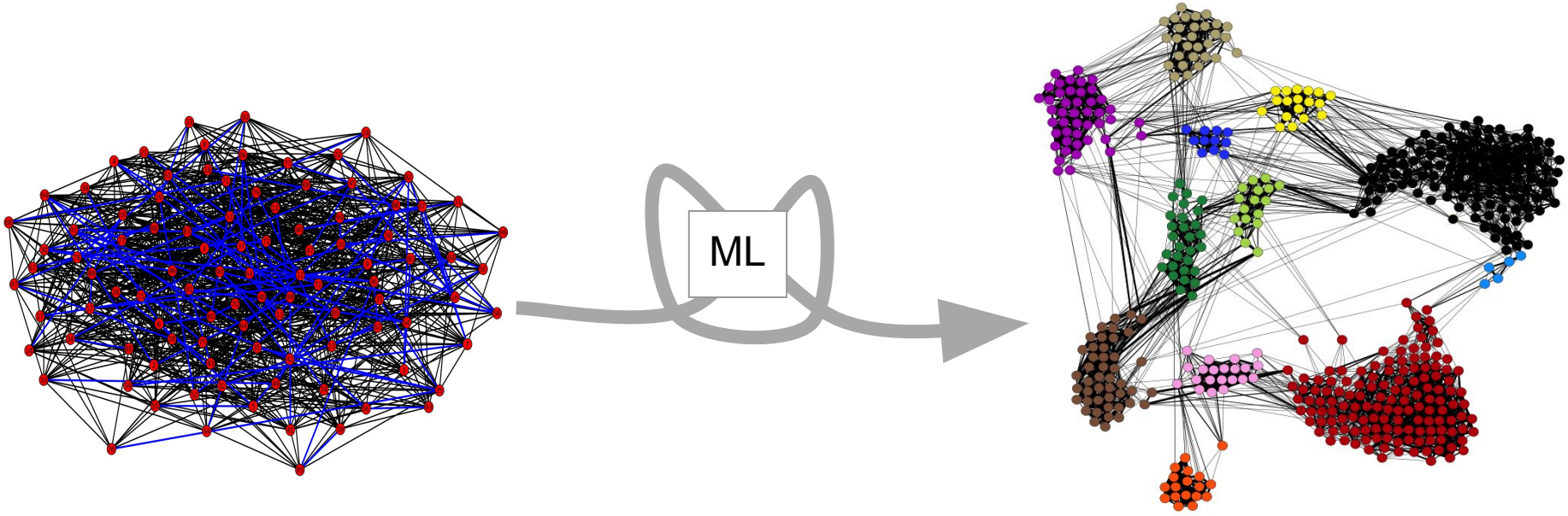
Scholarly
Academic Data Visualization & Analysis

Miguel Yanez (may2114), Michelle Tadmor (mdt2125), Yu Hsuan Shih (ys2898)

1

Motivation

Why do this?



How to give meaning to the chaos?

2

Dataset

Where did we get our data?

