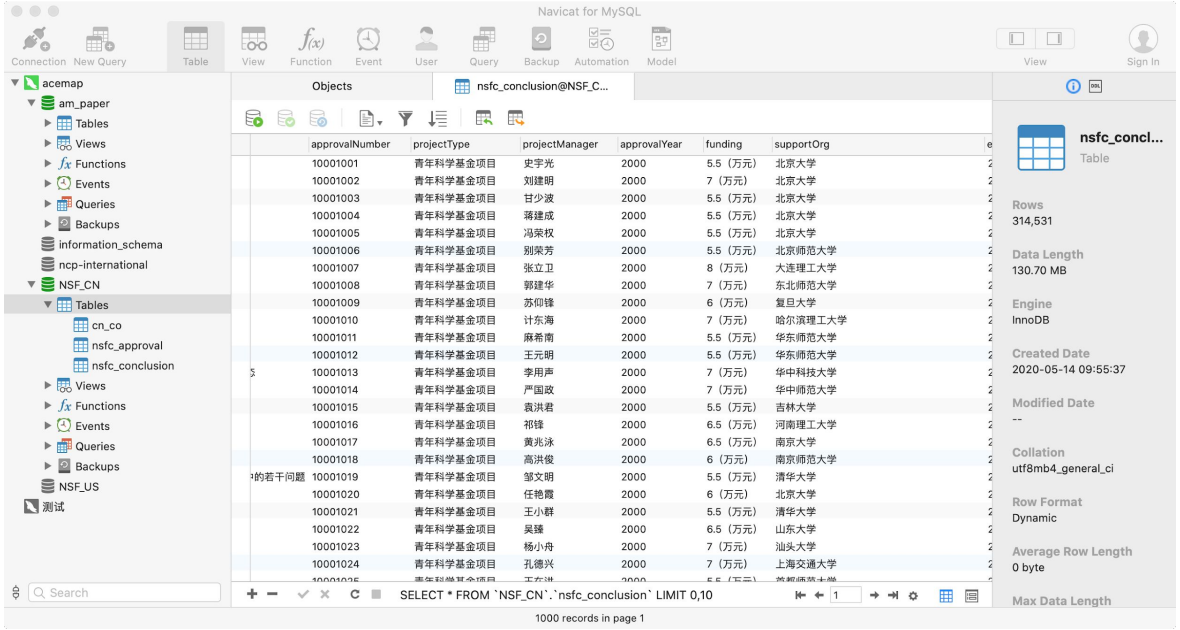
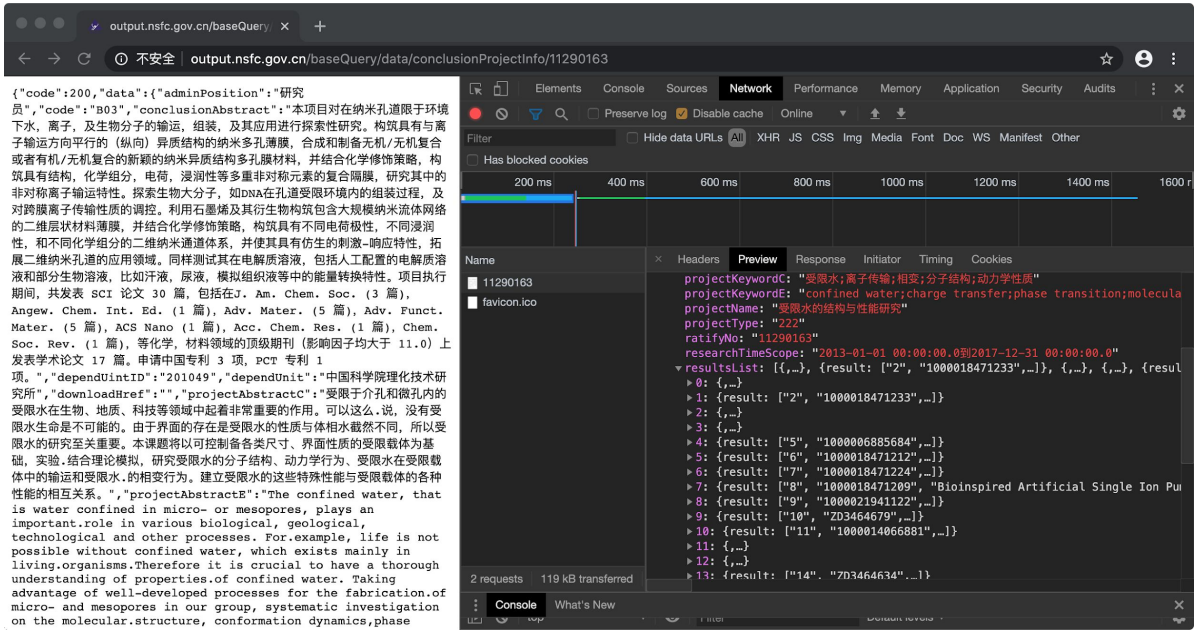
Mining collaboration in NSF grants

