Smart Recruitment

Finding

Potential Candidates
via GitHub Social Network

Rangsarid Pringwanid

Agenda

I. Business Values

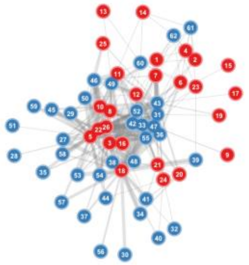
- Existing Business Pain point
- Business Problem Solving

II. Data Collections

- Data Gathering
- Data Pre-processing

III. Exploratory Analysis

- Network Analysis
- Geographic Search

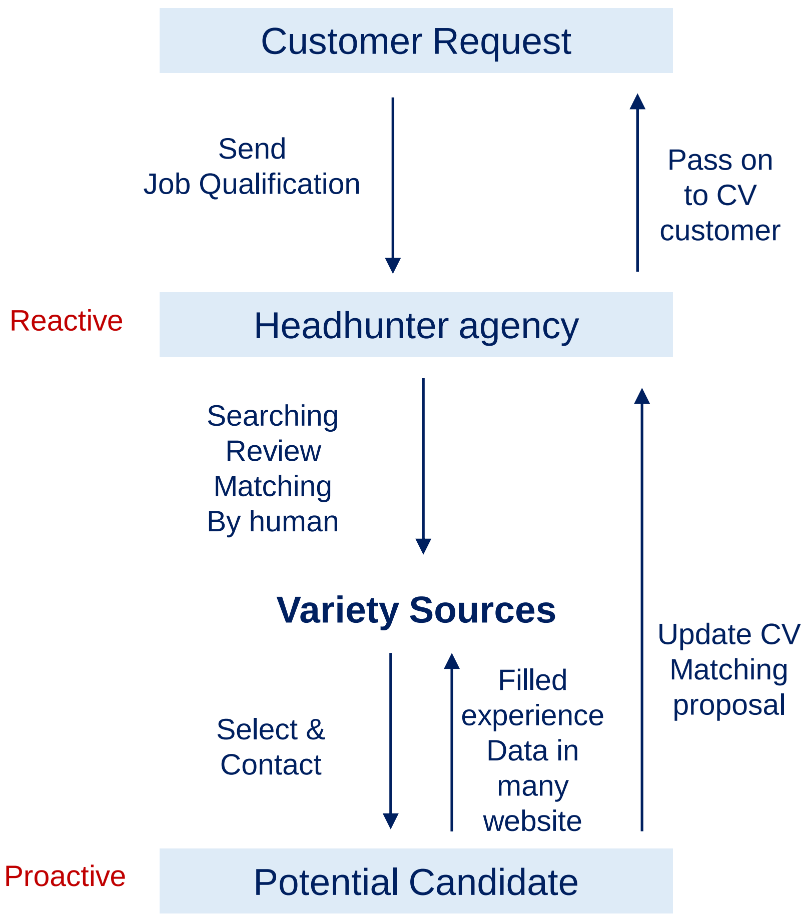


IV. Recommendation System

- Node Embedding technique
- Similarities

I. Business Concept

Business Pain point (Take-time, High-Cost, High-Human effort)



Customer Request

Send
Job Qualification

Pass on
to CV
customer

Reactive

Headhunter agency

Searching
Review
Matching
By human

Variety Sources

Update CV
Matching
proposal

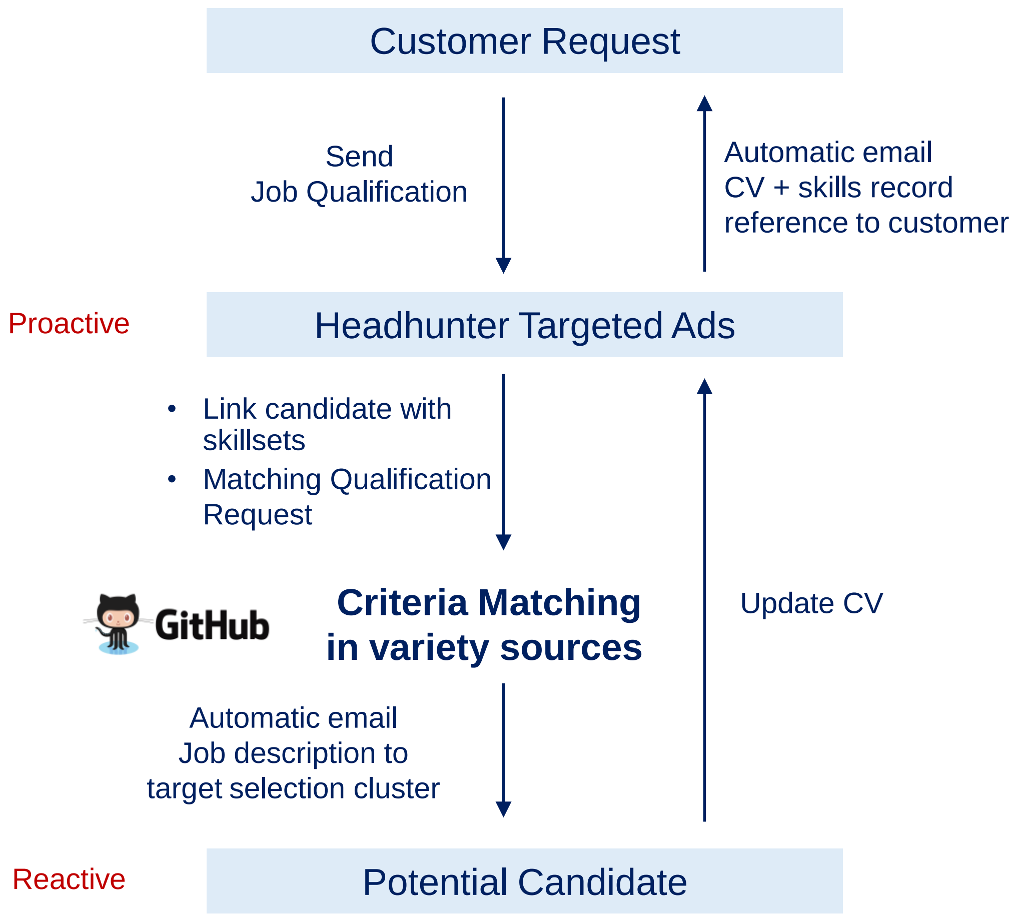
Select &
Contact

Filled
experience
Data in
many
website

Proactive

Potential Candidate

Business Solving (Less-time, Lower-Cost, Lower-effort, Automatic)



