Big Data Tutorial Assignments 6 and 7

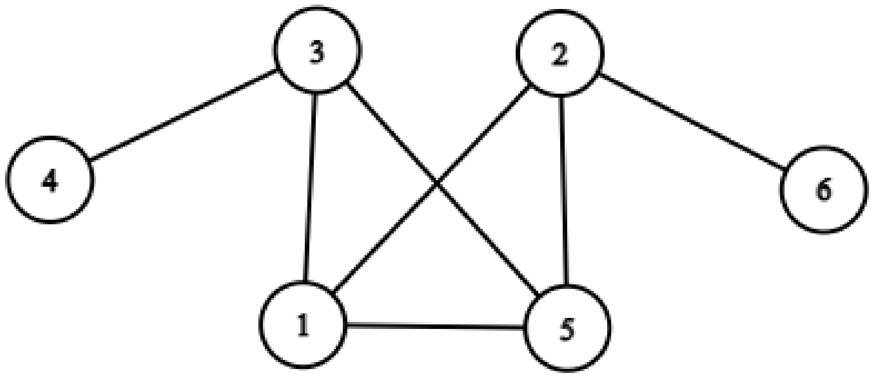
Marina Ernst
marinaernst@uni-koblenz.de

Institute for Web Science and Technologies
Universität Koblenz

Assignment 6

Recall

You are given a graph G. Which type of graph is it?



Weighted

Directed

Bipartite

Undirected

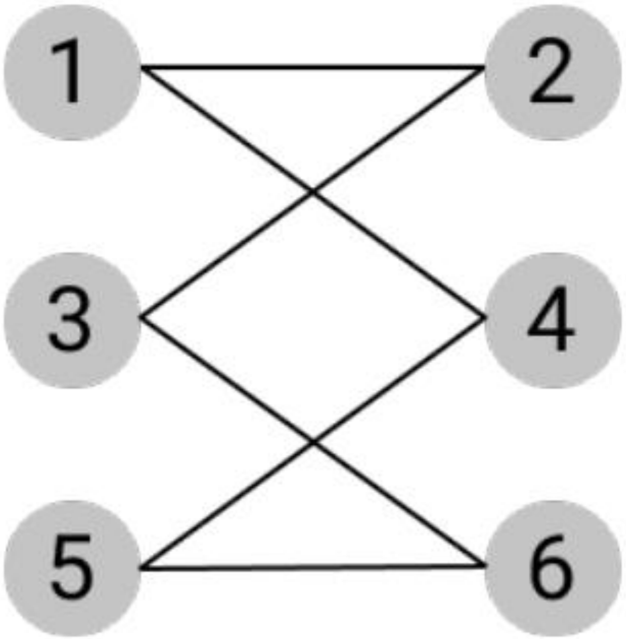
Unweighted

Unlabeled

Labeled

Recall

FYI: Bipartite graph



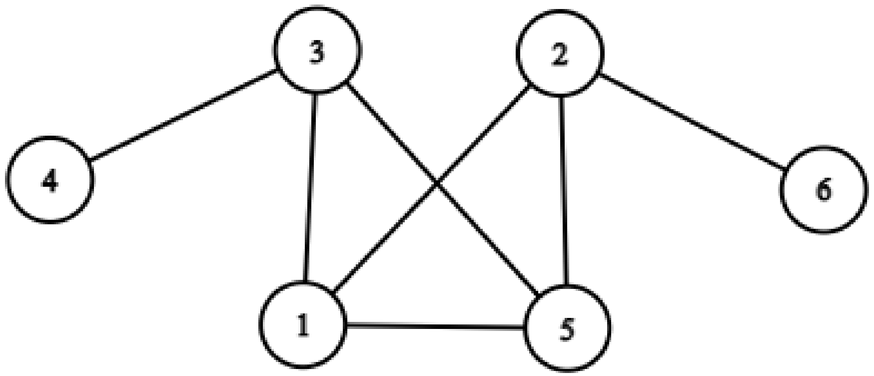
Bipartite graph definition:
Vertices have a disjoint split:

1. $\exists A, B \subsetneq V$
2. $V = A \cup B$
3. $A \cap B = \emptyset$
4. $A = 1, 3, 5$; $B = 2, 4, 6$
5. Such that all edges cross the disjoint set
6. $\forall e = (u, v) \in E : u \in A \wedge v \in B \vee u \in B \wedge v \in A$

Recall

universität koblenz
weiter:denken

You are given a graph G. Which type of graph is it?



