Machine Learning and Real-world Data

Example Sheet 4 - Zejia Yang

March 5, 2024

Graph algorithms

The basic idea is to use BFS twice. First, run BFS from an arbitrary vertex to find the most distant vertex with the longest distance. If not all vertices are visited, then return infinity (the graph is not connected). Otherwise, run BFS from that most distant vertex to find the longest path. The length of that longest path is the required diameter.

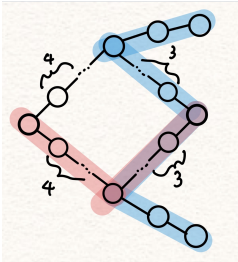


Figure 1: cycle

Graph 1 with cycles. If BFS starts at the leftmost vertex, then the farthest point is 7 units away. It gives the diameter: 7. However, the diameter is 10.

Betweenness centrality and Newman-Girvan method examples

For the remaining graph after I remove the highest betweenness centrality. Please see 2

(1)

1. nodes: 1.0 for all the nodes.
2. edges: 3.0 for all edges.

(2)

1. nodes: 1: 6.0, others: 0.0.
2. edges: 4.0 for all edges.

(3)

1. nodes: 1: 5.0 others: 0.0.
2. edges: $(1,2):3.0 \quad (1,3):3.0 \quad (1,4):4.0 \quad (1,5):4.0 \quad (2,3):1.0$

(4)

1. nodes: $1:0.0 \quad 2:3.0 \quad 3:0.0 \quad 4:3.0 \quad 5:0.0$
2. edges: $(1,2):4.0 \quad (2,3):2.0 \quad (2,4):4.0 \quad (3,4):2.0 \quad (4,5):4.0$

(5)

1. nodes: $1:0.0 \quad 2:3.0 \quad 3:0.0 \quad 4:3.0 \quad 5:0.0$
2. edges: $(1,2):4.0 \quad (2,3):6.0 \quad (3,4):3.0 \quad (4,5):1.0 \quad (3,5):3.0$

(6)

1. nodes: $2:5.0$, others : 0.0
2. edges: $(1,2):4.0 \quad (2,3):3.0 \quad (3,4):1.0 \quad (2,4):3.0 \quad (2,5):4.0$

(7)

1. nodes: $1:3.0 \quad 2:1.0 \quad 3:0 \quad 4:1.0 \quad 5:0.0$
2. edges: $(1,2):3.0 \quad (2,3):2.0 \quad (2,4):1.0 \quad (3,4):2.0 \quad (1,4):3.0 \quad (1,5):4.0$

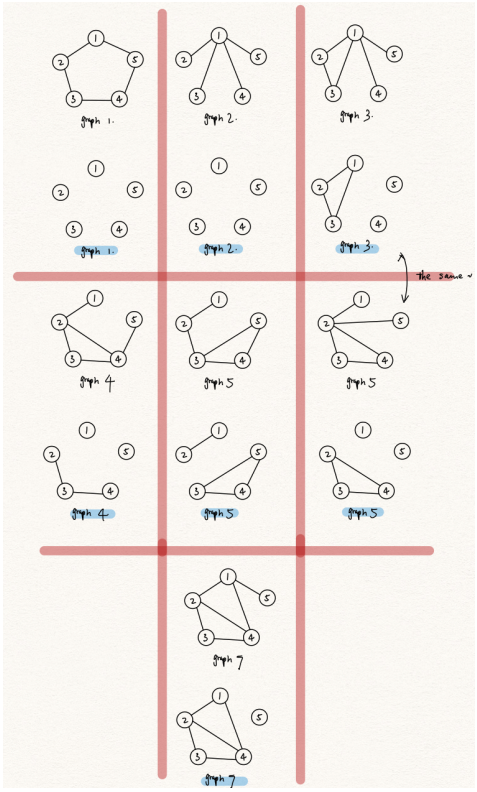


Figure 2: Betweenness Centrality

Random graphs and metrics

Erdős – Rényi

Undirected graph $G_{n,p} = (V, E)$. n is the number of vertices and p is the possibility of forming an edge $(u, v) \in E$.

1. degrees of nodes: binomial distribution since every edge has an equal probability of being chosen and there are at most $(n-1)$ edges incident to one vertex. As n grows larger, it tends to follow a Poisson distribution.
2. number of connected components: It depends on n and p. However, given that every pair of nodes has an equal possibility of being connected, the number of connected components is likely to be small (only 1). There exists a giant largest connected component.
3. (distribution of) shortest paths: It doesn't generate local clustering and triadic closures, so it's very likely that most of the paths have a distance around the maximum shortest path ($log(n)$). The maximum shortest path grows slowly with n.
4. extent of clustering: very low, and goes to zero as we increase p.

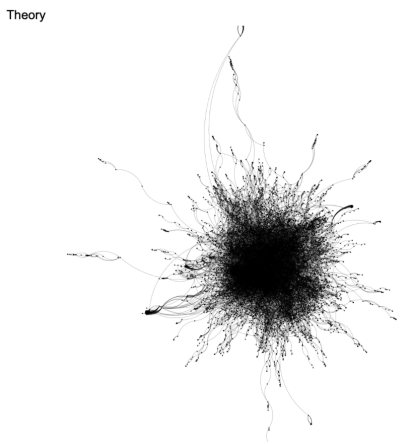
Watts-Strogatz

Undirected graph $G_{n,k,p} = (V, E)$. where n is $|V|$ and k is the initial degree of a node (an even integer) and p is a rewiring probability.

1. (distribution of) degrees of nodes: roughly Poisson distribution as n increases, since initially all nodes have the same degree k and every edge is equally likely to be replaced.
2. number of connected components: There is a high chance that it will be one because two clusterings are likely connected by one edge during rewiring.
3. (distribution of) shortest paths: It has short average path lengths due to local clustering and rewiring (connecting two clusterings in a shortcut).
4. extent of clustering: highly clustered due to initial connections with k nearest neighbours.

A collaboration network

An example collaboration graph. The property of a collaboration graph depends on the discipline.



1. (distribution of) degrees of nodes: From the example, the degrees of nodes concentrate on the range of small values, with an average degree of 5. In general, I think the distribution of degrees will be centered/skewed around the mean.
2. number of connected components: I expect it to have several weakly connected components since they represent collaborations within certain academic communities.
3. (distribution of) shortest paths: the same with Watts-Strogatz. The average short path length is short because of the local clustering and edges between clusters.

4. extent of clustering: highly clustered. Collaborators are likely to form triadic closures within a cluster.