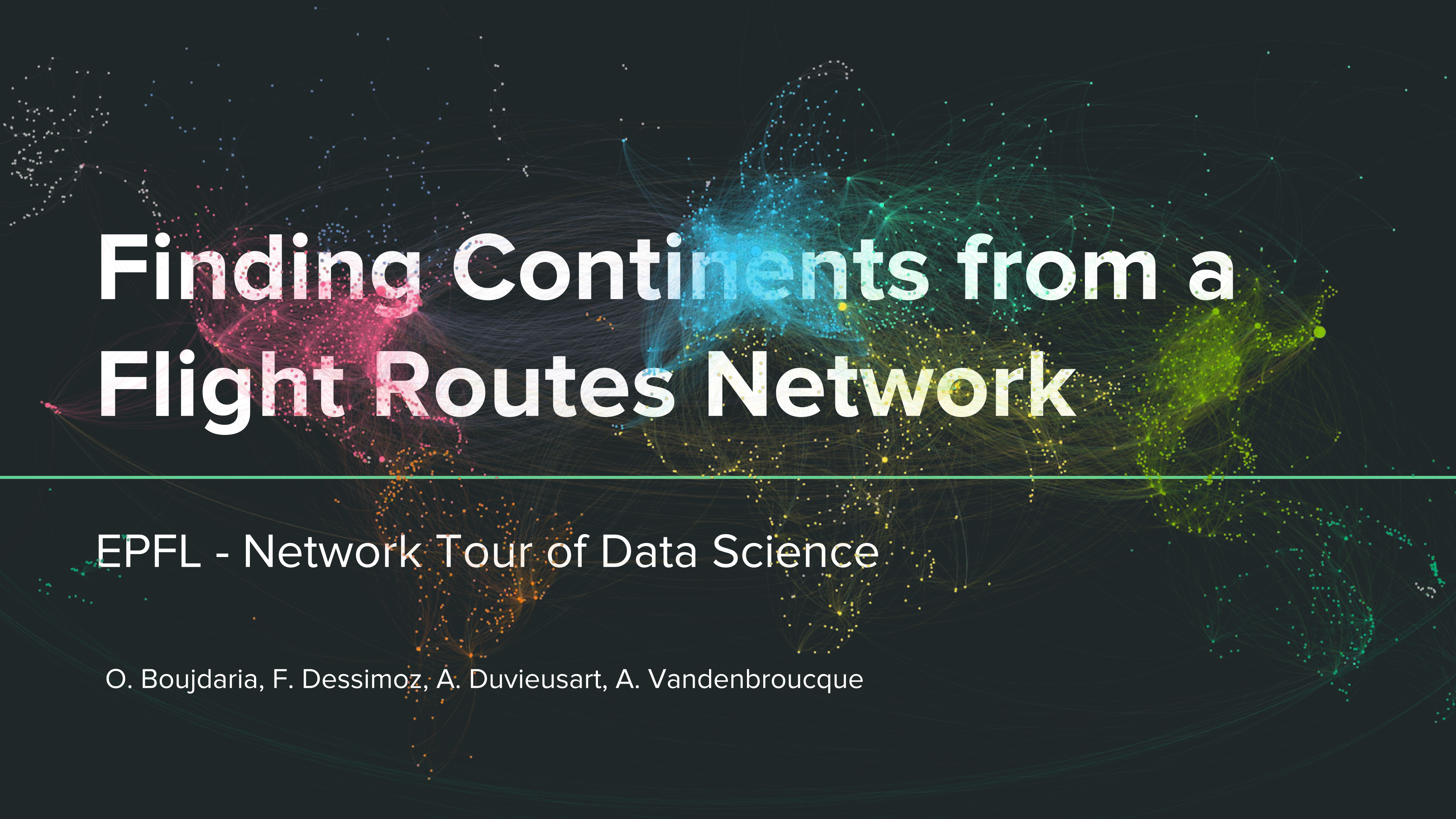


Finding Continents from a Flight Routes Network

EPFL - Network Tour of Data Science

O. Boujdaria, F. Dessimoz, A. Duvieusart, A. Vandenbroucque

Can we find continents from a graph of flight routes?

or more formally ...

Do continents form communities in the network?