Weighted

Directed

Bipartite

Undirected

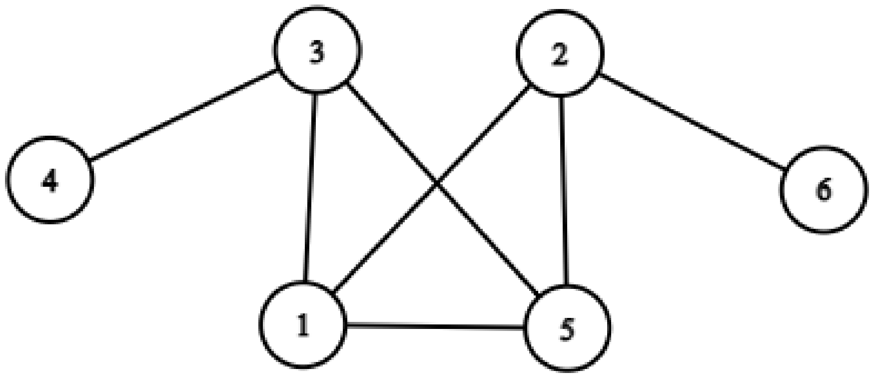
Unweighted

Unlabeled

Labeled

Recall

You are given the same Graph G as in the previous task. Calculate the diameter of the graph



$max(dist(u, v))$, for all $u, v \in V$, where $dist(u, v)$ is the distance between u and v.

Or: the maximum node eccentricity in G.

Or: the longest shortest path

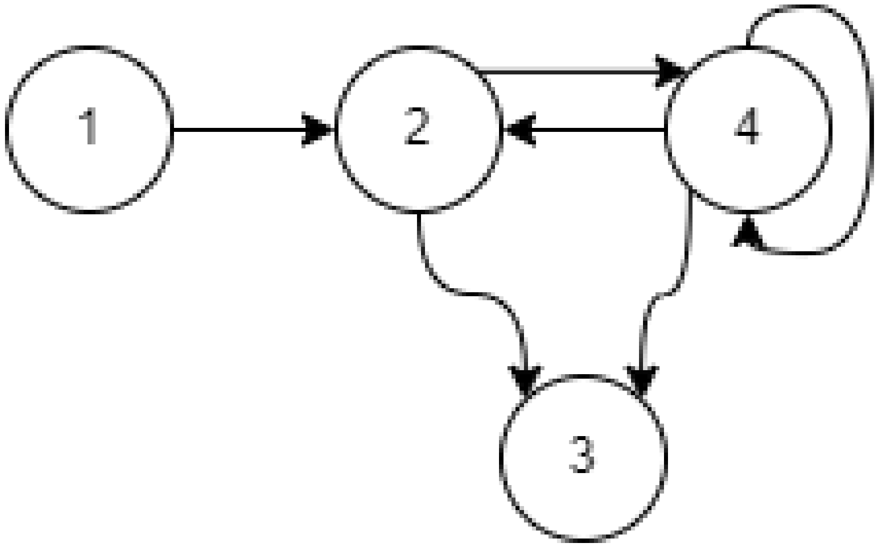
Diameter=4

{4,3,5,2,6} {4,3,1,2,5} {6,2,1,3,4} {6,2,5,3,4}

Recall

You are given the Adjacency Matrix A of graph G. Draw a graph based on this matrix.
Please do it on paper and then just upload the picture here.

