Clustering on wikipedia graphs
PGM 2024/2024

Melvin Gode*, Andrej Perovic* and Antoine Sicard*

Ecole Normale Supérieure Paris-Saclay, *all these authors contributed equally to this research

Introduction

Wikipedia, an online collaborative encyclopedia

• **Wikipedia** is one of the most visited websites in the word, with more 700 millions visits in 2022 and hosts, in its English version, over **7 million articles** on various topics ranging including cultures, art, geography, society and sciences *(Wikipedia 2024)*

• Highly valuable for education, research, documentation *(Head and Eisenberg 2010; Xiao and Askin 2012)*

• All articles on Wikipedia are **interconnected** through hyperlinks, forming a **highly complex graph structure** where nodes represent articles, and edges represent hyperlinks.

Wikipedia galaxy
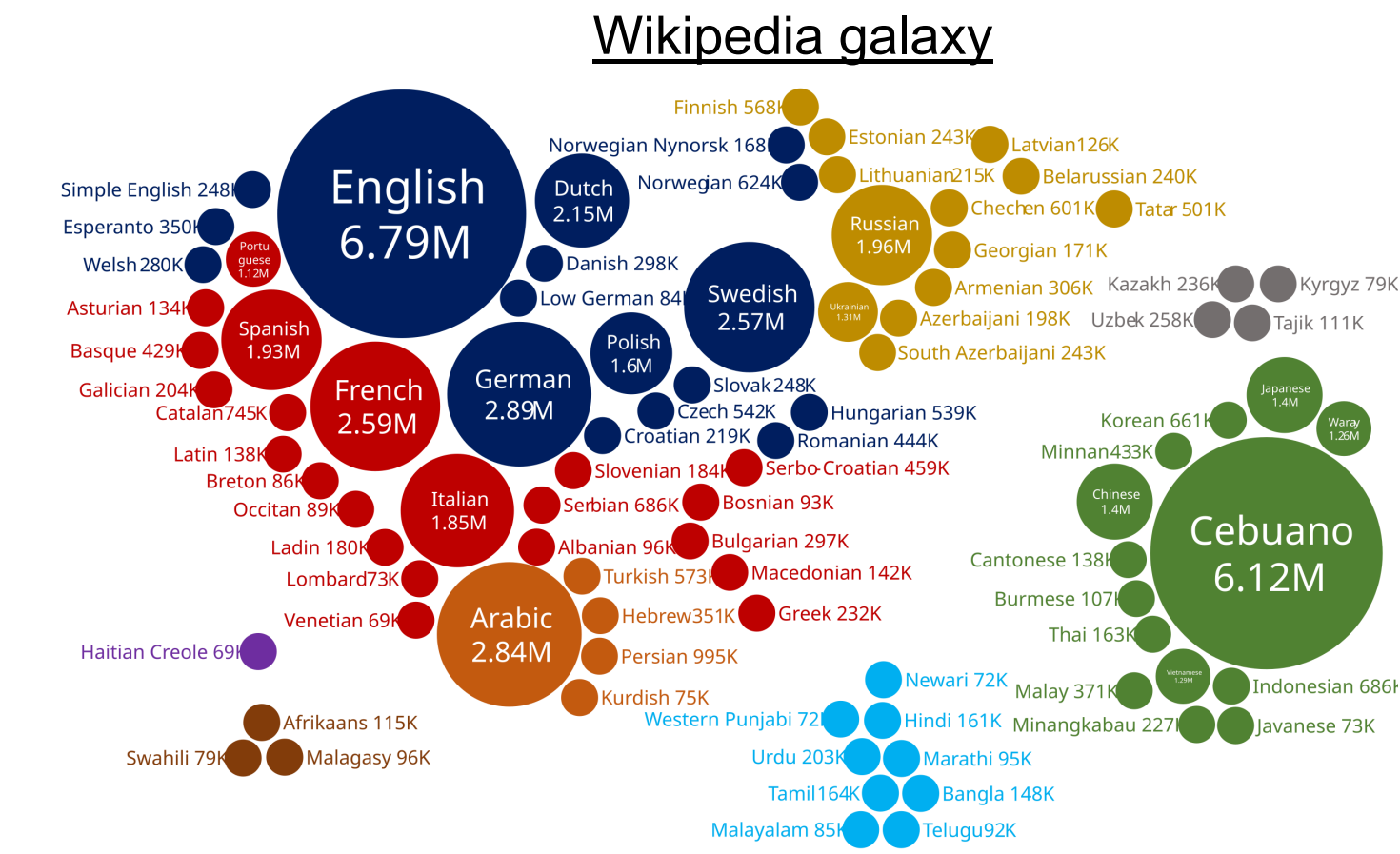
Clustering methods on graphs

