Network community detection on Wikipedia

Melvin Gode, Antoine Sicard and Andrej Perkovic

Ecole Normale Supérieure Paris-Saclay, Master MVA, PGM class

Introduction

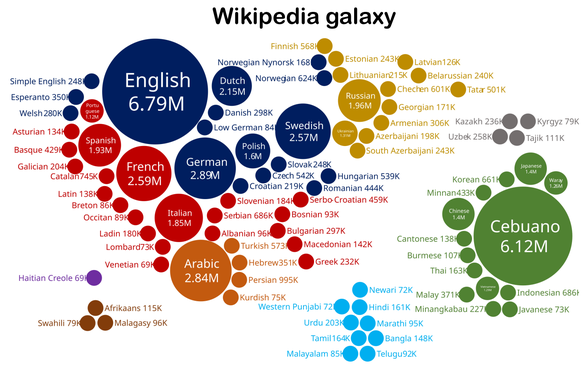
Wikipedia, an online collaborative encyclopedia

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

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

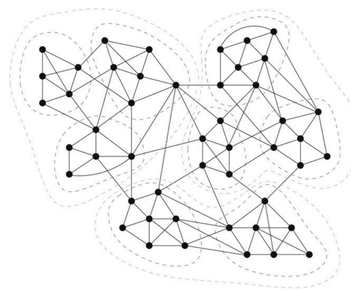
• All articles on Wikipedia are **interconnected** through hyperlinks, forming a **highly complex graph structure** where nodes represent articles, and edges represent hyperlinks

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 progressively merged into bigger clusters)

• Quality measures: **modularity** *(Newman 2011)*

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

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

Objective: implement and compare **different graph clustering approaches** and algorithms from *Schaeffer 2007* to extract insight in the **structure of Wikipedia** through **interpretable community detection**

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
→ Limitation to a **subset of 1,000 nodes and 4,628 edges**

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 iteratively the **edges with highest betweenness** to create clusters

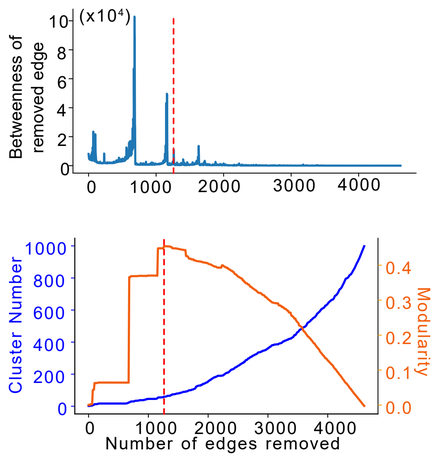
$$c_B(e) = \sum_{s,t \in V} \frac{\sigma(s,t \mid e)}{\sigma(s,t)} = \sum_{s,t \in V} \delta(s,t \mid e)$$

σ(s, t): number of shortest paths between nodes s and t
σ(s, t | e): number of those paths that contain edge e
δ(s, t | e) = σ(s,t|e)/σ(s,t) represents the pairwise dependencies of s and t on an intermediary edge e

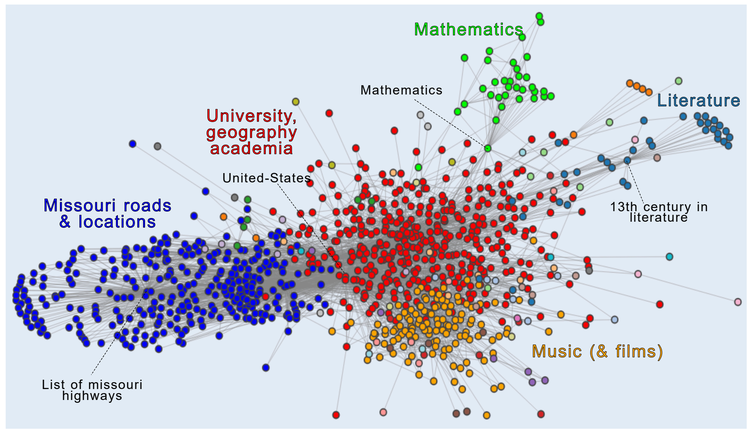
Two approaches to compute the betweenness:
1. Implementation of a first "naive" method (results not shown)
2. Effective implementation **adapted from Brandes, 2001** (result below)

*Complexity limited by a step in O(nodes x edges) → implementation of 2 **hyperparameters**: number of edges to remove at each step; number of sources to consider for betweenness computation*

When to stop clustering?



