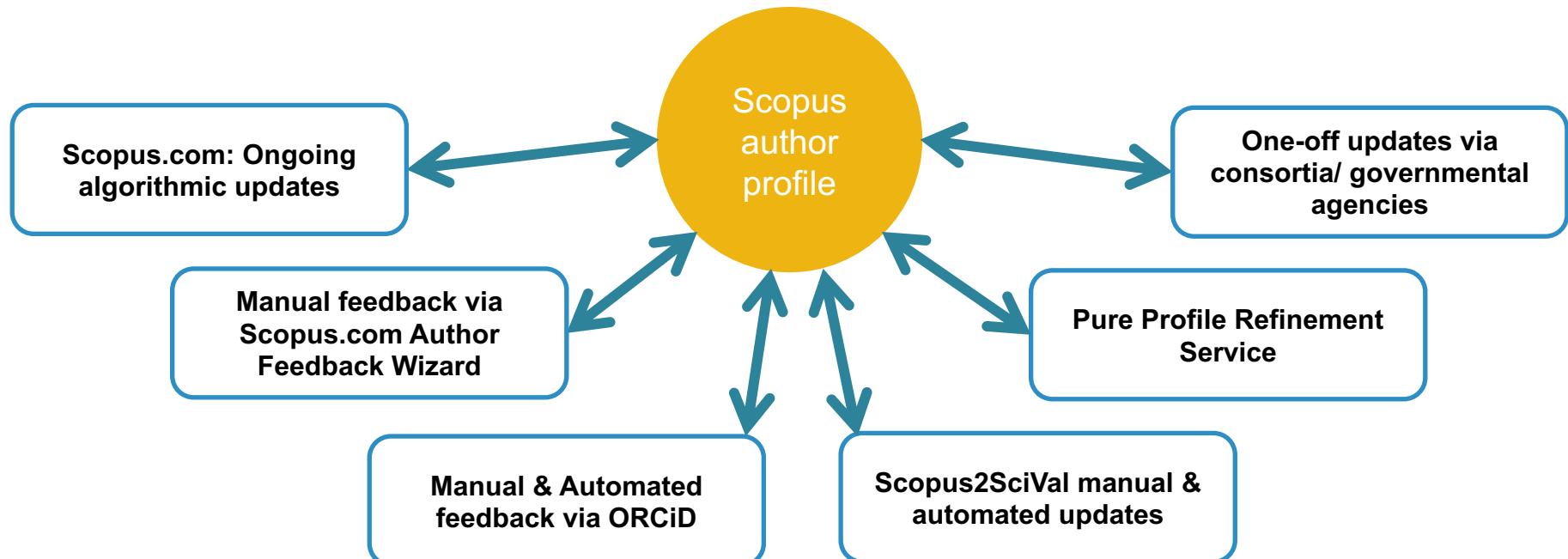
Scopus Data Model

By employing the state of the art disambiguation and deduplication algorithms, entities such as authors, institutes and cited document are disambiguated. This enables us to analyze trends and to track researchers and institutes.

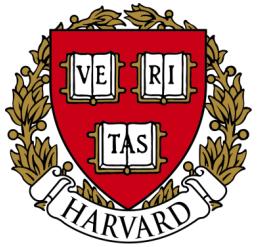


Author Disambiguation: AI algorithms enhanced with multi-level feedbacks

Scopus use a combination of automated and curated data to automatically build robust author profiles, which power the Elsevier Research Intelligence portfolio.



