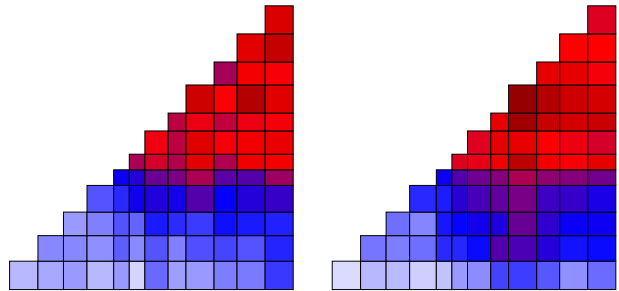
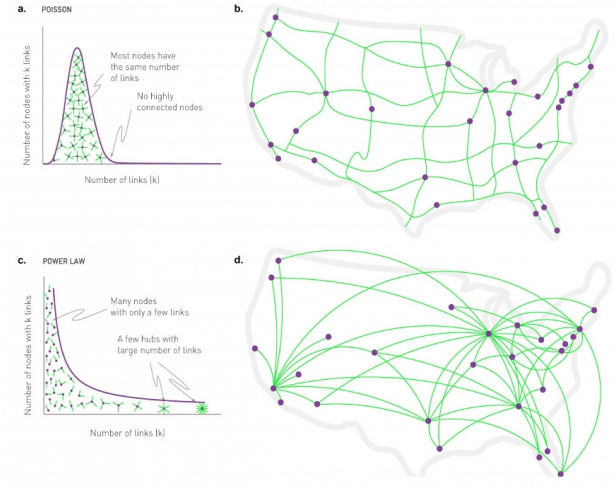
Modularity and Heavy-tailed Degree Distributions

Larry Wilson
Center for Communications Research - La Jolla

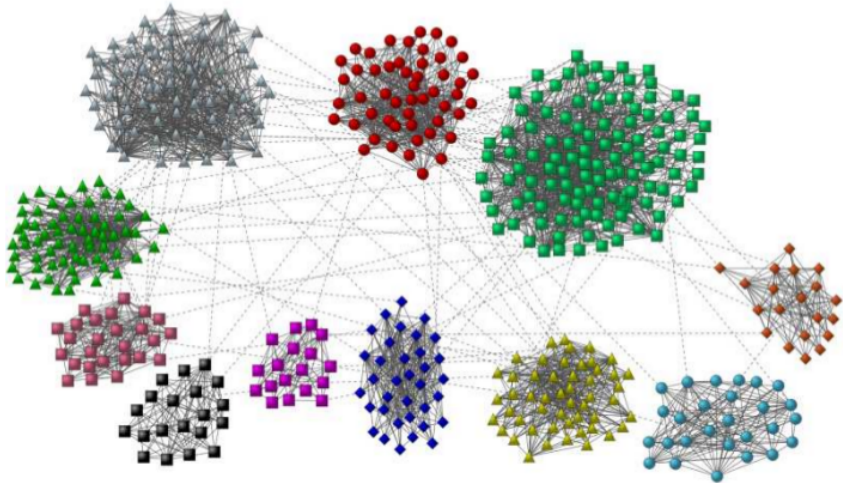
34th MCCCC
October 21–23, 2022



Graphs can have different degree distributions [Barabási, 2016]

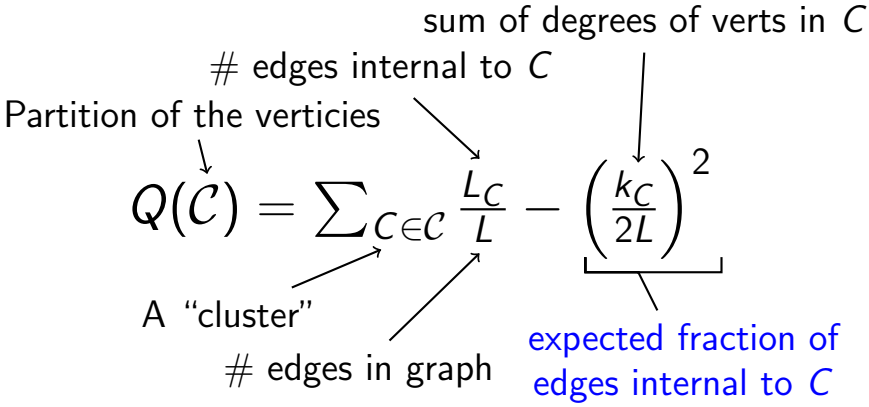


We experiment on LFR graphs, [Lancichinetti, Fortunato, Radicchi, 2008]



In our experiments, 1000 vertices, degree power law exponent $\gamma = 2.5$, at least 20 communities, community size power law exponent 2.0, mixing parameter $\mu = 0.5$.

