Community Detection Implementation on GitHub

Mehmet Furkan Sahin
Oguz Demir
Ilteris Tabak
Suleyman Ozulku

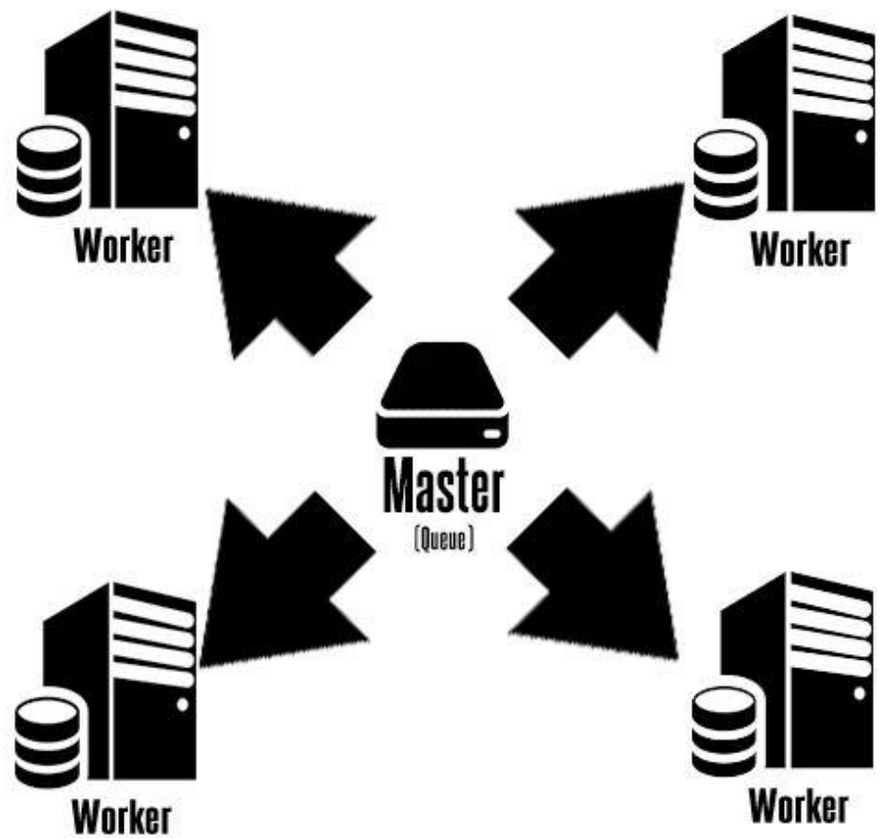
Social Networks are everywhere!

- Real world networks
- Communities
-  is ❤️
- Coder networks

Data

- Github API is 💔
  - 5000 reqs/hour
  - Connection is expensive
-  DB

- BFS Walker
- Distributed manner
- Open for anyone
- 80K people data
  - {login, company, followers, followings, languages, organizations}



Algorithm

- Real world networks are sparse
- `Walktrap`
- Computing communities in large networks using random walks, Pascal Pons, Matthieu Latapy, 2005.
- O(M*N^2) run time
- For 80k nodes, 600k edges ~more than 1 day

Algorithm Details

- Short walks stay in the community `[3,6]: t`
- Probability of being in node `k` for a node `i` in `t` steps is

$$P_{ik}^{t}$$

- `P`: weight matrix

Algorithm Details

- `r` distance [Euclidean distance]:

- Similarity of character of two vertices `<i,j>` in the graph

- Probability of being in node `k` in `t` steps for each of them

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik}^{t} - P_{jk}^{t})^2}{d(k)}}$$

Algorithm Details

```
for n-1 :
choose two most similar communities C1 and C2 based on
r distance (Ward's method)

merge these two communities into a new community C3 =
C1 ∪ C2

update the distances between communities
```

Algorithm Details

- In each iteration;