Customer Request

Send
Job Qualification

Automatic email
CV + skills record
reference to customer

Proactive

Headhunter Targeted Ads

- Link candidate with skillsets
- Matching Qualification Request

**Criteria Matching
in variety sources**

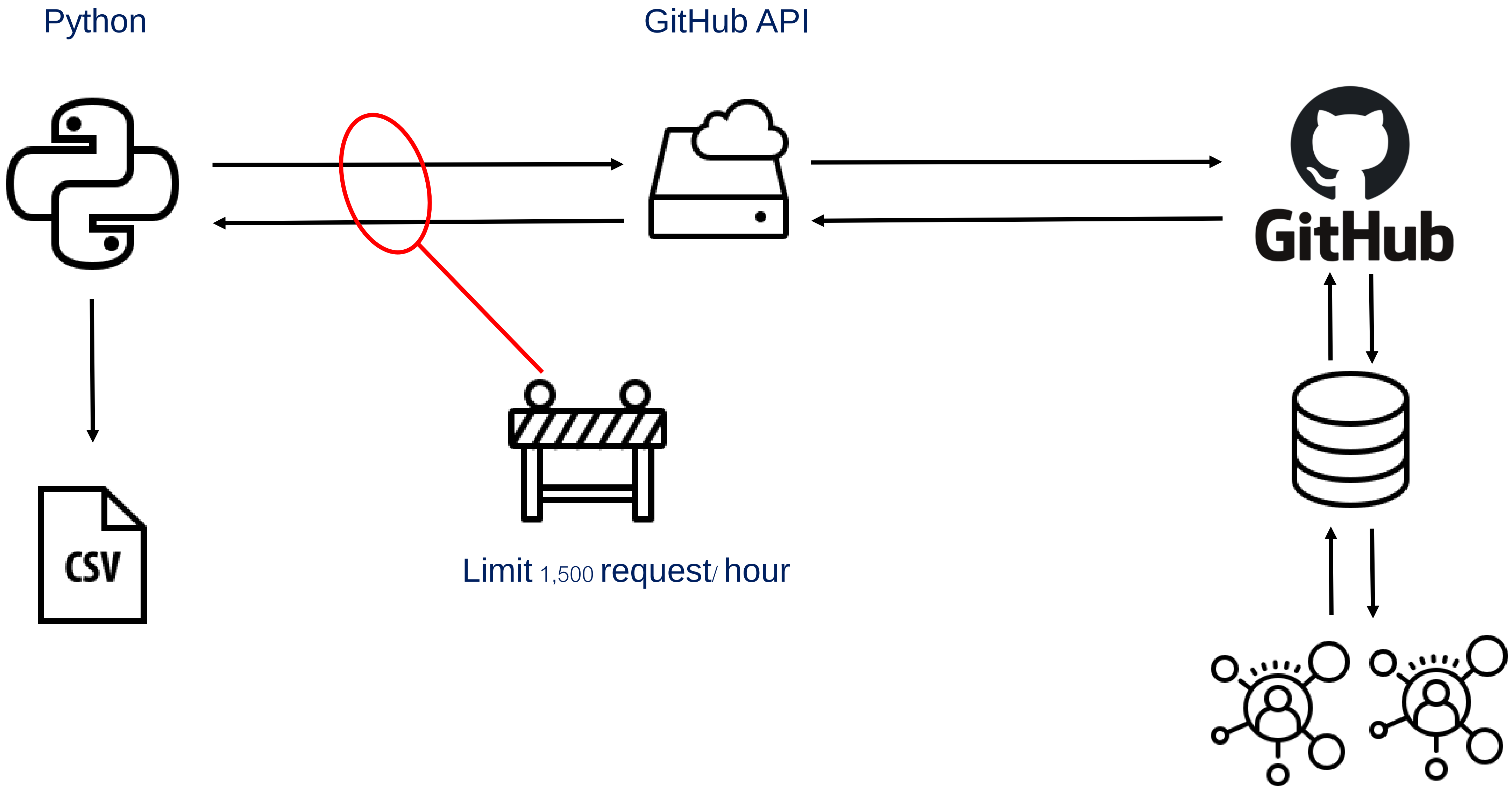
Update CV

Automatic email
Job description to
target selection cluster

Reactive

Potential Candidate

II. Data Collections (Data Gathering)

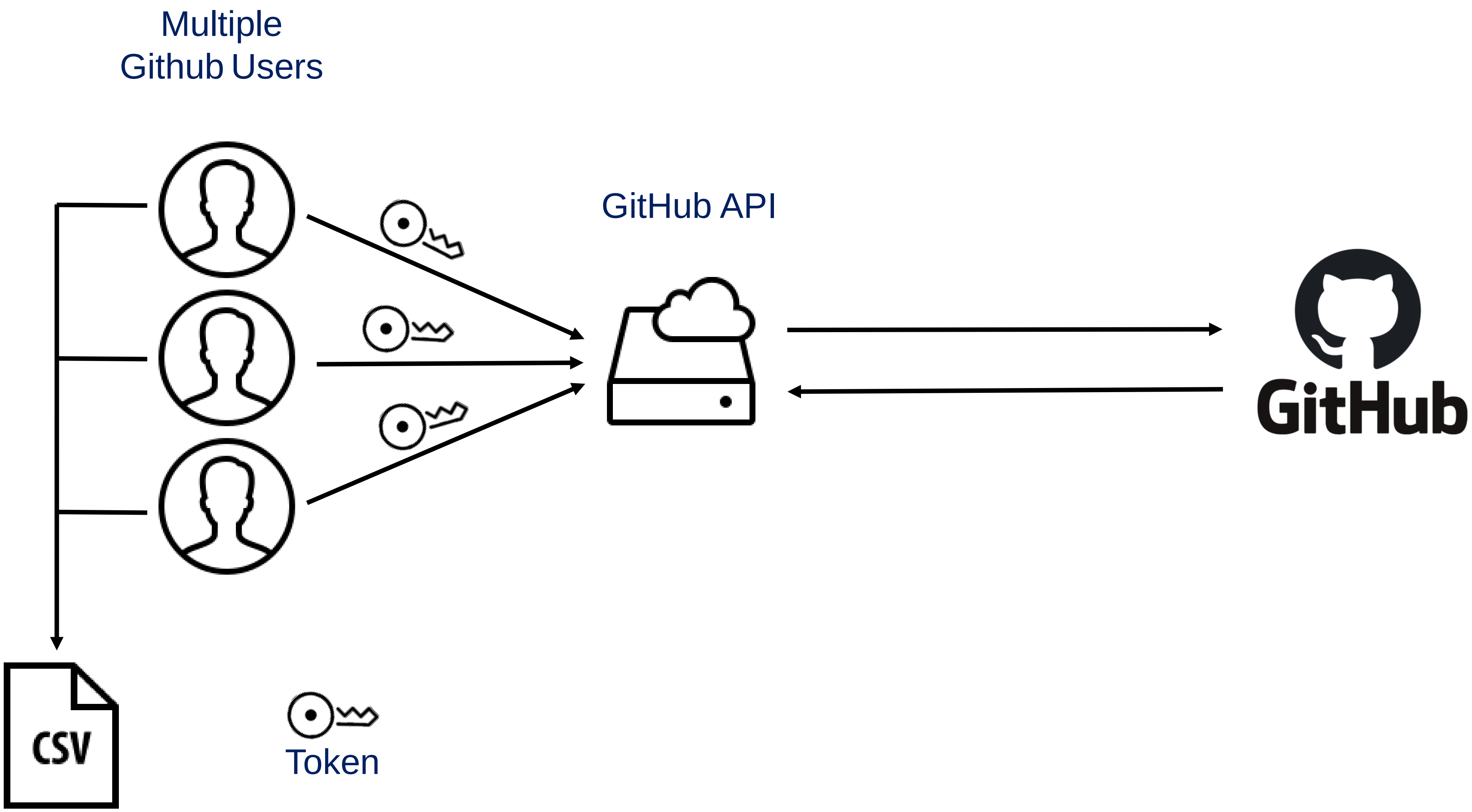


Python

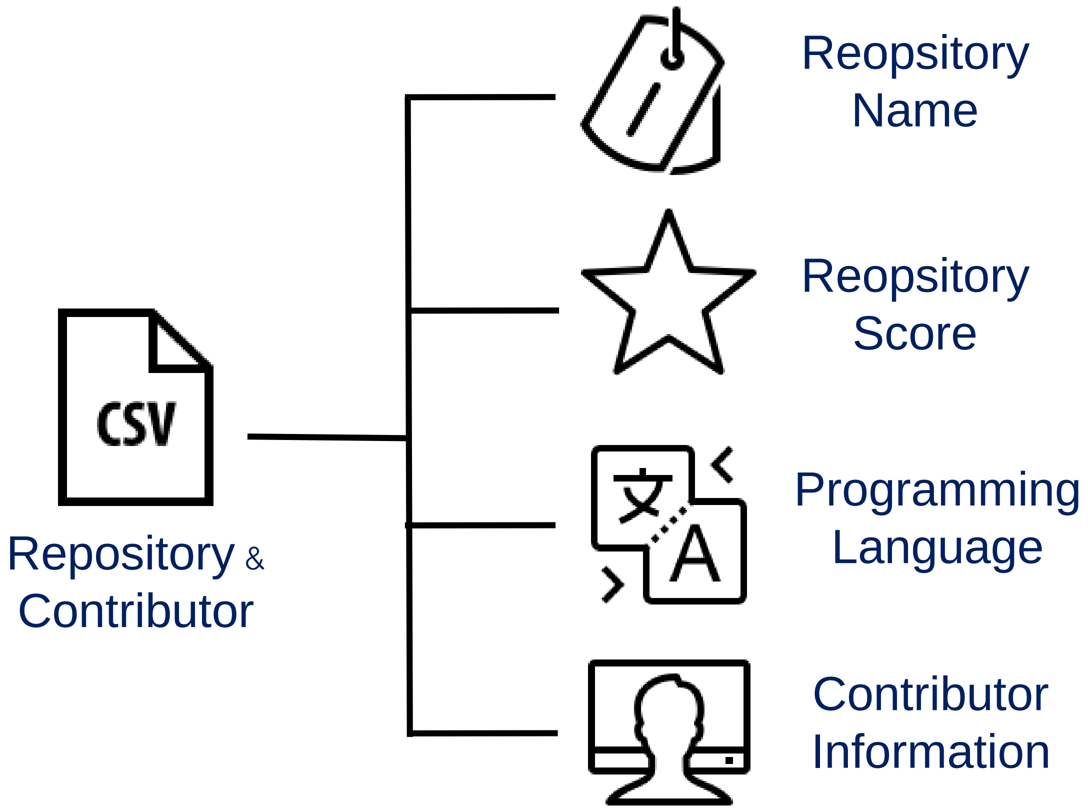
GitHub API

Limit 1,500 request/ hour

II. Data Collections (Data Gathering)



II. Data Collections (Data Gathering)



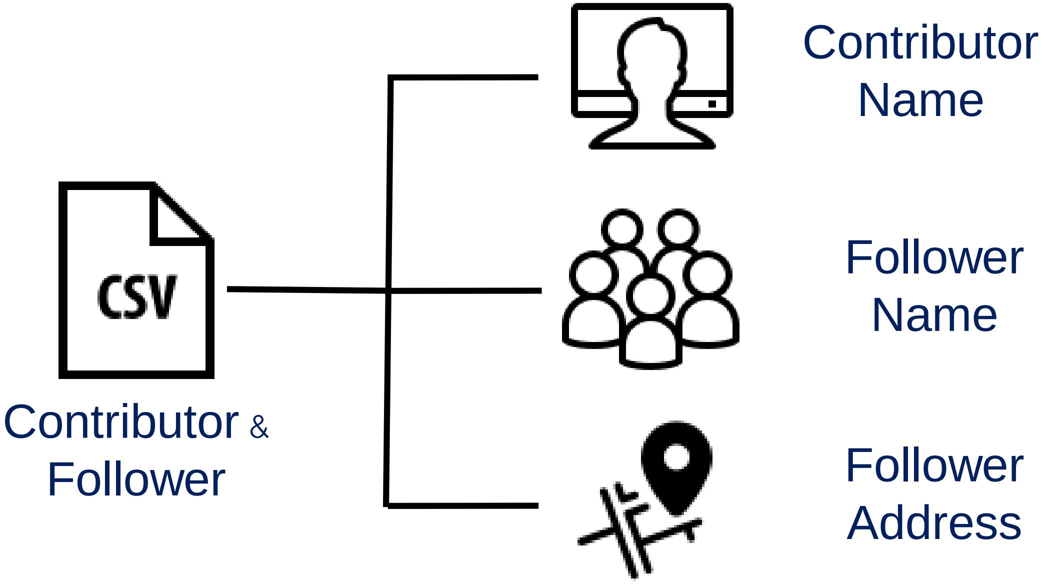
Repository & Contributor

Reopsitory Name

Reopsitory Score

Programming Language

Contributor Information



Contributor & Follower

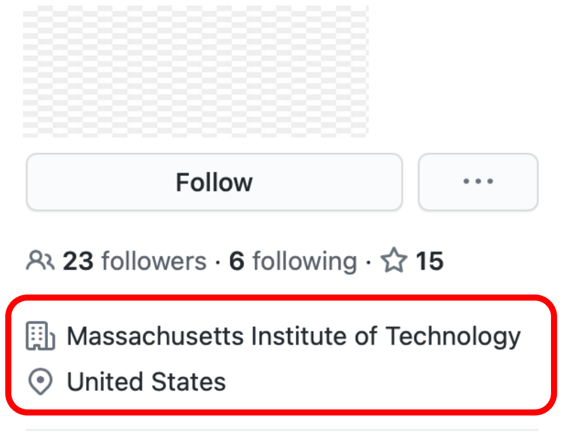
Contributor Name

Follower Name

Follower Address

II. Data Cleansing (Data Preprocessing)

GitHub user location



Follow

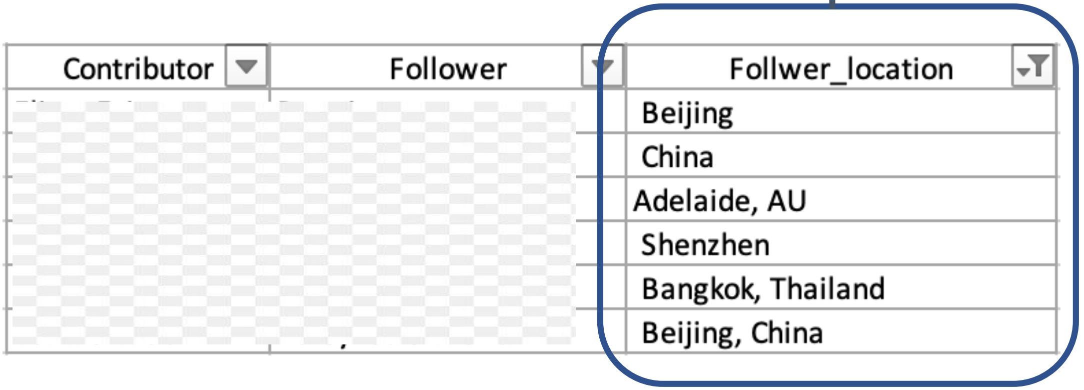
23 followers · 6 following · 15

Massachusetts Institute of Technology

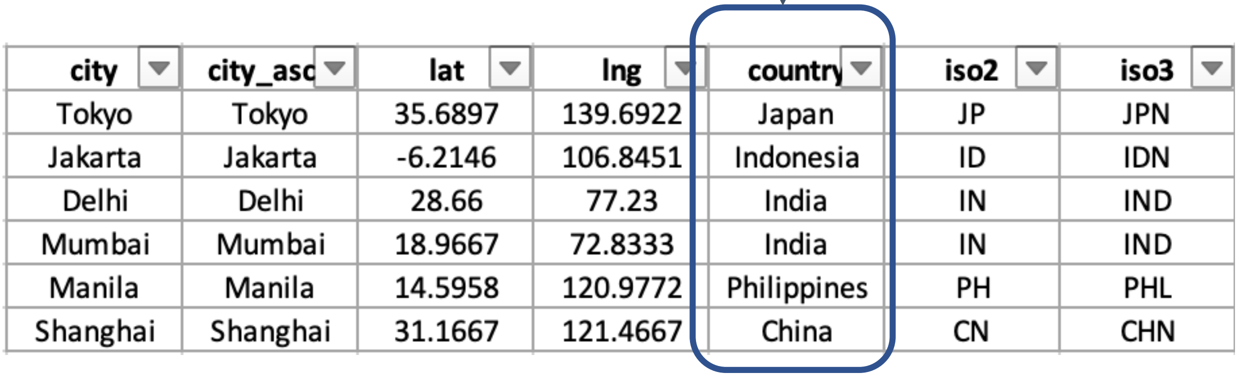
United States

simplemaps

Interactive Maps & Data



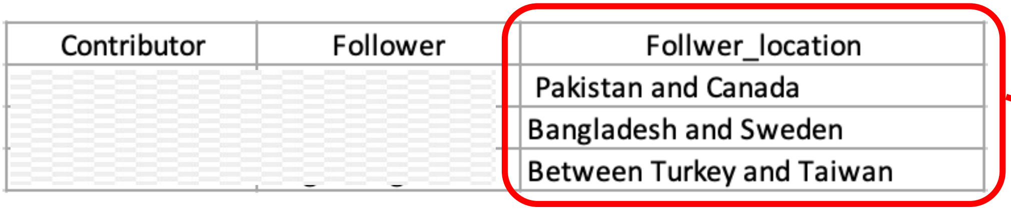
Contributor	Follower	Follwer_location
[redacted]	[redacted]	Beijing
[redacted]	[redacted]	China
[redacted]	[redacted]	Adelaide, AU