Modularity [Newman, Girvan, 2004] is a frequently used measure of the quality of a partition of the vertices



Random model: Degrees preserved, edges connected randomly

To deal with the "resolution limit" problem, we add a resolution into the modularity (and present a reformulation)

$r \in (0; 1]$, the "resolution"

$$Q_r(\mathcal{C}) = \sum_{C \in \mathcal{C}} r \frac{L_C}{L} - \left(\frac{k_C}{2L}\right)^2$$

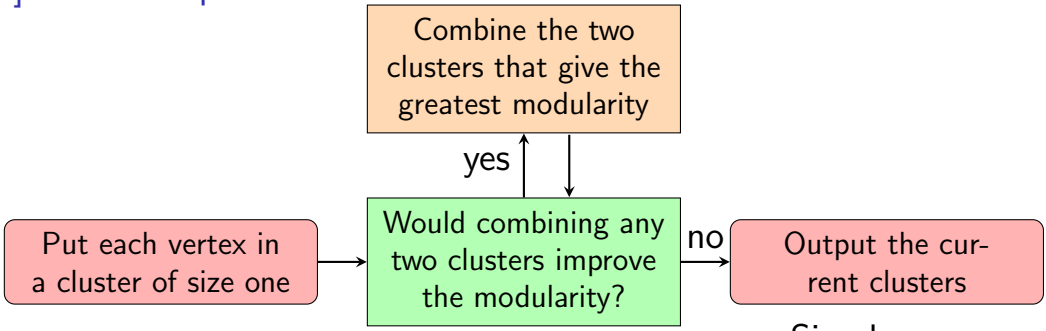
1 if in same cluster, 0 else

degree of vertex v

$$Q_r(\mathcal{C}) = \frac{1}{2L} \sum_v \sum_w C_{vw} \left(r \cdot A_{vw} - \frac{k_v k_w}{2L}\right)$$

1 if connected by an edge, 0 else

We use the "Greedy modularity" hillclimb [Clauset, Newman, Moore, 2004] to learn a partition of the vertices



This is simpler than the Louvain algorithm [Blondel, Guillaume, Lambiotte, Lefebvre, 2008] which has a better run-time and perhaps better clustering performance

Simpler may imply greater dependence on the quality of the score

We measure the performance of clustering via the Matthews Correlation Coefficient [Matthews, 1975] on whether pairs of vertices are clustered together or not correctly

	Pair in a recovered cluster together	Pair not in a recovered cluster together
Pair in a planted cluster together	A	B
Pair not in a planted cluster together	C	D