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$



Recall

universität koblenz
weiter:denken

Scale-free vs Random graph

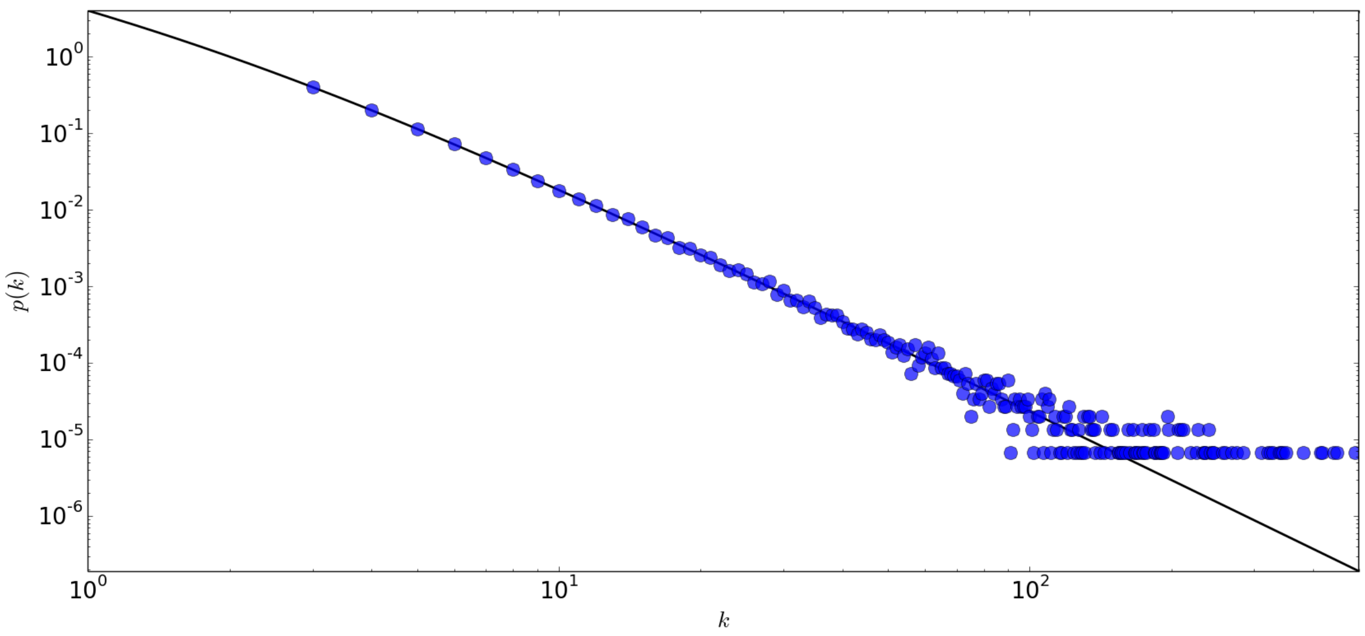
Not answered

Which of the following statements are true for scale-free graph and which for the random one?

	Scale-free graph	Random graph
Follows the power law	☐	☐
Pre-given number of nodes	☐	☐
Vast majority of nodes has only a few connections	☐	☐
Hubs are NOT present	☐	☐
Small world phenomena	☐	☐
Harder to partition	☐	☐

Recall

Scale-free network



https://en.wikipedia.org/wiki/Scale-free_network

Recall

Scale-free vs Random graph

Model solution

Which of the following statements are true for scale-free graph and which for the random one?

	Scale-free graph	Random graph
Follows the power law	☑	☐
Pre-given number of nodes	☐	☑
Vast majority of nodes has only a few connections	☑	☐
Hubs are NOT present	☐	☑
Small world phenomena	☑	☐
Harder to partition	☑	☐

Recall

universität koblenz
weiter:denken

Pregel

Not answered

Which of the following statements are true for Pregel

Unanswered	Right	Wrong	
☑	☐	☐	The processing happens In-Memory
☑	☐	☐	A vertex contains information about itself and the incoming edges
☑	☐	☐	The computation is described in terms of vertices, edges and a sequence of super-steps
☑	☐	☐	The computation is described in terms of vertices, edges and a sequence of super-steps
☑	☐	☐	Build on top of Hadoop
☑	☐	☐	Low scalability

Recall

Pregel

Model solution

Which of the following statements are true for Pregel

Unanswered	Right	Wrong	
☐	☑	☐	The processing happens In-Memory
☐	☐	☑	A vertex contains information about itself and the incoming edges
☐	☑	☐	The computation is described in terms of vertices, edges and a sequence of super-steps
☐	☑	☐	The computation is described in terms of vertices, edges and a sequence of super-steps
☐	☐	☑	Build on top of Hadoop
☐	☐	☑	Low scalability

Recall

universität koblenz
weiter:denken

Giraph vs Spark GraphX

Not answered

	Giraph	Spark GraphX
Computation in memory	☐	☐
Used by Alibaba	☐	☐
Adopts a vertex-cut approach to distributed graph partitioning	☐	☐
Built on top of Hadoop	☐	☐
Used by Facebook, LinkedIn, etc	☐	☐