Who uses Scopus Data? Examples



Volkswagen



SANOFI



MAX-PLANCK-GESELLSCHAFT



AstraZeneca



SIEMENS

© 2013 SIEMENS AG

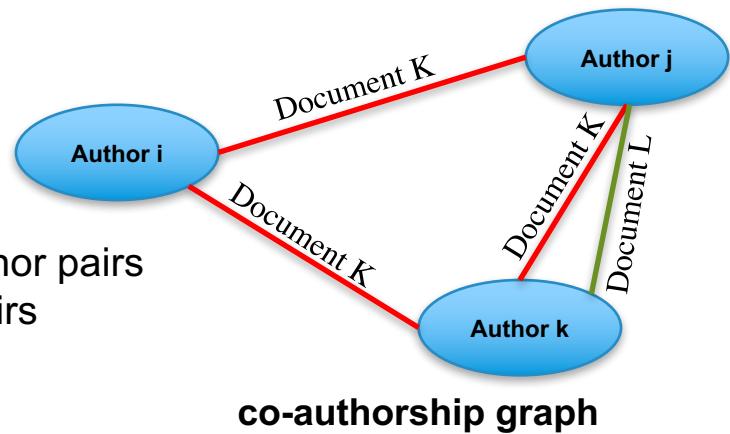


European Research Council



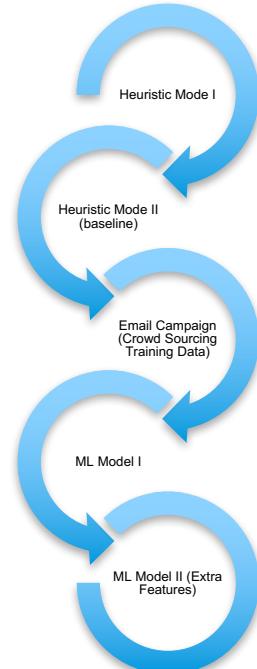
Co-authorship Graphs

- Based on disambiguated records, we can track all the publication, co-authors, affiliation, corresponding authors, citing authors, cited authors ... for each author
- Authors will be modeled as vertices and each document co-published represents an edge
- Current Analysis is based on:
 - 65.2M documents (208M Authorships)
 - 33.8M Authors (9.1M published within Elsevier)
 - 4.6B co-authorships based on 382M unique author pairs
 - 123M unique correspondence ordered author pairs



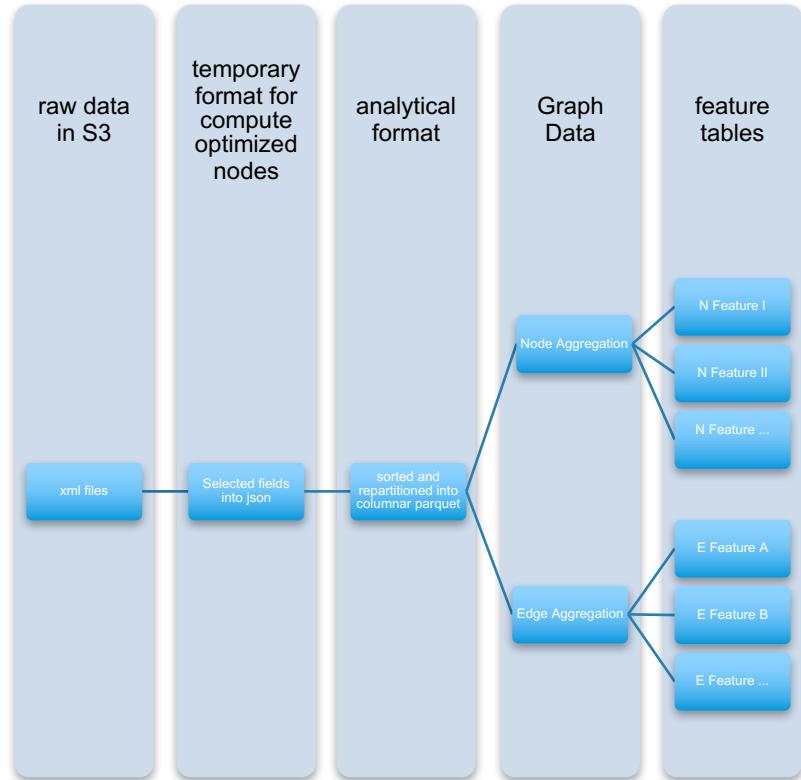
Problem Statement and Modeling

- **Problem**
 - For each author, find the most likely mentor among co-authors (explicit clues exist)
- **Solution**
 - most likely: ranking
 - being mentor: classification
 - CLASSIFICATION INPUT:
 - Approximate into a pairwise model as opposed to a model based on full-set
 - Approximate with an aggregated graph
 - TRAIN THE CLASSIFIER: Even for a simple pairwise model we need lots of training data to account for variations in different fields, years, countries,...
 - Start from simple heuristic common-sense models (cannot overfit) and improve it gradually
 - » Can predict accurately based on given clues/features, but cannot properly balance the interaction and proper weight for each clue
 - When satisfied, verify the early predictions via crowd sourcing and collect precise data for ML training

