Introduction to Community Detection

CS 598: Computational Scientometrics

January 23, 2025

Tandy Warnow

Scientometrics – some questions

- How is the scientific research community organized?
  - What ideas are researchers focusing on?
  - What authors are working on these ideas?
  - Depth vs breadth of impact (influence)
  - Who is driving the research?
- How does the research community change over time?
- What is the nature of interdisciplinary research, and how is it displayed in the literature?

These questions are sometimes investigated using clustering methods applied to large networks, such as citation graphs, authorship graphs, etc.

Rest of today

- High-level comments about clustering methods and community detection
- Discussion of some papers
  - Feng et al. (Survey of community search over big graphs)
  - Fortunato & Barthelemy (Resolution limit in community detection)
  - Von Luxburg et al. (Clustering: Science or Art?)
  - Traag et al. (From Louvain to Leiden, guaranteeing well-connected clusters)
  - Vaca-Ramirez and Peixoto (Evaluating fit of SBM networks to real world networks)
- Brief comments about what our research group has found

Preliminary discussion

- Topics:
  - How do you pick which papers to cite?
  - How do others pick papers?
- What does this tell you about interpretation of citation counts?
  - Who wins?
  - Does more citations mean better research?
  - Can you learn something about “communities” of authors and/or research papers?

Clustering vs Community Detection

- Clustering tends to be similar to “PCA” -- finding clearly separated groups, generally large ones
- Community detection often aims to find smaller groups, perhaps not so well separated from the rest of the graph
- Community detection and clustering can be considered the same problem when the input is just a graph

Clustering methods: issues to consider

- Do you know how many clusters you want to compute?
- What is the input? What is the output?
  - Input: just a graph? a weighted graph? or a distance matrix? Or, a matrix with extra information about nodes/edges (an “attributed graph”)?
  - Output: a partition of the nodes into disjoint sets? or overlapping sets? Or a “soft clustering” where each node is a member of each cluster with some probability?
- Does the clustering method attempt to solve an optimization problem? Is the problem NP-hard?
- What can you prove about the optimization problem and/or method?

Clustering methods: other issues to consider

- What techniques does the method use?
- Is there a type of network that the method is designed for?
- What studies have been performed using the method, and what did they show? (Note dataset properties, such as size, density, etc.)

Modularity optimization

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right],$$

Maximize the sum of the modularity scores per cluster

Assume the clustering has m clusters:

L = total number of edges in network
l_s = number of edges in cluster s
d_s = total degree of nodes in cluster s (including edges that leave cluster)

Modularity optimization

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right],$$

Notes:

- NP-hard
- Very popular – used all over
- Louvain is a popular software

Maximize the sum of the modularity scores per cluster

Assume the clustering has m clusters:

L = total number of edges in network
l_s = number of edges in cluster s
d_s = total degree of nodes in cluster s (including edges that leave cluster)

Selected papers

1. https://proceedings.mlr.press/v27/luxburg12a/luxburg12a.pdf. von Luxburg et al., JMLR 2012, “Clustering: Art or Science?”
2. https://doi.org/10.1007/s00778-019-00556-x. Fang et al., VLDB 2020, A survey of community search over big graphs
3. https://www.pnas.org/doi/full/10.1073/pnas.0605965104. Fortunato & Barthelemy, PNAS 2007, “Resolution limit in community detection”
4. https://www.nature.com/articles/s41598-019-41695-z Traag et al., Scientific Reports 2019, “From Louvain to Leiden: Guaranteeing well-connected clusters”
5. https://doi.org/10.1103/PhysRevE.105.054311 Vaca-Ramirez and Peixoto, Phys Rev E 2022, “Systematic assessment of the quality of fit of the stochastic block model for empirical networks”

Von Luxburg et al. Clustering: Science or Art?

- Specifically examines evaluation on artificial datasets, on classification benchmark datasets, and on real world datasets.
  - Example given of failure for benchmarks to validly evaluate accuracy (e.g., classifying images that may be B&W or color, and may contain cars).
- Statistical stability and other meta-criteria
- Proposes description of “clustering problems” according to different dimensions
  - Exploratory vs confirmatory
  - Qualitative vs quantitative
  - Unsupervised vs supervised

Von Luxburg et al. Clustering: Science or Art?