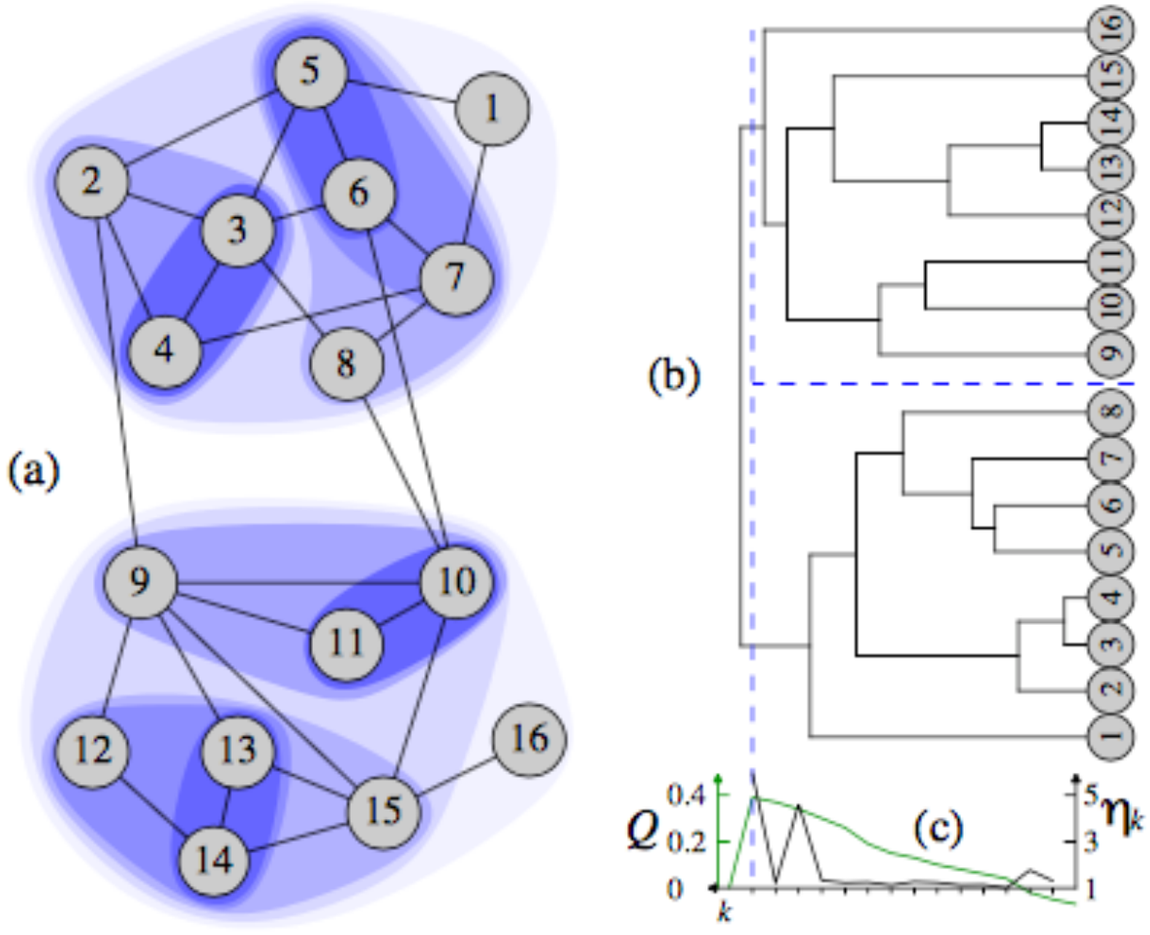
$$\sigma_k = \frac{1}{n} \sum_{C \in \mathcal{P}_k} \sum_{i \in C} r_{iC}^2$$

- Maximize;