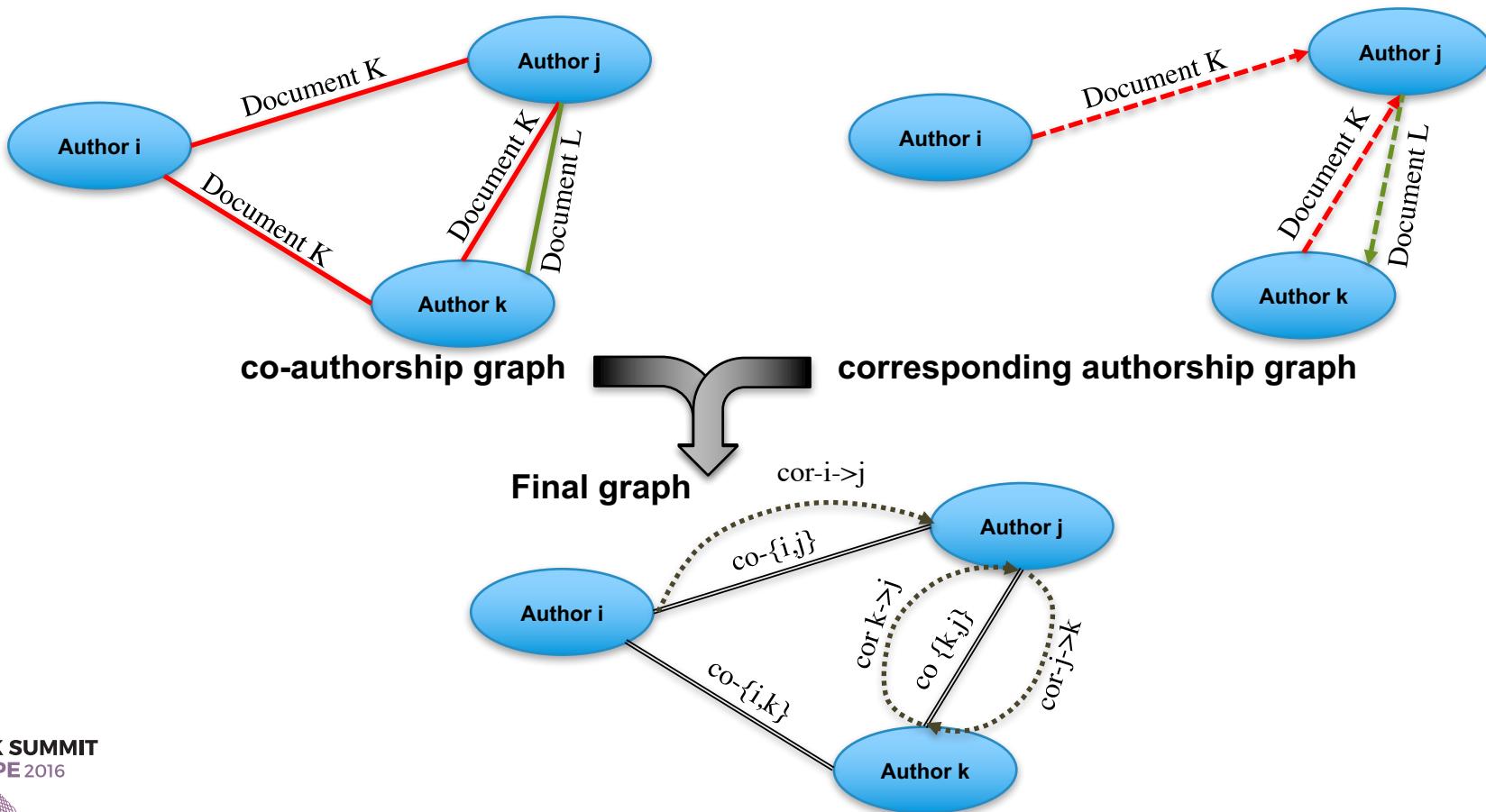


Spark Pipeline in AWS: ETL and Aggregate

- Feature are primarily built on dual summary statistics (around nodes for nodal features or around all edges between two nodes for edge features)
- Normalizing data is superbly important. For an article with n authors, following weights are adopted:
 - $1/n$: authorship
 - $1/(n*(n-1))$: co-authorship
 - $1/(n-1)$: for corresponding authorship
- Non-numeric features are aggregated primarily into
 - most common value (and its frequency) for nodes
 - overlap-share/cosine-similarity for edges



Mixing co-authorship and corresponding authorship



Feature Examples

- Number/weighted-sum of
 - authorships
 - co-authorships
 - corresponding authorships
- First/Last/AVG year
- Publication Milestone Years (to simplify time-series):
 - year reaching 3,5,10 publications (robustness against disambiguation and variations in academic fields)
- PageRank of Author in correspondence graph or citation graphs
- Number of co-authors/corresponding authors
- HIndex, Citation metrics
- Continent/Country, current affiliation(s)
- Email Domain(s)
- AVG author position in authorship sequences (from 0-1)
- Node Content representation by:
 - Journal frequency vector
 - Subject-area frequency vector

Applying Pairwise Model

1. Apply mentorship to each edge

```
graph=GraphFrame(vertices, edges)
paths =graph.find("(a1)-[e]->(a2)")
paths.rdd.map(mentorship).toDF().write.format('parquet').mode('overwrite').saveAsTable(...)
```

2. Rank co-authors based on their mentorship score and pick up the best score as the candidate mentor

3. Create the mentorship graph (batch process):

1. Look for loops of length L ($L=2,3,4,\dots$)
2. Break loops by the weakest link and replace that link with the next mentor candidate
3. Go back to step 2 and if no loop found increase L

4. Create academic family-trees (for select authors):

can locally break loops of bigger lengths (such as $L=7$)

Validations via Crowd Sourcing

- **Emails sent to opt-in (Mendeley/Scopus) User to evaluate our predictions and collect data for supervised ML model**
- **A/B/C testing for Email templates to optimize open/click rate**
