FINDING COMMUNITIES IN GRAPHS WITH THE GIRVAN-NEWMAN ALGORITHM

Aidan Roessler, Jake Jasmer, Tony Ni, Yang Tan, Advised by Layla Oesper

WHAT ARE COMMUNITIES?

Clusters of nodes in a graph that have stronger internal connections than external connections

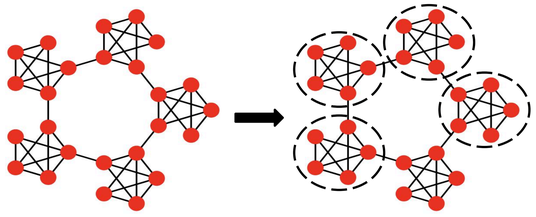


Figure 1. Caveman graph with hand labeled communities [1]

GIRVAN-NEWMAN ALGORITHM

1. Calculate the betweenness for all edges in the graph
2. Remove the edge with the highest betweenness
3. Store a list of set representations of the current connected components (communities)
4. Recalculate betweennesses for all edges affected by the removal
5. Repeat from step 2 until no edges remain.
6. Return the iteration of communities with the highest modularity [2]

- Classic **divisive** community detection algorithm
- **Edge Betweenness:** the amount of shortest paths between any node and any other given node that pass through an edge
- Uses **BFS** for finding all shortest paths (needed to calculate betweenness) in unweighted graphs and **Dijkstra's Algorithm** in weighted graphs
- **Runtime:** $O(|V||E|^2)$ for unweighted graphs and $O(|E|^2|V|log|V|)$ for weighted graphs

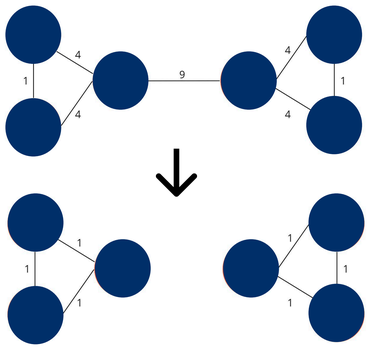


Figure 2. Example unweighted graph with edges labeled with their edge betweenness [3]

Figure 3. Example unweighted graph with edges labeled with their edge betweenness after deleting one edge [3]

MODULARITY

- Measures whether the **weight of edges** between nodes in a community **is stronger than** if they were **distributed randomly**
- Ranges from -1 to 1
- 0 indicates that the edge weights are equal to the random distribution

$$Q = \frac{1}{2m} \sum_{u,v} \left[A_{uv} - \frac{k_u k_v}{2m} \right] \delta(c_u, c_v) \quad [4]$$

$2m$ = total number of half edges

A_{uv} = the actual value of the edge weight for the edge (u, v)

$\frac{k_u k_v}{2m}$ = expected edge weight for the edge (u, v) (probability of two half edges being connected)

$\delta(c_u, c_v)$ = the Kronecker delta function. Evaluates to 1 if vertex u and v are in the same community and 0 otherwise

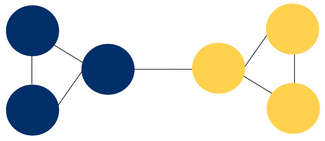


Figure 4. Example unweighted graph with communities that lead to **high modularity** [3]

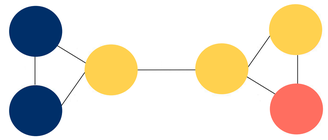


Figure 5. Example unweighted graph with communities that lead to **low modularity** [3]

KARATE CLUB RESULTS

A classic community detection graph is Zachary's Karate Club graph. It **models the relationships of a karate club that split into two new clubs** [5].