Clustering result



1.2. Divisive clustering using stochastic simulation

• **Simulations of random walks** by alternating between **expansion** and **inflation**

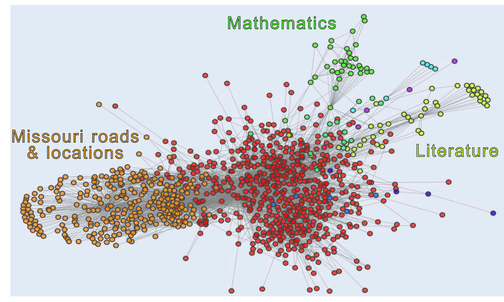
$T_1 = M$ Normalized adjacency

$T_{2i} = \mathrm{Exp}_{e(i)}(T_{2i-1})$ **Expansion**: simulating stochastic flow

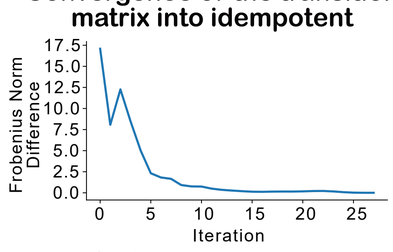
$T_{2i+1} = \Gamma_{r(i)}(T_{2i})$ **Inflation**: strengthens strong connections and diminishes weak ones

$e(i) = 2$ usually $r(i)$ input parameter

Clustering result (highly optimized library)
library: micans.org/mcl

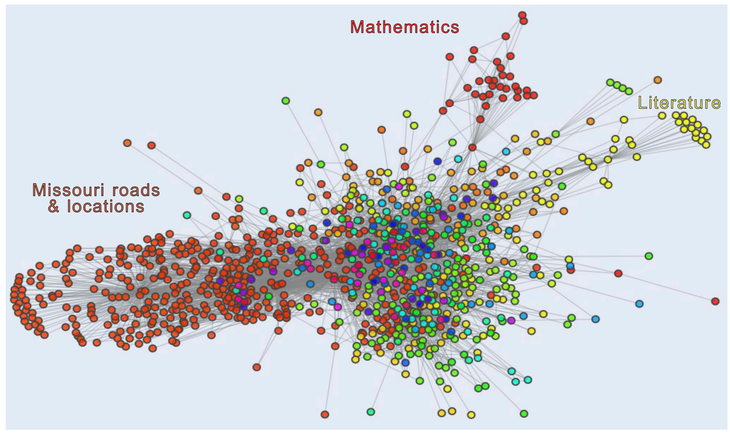


Convergence of the transition matrix into idempotent



• Very **fast**
• Python implementation leads to degenerate cases for a large number of nodes and a sparse adjacency matrix
• Very **fragile** susceptible to floating point precision, numerical instability

Clustering result (our implementation)



→ Both top-down cluster division methods give **meaninful clusters that effectively capture and reflect well-defined topics**

NB: cluster main topics are determined by looking at most frequent words in article names (automated wordclouds + hand analysis)

2. Bottom-up cluster aggregation

• Nodes start in individual clusters which are then **iteratively merged** based on a pairing criterion

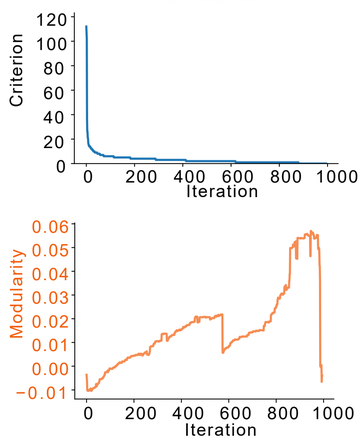
• The **criterion** is outside neighborhood:

$$r(c_i, c_j) = \frac{\langle \Gamma_{\mathrm{out}}(c_i); \Gamma_{\mathrm{out}}(c_j) \rangle}{|c_i| \times |c_j|}$$