...with comparison between China & US

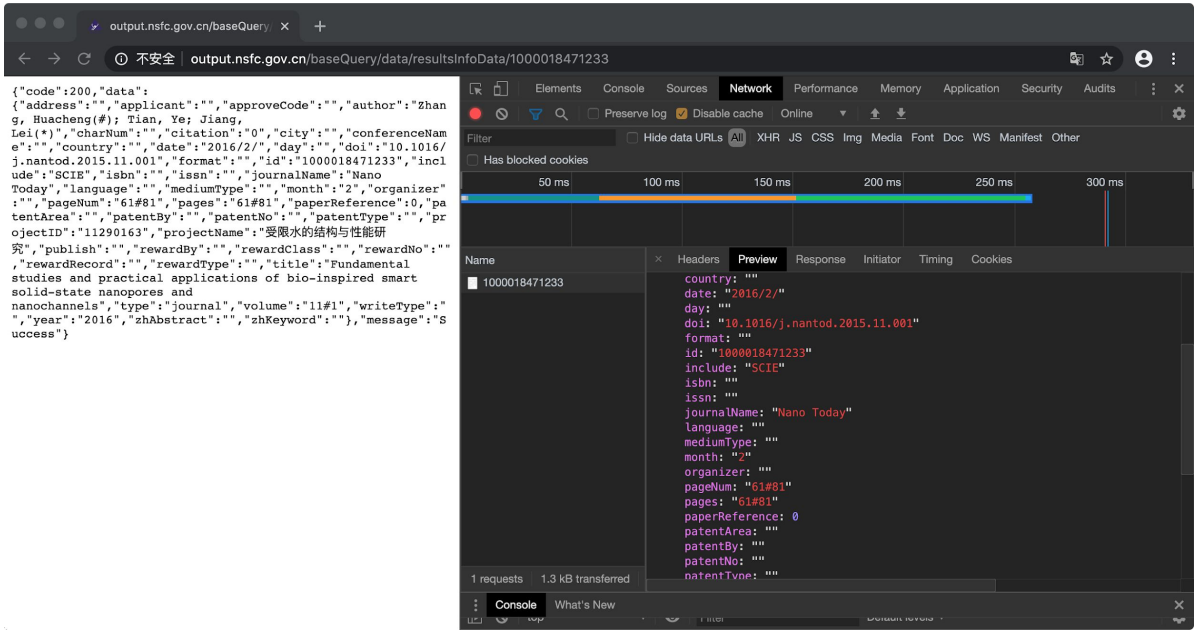
Crawler


Navicat for MySQL
Connection
New Query
Table
View
Function
Event
User
Query
Backup
Automation
Model
View
Sign In
acemap
am_paper
Tables
Views
Functions
Events
Queries
Backups
information_schema
ncp-international
NSF_CN
Tables
cn_co
nsfc_approval
nsfc_conclusion
Views
Functions
Events
Queries
Backups
NSF_US
测试
Search
Objects
nsfc_conclusion@NSF_C...
approvalNumber
projectType
projectManager
approvalYear
funding
supportOrg
10001001 青年科学基金项目 史宇光 2000 5.5（万元） 北京大学
10001002 青年科学基金项目 刘建明 2000 7（万元） 北京大学
10001003 青年科学基金项目 甘少波 2000 5.5（万元） 北京大学
10001004 青年科学基金项目 蒋建成 2000 5.5（万元） 北京大学
10001005 青年科学基金项目 冯荣权 2000 5.5（万元） 北京大学
10001006 青年科学基金项目 别荣芳 2000 5.5（万元） 北京师范大学
10001007 青年科学基金项目 张立卫 2000 8（万元） 大连理工大学
10001008 青年科学基金项目 郭建华 2000 7（万元） 东北师范大学
10001009 青年科学基金项目 苏仰锋 2000 6（万元） 复旦大学
10001010 青年科学基金项目 计东海 2000 7（万元） 哈尔滨理工大学
10001011 青年科学基金项目 麻希南 2000 5.5（万元） 华东师范大学
10001012 青年科学基金项目 王元明 2000 5.5（万元） 华东师范大学
10001013 青年科学基金项目 李用声 2000 7（万元） 华中科技大学
10001014 青年科学基金项目 严国政 2000 7（万元） 华中师范大学
10001015 青年科学基金项目 袁洪君 2000 5.5（万元） 吉林大学
10001016 青年科学基金项目 祁锋 2000 6.5（万元） 河南理工大学
10001017 青年科学基金项目 黄兆泳 2000 6.5（万元） 南京大学
10001018 青年科学基金项目 高洪俊 2000 6（万元） 南京师范大学
10001019 青年科学基金项目 邹文明 2000 5.5（万元） 清华大学
10001020 青年科学基金项目 任艳霞 2000 6（万元） 北京大学
10001021 青年科学基金项目 王小群 2000 5.5（万元） 清华大学
10001022 青年科学基金项目 吴臻 2000 6.5（万元） 山东大学
10001023 青年科学基金项目 杨小舟 2000 7（万元） 汕头大学
10001024 青年科学基金项目 孔德兴 2000 7（万元） 上海交通大学
SELECT * FROM `NSF_CN`.`nsfc_conclusion` LIMIT 0,10
1000 records in page 1
nsfc_concl...
Table
Rows
314,531
Data Length
130.70 MB
Engine
InnoDB
Created Date
2020-05-14 09:55:37
Modified Date
--
Collation
utf8mb4_general_ci
Row Format
Dynamic
Average Row Length
0 byte
Max Data Length