Recall

universität koblenz
weiter:denken

Giraph vs Spark GraphX

Model solution

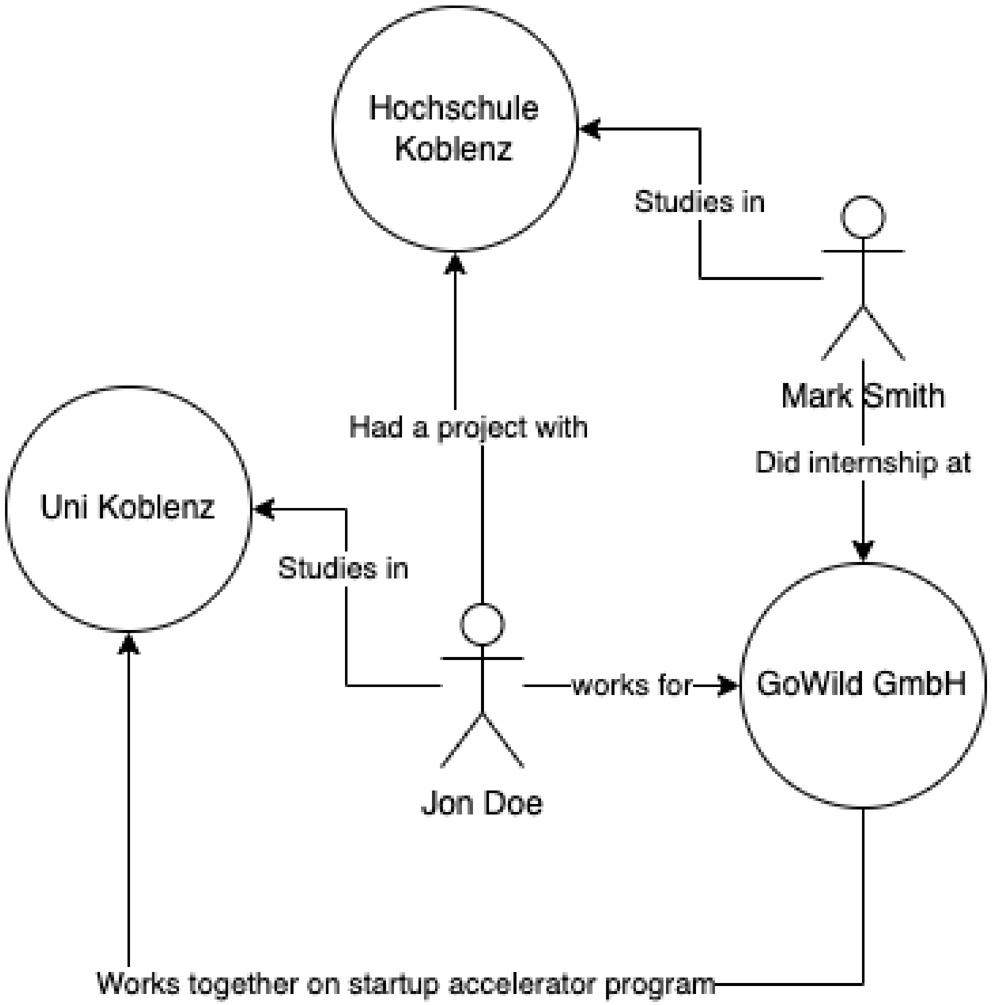
	Giraph	Spark GraphX
Computation in memory	☑	☑
Used by Alibaba	☐	☑
Adopts a vertex-cut approach to distributed graph partitioning	☐	☑
Built on top of Hadoop	☑	☐
Used by Facebook, LinkedIn, etc	☑	☐

Knowledge Questions

Graph data models

Carefully study "Knowledge Graphs" article from the further reading materials. There you will find different graph data models.

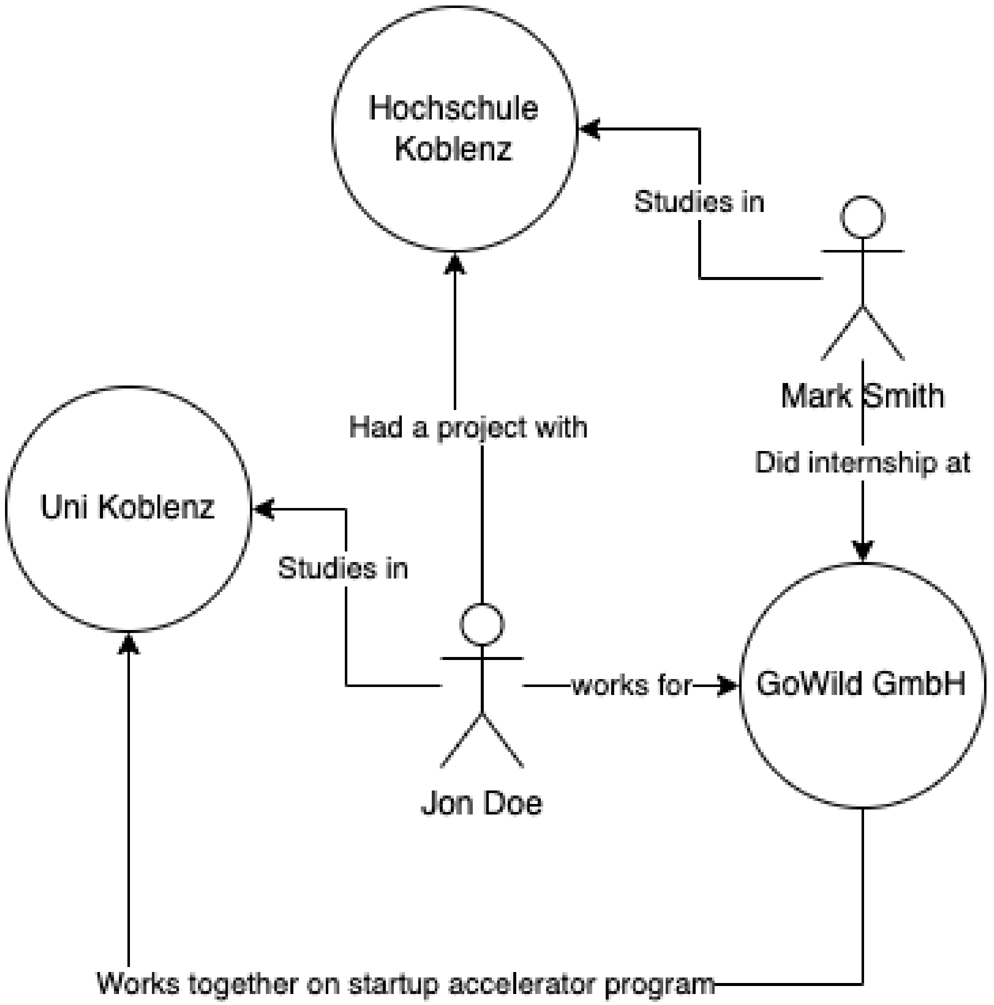
You are given a very simple graph. Which graph model is represented here? Explain your answer.