[redacted]	[redacted]	Shenzhen
[redacted]	[redacted]	Bangkok, Thailand
[redacted]	[redacted]	Beijing, China



city	city_asc	lat	lng	country	iso2	iso3
Tokyo	Tokyo	35.6897	139.6922	Japan	JP	JPN
Jakarta	Jakarta	-6.2146	106.8451	Indonesia	ID	IDN
Delhi	Delhi	28.66	77.23	India	IN	IND
Mumbai	Mumbai	18.9667	72.8333	India	IN	IND
Manila	Manila	14.5958	120.9772	Philippines	PH	PHL
Shanghai	Shanghai	31.1667	121.4667	China	CN	CHN

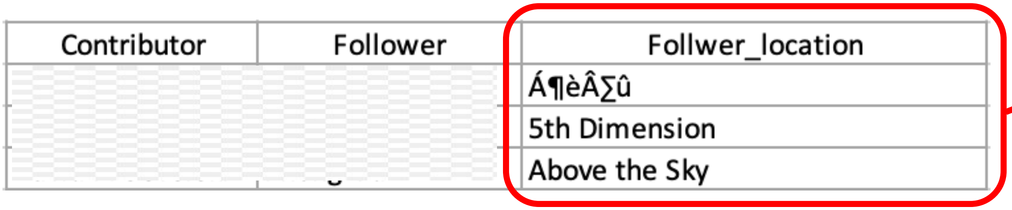
II. Data Cleansing (Data Preprocessing)

Ambiguous location



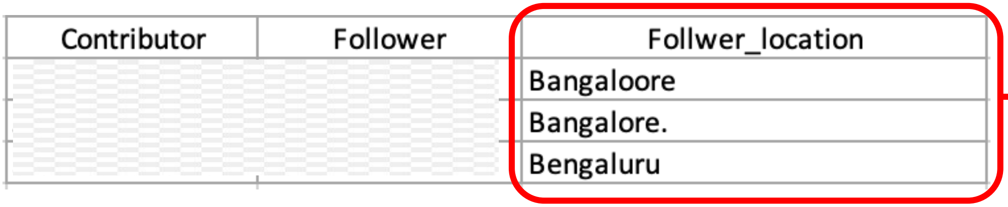
Contributor	Follower	Follwer_location
[redacted]	[redacted]	Pakistan and Canada
[redacted]	[redacted]	Bangladesh and Sweden
[redacted]	[redacted]	Between Turkey and Taiwan

Incorrect location



Contributor	Follower	Follwer_location
[redacted]	[redacted]	Á¶èÂ∑û
[redacted]	[redacted]	5th Dimension
[redacted]	[redacted]	Above the Sky

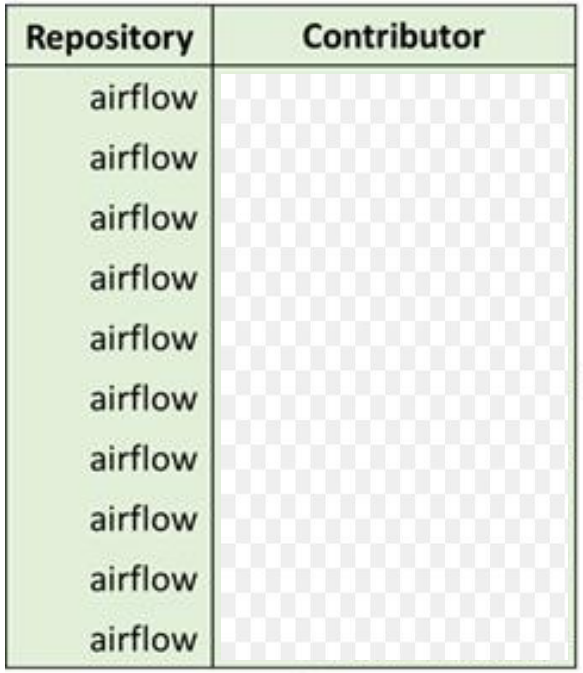
Typo and multiple spell



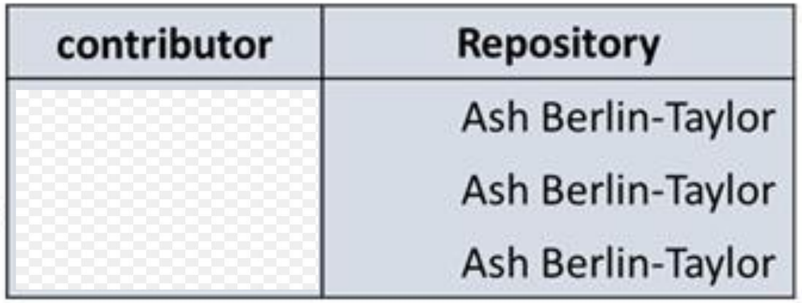
Contributor	Follower	Follwer_location
[redacted]	[redacted]	Bangaloore
[redacted]	[redacted]	Bangalore.
[redacted]	[redacted]	Bengaluru

Bangalore

III. Exploratory Analysis



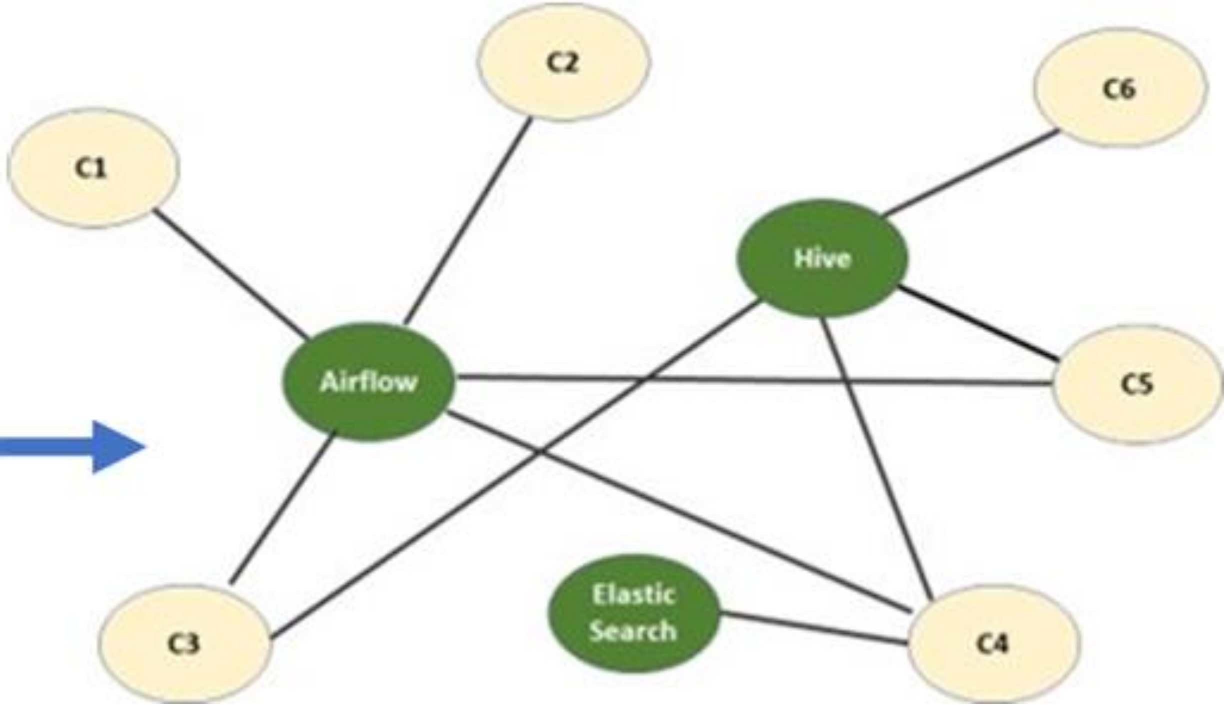
Repository	Contributor
airflow	
airflow	
airflow	
airflow	
airflow	
airflow	
airflow	
airflow	
airflow	
airflow	