Outline

I - Graph Properties and Creation

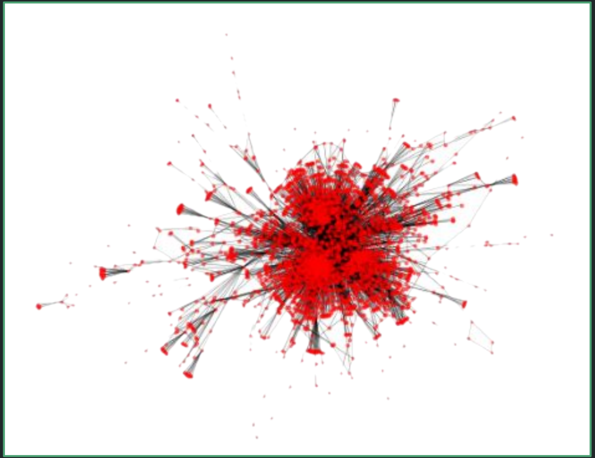
II - Community Detection

a. Spectral Clustering
b. Girvan-Newman Algorithm
c. Greedy Modularity Maximization
d. Louvain Algorithm

III - Comparison

IV - Conclusion

Creation of the graph

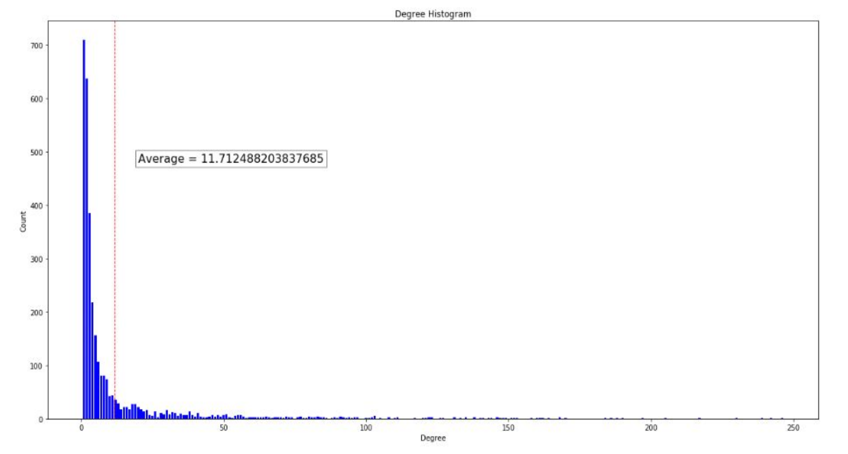


Number of Edges	67,663
Number of Nodes	3,321

- Represents airports and flights from different airlines
- Merged with dataset of airport locations
- Retrieved largest connected component for the rest of the project