Knowledge Questions

Directed Edge-labelled Graphs.
(Multi-relational graph)

Why not Heterogeneous?



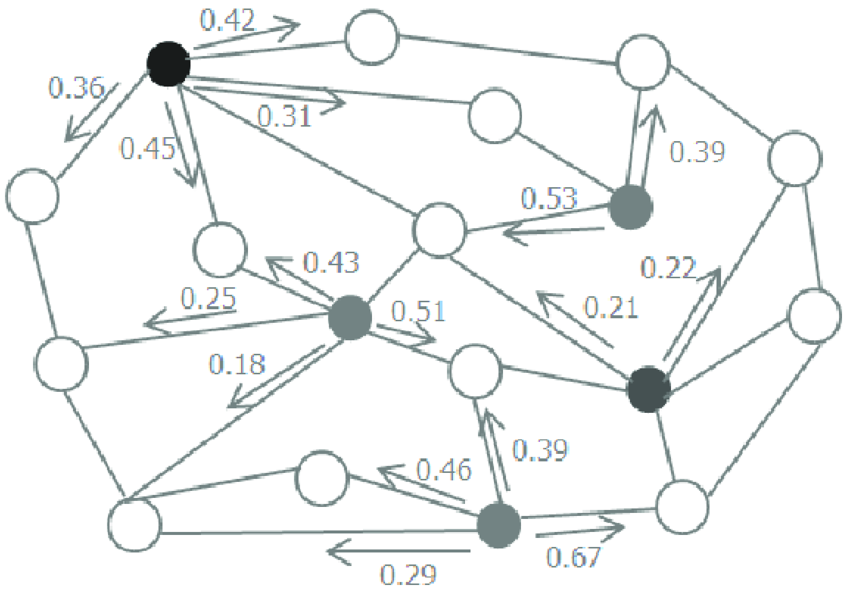
Knowledge Questions

universität koblenz
weiter:denken

Label propagation

Based on further reading materials (Graph at Facebook) what is the label propagation? How does it work?

Label propagation is an iterative graph algorithm that infers unlabeled data from labeled data. The basic idea is that during each iteration of the algorithm, every vertex propagates its probabilistic labels to its neighboring vertices, collects the labels from its neighbors, and calculates new probabilities for its labels.



https://www.researchgate.net/publication/324011683_Semi-Supervised_Classification_of_Hyperspectral_Images_Based_on_Extended_Label_Propagation_and_Rolling_Guidance_Filtering

Knowledge Questions

BFS

Carefully study the paper Graph structure in the web.
Explain what BFS is. And what is it used for in the context of this study?

"A breadth-first search (BFS) on a directed graph begins at a node u of the graph, and proceeds to build up the set of nodes reachable from u in a series of layers. Layer 1 consists of all nodes that are pointed to by an arc from u. Layer k consists of all nodes to which there is an arc from some vertex in layer k-1, but are not in any earlier layer.