contributor	Repository
	Ash Berlin-Taylor
	Ash Berlin-Taylor
	Ash Berlin-Taylor

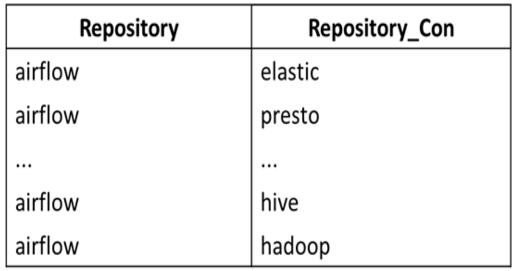
Bipartite

Graph

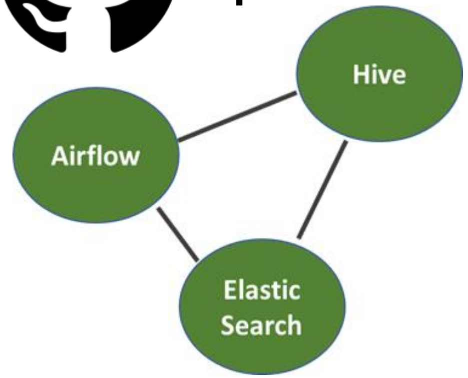


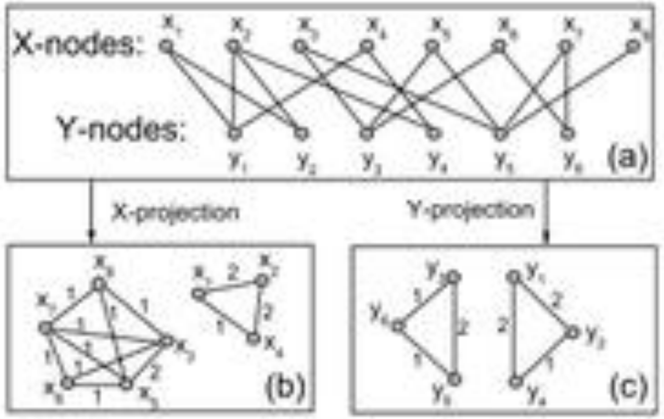
III. Exploratory Analysis

Projected Graph



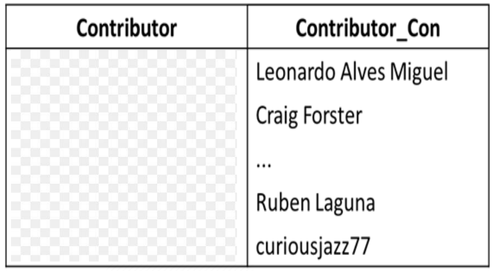
Repository	Repository_Con
airflow	elastic
airflow	presto
…	…
airflow	hive
airflow	hadoop



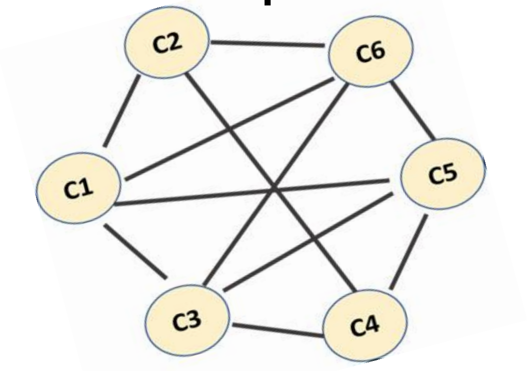


X-nodes:
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
Y-nodes:
y_1 y_2 y_3 y_4 y_5 y_6
(a)
X-projection
Y-projection
(b)
(c)

FIG. 1. Illustration of a bipartite network (a), as well as its X projection (b) and Y projection (c). The edge weight in (b) and (c) is set as the number of common neighbors in Y and X, respectively.



Contributor	Contributor_Con
	Leonardo Alves Miguel
	Craig Forster
	…
	Ruben Laguna
	curiousjazz77



III. Exploratory Analysis

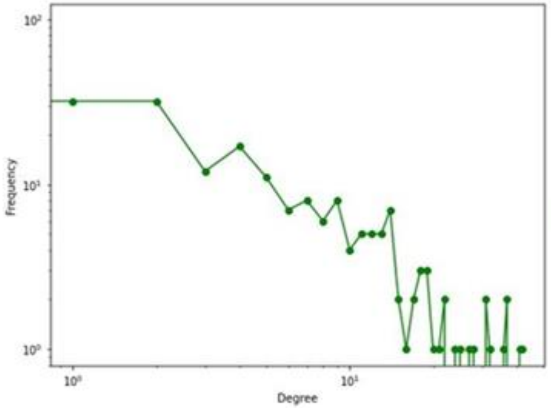
Overview

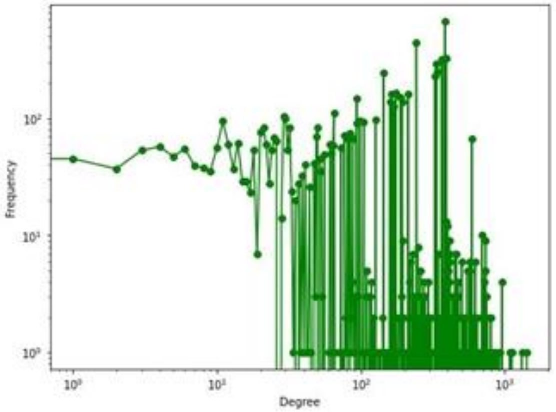
Repository

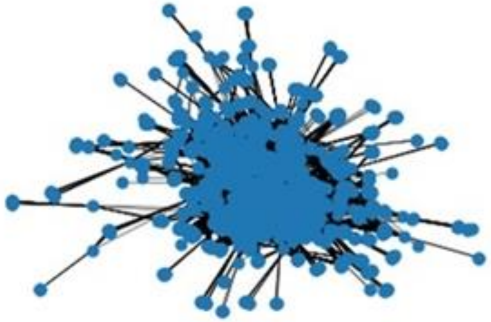
Average Degree : 5.12
Density : 0.018
Assortivity : -0.095
Average Clustering : 0.29

Contributor

Average Degree : 198.76
Density : 0.024
Assortivity : 0.482
Average Clustering : 0.96