Crawler


output.nsfc.gov.cn/baseQuery
不安全 | output.nsfc.gov.cn/baseQuery/data/conclusionProjectInfo/11290163
{"code":200,"data":{"adminPosition":"研究员","code":"B03","conclusionAbstract":"本项目对在纳米孔道限于环境下水，离子，及生物分子的输运，组装，及其应用进行探索性研究。构筑具有与离子输运方向平行的（纵向）异质结构的纳米多孔薄膜，合成和制备无机/无机复合或者有机/无机复合的新颖的纳米异质结构多孔膜材料，并结合化学修饰策略，构筑具有结构，化学组分，电荷，浸润性等多重非对称元素的复合隔膜，研究其中的非对称离子输运特性。探索生物大分子，如DNA在孔道受限环境内的组装过程，及对跨膜离子传输性质的调控。利用石墨烯及其衍生物构筑包含大规模纳米流体网络的二维层状材料薄膜，并结合化学修饰策略，构筑具有不同电荷极性，不同浸润性，和不同化学组分的二维纳米通道体系，并使其具有仿生的刺激-响应特性，拓展二维纳米孔道的应用领域。同样测试其在电解质溶液，包括人工配置的电解质溶液和部分生物溶液，比如汗液，尿液，模拟组织液等中的能量转换特性。项目执行期间，共发表 SCI 论文 30 篇，包括在J. Am. Chem. Soc.（3 篇），Angew. Chem. Int. Ed.（1 篇），Adv. Mater.（5 篇），Adv. Funct. Mater.（5 篇），ACS Nano（1 篇），Acc. Chem. Res.（1 篇），Chem. Soc. Rev.（1 篇），等化学，材料领域的顶级期刊（影响因子均大于 11.0）上发表学术论文 17 篇。申请中国专利 3 项，PCT 专利 1 项。","dependUintID":"201049","dependUnit":"中国科学院理化技术研究所","downloadHref":"","projectAbstractC":"受限于介孔和微孔内的受限水在生物、地质、科技等领域中起着非常重要的作用。可以这么.说，没有受限水生命是不可能的。由于界面的存在是受限水的性质与体相水截然不同，所以受限水的研究至关重要。本课题将以可控制备各类尺寸、界面性质的受限载体为基础，实验.结合理论模拟，研究受限水的分子结构、动力学行为、受限水在受限载体中的输运和受限水.的相变行为。建立受限水的这些特殊性能与受限载体的各种性能的相互关系。","projectAbstractE":"The confined water, that is water confined in micro- or mesopores, plays an important.role in various biological, geological, technological and other processes. For.example, life is not possible without confined water, which exists mainly in living.organisms.Therefore it is crucial to have a thorough understanding of properties.of confined water. Taking advantage of well-developed processes for the fabrication.of micro- and mesopores in our group, systematic investigation on the molecular.structure, conformation dynamics,phase
Elements Console Sources Network Performance Memory Application Security Audits
Preserve log Disable cache Online
Filter Hide data URLs All XHR JS CSS Img Media Font Doc WS Manifest Other
Has blocked cookies
200 ms 400 ms 600 ms 800 ms 1000 ms 1200 ms 1400 ms
Name
11290163
favicon.ico
2 requests 119 kB transferred
Headers Preview Response Initiator Timing Cookies
projectKeywordC: "受限水;离子传输;相变;分子结构;动力学性质"
projectKeywordE: "confined water;charge transfer;phase transition;molecula
projectName: "受限水的结构与性能研究"
projectType: "222"
ratifyNo: "11290163"
researchTimeScope: "2013-01-01 00:00:00.0到2017-12-31 00:00:00.0"
resultsList: [{,…}, {result: ["2", "1000018471233",…]}, {,…}, {,…}, {resul
0: {,…}
1: {result: ["2", "1000018471233",…]}
2: {,…}
3: {,…}
4: {result: ["5", "1000006885684",…]}
5: {result: ["6", "1000018471212",…]}
6: {result: ["7", "1000018471224",…]}
7: {result: ["8", "1000018471209", "Bioinspired Artificial Single Ion Pu
8: {result: ["9", "1000021941122",…]}
9: {result: ["10", "ZD3464679",…]}
10: {result: ["11", "1000014066881",…]}
11: {,…}
12: {,…}
13: {result: ["14", "ZD3464634",…]}
Console What's New