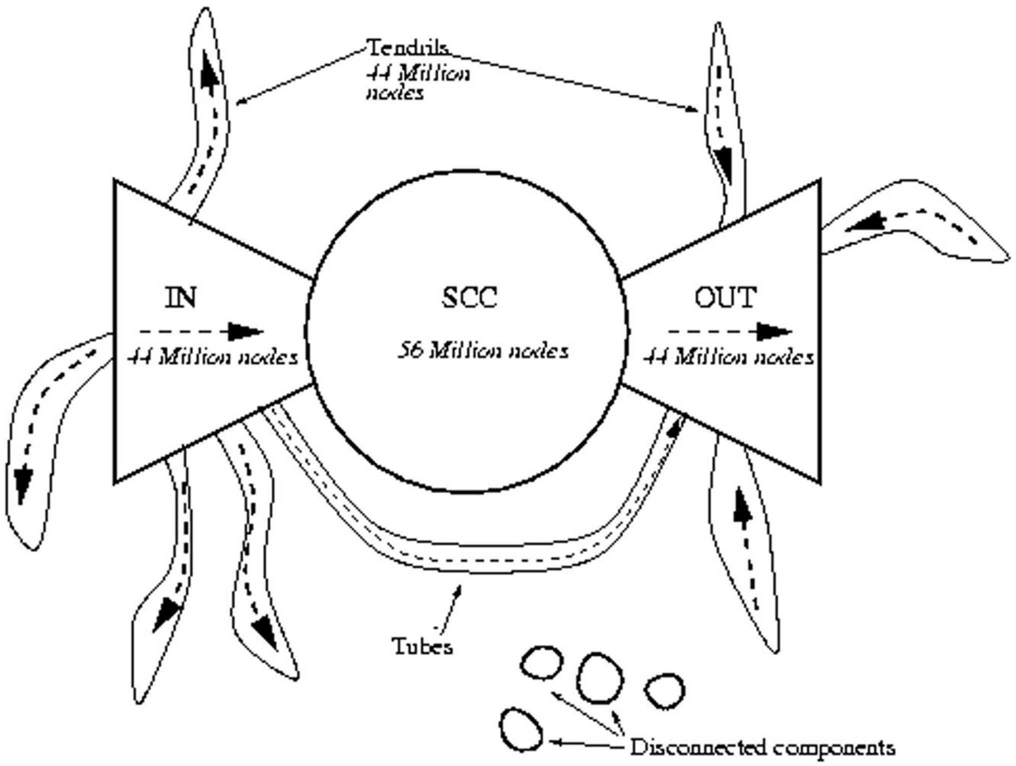
BFS algorithm is used to search a tree or graph data structure for a node that meets a set of criteria.

Knowledge Questions

BFS

Carefully study the paper Graph structure in the web.
Explain what BFS is. And what is it used for in the context of this study?

"we use the BFS runs to estimate the positions of the remaining nodes"