- **Clicks lead to a submit page to avoid random clicks**

Mendeley

Discover relevant research and save time. [View this email in your browser.](#)

Dear Reza Karimi,

To improve the quality of our services on Mendeley and Scopus, it is useful to have mentor and mentee relationships. With this information we will be able to create more relevant publication alerts and improve our suggestions for whom to follow on Mendeley.

Based on your publications it looks like among your co-authors, [REDACTED] who has published on behalf of [REDACTED], has acted as your primary academic advisor.

If you would like to help us improve the alerts and recommendations offered to you, we would like to know which of the following four options you think best fits your professional relationship with him/her:

A. Your primary advisor.
B. One of several primary advisors.
C. Not a primary advisor, but one of your advisors.
D. Not your advisor.

Follow us [Twitter](#) [Facebook](#)

Contact Us Support

This message has been sent to r.karimi@elsevier.com from Elsevier Communications on behalf of Mendeley.

If you no longer wish to receive messages of this nature from us in the future, please [click here](#).

Visit the [Elsevier Preference Center](#) to manage more of your communication preferences with us.

Copyright © 2016 Elsevier B.V. All rights reserved. | [Elsevier Privacy Policy](#)

Elsevier B.V. Registered Office: Radarweg 29, 1043 NX Amsterdam, The Netherlands. Reg. No. 33156992 – Netherlands. VAT No. NL 005033019801.

ELSEVIER + Mendeley Primary Academic Advisor

Please confirm the option presented here is the one which best fits your professional relationship with this author. If this is correct, please click the Submit button.

About you
Work Email Address*
r.karimi@elsevier.com

Please confirm that the following option is the one which best fits your professional relationship with this author.*
Your primary advisor

Submit

Mendeley

Discover relevant research and save time. [View this email in your browser.](#)

Thank you for your feedback. Your input is vital in helping us improve our products.