• The article **"Graph clustering"** by *Schaeffer (2007)* provides an overview of methods for graph clustering.
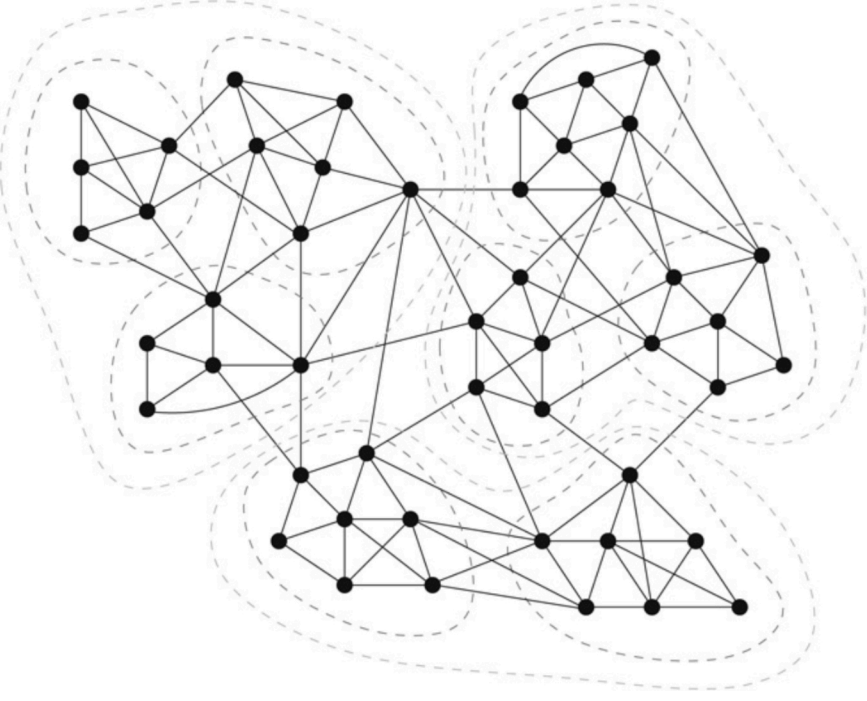
• Local clustering methods

• Global clustering methods:

- **Divisive clustering** (edges are progressively removed to form clusters, e.g. spectral methods, betweenness, voltage methods, random walks)

- **Agglomerative clustering** (clusters are proggressively merged into bigger clusters)

• Quality measures: **modularity**

$$\mathcal{M} = \sum_{c=1}^{K} \left[\frac{L_c}{m} - \left(\frac{k_c}{2m}\right)^2 \right]$$

L_c : number of edges in cluster c
k_c : total node-degree in cluster c
m : total number of edges in graph G
K : number of clusters

Objective: analyze and better understand this complexity enlightening the relationships and connections within **Wikipedia dense network** adapting **several graph clustering methods** developed in *Schaeffer, 2007*

Developed approach and results

Analyze a dataset published by Stanford *(https://snap.stanford.edu/data/wiki-topcats.html)*, which represents the largest connected component of English-language Wikipedia articles.