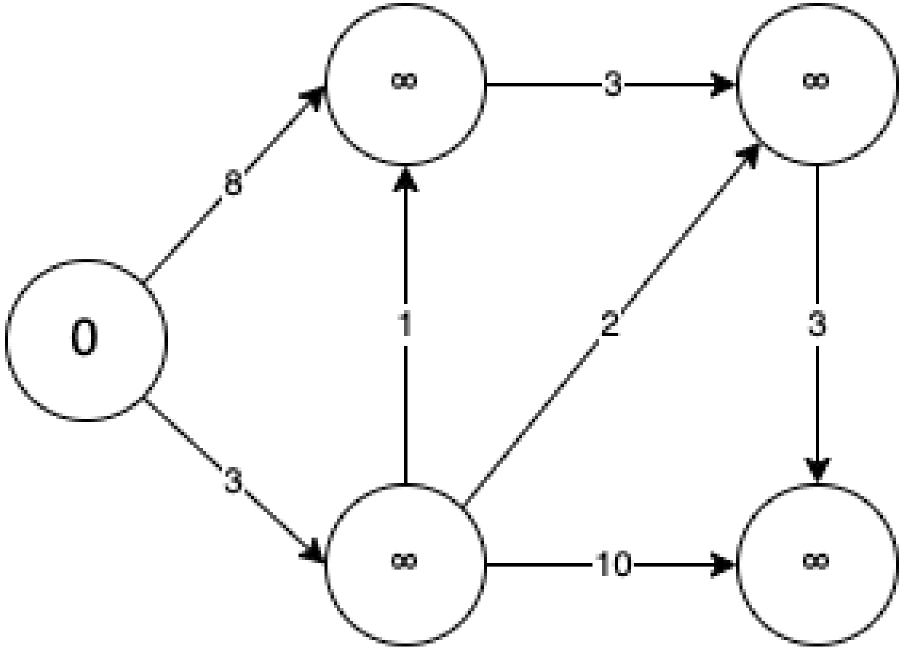


› Assignment 7

Recall

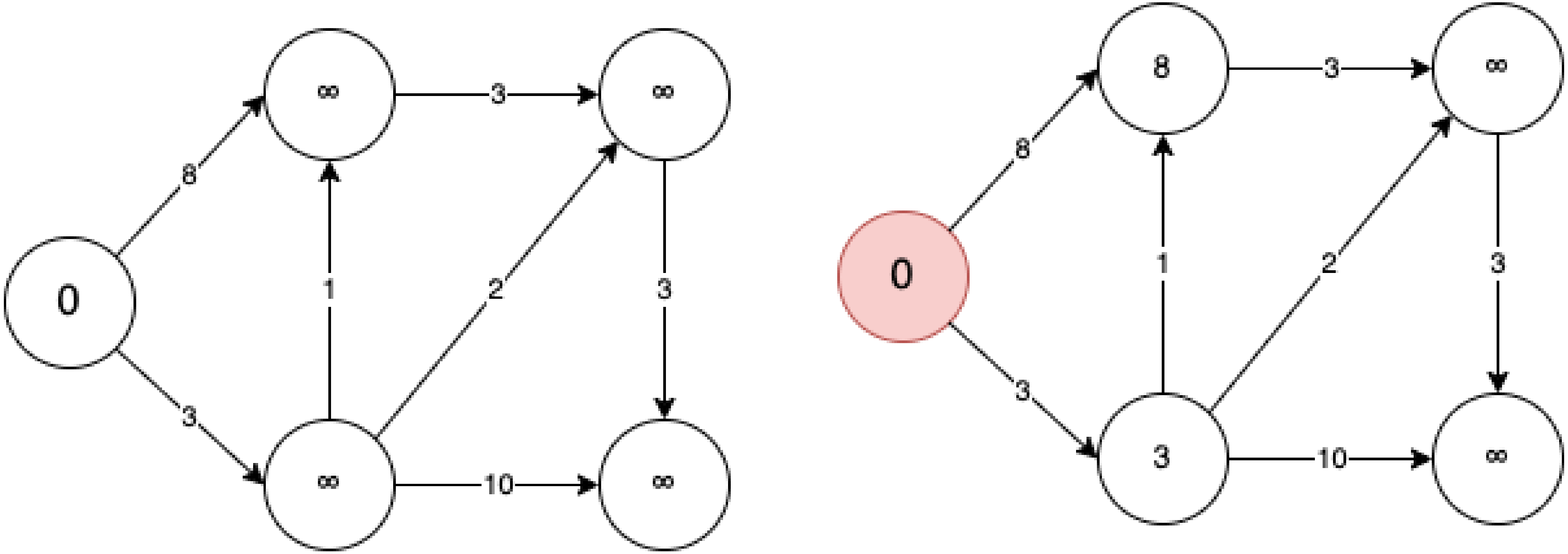
Dijkstra‘s algorithm

You're given a weighted graph.
Please perform Dijkstra‘s algorithm on it. Remember, that for every step, you should redraw the graph (see the examples from the lecture)