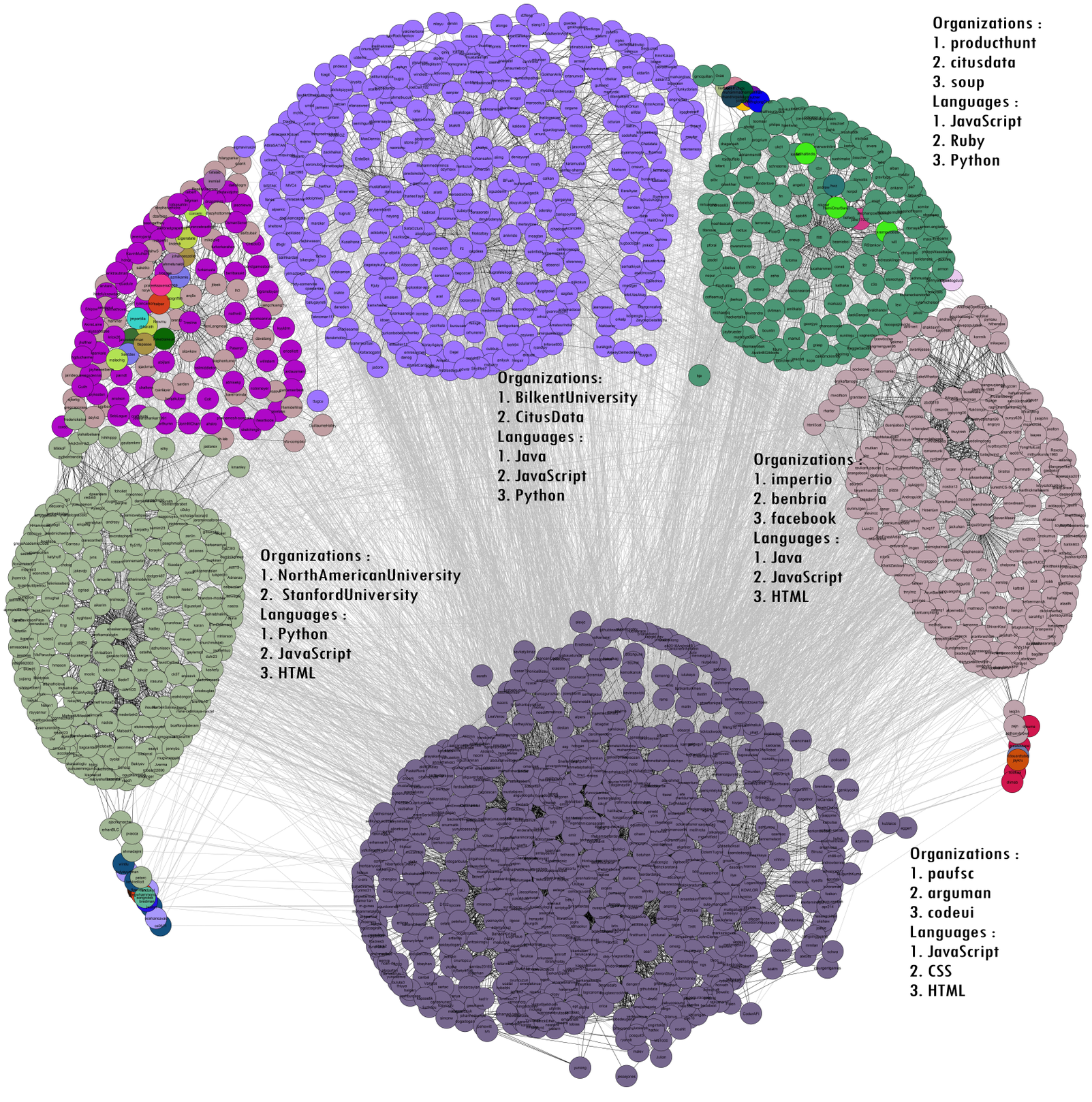
$$\eta_k = \frac{\Delta \sigma_k}{\Delta \sigma_{k-1}} = \frac{\sigma_{k+1} - \sigma_k}{\sigma_k - \sigma_{k-1}}$$

Algorithm Details

- Output is the communities in dendrogram
- maximum η_k is reached with 2 communities



Output



Organizations :
1. producthunt
2. citusdata
3. soup
Languages :
1. JavaScript
2. Ruby
3. Python

Organizations:
1. BilkentUniversity
2. CitusData
Languages :
1. Java
2. JavaScript
3. Python

Organizations :
1. impertio
2. benbria
3. facebook
Languages :
1. Java
2. JavaScript
3. HTML

Organizations :
1. NorthAmericanUniversity
2. StanfordUniversity
Languages :
1. Python
2. JavaScript
3. HTML

Organizations :
1. paufsc
2. arguman
3. codeui
Languages :
1. JavaScript
2. CSS
3. HTML

Future Work

- Dimension increase on relations
  - Repository, issues, stars, etc.
- More analysis with new dimensions

Thank you.