- *"We argue that clustering should not be treated as an application-independent mathematical problem, but should always be studied in the context of its end-use."*
- *"In this paper we do not really care how a clustering algorithm works, as long as it achieves the goal we have set."*

Von Luxburg et al. Clustering: Science or Art?

- *If clustering researchers want real impact in applications, then it is time to step back from a purely mathematical and algorithmic point of view.*
- *What is missing is not "better" clustering algorithms but a problem-centric perspective in order to devise meaningful evaluation procedures*

Fang et al., “A survey of community search over big graphs” (VLDB 2020)

- Community Search (CS): Input: network N and vertex v, and the objective is a connected community (cluster) containing v that optimizes some “cohesiveness” criterion, such as:
    - The community is a k-core
    - The community is a k-truss (every edge is in at least k-2 triangles)
    - The community is a k-clique
    - The community is a k-ECC (has edge-connectivity at least k)
- They pay substantial attention to computational requirements for different CS methods and problems.
- CS is not the same as Community Detection (CD), which produces a disjoint clustering where each cluster is a valid community. They argue CS is better than CD. (However…)

Fortunato & Barthélemy

- Basic observation is that Modularity Optimization is subject to a resolution limit (will not find “small modules”, although these will be ”valid”)

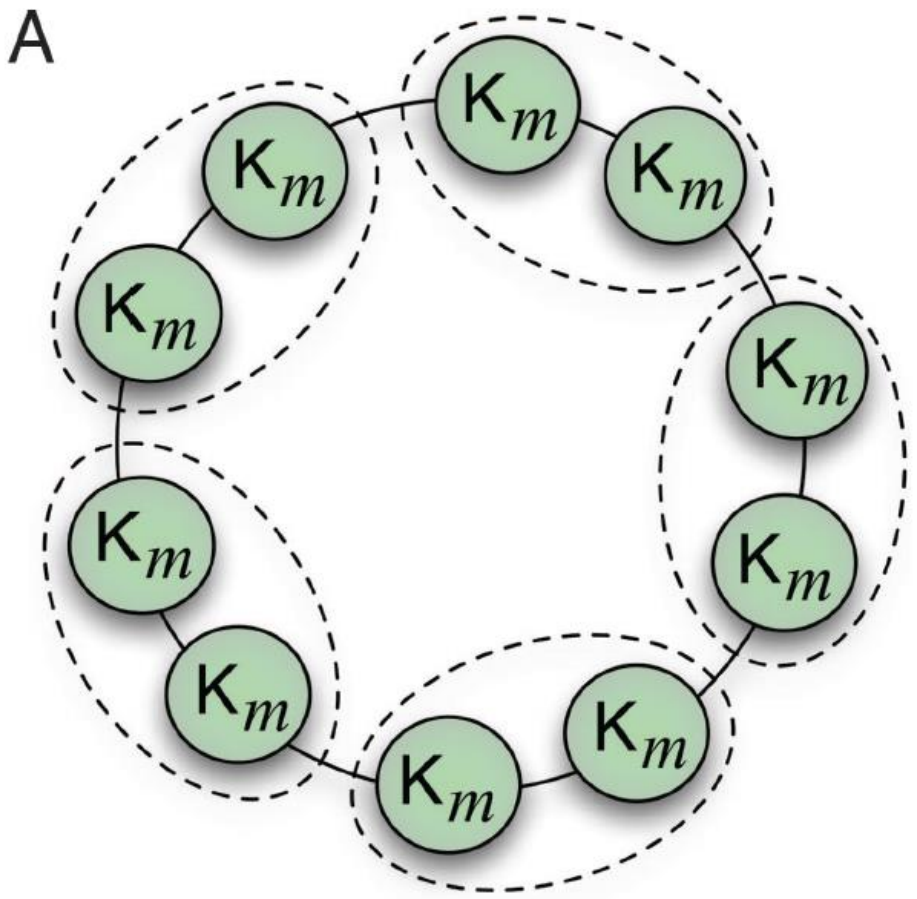


Figure 3(A) from Fortunato & Barthélemy

Modularity optimization

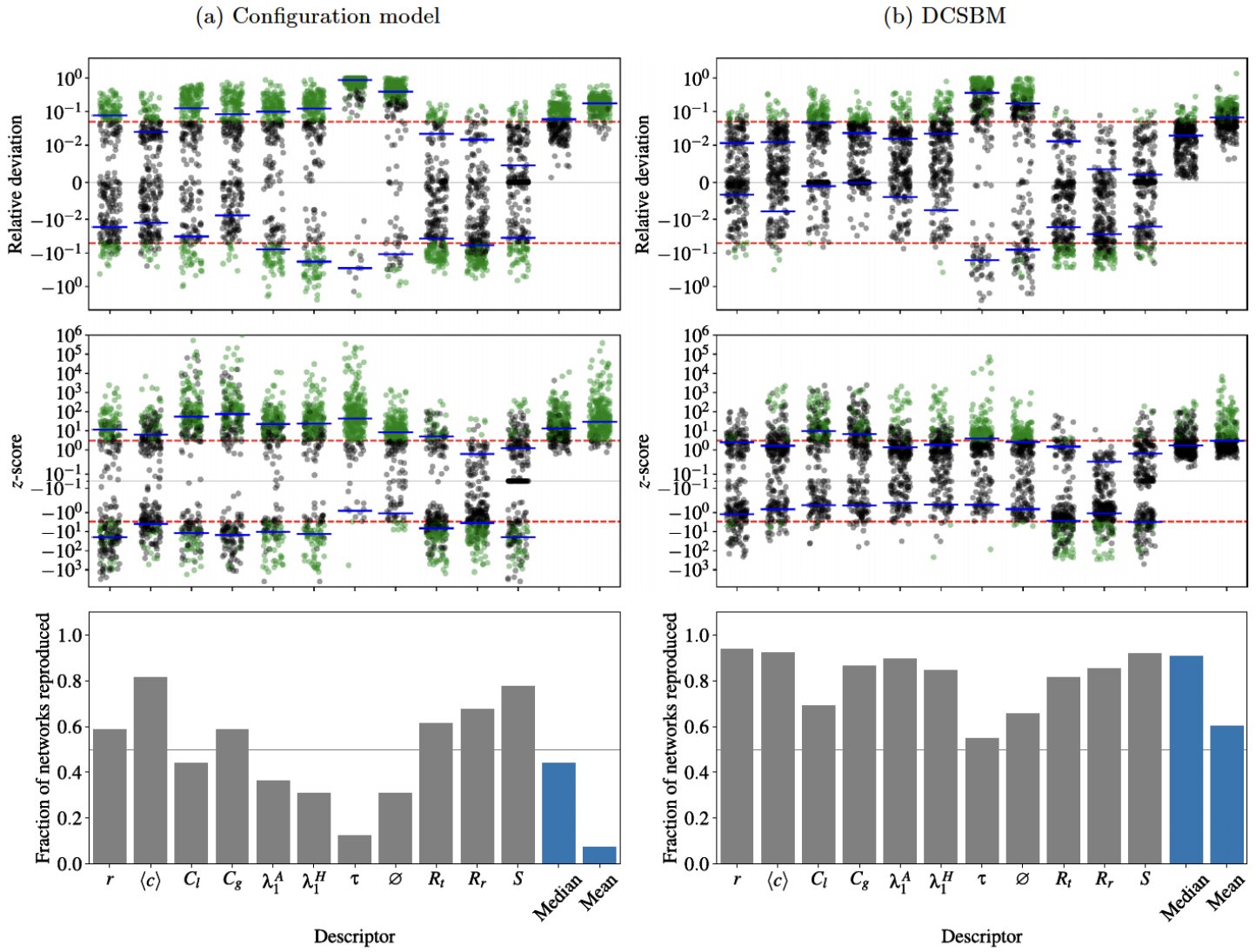
- Modularity optimization seeks a partition of the vertices into disjoint sets (i.e., clusters) so that the sum of the "modularity" scores of each cluster is maximized
- Any cluster with positive modularity is considered "valid"
- Fortunato and Barthelemy proved that optimizing modularity has a resolution limit – clusters below a threshold will not be found
- Hence, they said modularity optimization was not a sufficient clustering method.
- See https://en.wikipedia.org/wiki/Modularity_(networks) for additional references.

Traag et al., “From Louvain to Leiden”

- Traag et al. showed that Louvain, optimizing modularity, could produce disconnected clusters
- They proposed the Leiden heuristic for modularity and showed it was guaranteed to never produce disconnected clusters
- They also proposed a new clustering criterion: optimizing under the Constant Potts Model (CPM), and proved an optimal CPM clustering would produce “well-connected clusters” (i.e., clusters without small edge cuts).