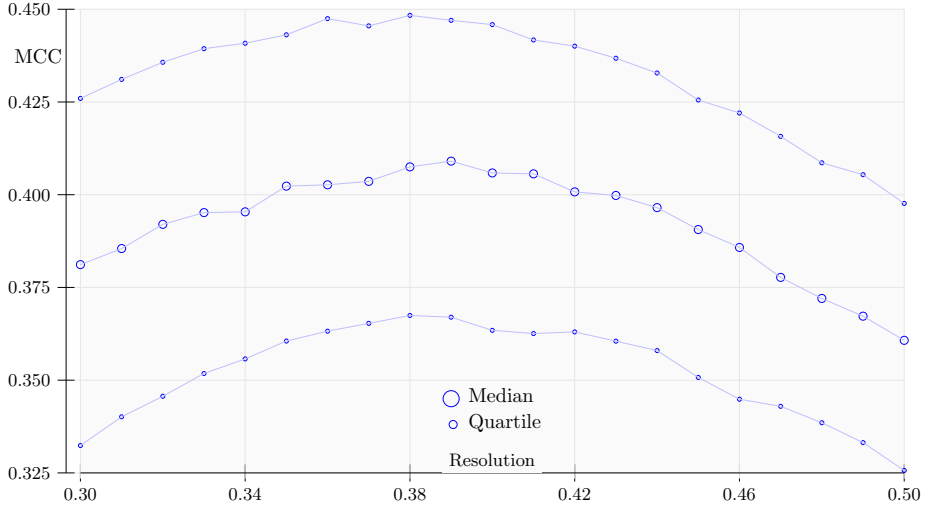
Much smaller class

MCC

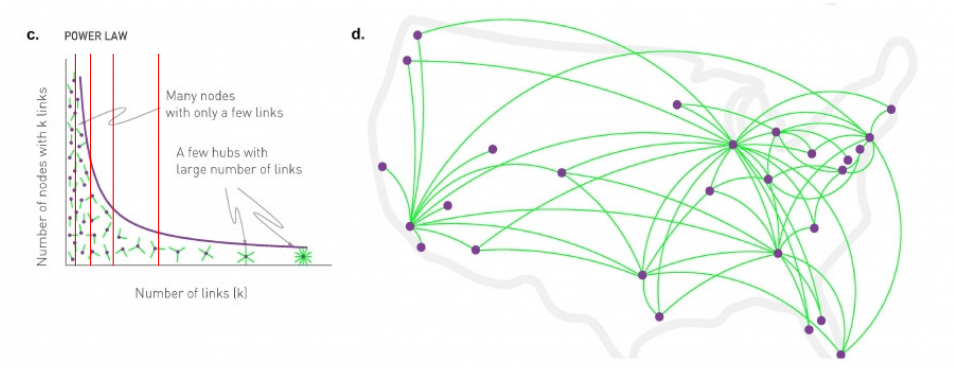
MCC recommended for imbalanced class sizes

score $\in [-1; 1]$

We set the resolution by maximizing the median MCC over 1001 graphs (one hillclimb per graph)

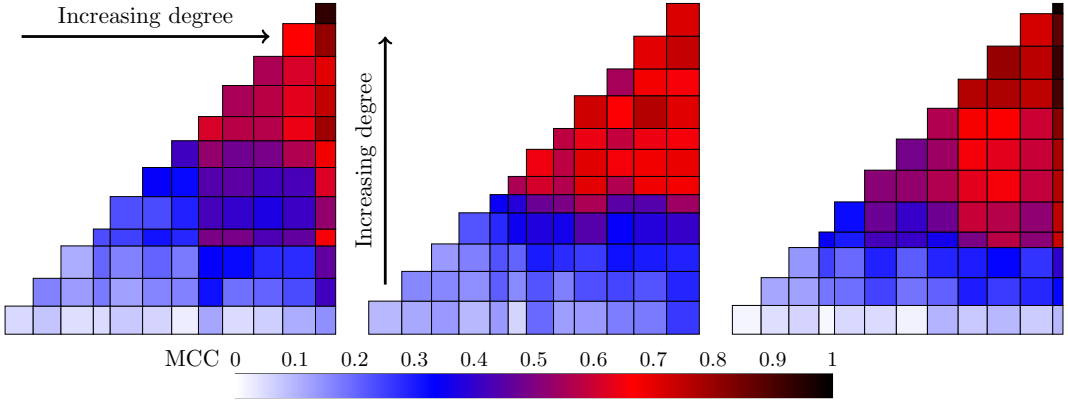


In depicting results, we slice the vertices by degree



Keep including the vertices of the next highest degree until that would push us over 100 vertices in the slice.

Performance is poor on pairs that include a low degree vertex



Median MCC for recovered clusters	All pairs	Low-high pairs
	0.4091	0.2115

(Low-high pairs: one vertex of degree ≤ 20, one of degree ≥ 40)

We propose *flat modularity* which arises from modularity by changing the random model

(global) average vertex degree

$$Q^\flat(\mathcal{C}) = \sum_{C \in \mathcal{C}} \frac{L_C}{L} - \underbrace{\left(\frac{|C|\widehat{k}}{2L}\right)^2}$$