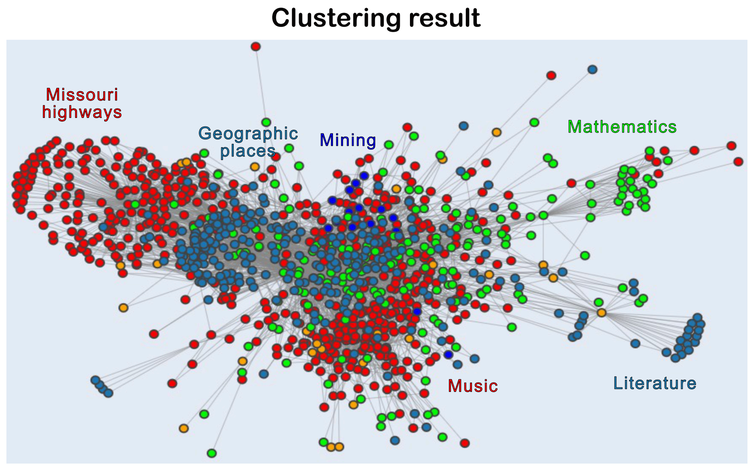
similarity measure; cluster size normalization

with $\Gamma_{out}(c) = \bigcup_{x \in c} \{y | e_{x,y} \in E, y \notin c\}$

Criteria



Clustering result



• **Complexity:** *O(nodes x edges)*

→This algorithm has a **fast runtime** (~5s on the data) but produces **less clearly separated clusters** (associated with lower modularity)

3. Method comparison

• Various approaches with **forces** and **weaknesses**

	Top-down cluster division		Bottom-up cluster aggregation
	Betweenness	Stochastic simulation	
Criterion used	Betweenness of removed edges	Idempotence of stochastic matrix	Outside neighborhood
Modularity	0.43	0.33	0.06
Meaningful clusters?	+++	++ (but +++ with the optimized library)	++
Run time	12 minutes	1.6 s	5 s

Conclusion, discussion and perspectives

• **Meaningful clustering** of **Wikipedia articles**:
- Clusters reflect **well-defined topics**, with only **limited overlap** between them: Mathematics, Literature, Music, Missouri, Geography, etc.
- Clusters with **thematic proximity** exhibit **strong connections**, while those with dissimilar topics are weakly connected
- **Hierarchical organization**: one general article is referencing to more specific articles and out-of-cluster articles mainly refer to this general

• Important faced **drawback**: algorithm **complexity** + **limited computing resource** which prevented us from analyzing the whole dataset
→ We only worked on a graph subset

→ Next: **scale up** our experiments to get a more macroscopic picture of the whole Wikipedia encyclopedia

→ Next: exploit the **advantage of hierarchical clustering** (granularity, dendograms, different scales = valuable semantic insight?)

→ Next: **more complex techniques**, finite mixture models and stochastic block models

References

Main studied article:

- Schaeffer, Satu Elisa. 2007. « Graph Clustering ». *Computer Science Review* 1 (1): 27-64. https://doi.org/10.1016/j.cosrev.2007.05.001.

Other references:

- Brandes, Ulrik. 2001. « A Faster Algorithm for Betweenness Centrality ». *The Journal of Mathematical Sociology* 25 (2): 163-77. https://doi.org/10.1080/0022250X.2001.9990249.

- Head, Alison J., et Michael B. Eisenberg. 2010. « How Today's College Students Use Wikipedia for Course-Related Research ». *First Monday*, février. https://doi.org/10.5210/fm.v15i3.2830.

- Newman, M. E. J. 2011. « Networks: An Introduction ». In , Oxford University Press, 224.

- Xiao, Lu, et Nicole Askin. 2012. « Deliberation in Wikipedia: Rationales in Article Deletion Discussions ». *Proceedings of the American Society for Information Science and Technology* 49 (1): 1-4. https://doi.org/10.1002/meet.14504901234.