Crawler


output.nsfc.gov.cn/baseQuery
不安全 | output.nsfc.gov.cn/baseQuery/data/resultsInfoData/1000018471233
{"code":200,"data":{"address":"","applicant":"","approveCode":"","author":"Zhang, Huacheng(#); Tian, Ye; Jiang, Lei(*)","charNum":"","citation":"0","city":"","conferenceName":"","country":"","date":"2016/2/","day":"","doi":"10.1016/j.nantod.2015.11.001","format":"","id":"1000018471233","include":"SCIE","isbn":"","issn":"","journalName":"Nano Today","language":"","mediumType":"","month":"2","organizer":"","pageNum":"61#81","pages":"61#81","paperReference":0,"patentArea":"","patentBy":"","patentNo":"","patentType":"","projectID":"11290163","projectName":"受限水的结构与性能研究","publish":"","rewardBy":"","rewardClass":"","rewardNo":"","rewardRecord":"","rewardType":"","title":"Fundamental studies and practical applications of bio-inspired smart solid-state nanopores and nanochannels","type":"journal","volume":"11#1","writeType":"","year":"2016","zhAbstract":"","zhKeyword":""},"message":"Success"}
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Audits
Preserve log
Disable cache
Online
Filter
Hide data URLs
All
XHR
JS
CSS
Img
Media
Font
Doc
WS
Manifest
Other
Has blocked cookies
50 ms
100 ms
150 ms
200 ms
250 ms
300 ms
Name
1000018471233
Headers
Preview
Response
Initiator
Timing
Cookies
country: ""
date: "2016/2/"
day: ""
doi: "10.1016/j.nantod.2015.11.001"
format: ""
id: "1000018471233"
include: "SCIE"
isbn: ""
issn: ""
journalName: "Nano Today"
language: ""
mediumType: ""
month: "2"
organizer: ""
pageNum: "61#81"
pages: "61#81"
paperReference: 0
patentArea: ""
patentBy: ""
patentNo: ""
patentType: ""
1 requests
1.3 kB transferred
Console
What's New

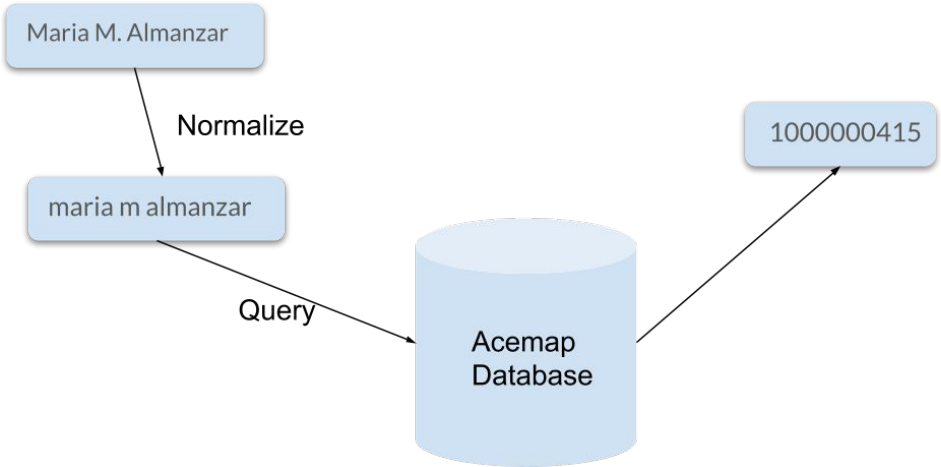
Author Mapping

Mapping from name to Acemap-ID: **Name** (Maria M. Almanzar) -> **AuthorID** (1000000415)

- Large volume of data: **91,458,238** authors in total, took almost **2 min** to traverse, over **8 GB** dump;
- Confusion between duplicate names: more than **10** authors with name ‘A. A. A. Mohamed’;
- Ambiguity rising from degeneracy: ‘张三’ or ‘San Zhang’ or ‘San ZHANG’ or ‘Zhang San’;
- Multiple Institutions: Geoffrey Hinton worked for Google and the University of Toronto at the same time;
- Typo: ‘Keith Ross’ (NYU Professor) and ‘Keith Rose’ (surgeon) and ‘Keith Ros’ (nobody);

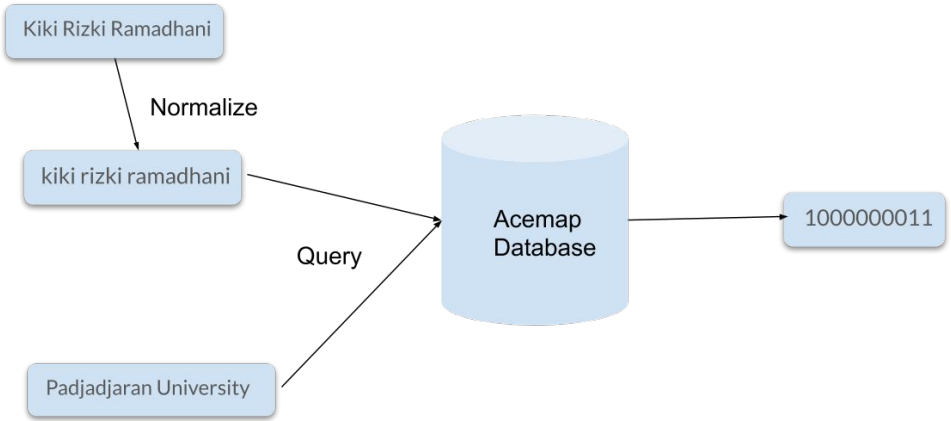
Author Mapping

Strategy One: Direct Matching of Name.