Properties of the graph

- Unweighted
- Undirected



Graph Density	0.3%
Average Clustering Coefficient	0.49
Diameter	12

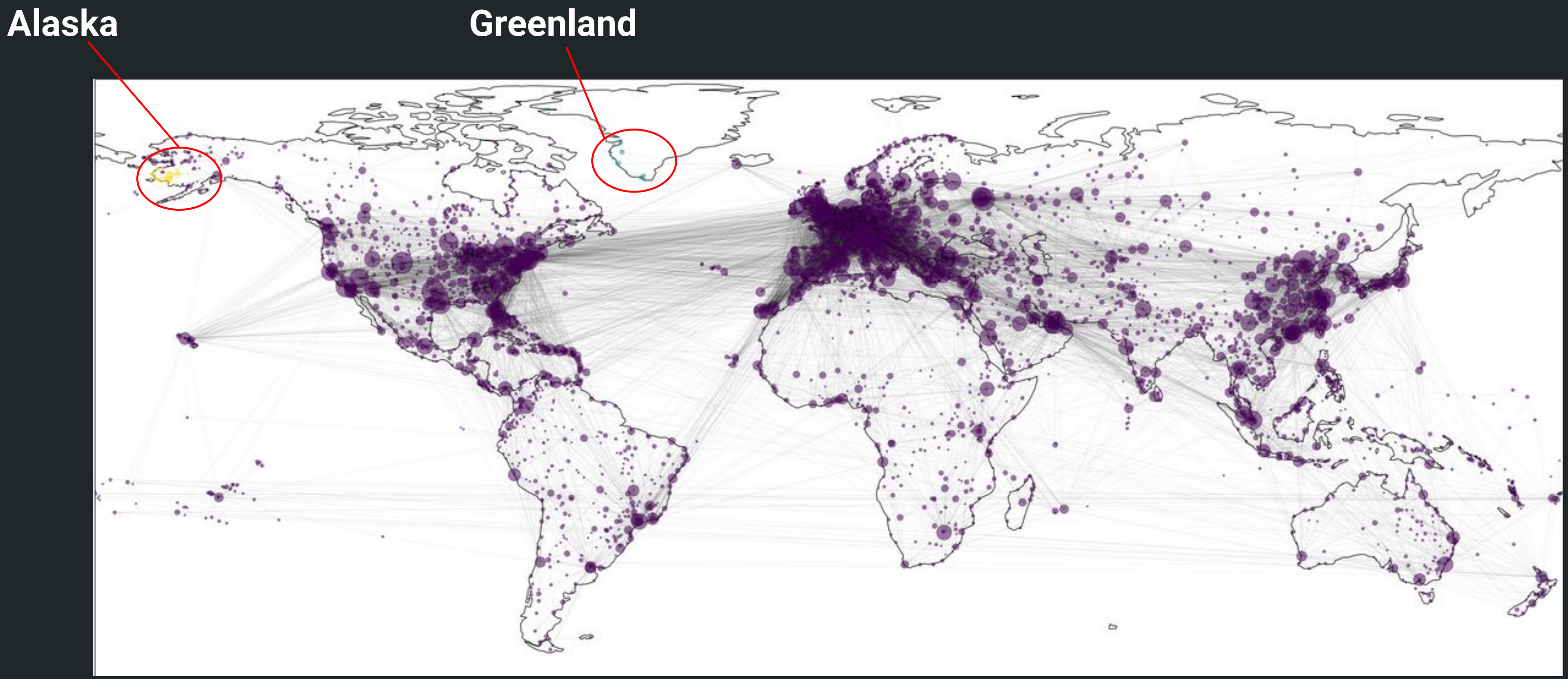
Community Detection

Spectral Clustering

$$RatioCut(A_1, A_2, ...A_k) = \sum_{i=1}^{k} \frac{Cut(A_i, \bar{A}_i)}{|A_i|}$$

- Relaxed formulation of the RatioCut optimization problem
- Need to compute the top *k* eigenvectors of the graph Laplacian
- This forms an embedding in *k* dimensions for the nodes
- Use K-Means algorithm on the embeddings to find cluster assignments

Results for Spectral Clustering



Results for Spectral Clustering

- Detects small clusters (Alaska, Greenland)

- The method works well when there are *clear* clusters in the graph

- Results make sense since for example, Europe and America are not distinct communities since there many inter-connections