Contact Us Support

This message has been sent to r.karimi@elsevier.com from Elsevier Communications on behalf of Mendeley.

Visit the [Elsevier Preference Center](#) to manage more of your communication preferences with us.

Copyright © 2016 Elsevier B.V. All rights reserved. | [Elsevier Privacy Policy](#)

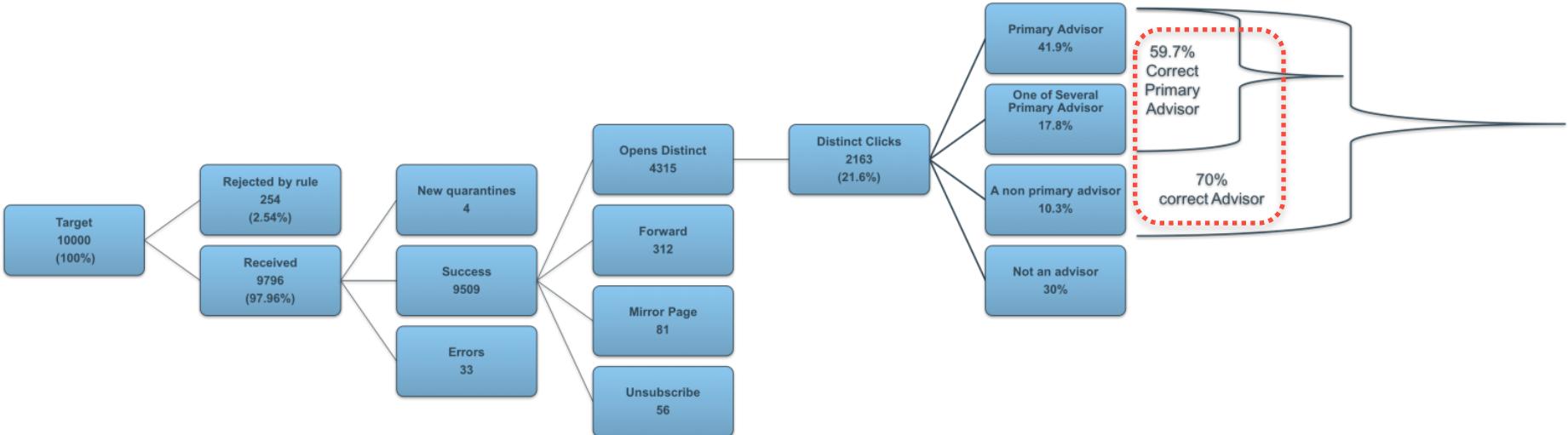
Elsevier B.V. Registered Office: Radarweg 29, 1043 NX Amsterdam, The Netherlands. Reg. No. 33156992 – Netherlands. VAT No. NL 005033019801.

Email Clicked

Submit Page

Thank you email

Prediction Accuracy



- **Funnel:**
 - 10000 randomly selected recipients
 - 4315 recipients opened their email
 - 1742 recipients clicked on provided choices
 - 413 recipients submitted a choice
- **Click vs Submit**
 - 23% of those who clicked, submitted
 - 94% of submission and click choices were identical, the rest is primarily a switch to a more conservative choice

Academic Family-Tree

- **Academic family-trees are important:**
 - Recommendations such as article/people in Mendeley
 - Be aware of conflict of interests: for example in reviewer selection or funding panelist
- **Special features of academic family-trees:**
 - Low connectivity (every one has only (one) mentor)
 - A structure growing with time like snow crystals
- **Given the simplicity of the mentorship graph when it is cached, Spark can act as a back-end for instantaneous subgraph creation or other non-batch analysis**



Academic progeny of Prof. ... in MIT
(intentionally blurred screenshot to minimize personal impact on names displayed)

(Big) Graph Visualization (by Spark)

- To visualize (big) graphs in Spark application we need two components:
 1. **Sub-Graph creation:** it would not be possible to visualize all edge/nodes, especially for highly connected graphs such as:
 1. co-authorships
 2. journal to journal citation
 3. institute to institute citation

To select a limited set of node/edges, filter or GraphFrame queries can help. They work great, if the sub-graph can be obtained by some global filters (such as n-top strongest edges). However, we had to make substantial development to cover sub-graphs centered around a given node. The local sub-graph creation was extended by our library via customized development in line with data-frame operations.