Author Mapping

Strategy Two: Direct Matching of Name + Affiliation.

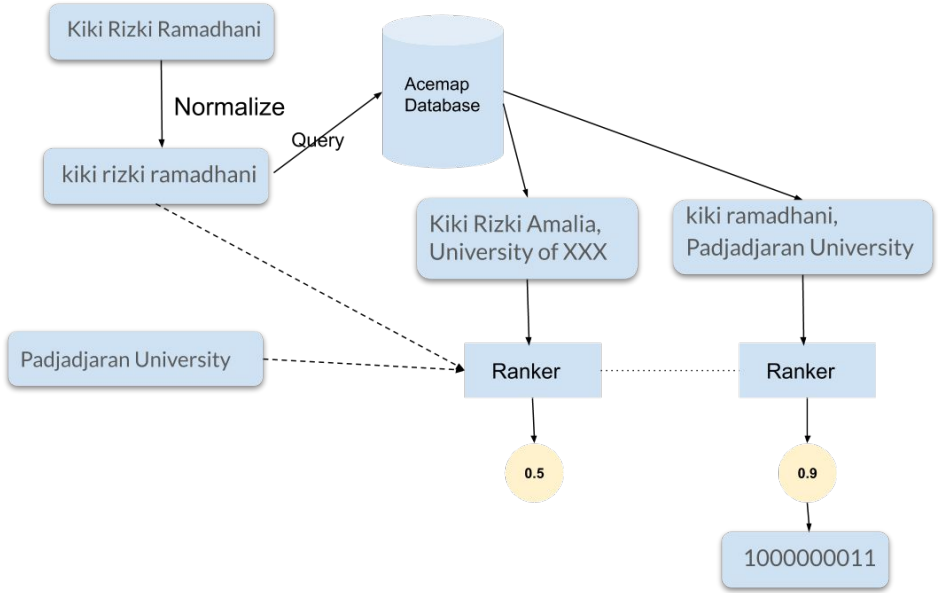


Author Mapping

Strategy Three: similarity ranking of Name + Affiliation (based on Levenshtein distance).

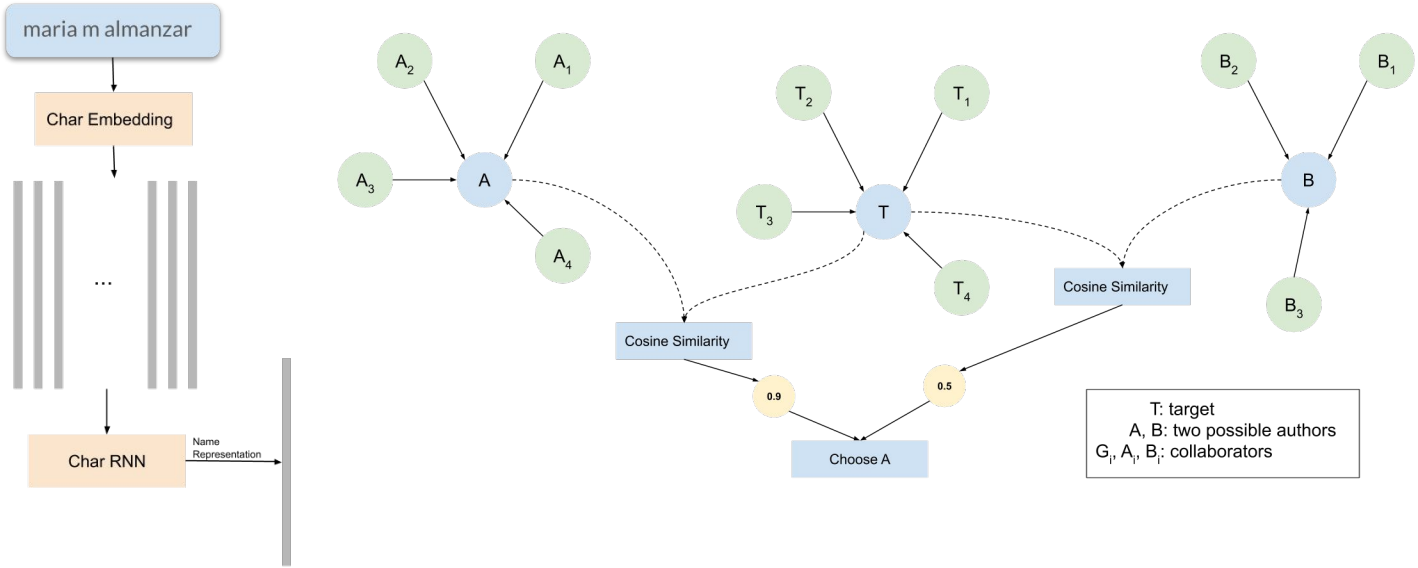
- The Levenshtein distance between two string is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$



Author Mapping

Strategy Four: similarity ranking of Name + Collaborators (based on Graph Neural Network).