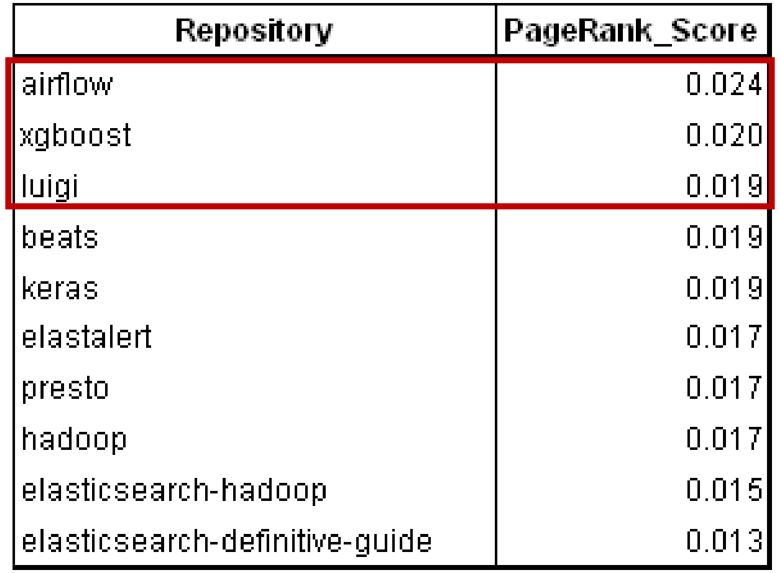




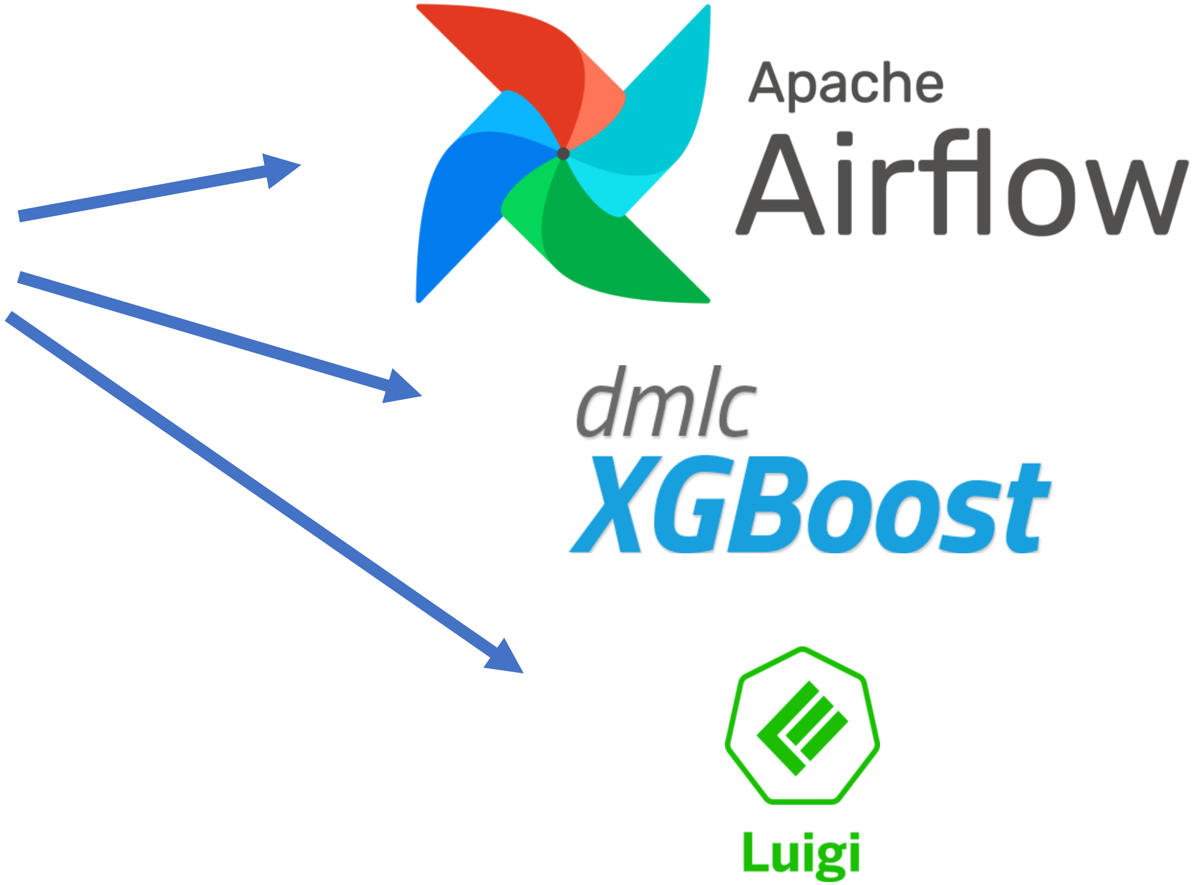


III. Exploratory Analysis

Influential Repositories

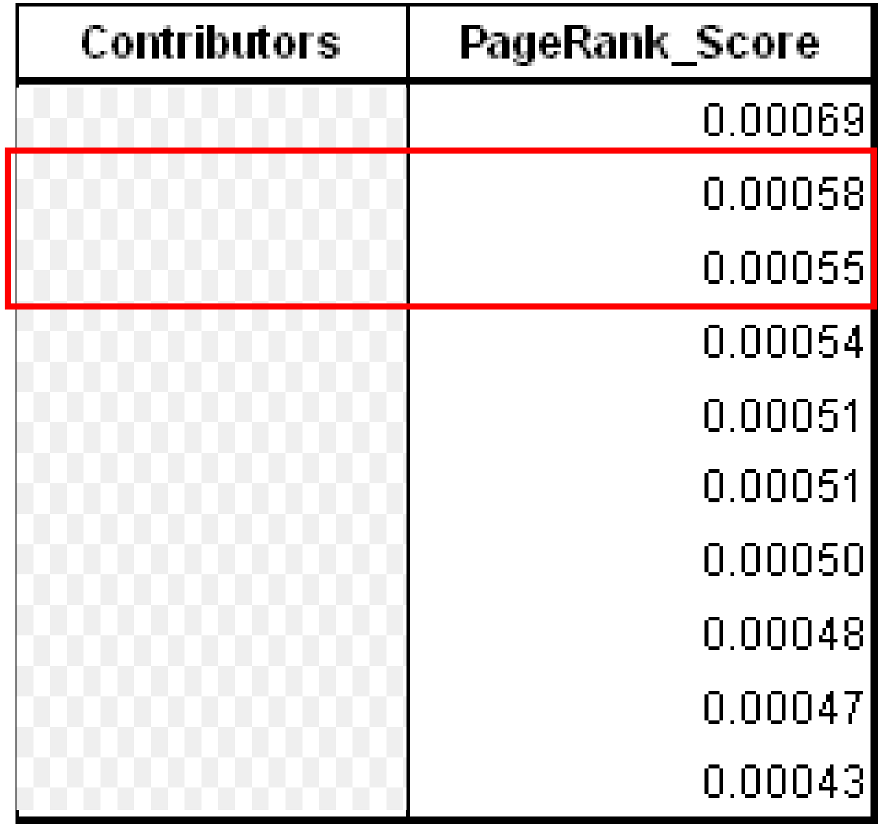


Repository	PageRank_Score
airflow	0.024
xgboost	0.020
luigi	0.019
beats	0.019
keras	0.019
elastalert	0.017
presto	0.017
hadoop	0.017
elasticsearch-hadoop	0.015
elasticsearch-definitive-guide	0.013

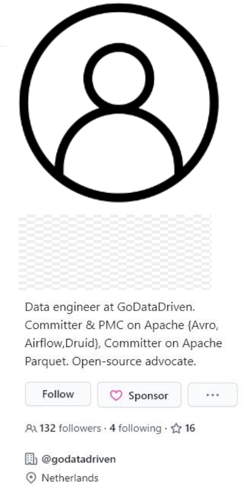


III. Exploratory Analysis

Influential Contributors



Contributors	PageRank_Score
[redacted]	0.00069
[redacted]	0.00058
[redacted]	0.00055
[redacted]	0.00054
[redacted]	0.00051
[redacted]	0.00051
[redacted]	0.00050
[redacted]	0.00048
[redacted]	0.00047
[redacted]	0.00043



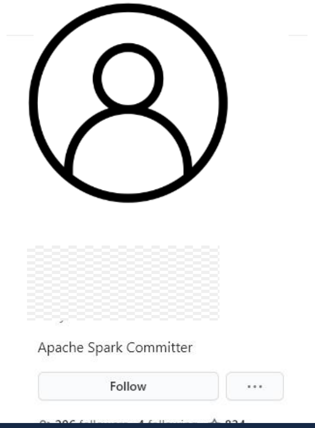
[redacted]

Data engineer at GoDataDriven. Committer & PMC on Apache (Avro, Airflow,Druid), Committer on Apache Parquet. Open-source advocate.

132 followers · 4 following · ☆ 16

@godatadriven

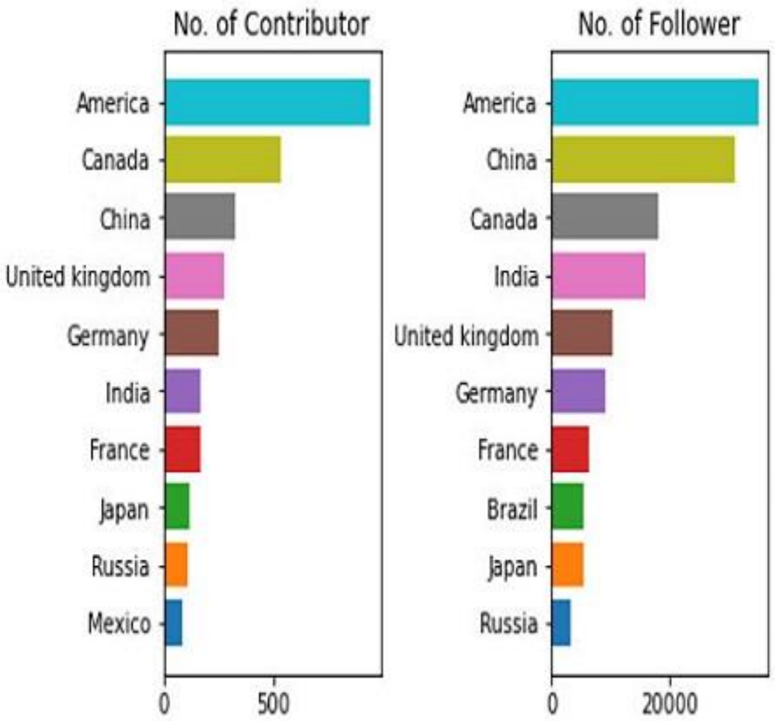
Netherlands

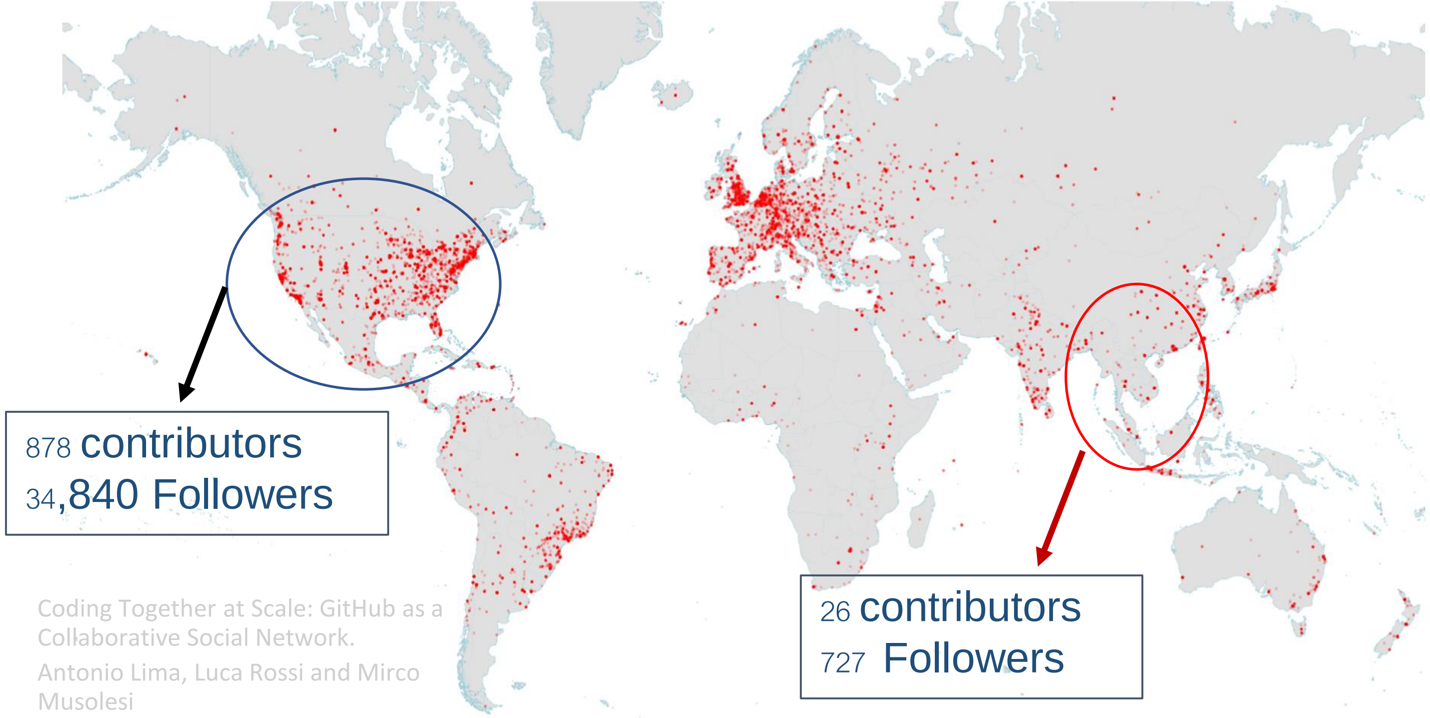


[redacted]