2. **Visualization library:** Adopted D3.js and displayHTML to visualize (Edges, Vertices) Dataframes interactively. The library recognizes following elements:

- Directed Graphs (such as citation) including self referral edges or non-directed graphs (such as co-authorship) displayed with or without arrows
- size, colors based on continuous/categorical data, line type for edges and nodes
- Line types

Visualization Example I: Global filter

```
> Esize='link_share'  
limit_top_global=250  
this_edges_top=edges.filter('src<>dst').sort(Esize, ascending=False)  
graph_scale_top=link_scale( this_edges_top, central=False)  
this_json_top=json_map(vertices, this_edges_top, Vgroup='country',
```

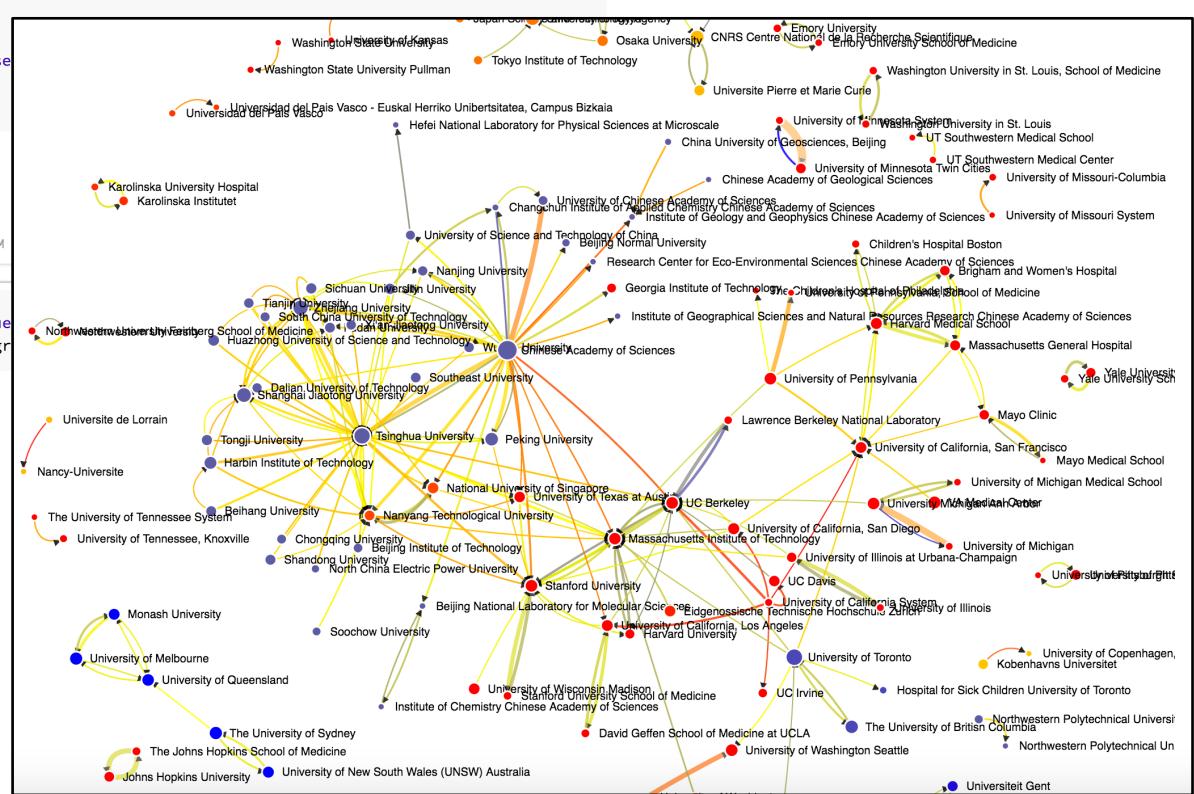
► (7) Spark Jobs

```
#Link= 250 , #Source_Nodes= 123 , #Destination_Nodes= 111
```

```
graph_scale= 31.1084468416
```