Girvan-Newman Algorithm

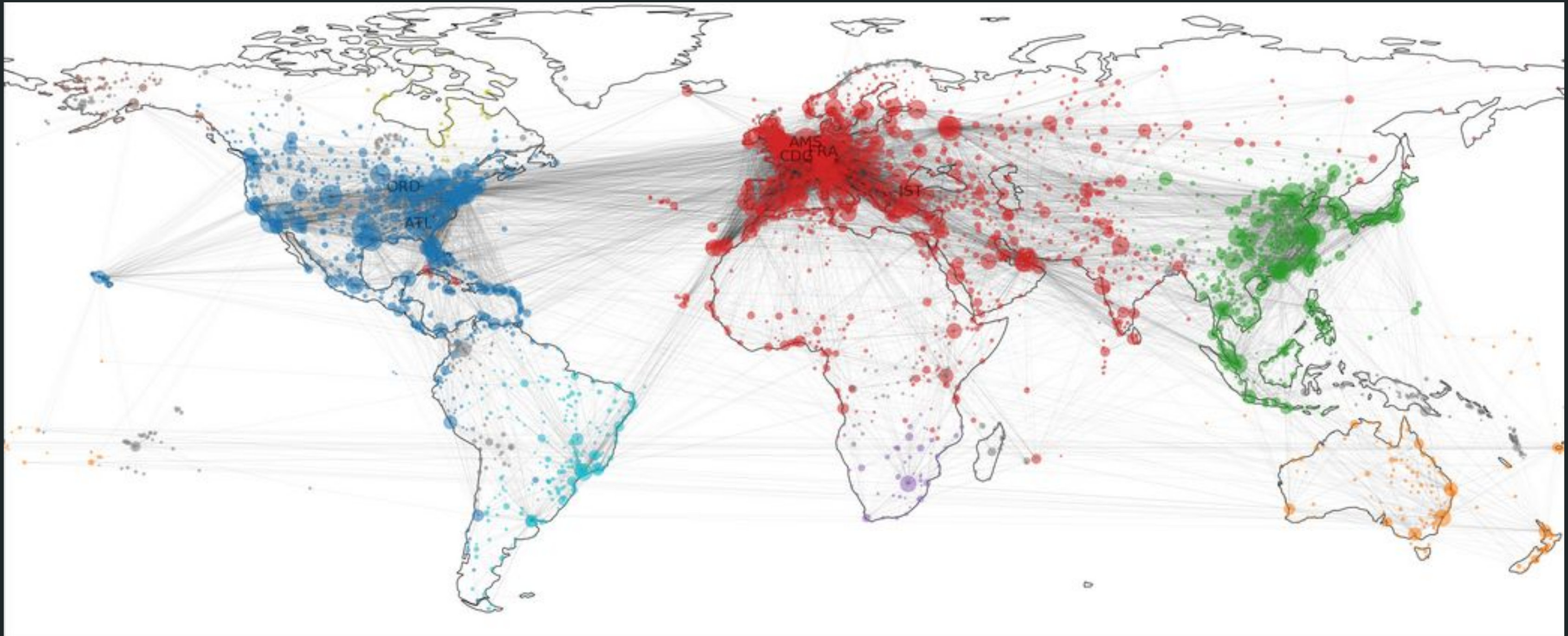
- Idea : edges appearing in many shortest paths are inter-community edges

- Compute shortest paths between all pairs of nodes and label each edge with the number of shortest paths they are a part of (i.e. their betweenness)

- Remove the edge with highest betweenness centrality and iterate until we get 2 separate graphs

- Computationally costly → use a randomized version by sampling edges

Results Girvan-Newman



Results Girvan-Newman

- This method detects continents well but also detects many small communities in the process

- Running time is really high, but we can get a 2x speed-up with the randomized version

Modularity Maximization

Modularity

- The goal is to have a measure of quality for the partition of a network into communities

$$Q = \frac{1}{2|E|} \sum_{i,j \in V} \left(A_{ij} - \frac{d_i d_j}{2|E|} \right) \delta_{C_iCj}$$