1. Top-down cluster division

1.1. Divisive clustering using edge betweenness

• **Edge betweenness**: fraction of shortest paths passing through this edge between all pairs of nodes

• **Principle**: remove iterately the **edges with highest betweenness** to create clusters

$$c_B(e) = \sum_{s,t \in V} \frac{\sigma(s,t \mid e)}{\sigma(s,t)} = \sum_{s,t \in V} \delta(s,t \mid e) = \sum_{s \in V} \delta(s \mid e)$$

L_c : number of edges in cluster c
k_c : total node-degree in cluster c
m : total number of edges in graph G
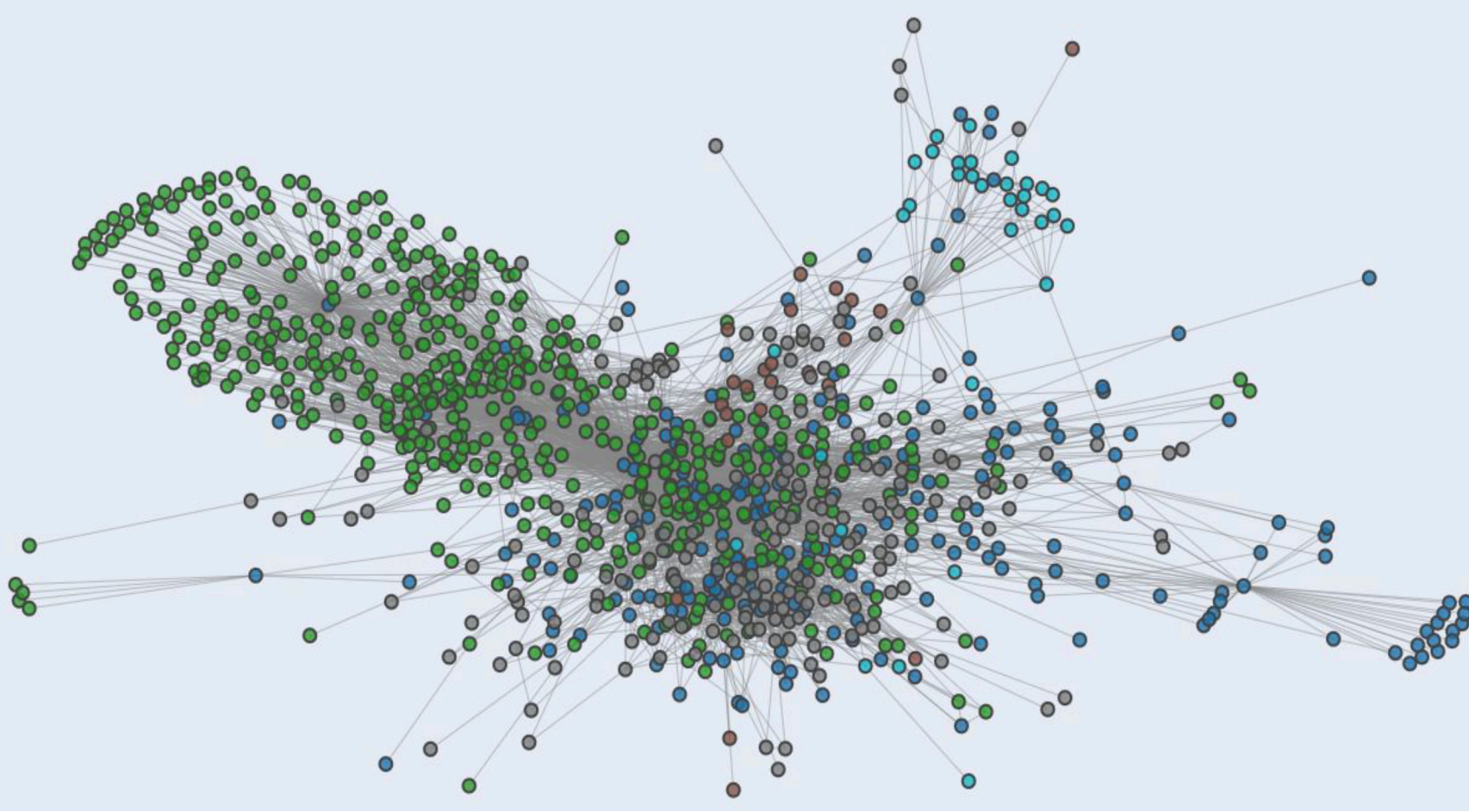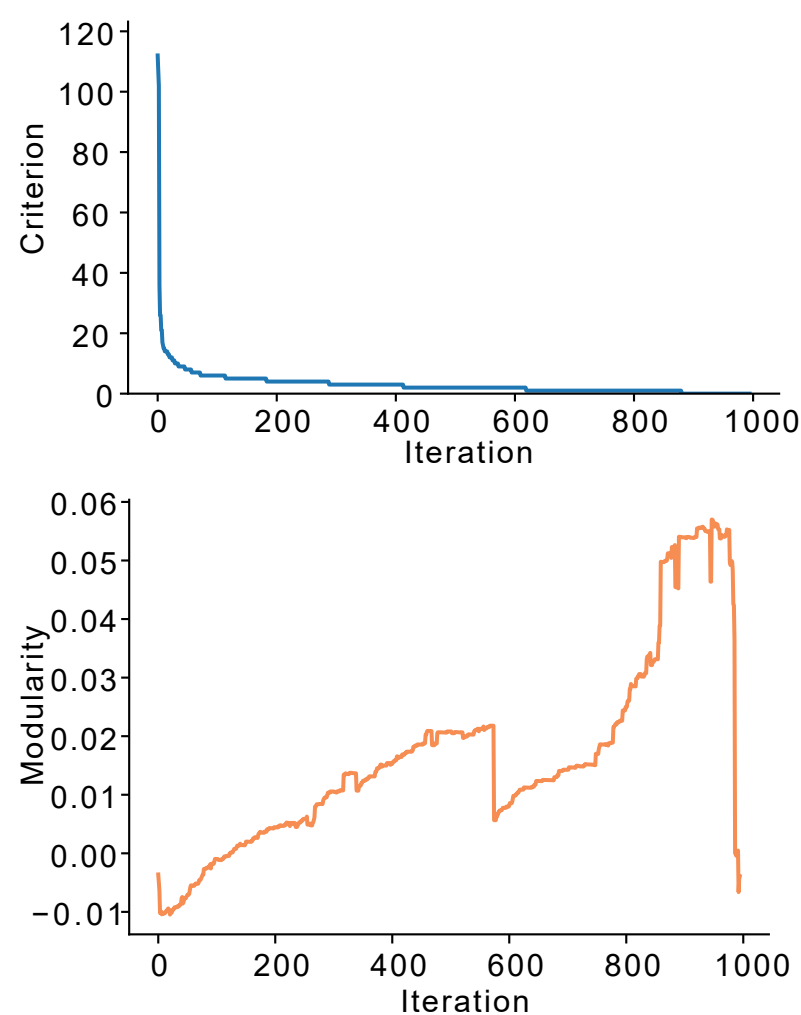K : number of clusters