Apache Spark Committer

III. Exploratory Analysis (Country Analysis)





878 contributors
34,840 Followers

26 contributors
727 Followers

Coding Together at Scale: GitHub as a Collaborative Social Network.
Antonio Lima, Luca Rossi and Mirco Musolesi

The top 10 rank of contributors and followers :
- Majority of users is located in North America and in Europe
- The leading countries are the United States (USA) and China, Canada on both graph

III. Exploratory Analysis(Country Analysis)

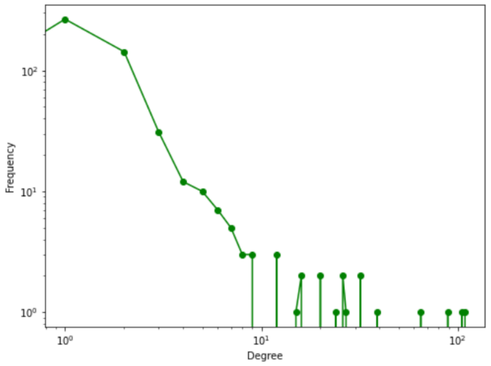
Overview with Thailand

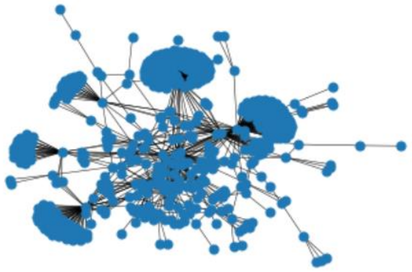
Repository

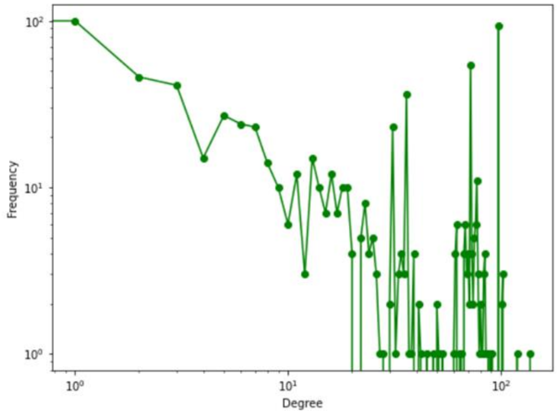
Users (Countributor + Follower)

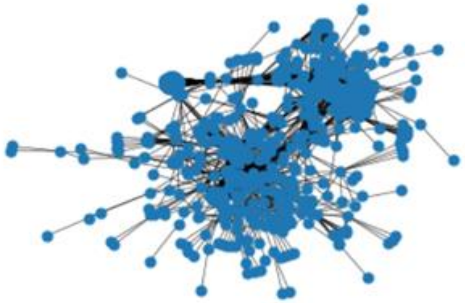
Average Degree : 3.14
Density : 0.006
Assortivity : -0.452
Average Clustering : 0.26

Average Degree : 34.32
Density : 0.006
Assortivity : 0.788
Average Clustering : 0.68



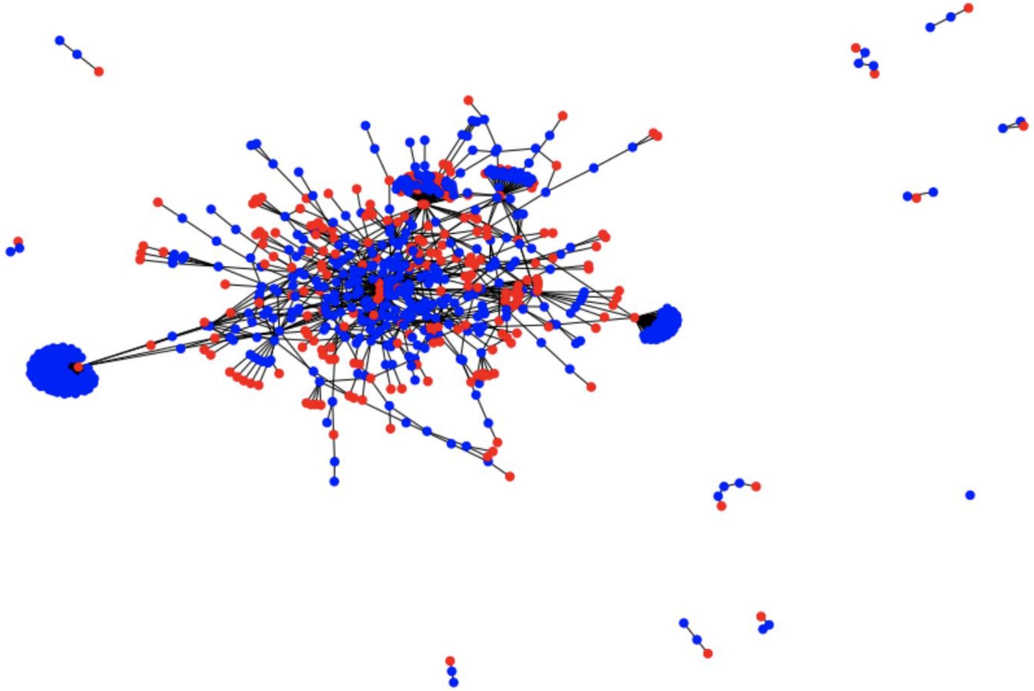




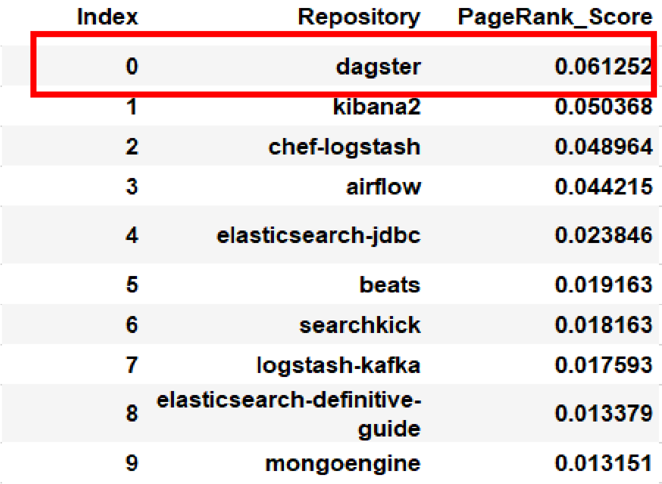


III. Exploratory Analysis (Country Analysis)

Number of Repository Nodes : 499
Number of User Nodes : 753
Number of Edges : 324

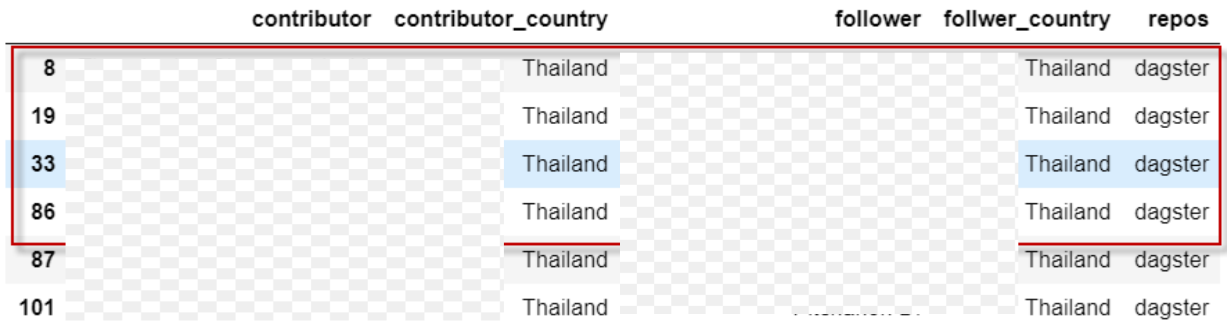


Red : Users in Thailand
Bule : Users in Other Country



Index	Repository	PageRank_Score
0	dagster	0.061252
1	kibana2	0.050368
2	chef-logstash	0.048964
3	airflow	0.044215
4	elasticsearch-jdbc	0.023846
5	beats	0.019163
6	searchkick	0.018163
7	logstash-kafka	0.017593
8	elasticsearch-definitive-guide	0.013379
9	mongoengine	0.013151

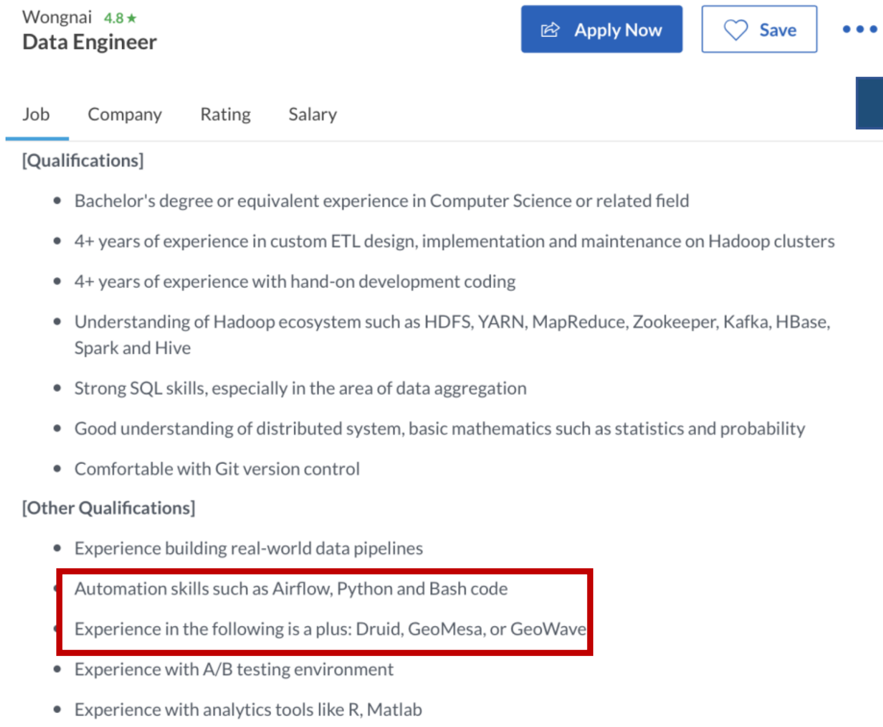
DAGSTER



	contributor	contributor_country	follower	follwer_country	repos
8	[redacted]	Thailand	[redacted]	Thailand	dagster
19	[redacted]	Thailand	[redacted]	Thailand	dagster
33	[redacted]	Thailand	[redacted]	Thailand	dagster
86	[redacted]	Thailand	[redacted]	Thailand	dagster
87	[redacted]	Thailand	[redacted]	Thailand	dagster
101	[redacted]	Thailand	[redacted]	Thailand	dagster