2,092,356
Papers

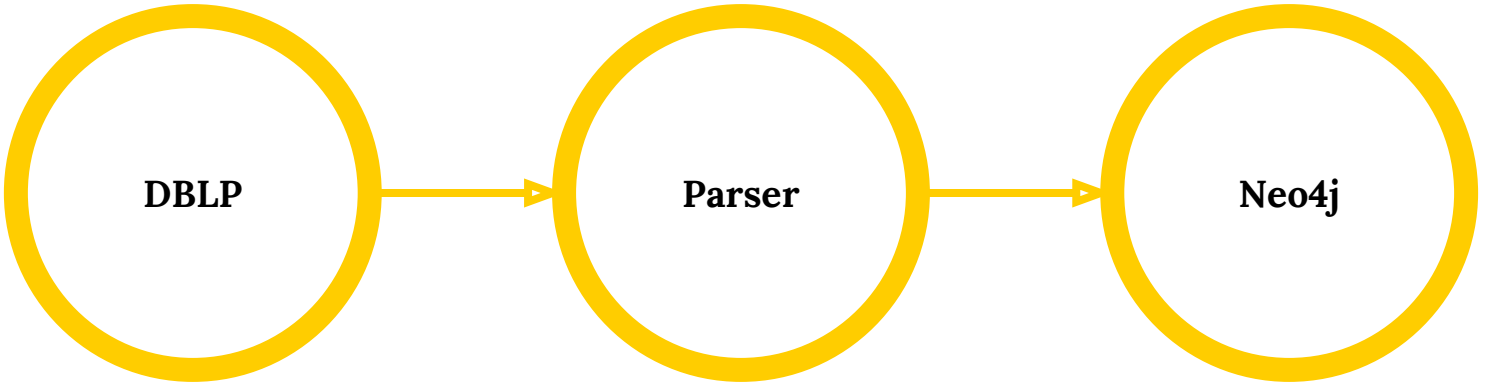
8,024,869
Citations

1,712,433
Authors

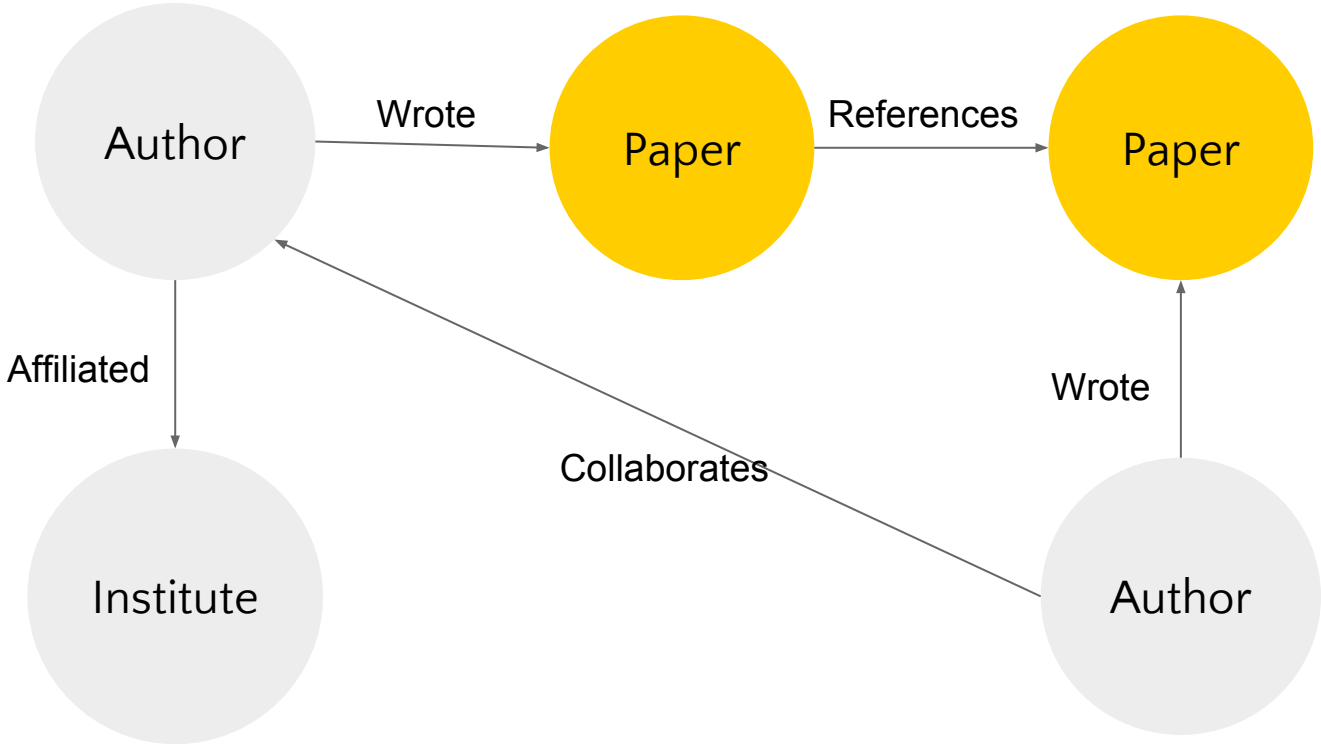
4,258,615
Collaborations

dblp
computer science bibliography

Feature Extraction Process



Extracted Relationships



3

Architecture

What approach did we take?