Author Mapping

Design of Experiments:

- Use authors from NSF_US as dataset, randomly extract 10% of it as test set;
- Manually create typos at rate **α** in test set (insertions, deletions or substitutions);
- Set a drop rate **β** of affiliation and collaborators;

Author Mapping

	Training Time	Inference Time	Acc*	Acc**
Name	-	0.01 s/item	0.93	0.42
Name + Affiliation	-	0.02 s/item	0.98	0.53
Levenshtein	-	0.42 s/item	**0.99**	0.89
GNN	2 hours	1.98 s/item	0.97	**0.95**

* : with $\alpha = 0$ and $\beta = 0$
**: with $\alpha = 0.2$ and $\beta = 0.8$

Author Mapping

Deployment:

- We will deploy our algorithms on Acemap server and provide an API for users;

Contribution Ⅰ: NSF-CN DataBase

Intuition and Contribution:

- Acemap NSF-CN database contains the principal investigator and the grant information;
- We built a table (NSF_CN.grants), using ‘grants.projectManager’ as foreign key;
- We built a table (NSF_CN.cn_nnsf_participants),

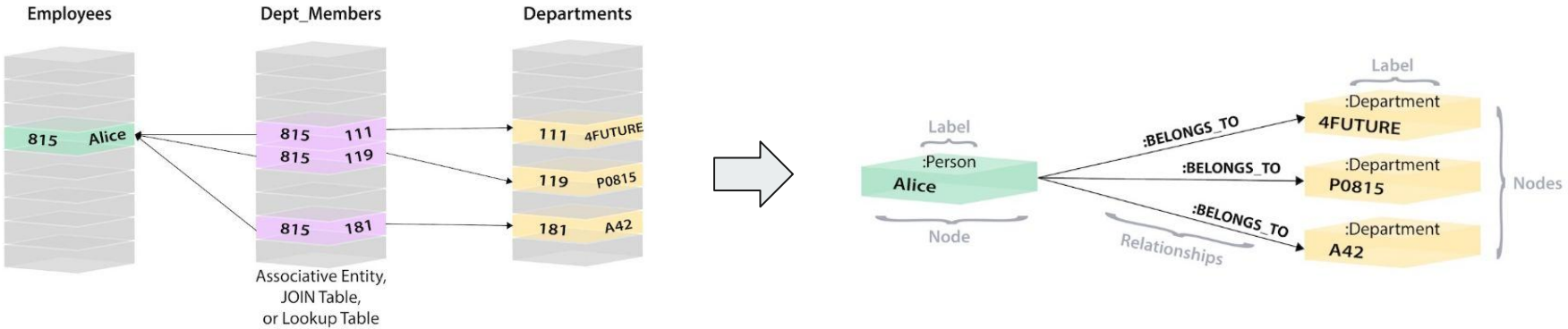
Contribution Ⅱ: NSF-US Grant Table

Intuition and Contribution:

- Acemap NSF-US database contains the principal investigator and the grant information;
- We built a table (NSF_US.grants) providing detailed information about grants;
- We built a table (NSF_US.participants) providing all investigators in the grants.

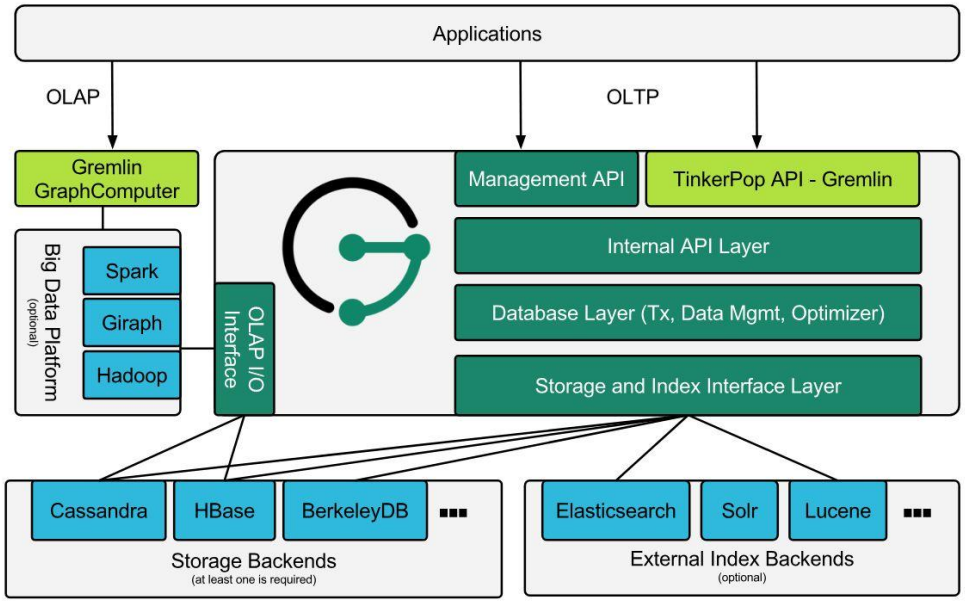
Graph Database

- A type of NoSQL database.
- Uses graph structures for semantic queries with nodes, edges, and properties to represent and store data.
- Flexible & Extensible.