expected fraction of edges internal to C

Random model: # edges preserved, edges connected randomly

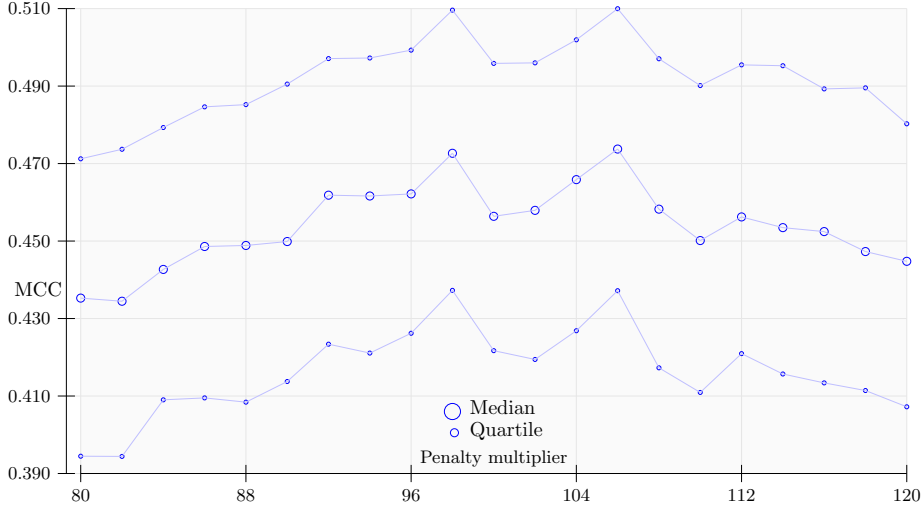
We have relaxed the constraint of preserving vertex degrees

We add in resolution and reformulate

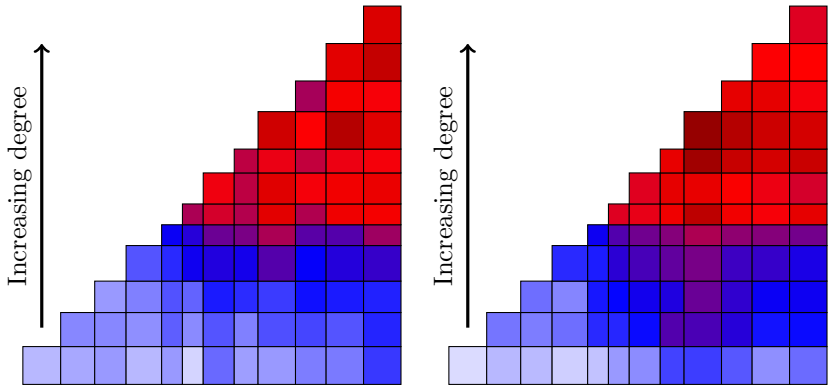
$$
\begin{aligned}
Q_r^{\flat}(\mathcal{C}) &= \frac{1}{2L}\sum_v\sum_w C_{vw}\left(r\cdot A_{vw} - \frac{\widehat{k}\widehat{k}}{2L}\right)\\
\frac{1}{r}Q_r^{\flat}(\mathcal{C}) &= \frac{1}{2L}\sum_v\sum_w C_{vw}\left(A_{vw} - \frac{1}{r}\cdot\frac{\widehat{k}\widehat{k}}{2L}\right)\\
Q_R^{\flat}(\mathcal{C}) &= \frac{1}{2L}\sum_v\sum_w C_{vw}\left(A_{vw} - R\cdot\frac{1}{2L}\right)
\end{aligned}
$$

We call R the "penalty multiplier"

We set the penalty multiplier by maximizing the median MCC over 1001 graphs (one hillclimb per graph)



Anecdotal evidence that switching to flat modularity improves clustering of low degree vertices



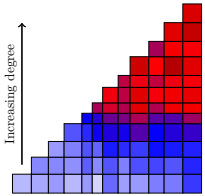
Performance on the median performance graph for modularity: modularity ($r = 0.39$) left and flat modularity ($R = 98$) right.

Statistical evidence

		All pairs		Low-high pairs	
μ	γ	Q_r	$Q_R^\flat$	Q_r	$Q_R^\flat$
0.5	2.5	0.4091	0.4727	0.2115	0.2661
0.5	3.0	0.3762	0.4455	0.2154	0.2702
0.5	3.5	0.3462	0.4210	0.2224	0.2701
0.6	2.5	0.1072	0.1880	0.0271	0.0468
0.6	3.0	0.0949	0.1700	0.0334	0.0535
0.6	3.5	0.0861	0.1529	0.0363	0.0588

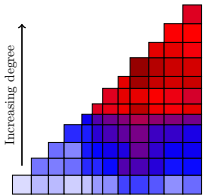
Median MCC between recovered clusters and planted clusters on all pairs and pairs with one vertex of degree at most 20 and one of degree at least 40; larger MCC is better. The recovered clusters are found via Greedy Modularity using either Q_r or $Q_R^\flat$ with r and R chosen to optimize the median for all pairs.

The switch to flat modularity improves clustering overall and particularly for low degree vertices



Low degree vertices seem to be harder to cluster

The hill-climbing score may exacerbate this



Flat modularity improves on modularity

This change may simplify the implementation

- There's no need to track the vertex degrees

Any questions?