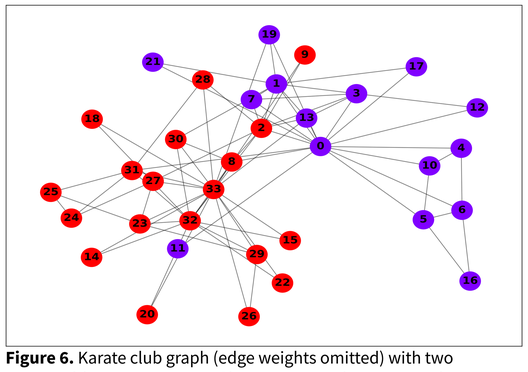
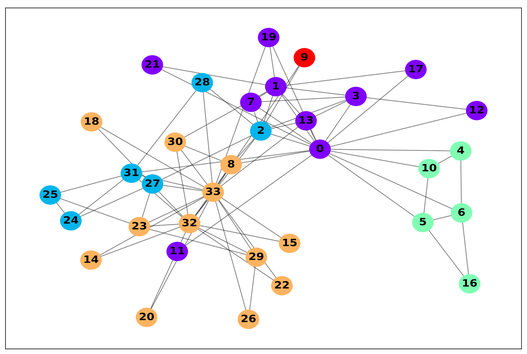


Figure 6. Karate club graph (edge weights omitted) with two communities labeled by our Girvan-Newman implementation



Karate club graph (edge weights omitted) with communities that maximize modularity according to our Girvan-Newman implementation

- Like Girvan and Newman's original implementation, when the max communities was restricted to two, we only misclassified individual #2 [2]
- The graph with the highest modularity (0.38) actually contains five communities
- This doesn't match the real world communities since there are closer relationships within the two

OTHER RESULTS

Dataset	$\|V\|$	$\|E\|$	Algorithm	Communities Found	Time (s)	Peak Memory (mb)	Modularity
Karate Club	34	78	Girvan-Newman	5	0.27	0.07	0.38
			Louvain	**3**	**0.11**	0.07	0.43
			BVNS	5	14.54	**0.03**	**0.44**
College Football	115	613	Girvan-Newman	**10**	28.08	29.19	**0.60**
			Louvain	5	**3.06**	0.20	0.58
			BVNS	15	276.46	**0.04**	0.49
Urban Movement	360	2925	Girvan-Newman	5	19592.44	1.40	**0.67**
			Louvain	4	**557.48**	1.35	0.65
			BVNS	26	15581.97	**0.41**	0.27
C. Elegans Neurons	473	5025	Girvan-Newman	69	8115.11	30.25	0.35
			Louvain	**10**	**265.79**	2.23	**0.53**
			BVNS	27	7845.69	**0.06**	0.25

Table 1. An overview of our results for using each algorithm on choice datasets.

- **Louvain:** uses a top down approach to merge nodes into communities that maximize modularity
- **Basic Variable Neighborhood Search (BVNS):** randomization approach to simulate gradient descent for forming communities with maximum modularity

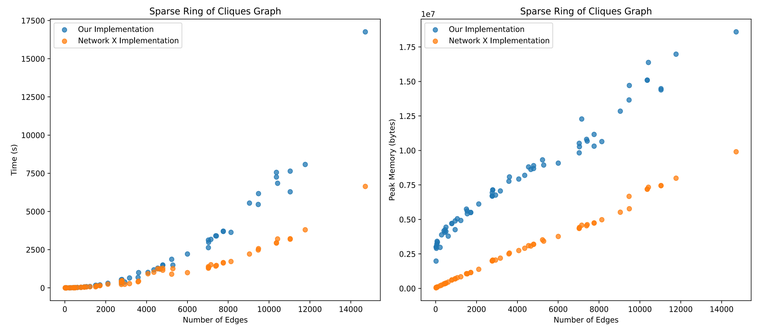


Figure 8. Results from running our implementation of Grivan-Newman and Network X's on a sparse ring of cliques graph, were sparse means a graph density range of [0.08, 0.12]

SIGNIFICANCE

- **Girvan-Newman works best on sparse graphs**
- Since its runtime is bounded by |E| and Louvain and BVNS's runtimes aren't, it performs much worse then them with high edge counts
- Modularity is a good "source of truth" metric to use across very different algorithms, but depending on the algorithm and application, it often doesn't reflect real world communities
- Because the definition of communities varies vastly across applications, **when it comes to community detection, there is no one size fits all algorithm**

WORKS CITED

[1] E. W. Weisstein. Caveman Graph. Publisher: Wolfram Research, Inc.
[2] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12):7821–7826, June 2002.
[3] Girvan-Newman algorithm | Memgraph's Guide for NetworkX library.
[4] Modularity (networks), Wikipedia, Nov. 2024.
[5] W. W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. Journal of Anthropological Research, 33(4):452–473, 1977. Publisher: [University of New Mexico, University of Chicago Press].