The Tools

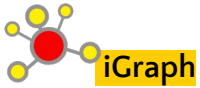
To store our dataset

To create visualizations

To communicate with dataset

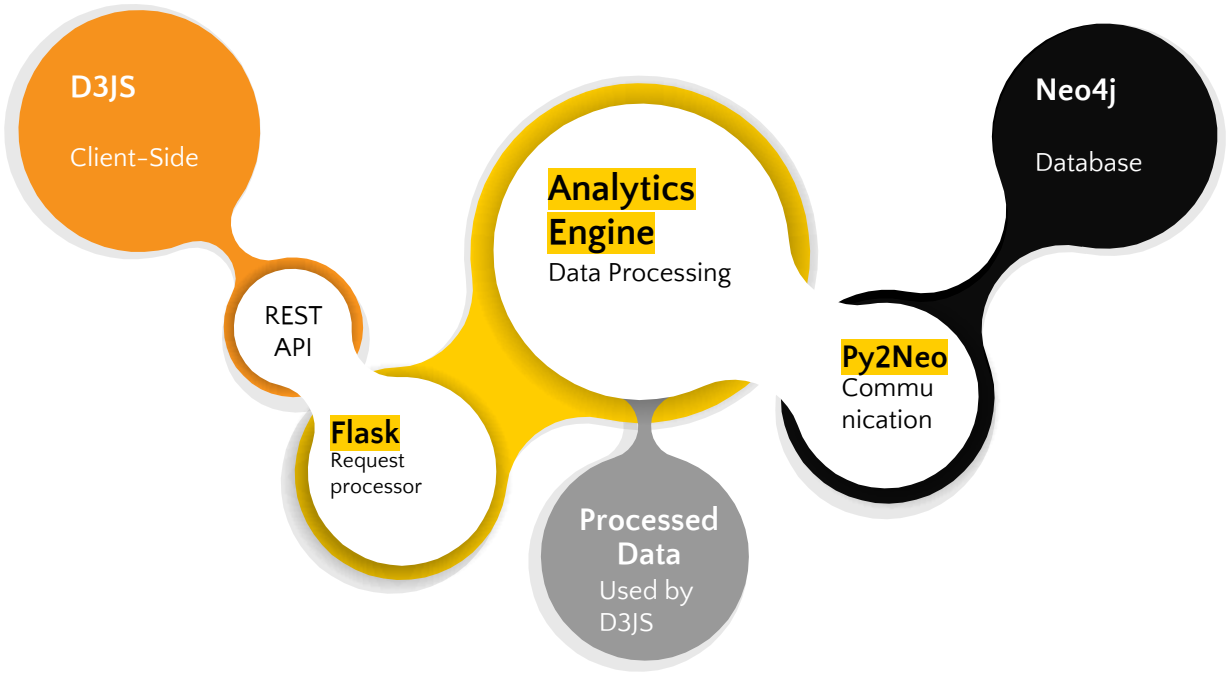
To run our web app

To host our ecosystem



To perform analysis

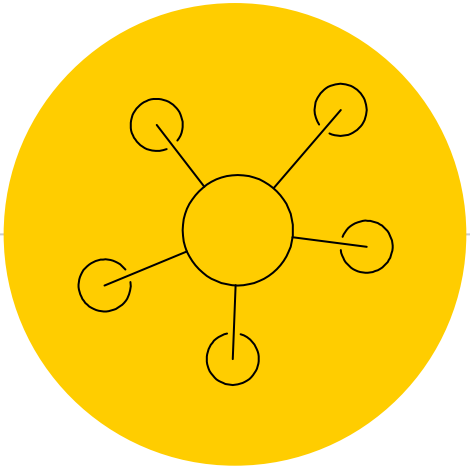
Our Setup


D3JS
Client-Side
REST
API
Flask
Request
processor
Analytics
Engine
Data Processing
Processed
Data
Used by
D3JS
Py2Neo
Commu
nication
Neo4j
Database

4

Algorithms

A bit of background...