IV. Recommendation System

Job Description



Wongnai 4.8★

Data Engineer

Apply Now | Save | •••

Job | Company | Rating | Salary

[Qualifications]

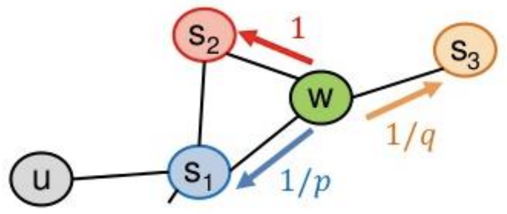
- Bachelor's degree or equivalent experience in Computer Science or related field
- 4+ years of experience in custom ETL design, implementation and maintenance on Hadoop clusters
- 4+ years of experience with hand-on development coding
- Understanding of Hadoop ecosystem such as HDFS, YARN, MapReduce, Zookeeper, Kafka, HBase, Spark and Hive
- Strong SQL skills, especially in the area of data aggregation
- Good understanding of distributed system, basic mathematics such as statistics and probability
- Comfortable with Git version control

[Other Qualifications]

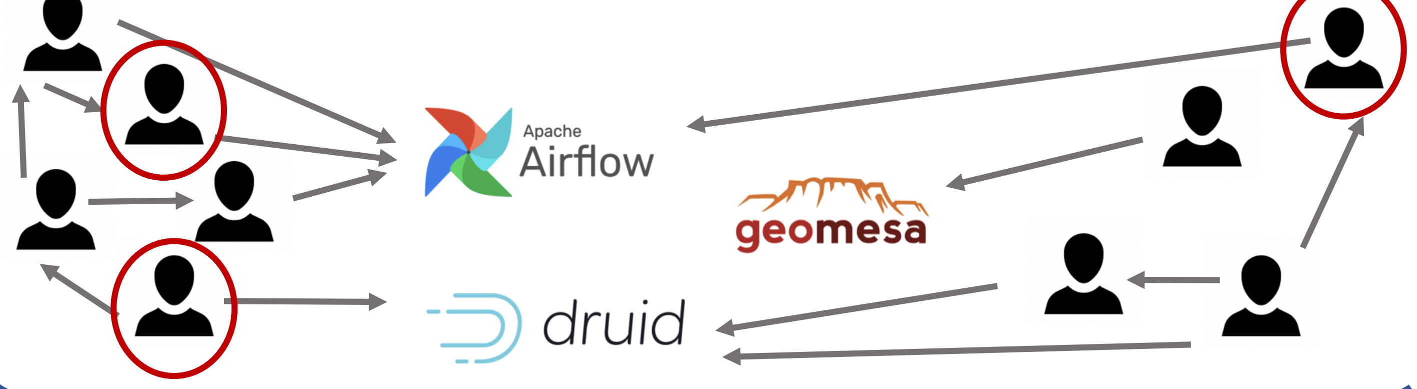
- Experience building real-world data pipelines
- Automation skills such as Airflow, Python and Bash code
- Experience in the following is a plus: Druid, GeoMesa, or GeoWave
- Experience with A/B testing environment
- Experience with analytics tools like R, Matlab

Skill Extraction

Network Modeling



Recommendation Potential Candidates



IV. Recommendation System

Node2Vec

A node embedding model by extending Skip-gram architecture to networks is used to satisfy the exploration-exploitation trade-off with random walk sampling to explore neighborhoods in local and global structure.

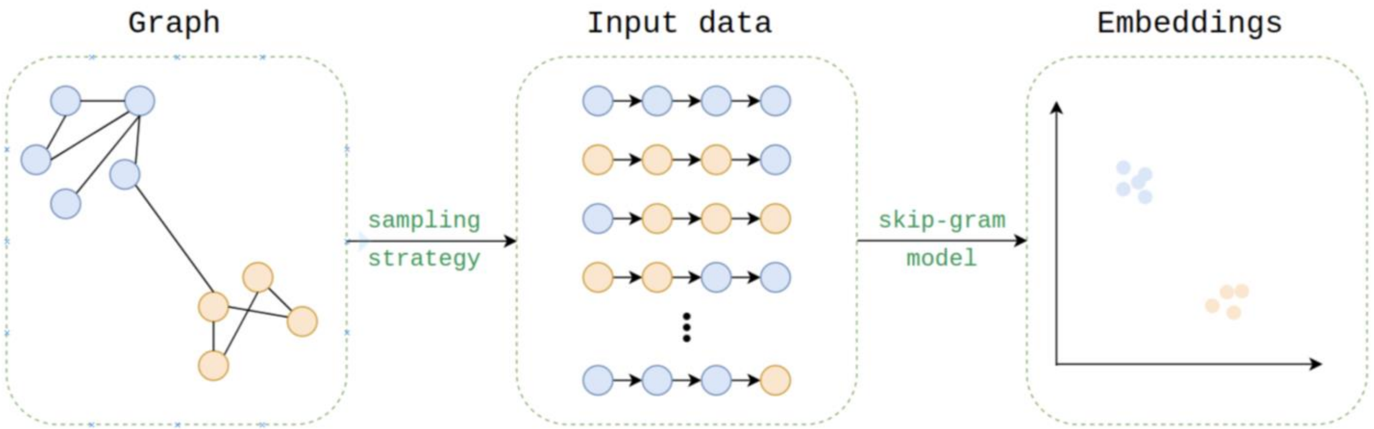
Concept:

- Breadth-First Sampling (BFS): Focus on local neighborhoods.
- Depth-First Sampling (DFS): Focus on global neighborhoods.

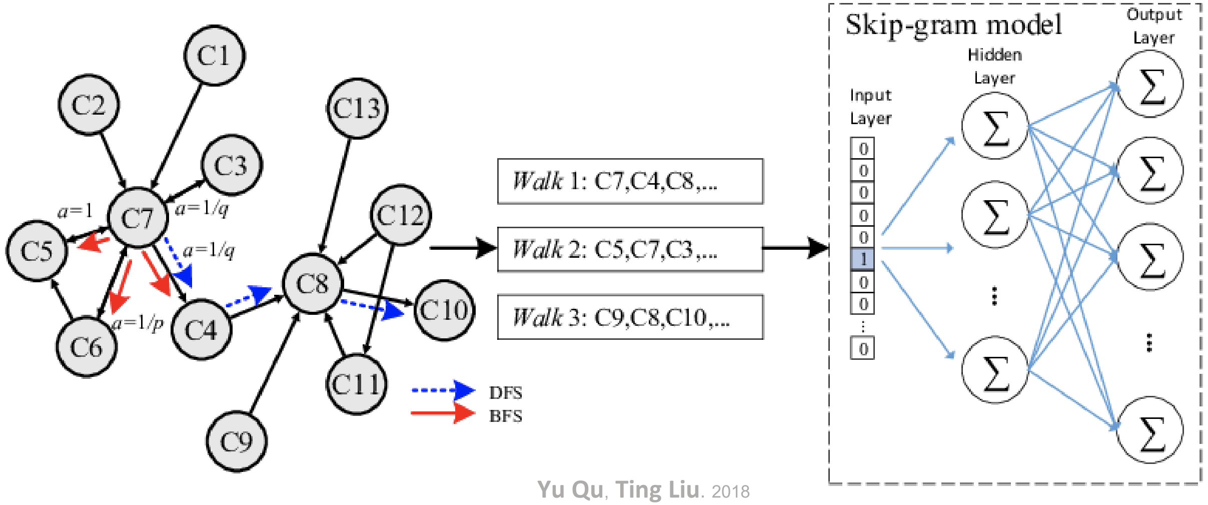
Parameters:

P,Q for model transition probabilies.

- P is return parameter
- Q is "walk away" parameter



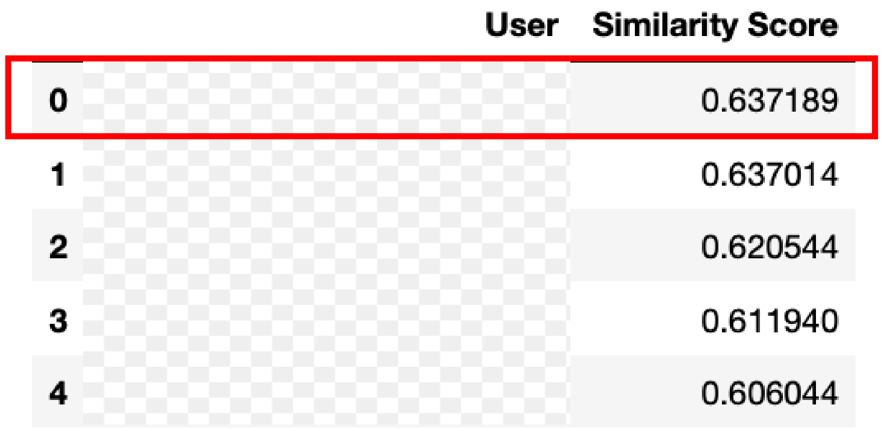
Node2vec embedding process



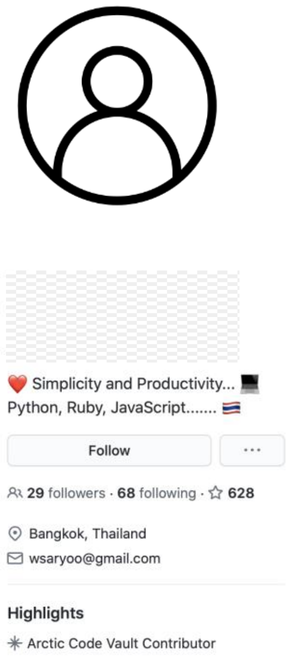
Yu Qu, Ting Liu. 2018

IV. Recommendation System

Airflow Potential Candidates



	User	Similarity Score
0	[redacted]	0.637189
1	[redacted]	0.637014
2	[redacted]	0.620544
3	[redacted]	0.611940
4	[redacted]	0.606044



[redacted]

❤️ Simplicity and Productivity... 💻 Python, Ruby, JavaScript....... 🇹🇭

29 followers · 68 following · ☆ 628

Bangkok, Thailand

wsaryoo@gmail.com

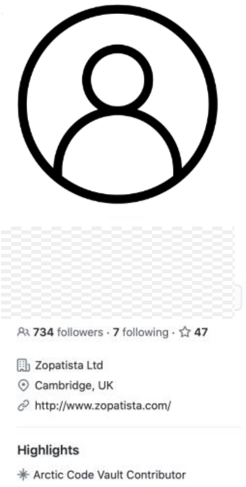
Highlights

Arctic Code Vault Contributor

Follows

Apache Airflow

Contributor



[redacted]

734 followers · 7 following · ☆ 47

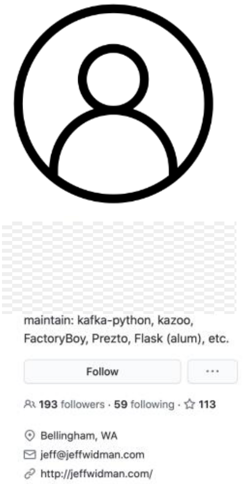
Zopatista Ltd

Cambridge, UK

http://www.zopatista.com/

Highlights

Arctic Code Vault Contributor



[redacted]

maintain: kafka-python, kazoo, FactoryBoy, Prezto, Flask (alum), etc.