Two approaches to compute the betweenness:
• Implementation of a "brut-force" method (complexity $O(n^3)$)
• Implementation **based on Brandes, 2001** (complexity $O(nm)$), based on dependency accumulation

→ Limitation to a subset of 1,000 nodes

When to stop clustering?

Optimal clustering result
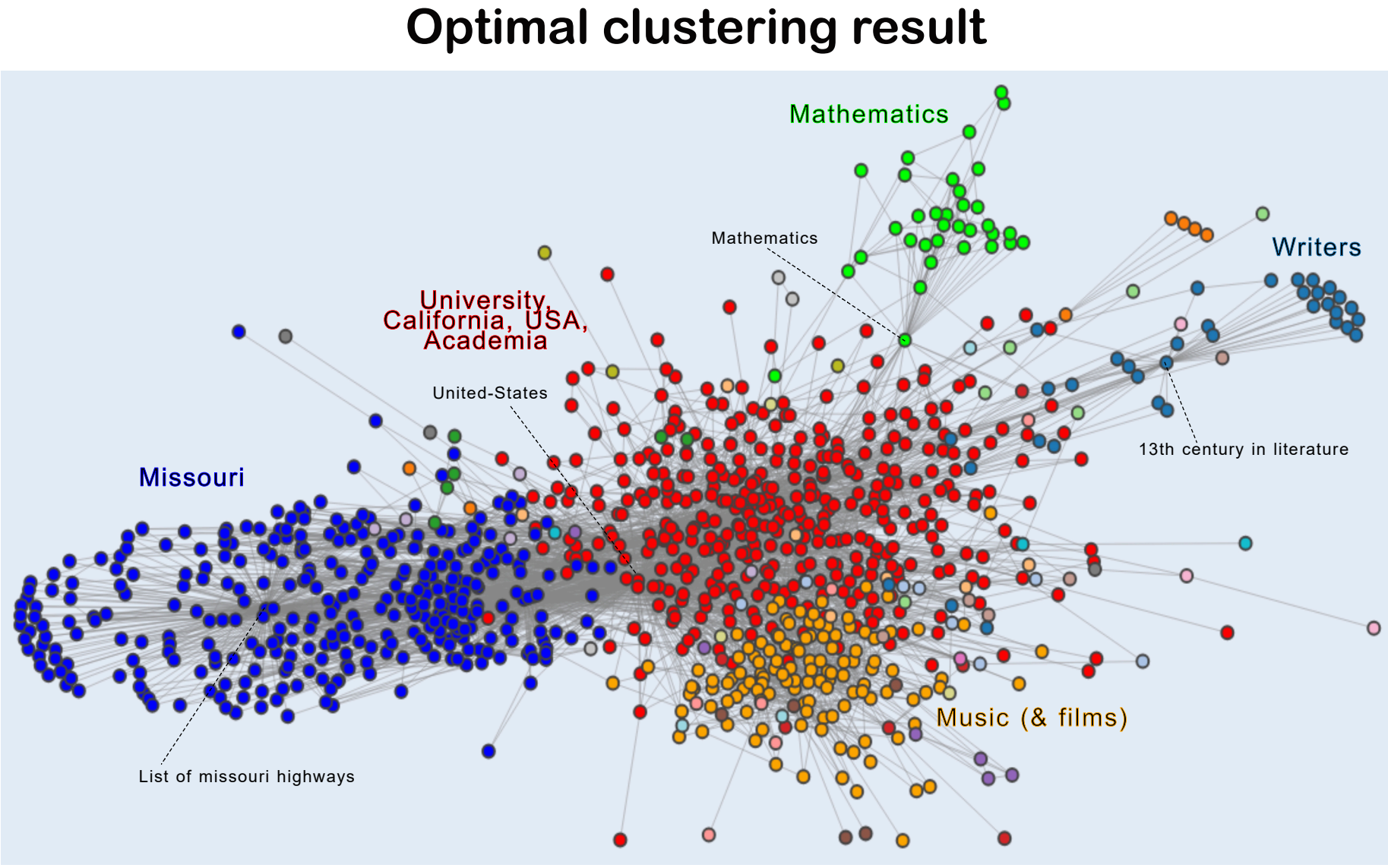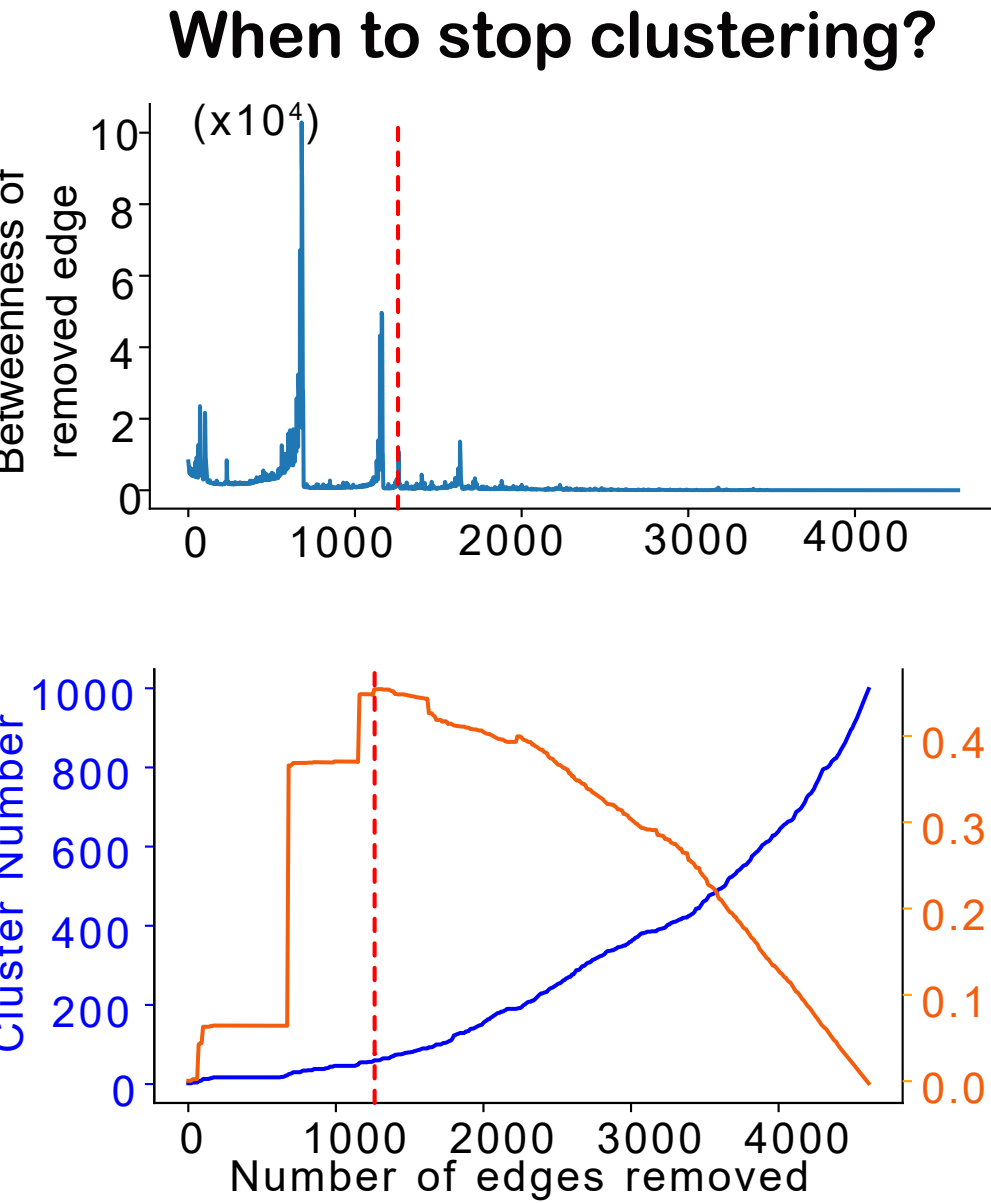
1.2. Divisive clustering using random walks

→ Both top-down cluster division methods give **meaninful clusters**

2. Bottom-up cluster aggregation

→

Conclusion and perspectives

• **First opto-chemical tool** to **selectively** target NMDARs containing **two GluN2B subunits** (2B-diheteromers).

• **Decrease** of **synaptic 2B-dihet.** expression with age (GluN2B→GluN2A switch).

→ Next: *in vivo* implementation to address currently **debated physiological roles** of 2B-dihets and investigate their **therapeutic potential**.

Acknowledgements and references

1. Paoletti, P., Bellone, C. & Zhou, Q. *Nat Rev Neurosci* 14, 383-400 (2013).
2. von Engelhardt, J. et al. Neuron 60, 846-60 (2008).
3. Zhou, Q. & Sheng, M. Neuropharmacology 74, 69-75 (2013).
4. Kramer, R.H., Mourot, A. & Adesnik, H. *Nat Neurosci* 16, 816-23 (2013).
5. Mony, L., Zhu, S., Carvalho, S. & Paoletti, P. *EMBO J* 30, 3134-46 (2011).
6. Berlin, S. et al. Elife 5(2016).
7. Canales, A., et al. *Acc Chem Res* 51, 829-838 (2018).