Vaca-Ramirez and Peixoto: Evaluating SBMs

- Stochastic Block Models are probabilistic generative models that generate networks with ground truth clusters. They can also be used as methods to cluster networks.
- Vaca-Ramirez and Peixoto evaluate how well their SBM software (graph-tool) produces networks similar to given real-world networks.
- They compare SBMs to a very simple “configuration model”, and find SBMs are superior.



SBMs are better than the configuration model for network properties

What about cluster properties?

But SBMs have disconnected ground truth clusters!

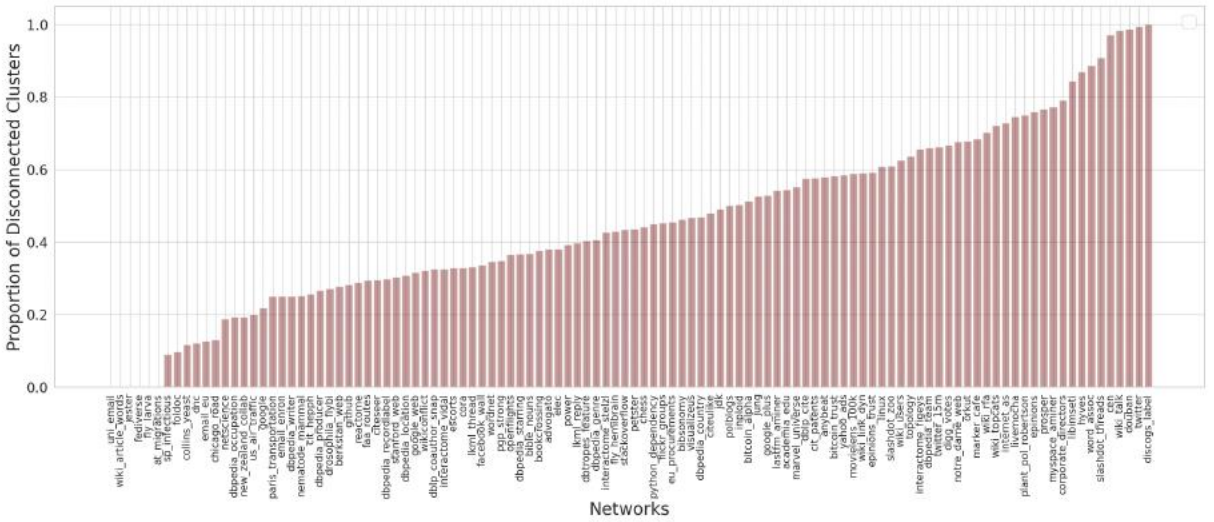


Fig. 2. Proportion of Disconnected clusters in SBM generated networks. The x-axis shows 110 SBM networks generated using parameters from real world networks clustered with the Leiden algorithm (training data). Since Leiden clusterings are guaranteed to be connected, this figure shows that SBM method failed to reproduce the connectivity of the real-world clusterings studied here.

Lahari Anne, The-Anh Vu –Le, Minhyuk Park, Tandy Warnow, and George Chacko

Complex Networks and Applications, 2024

Our research group (Chacko & Warnow)

- See http://tandy.cs.illinois.edu/bibliometrics.html
- We are currently looking at:
  - Developing better synthetic network generators so they reflect clustered real-world networks
  - Modifying clustering methods to ensure they produce well-connected clusters
  - Ensemble methods for community detection (clustering)
  - Theoretical properties about clustering
  - Agent-based models for generating synthetic networks and studying behaviour
  - Massively parallel implementation of clustering methods for ultra-large networks
  - Developing improved community search methods
  - Overlapping clusters, and complex multi-level clusterings
  - Using textual analysis in combination with graph-theoretic methods
  - Discovery using clustering and community search methods

Summary from today

- Clustering methods should be designed and evaluated within a context (Von Luxburg). Our context is Scientometrics.
- Despite decades of research, there are basic problems about clustering that are not yet adequately addressed
- Clustering methods are used in many applications, but using off-the-shelf clustering methods may lead to poor research– better methods are really needed
- Many methods do not scale well to large datasets, so better implementations are needed