Clustering

Girvan - Newman algorithm

Popular graph clustering algorithm published in 2002

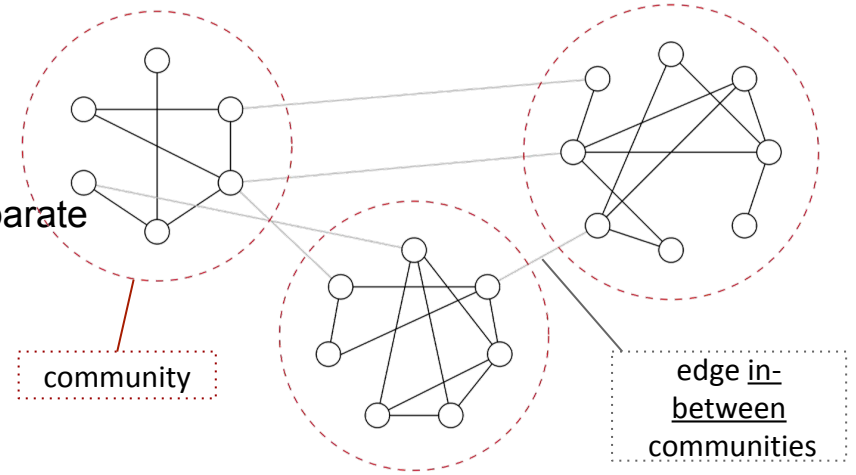
Community: connected graph component

Edge in-between score: measures the probability that a shortest path will travel through an edge.

Goal: remove spurious edges which connect otherwise separate communities.

GM Outline:

(1) Compute inbetween score for all edges (dijkstra)
(2) remove top scoring edges
(3) re-compute score for all edges and repeat (2)

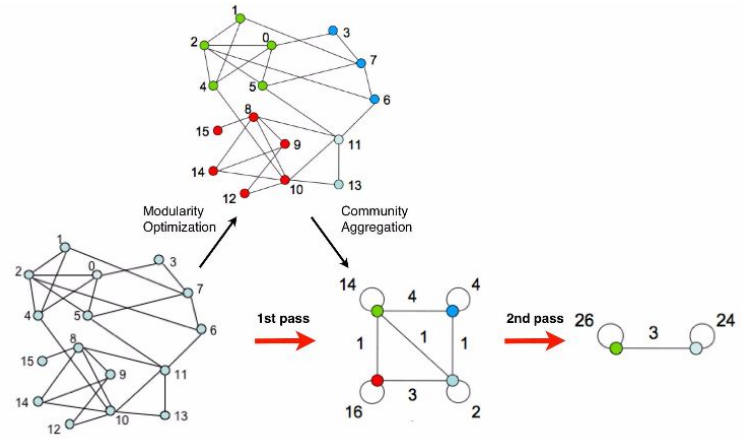


$$O(|V||E|^3)$$

Girvan M, Newman MEJ. Community structure in social and biological networks. 2002

Multi-Level

Modularity: quality score of a partition given a graph structure which compared the intra-community edge count with the expected number of edges for the graph. ML is one of many graph partitioning algorithms that seek to maximize this score.



Goal: Find a partition which maximizes the modularity score for a given graph.

Obstacles: search space is exponential, resolution Limit

Solution: **Multi-Level** - Approximate optimal partition using an iterative greedy passes

ML Outline:

(1) init: each node is a partition
(2) max-modularity: merge nodes with neighboring nodes
(3) aggregate based on the partitioning and repeat (2)