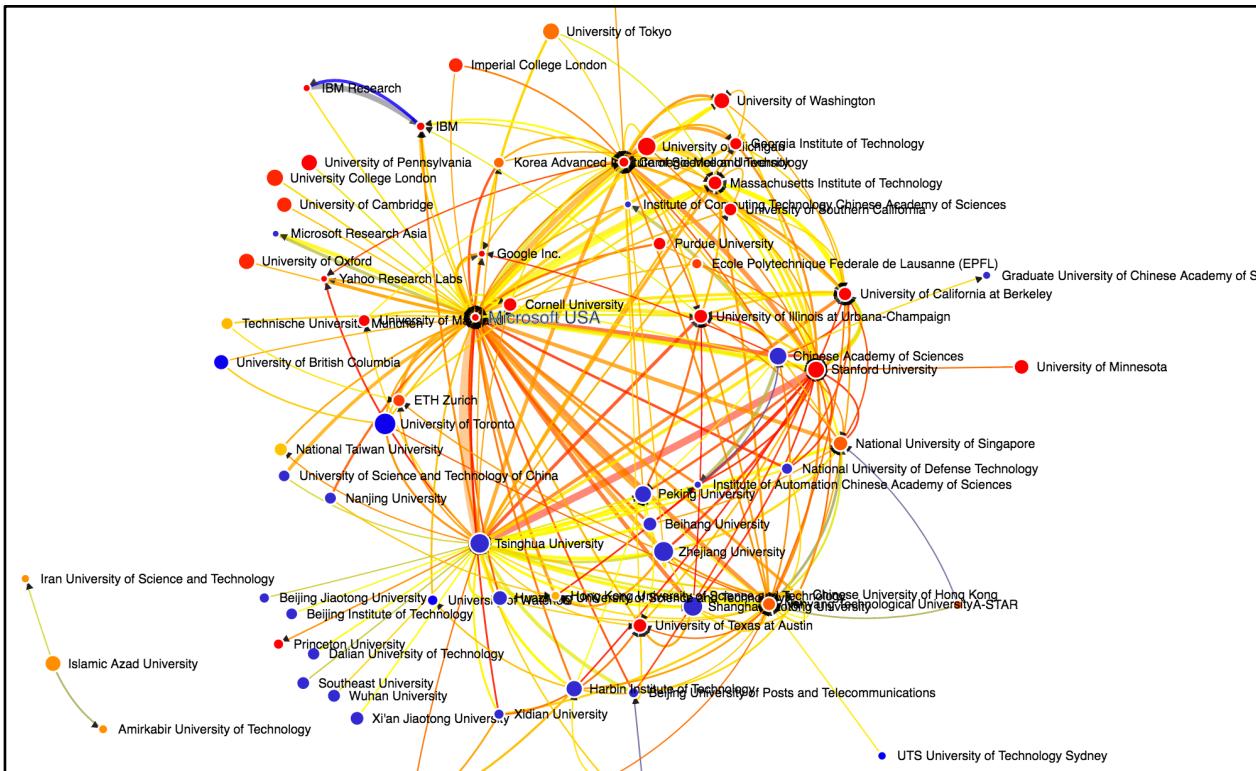
```
Command took 33.35 seconds -- by r.karimi@elsevier.com at 10/17/2016, 11:57:00 AM
```

```
> displayHTML(D3_Network_Custom(Arrow=True, Curved=True, Node_HM=True  
linkStrength=1, linkDistance=400/graph_scale_top*2, charge=-2000/graph
```

