Hierarchical Graph Clustering by Node Pair Sampling

Thomas Bonald* · Bertrand Charpentier* · Alexis Galland† · Alexandre Hollocou†

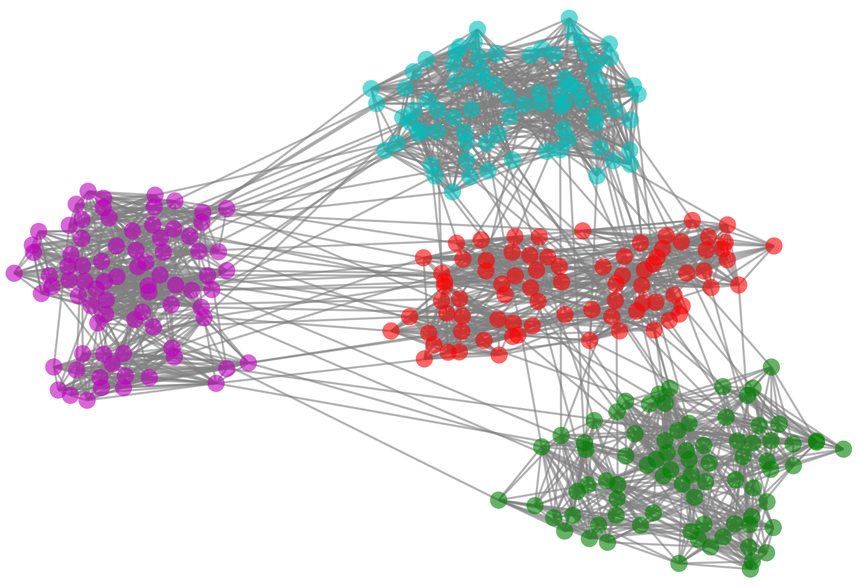
*Telecom ParisTech, France †Inria, France

Abstract

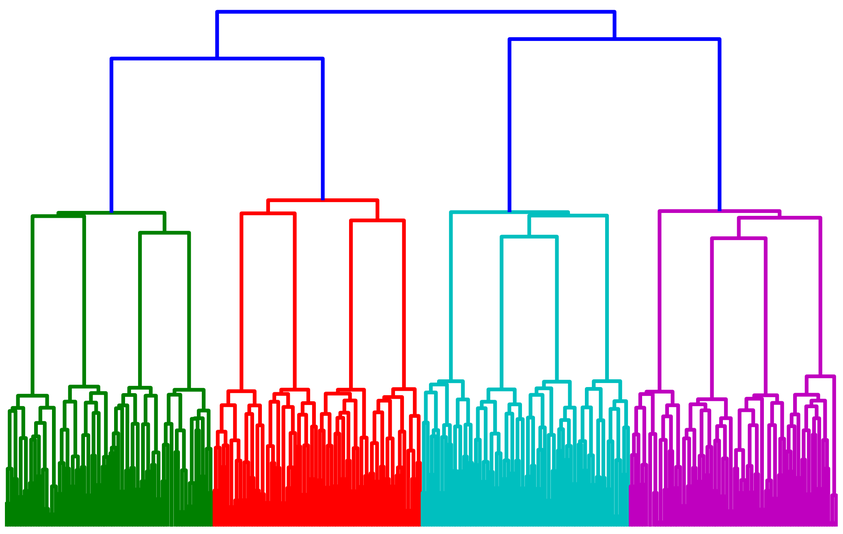
We present a novel hierarchical graph clustering algorithm inspired by modularity-based clustering techniques. The algorithm relies on the linkage induced by the probability of sampling node pairs. The output of the algorithm is a dendrogram, which reveals the multi-scale structure of the graph.

Motivation

Many graphs have a multi-scale structure:



This structure can be represented by a dendrogram:



(a) How to find a "good" hierarchical clustering?
(b) How to assess the quality of this clustering?

Node pair sampling

Graph with weighted adjacency matrix A
Node pair sampling:

$$p(i,j) = \frac{A_{ij}}{w} \quad \text{with } w = \sum_{i,j} A_{ij}$$

Node sampling:

$$p(i) = \sum_j p(i,j) = \frac{w_i}{w} \quad \text{with } w_i = \sum_j A_{ij}$$

Modularity and resolution

Modularity of clustering C at resolution γ:

$$Q_\gamma(C) = \sum_{i,j} (p(i,j) - \gamma p(i)p(j))\, \delta(C)_{ij}$$

Maximum resolution and linkage

Maximum resolution:

$$\gamma^+ = \max_{i,j} \frac{p(i,j)}{p(i)p(j)}$$

Linkage:

$$\sigma(i,j) = \frac{p(i,j)}{p(i)p(j)} = \frac{p(i|j)}{p(i)}$$

Update formula

$$\sigma(i \cup j, k) = \frac{p(i)}{p(i)+p(j)}\sigma(i,k) + \frac{p(j)}{p(i)+p(j)}\sigma(j,k)$$

Agglomerative algorithm

Repeat:

- Merge nodes $\textbf{arg max}_{i,j}\, \sigma(i,j)$
- Update σ

The search of the node pair to merge relies on the chain of nearest neighbors.

Experiments

Hierarchical clustering of the graph of Wikipedia for schools (subgraph of Wikipedia, with 4,589 articles and 106,644 links)

Top-10 clusters (among 100 clusters)

#	Size	Top articles
1	288	Scientific classification, Animal, Chordate, Bird
2	231	Iron, Oxygen, Electron, Hydrogen
3	196	England, Wales, Elizabeth II of the United Kingdom
4	164	Physics, Mathematics, Science, Albert Einstein
5	148	Portugal, Ethiopia, Mozambique,Madagascar
6	139	Washington, D.C., President of the United States
7	129	Earth, Sun, Astronomy, Star, Gravitation
8	127	Plant, Fruit, Sugar, Tea, Flower
9	104	Internet, Computer, Mass media
10	99	Jamaica, The Beatles, Hip hop music, Jazz, Piano

Top-10 subclusters of cluster #1

#	Size	Top pages
1	71	Dinosaur, Fossil, Reptile, Cretaceous, Jurassic
2	51	Binomial nomenclature, Bird, Insect
3	24	Mammal, Lion, Cheetah, Giraffe
4	22	Animal, Ant, Arthropod, Spider, Bee
5	18	Dog, Bat, Vampire, George Byron
6	16	Eagle, Golden Eagle, Bird of prey
7	16	Chordate, Parrot, Gull, Surtsey, Herring Gull
8	15	Feather, Extinct birds, Mount Rushmore, Cormorant
9	13	Miocene, Eocene,Beaver, Radhanite
10	12	Crow, Dove, Pigeon, Rock Pigeon

Bibliography

[1] Lambiotte, Delvenne & Barahona (2014)
Random walks, Markov processes and the multiscale modular organization of complex networks
IEEE Transactions on Network Science and Engineering