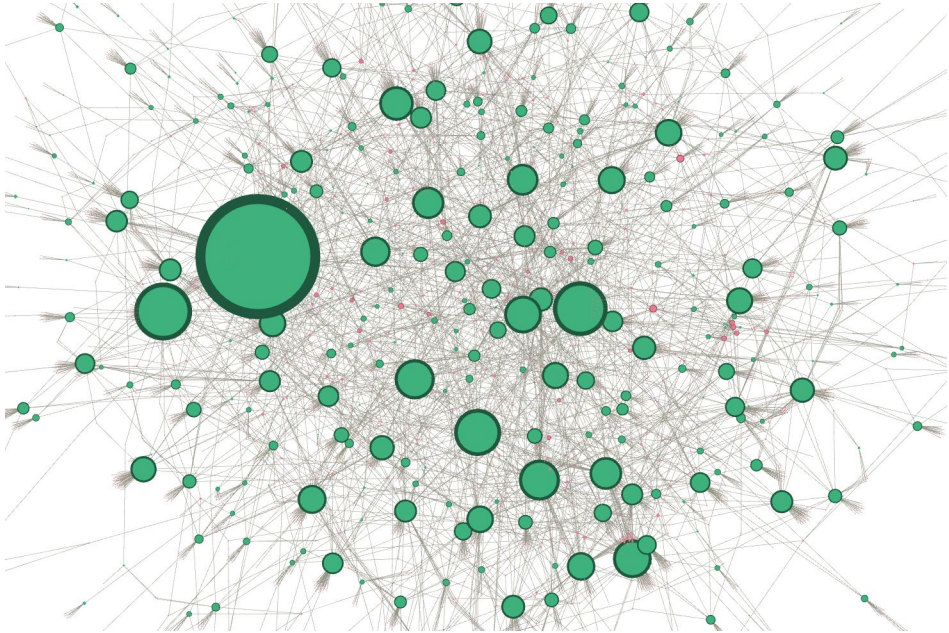


JanusGraph

- Deployment
  - Hbase as storage backend.
  - Solr as external index backend.
  - Gremlin as graph traversal language.
- Performance Test
  - 5.75s for finding the US scholar with most collaborations. .



Connect to Gephi

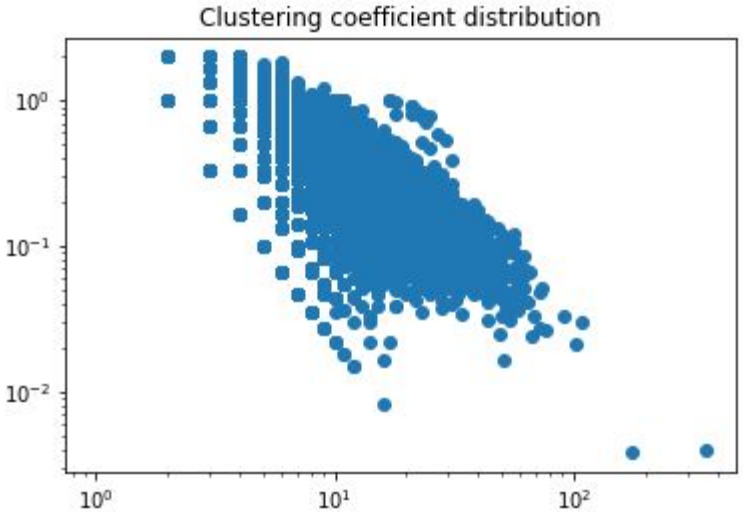


Complex Network Analysis

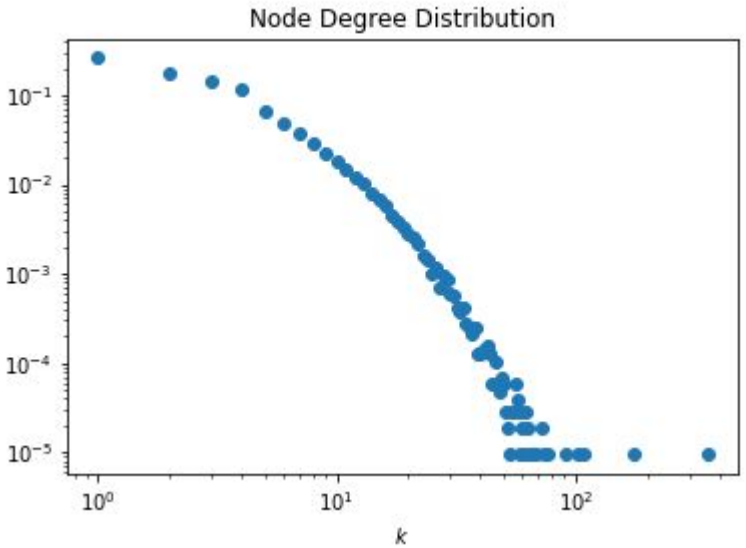
- Qualitative methods for understanding network characteristics
- Metrics we look at
  - Size
  - Degree distribution, average node degree
  - Diameter
  - Assortativity coefficient: Do connected nodes have similar degrees - rich club?
  - Clustering coefficient: Tendency for nodes to cluster together
  - Rich club coefficient: Are well-connected nodes connected?
- Cross-comparison between US, CN and US + CN