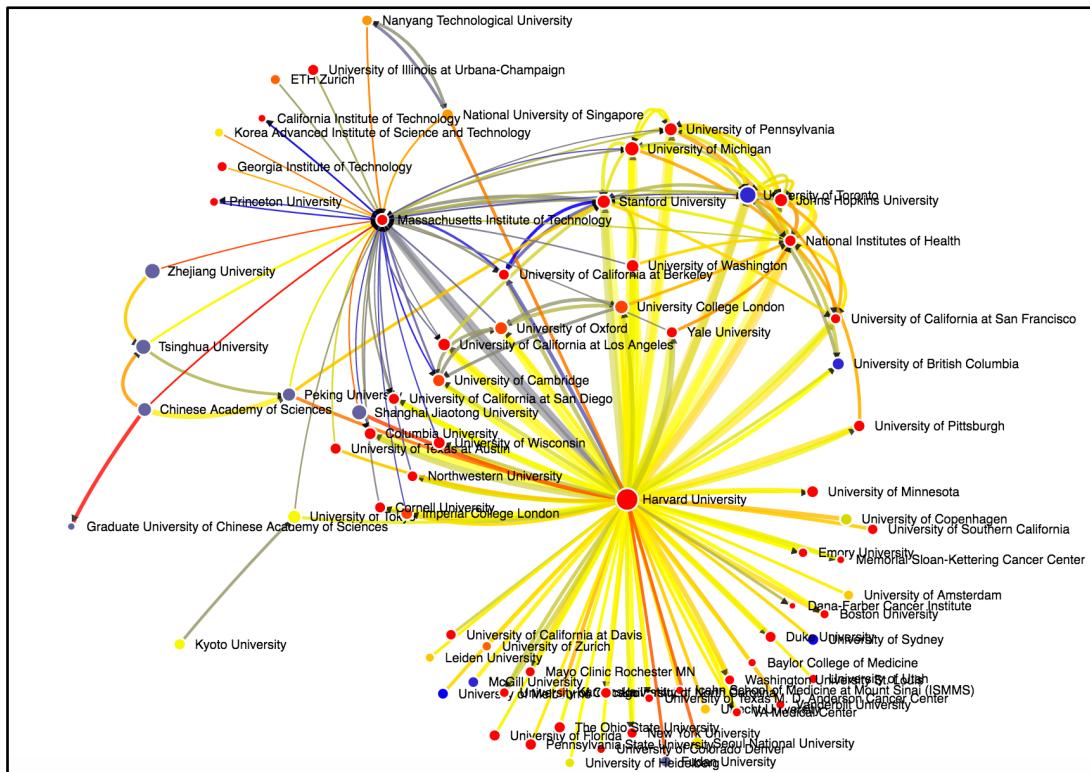
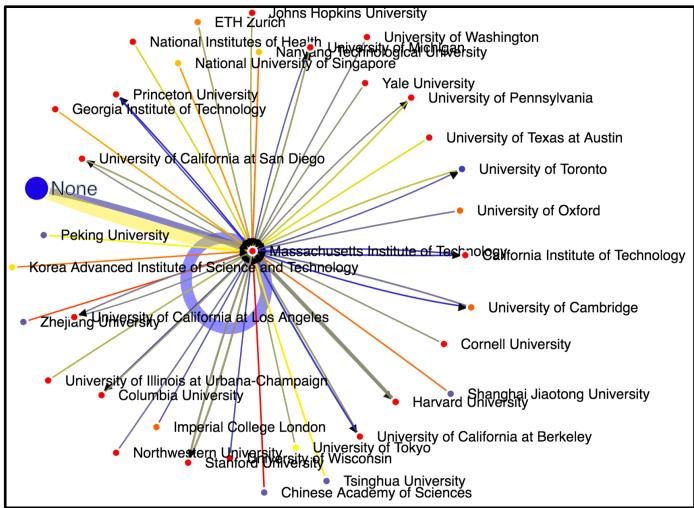


Top 250 Global non-self institute citations

Visualization Example II: Top Institute Citations in Computer Science



Visualization Example III: Isub-graph of top citations around MIT



THANK YOU.

Feel free to reach me for further information or if interested to join our team:

r.karimi@elsevier.com

Acknowledgment:

Bob Schijvenaars, Antonio Gulli, Darin McBeath, Daul Trevor,
and Elsevier members in BOS/Labs/Scopus teams