$$O(|V||E|)$$

V.D. Blondel et al. "*Fast unfolding of communities in large networks*," 2008

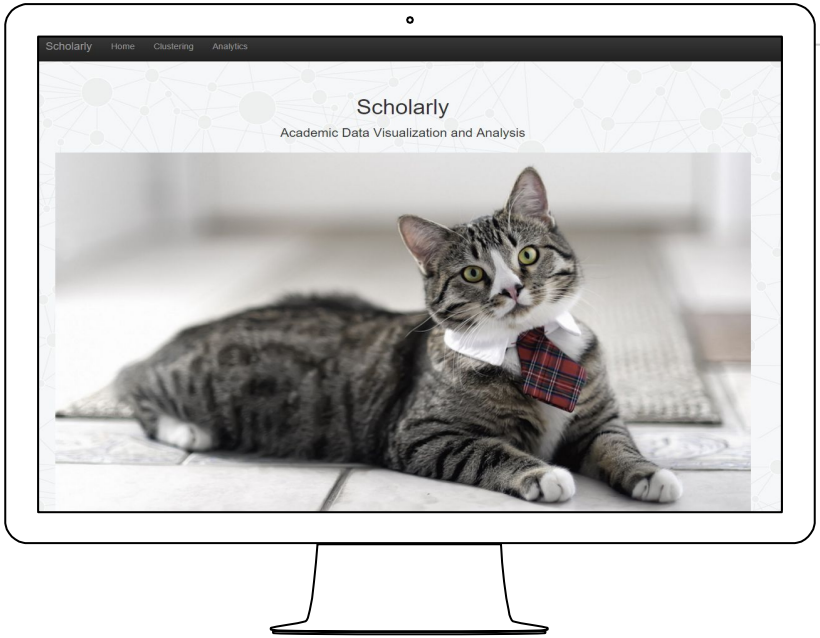
5

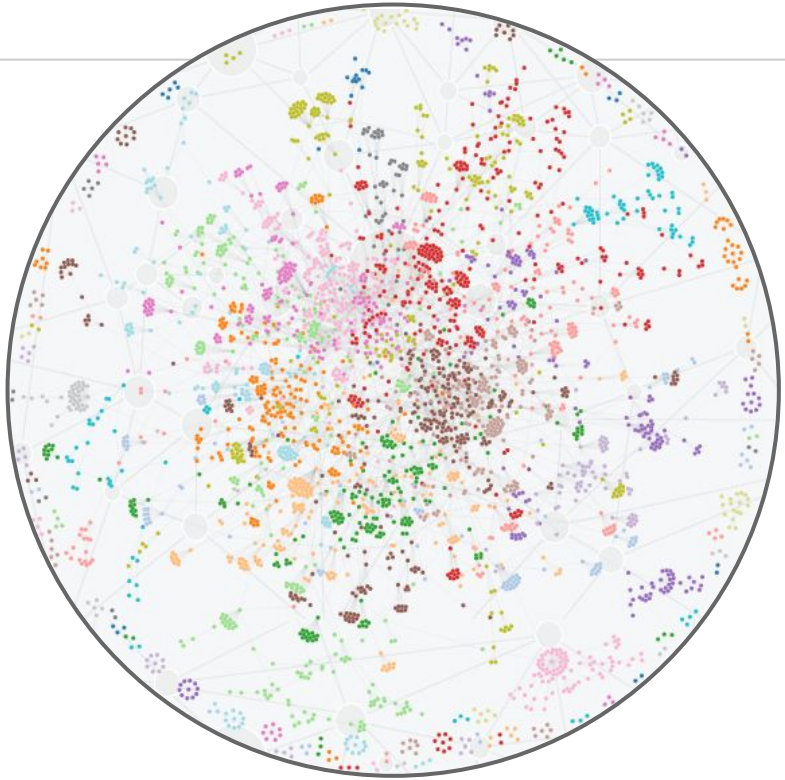
Visualizations

What's the best way to show our results?