Greedy Modularity Maximization Algorithm

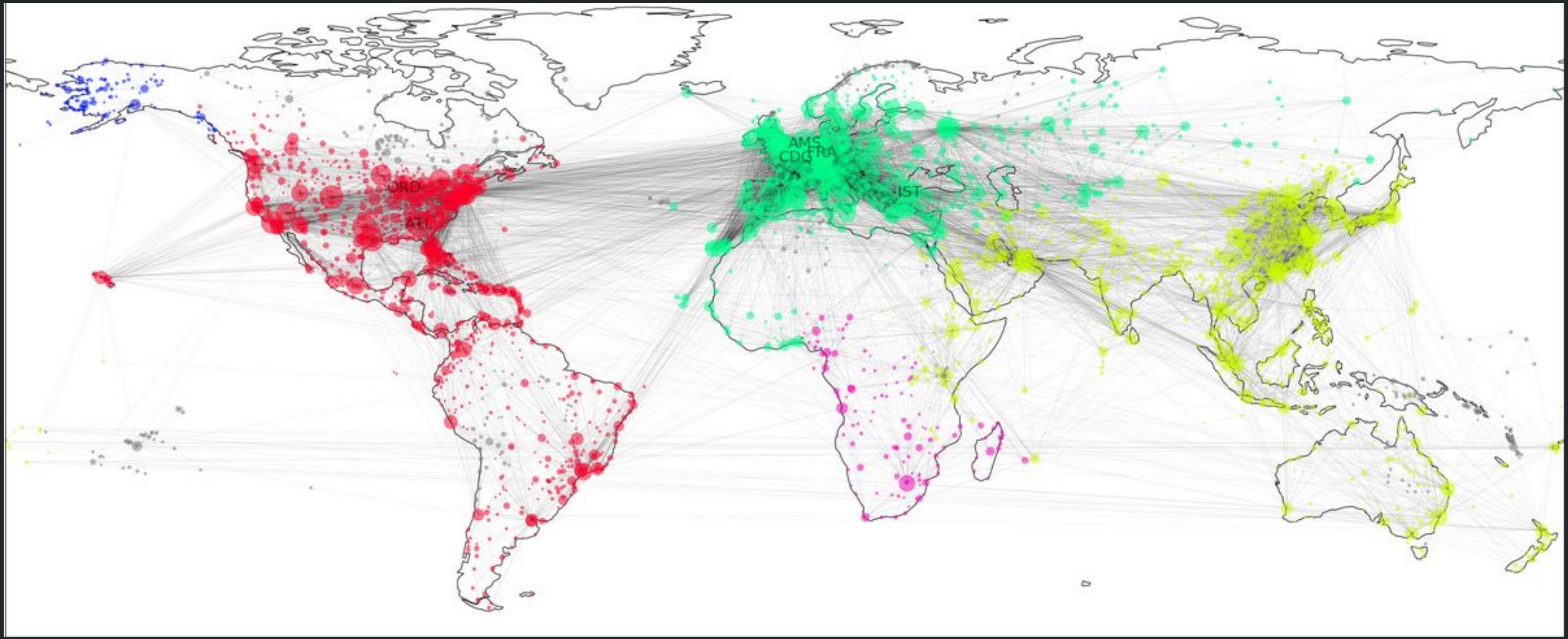
- Start with every node as a community

- At each step, find the pair of community that gives the highest gain in modularity when merged together

- Repeat until there is only one community left

- Return the partition of node into communities that gives the maximum modularity

Results for Greedy Modularity Maximization



Results for Greedy Modularity Maximization

- Can detect most continents, and rediscover communities found by Spectral Clustering

- Running time is a lot better than Girvan-Newman

- Greedy approach works well, but can we do better?