193 followers · 59 following · ☆ 113

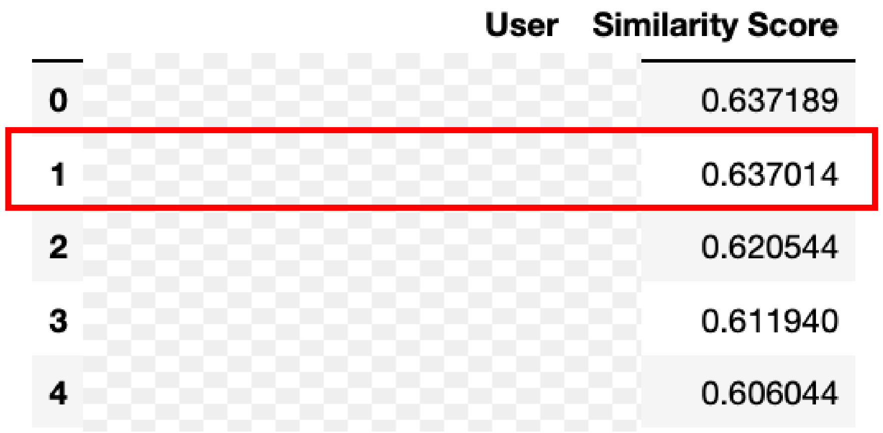
Bellingham, WA

jeff@jeffwidman.com

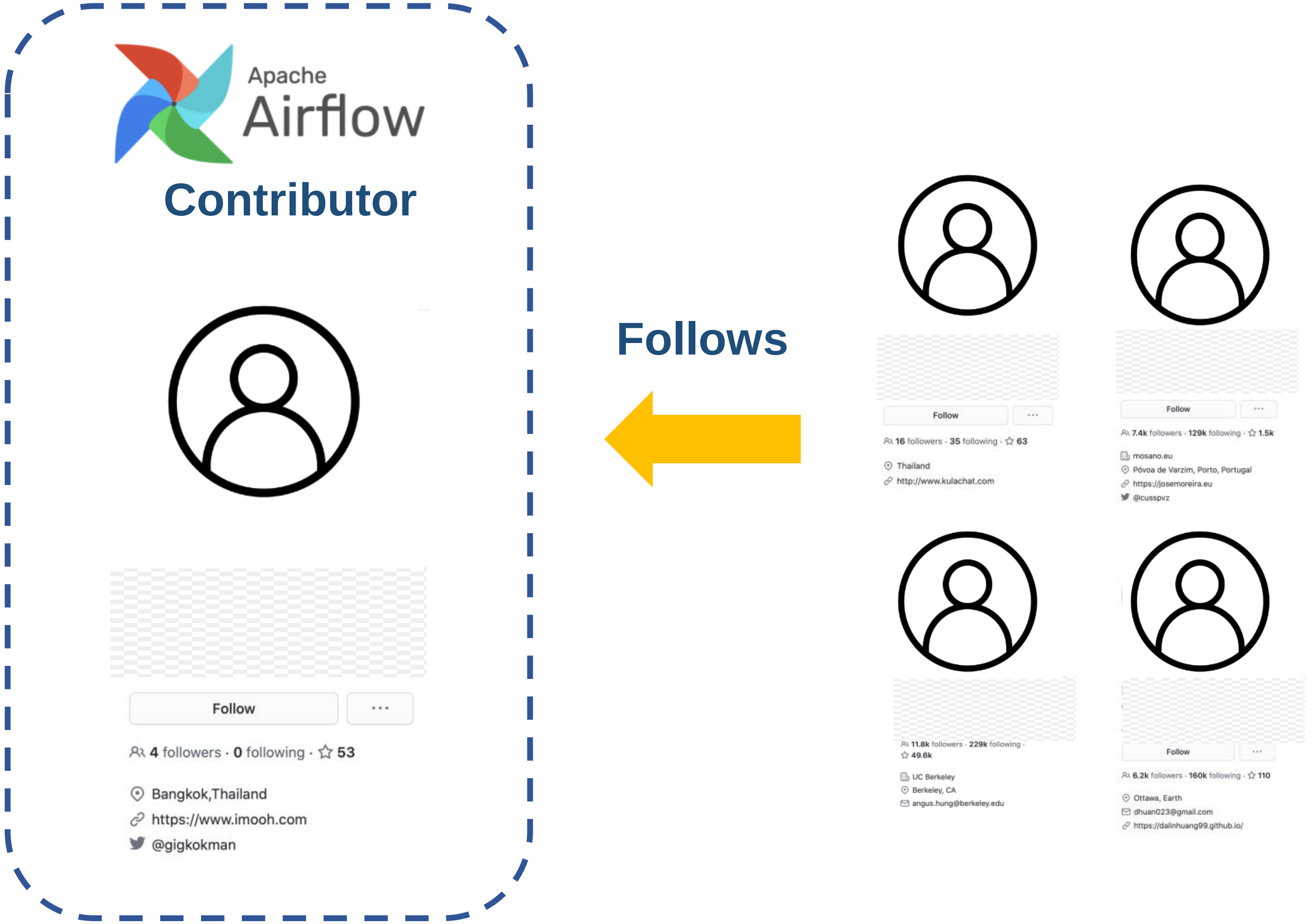
http://jeffwidman.com/

IV. Recommendation System

Airflow Potential Candidates



	User	Similarity Score
0	[redacted]	0.637189
1	[redacted]	0.637014
2	[redacted]	0.620544
3	[redacted]	0.611940
4	[redacted]	0.606044



Contributor

[redacted]

4 followers · 0 following · ☆ 53

Bangkok,Thailand
https://www.imooh.com
@gigkokman

Follows

[redacted]

16 followers · 35 following · ☆ 63

Thailand
http://www.kulachat.com

[redacted]

7.4k followers · 129k following · ☆ 1.5k

mosano.eu
Póvoa de Varzim, Porto, Portugal
https://josemoreira.eu
@cusspvz

[redacted]

11.8k followers · 229k following · ☆ 49.6k

UC Berkeley
Berkeley, CA
angus.hung@berkeley.edu

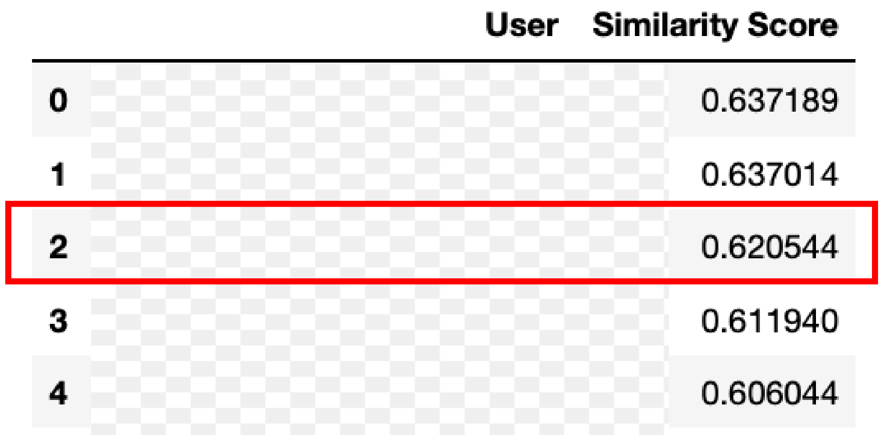
[redacted]

6.2k followers · 160k following · ☆ 110

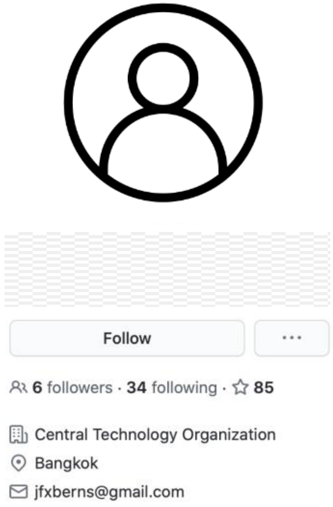
Ottawa, Earth
dhuan023@gmail.com
https://dalinhuang99.github.io/

IV. Recommendation System

Airflow Potential Candidates



	User	Similarity Score
0	[redacted]	0.637189
1	[redacted]	0.637014
2	[redacted]	0.620544
3	[redacted]	0.611940
4	[redacted]	0.606044



[redacted]

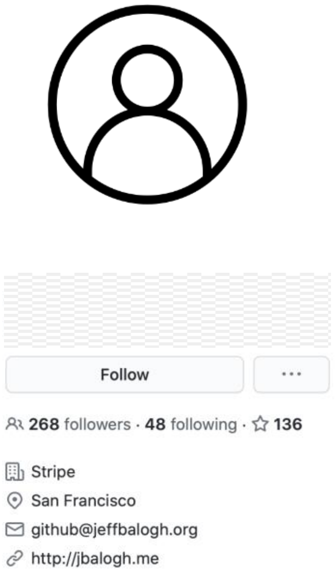
6 followers · 34 following · ☆ 85

Central Technology Organization
Bangkok
jfxberns@gmail.com

Follows

Apache Airflow

Contributor



[redacted]

268 followers · 48 following · ☆ 136

Stripe
San Francisco
github@jeffbalogh.org
http://jbalogh.me

Thank You