Demo

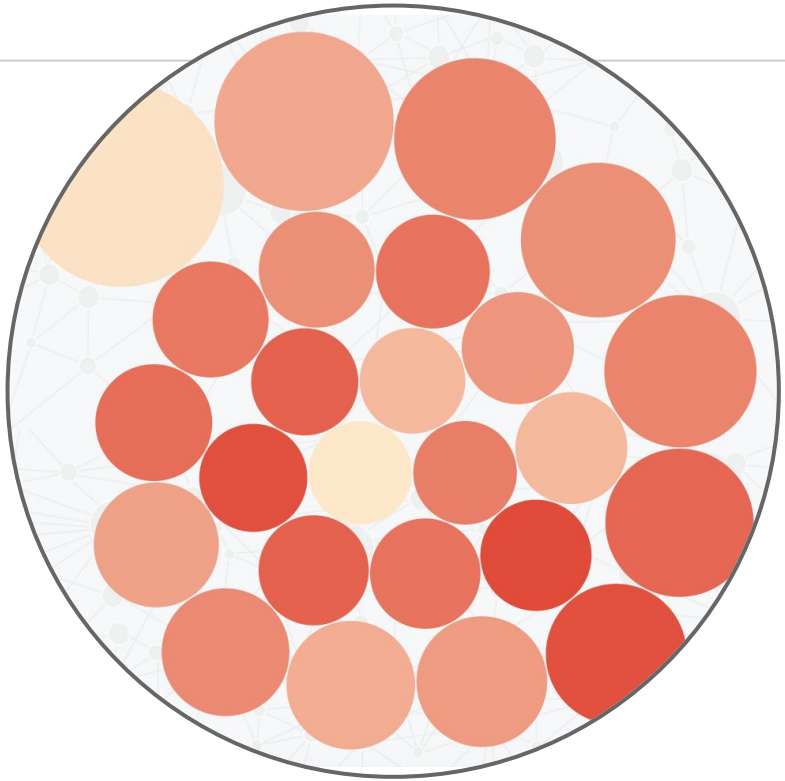
Let’s see it in action.





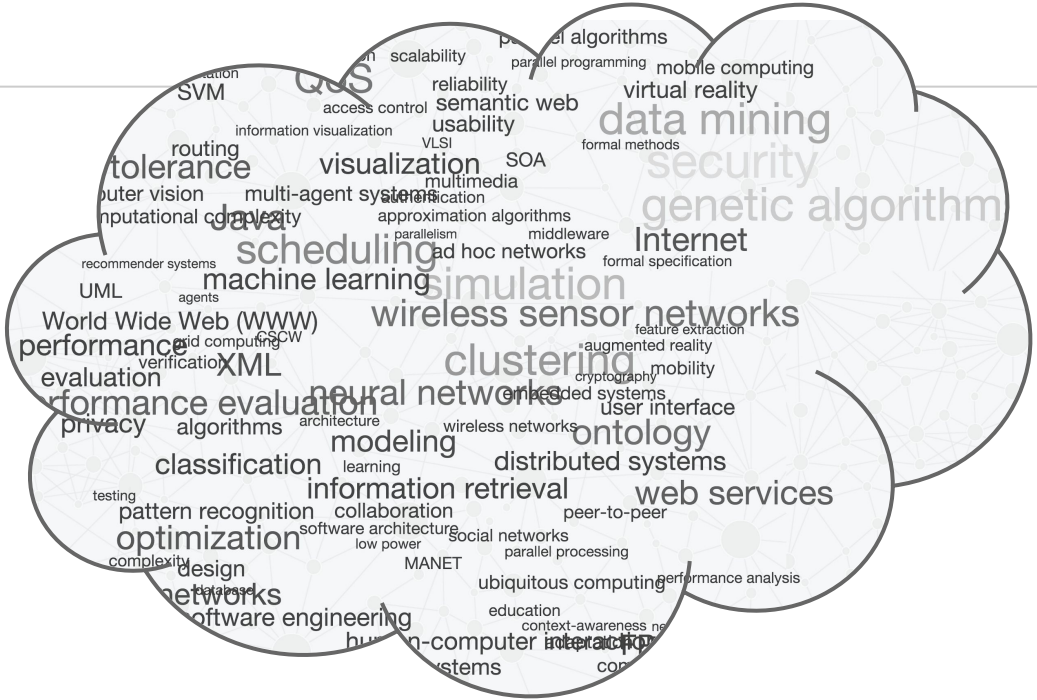
A Paper Community Cluster

- Nodes: Papers
- Colors: Clusters
- Edges: References



Bubble Chart

- Bubble
- Color
- Size
- Position



Word Cloud

- Commonly used words are larger and slightly faded in color.
- Less common words are smaller and darker.

Thanks!

Any questions ?

Source Code available
@
https://github.com/mayanez/Scholarly