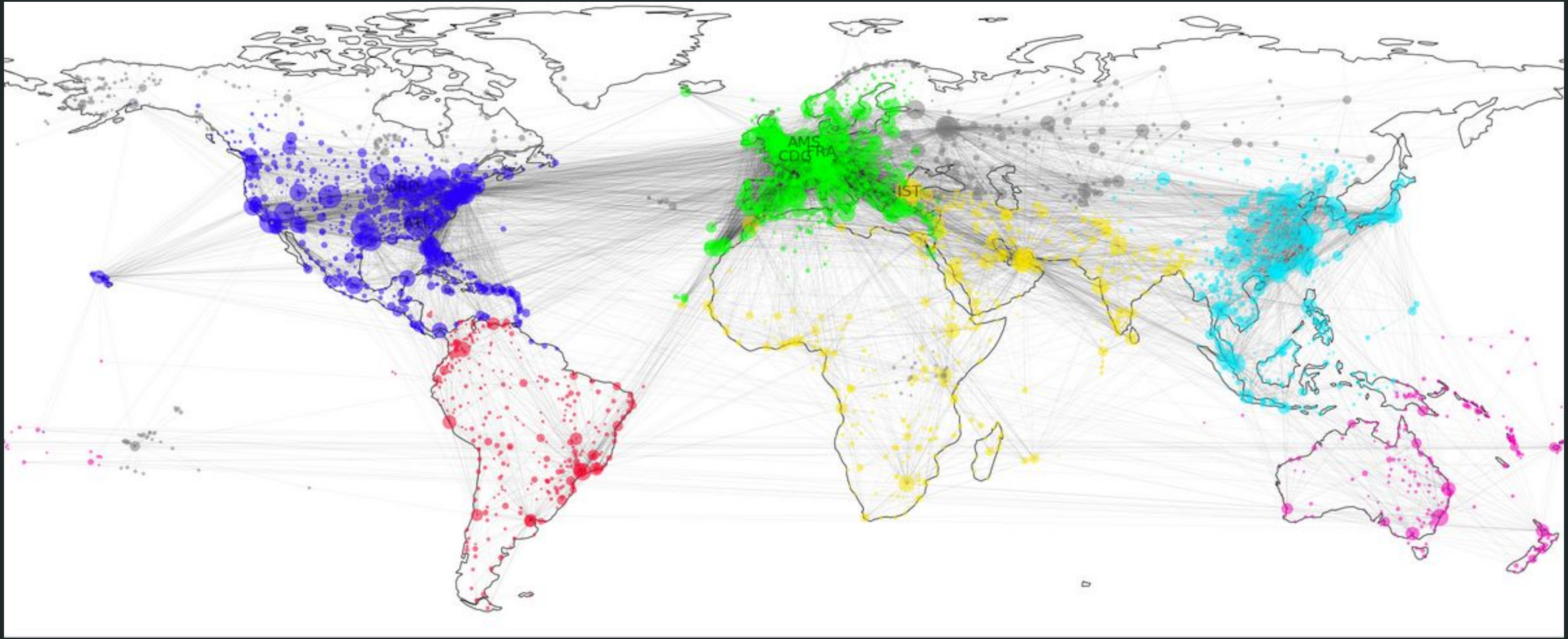
Louvain Algorithm

- **Idea : Start with each node as a community and iterate over the following two steps:**

1. **Iterates on each node in the network, removes it, and compute the change in the modularity if we place this node in the community of one of its neighbor**

2. **Construct a coarse grained network with the communities found in the first step, i.e treat each community found in the previous step as a new node.**

Results for Louvain

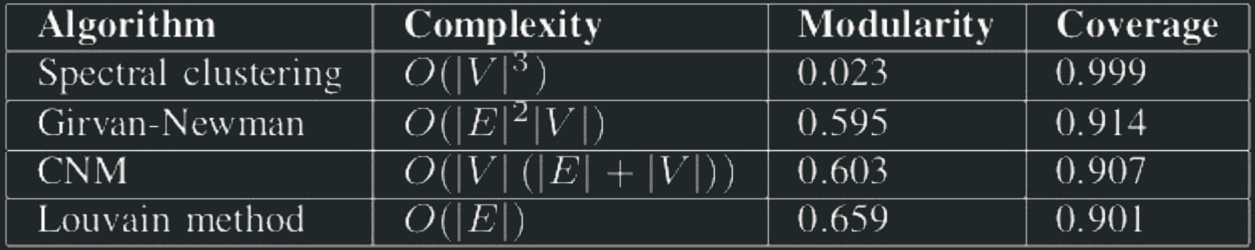


Results for Louvain

- The 6 main communities represent continents very well
- Slightly different from Greedy Modularity (see Asia and Oceania)
- Running time is a lot better than previous algorithms

Comparison

- Use two metrics to compare models: *Modularity* and *Coverage*
- The *coverage* of a partition is the ratio of the number of intra-community edges to the total number of edges in the graph



Algorithm	Complexity	Modularity	Coverage
Spectral clustering	$O(\|V\|^3)$	0.023	0.999
Girvan-Newman	$O(\|E\|^2\|V\|)$	0.595	0.914
CNM	$O(\|V\|\,(\|E\|+\|V\|))$	0.603	0.907
Louvain method	$O(\|E\|)$	0.659	0.901

Conclusion

- Many algorithms exist for community detection
- Their results depend on the graph structure
- Detecting continents was indeed possible, discovering also smaller structures at the same time
- Speeding up algorithms becomes important (for large scale networks)
- Community detection is becoming more and more important with large networks available today

References

- U. V. Luxburg, “*A tutorial on spectral clustering,*” Stat. Comput. , pp. 395–416, 2007

- M. Girvan and M. E. J. Newman, “*Community structure in social and biological networks,*” ArXiv.org pp. 2–3, 2001

- F. Botta and C. I. del Genio, “*Finding network communities using modularity density,*” ArXiv.org, pp. 1–3, 2016

- S. Papadopoulos, “*Community Detection in Social Media,*” Data Mining and Knowledge Discovery, Springer , pp 515–554 , 2012