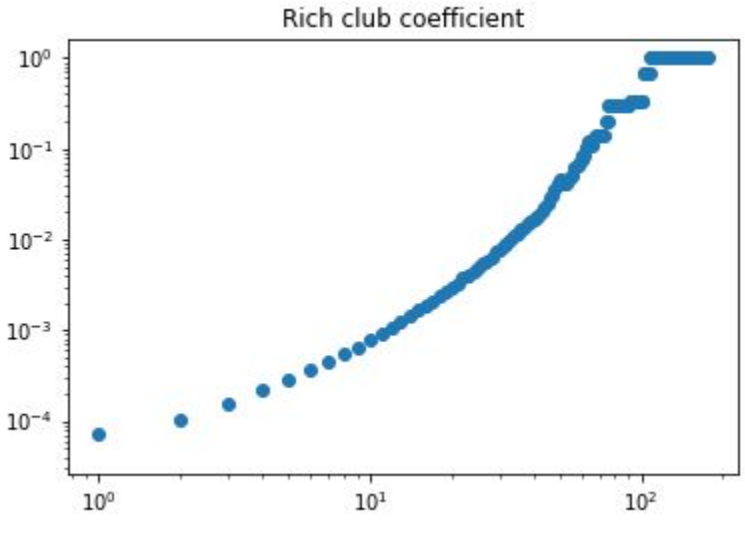
Example: US NSF

- Size: |V| = 104524, |E| = 235320
- Average degree: 4.50
- (Pseudo) Diameter: 20
- Assortativity: 0.07
  - Close to 0, not assortative
- Average clustering coefficient: 0.46 (High)



Example: US NSF

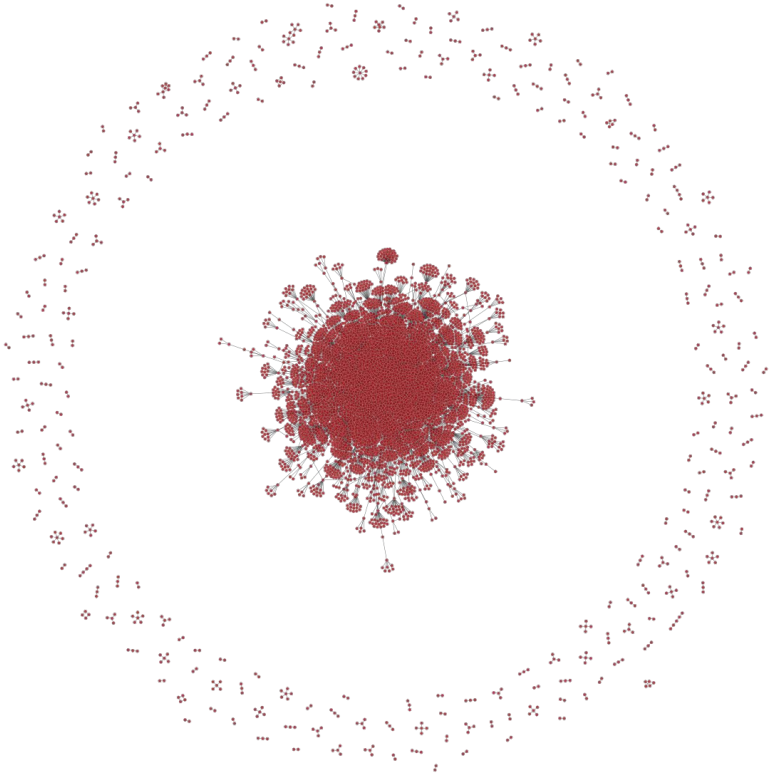




Does NOT follow power law

Cross-comparison

Data	Diameter	Assortativity	Average clustering coefficient	Rich club coefficient
US	20	0.07	0.45	< 1
CN	18	0.13	0.56	< 1
CN + US	15	-0.25	0.32	< 1



The Final CN-US collaboration (not coronavirus)

Thanks

Notable Metrics

- Assortativity
  - Two connected nodes imply their similarity.
  - Assortativity coefficient, with regard to a property, measures the overall similarity on this property between two connected nodes?
  - Actually Pearson'r, so falls in [-1, 1]
- Clustering coefficient
  - For each node, the number of edges in its neighborhood / the number of edges if its neighborhood is complete
  - Understood w.r.t a comparable random graph
- Rich club coefficient
  - For degree k, the number of edges in the subgraph formed by nodes with degree >= k / complete graph

Computing the Metrics

- Select a network-analysis framework
  - Strike the right balance between speed and ease of use
  - Contenders: networkx, igraph, graph-tool, SNAP, networkkit
- Networkx
  - The most-feature complete
  - Written in pure Python = extremely slow
  - Benchmarks[1] show that it is 10 times slower than the *slowest* library
- Graph-tool
  - C++ with Python interface = fast (like Numpy)
  - We have to implement the rich club coefficient ourselves
- Why not Gephi
  - Not well-maintained
  - Lack many metrics

1 https://www.timlrx.com/2020/05/10/benchmark-of-popular-graph-network-packages-v2/