Recall

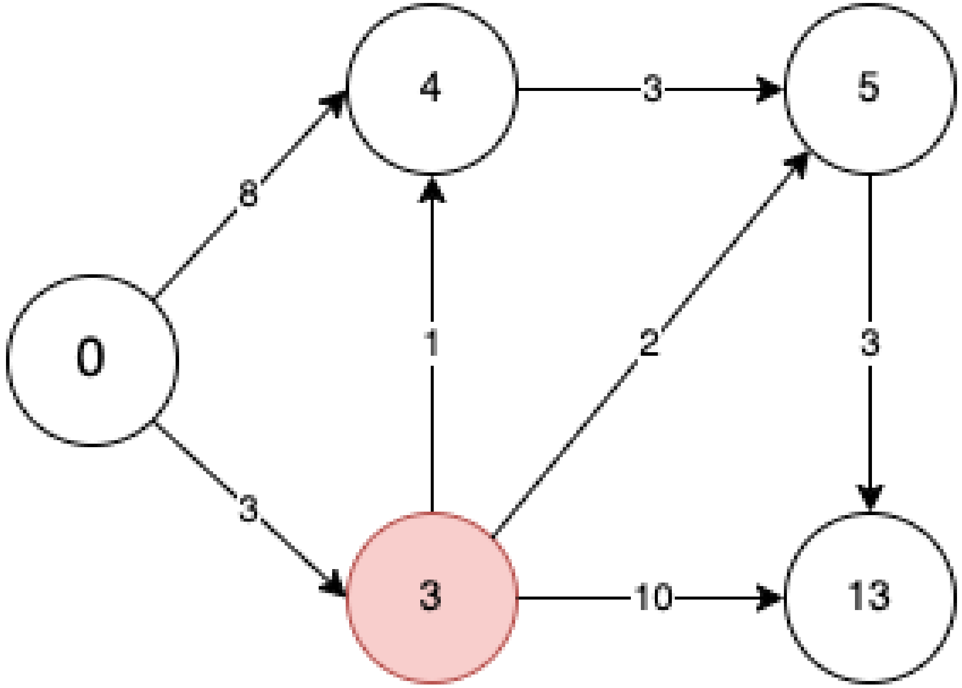
Dijkstra‘s algorithm

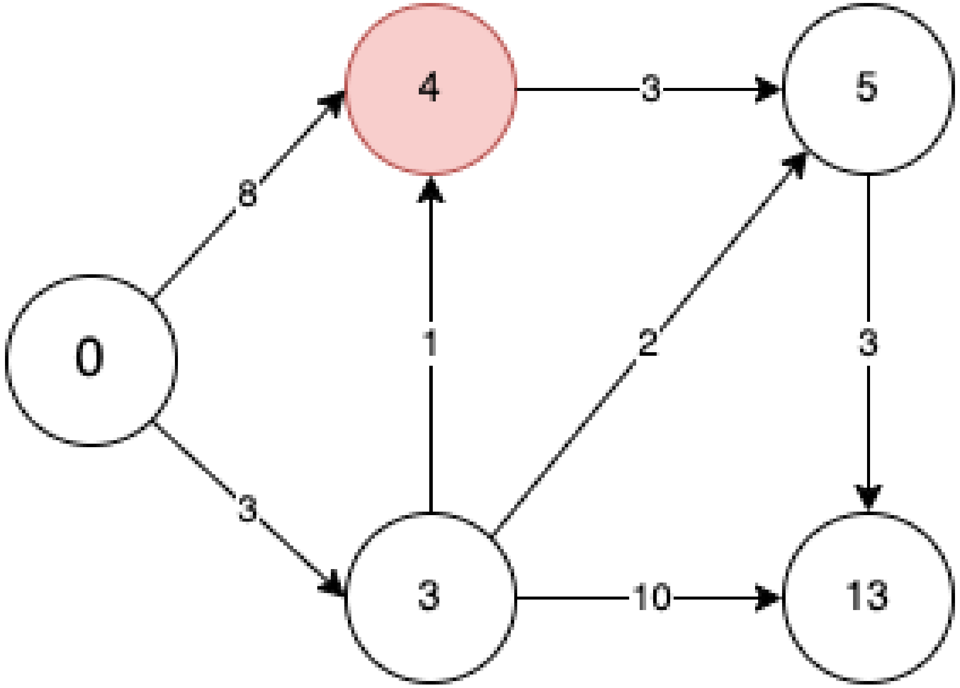


Recall

universität koblenz
weiter:denken

Dijkstra‘s algorithm

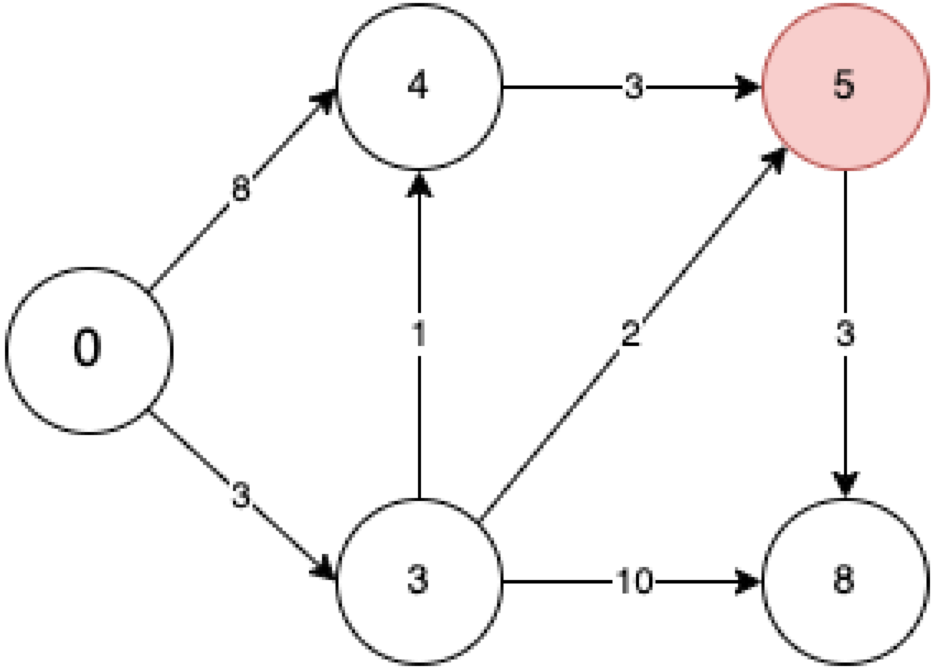


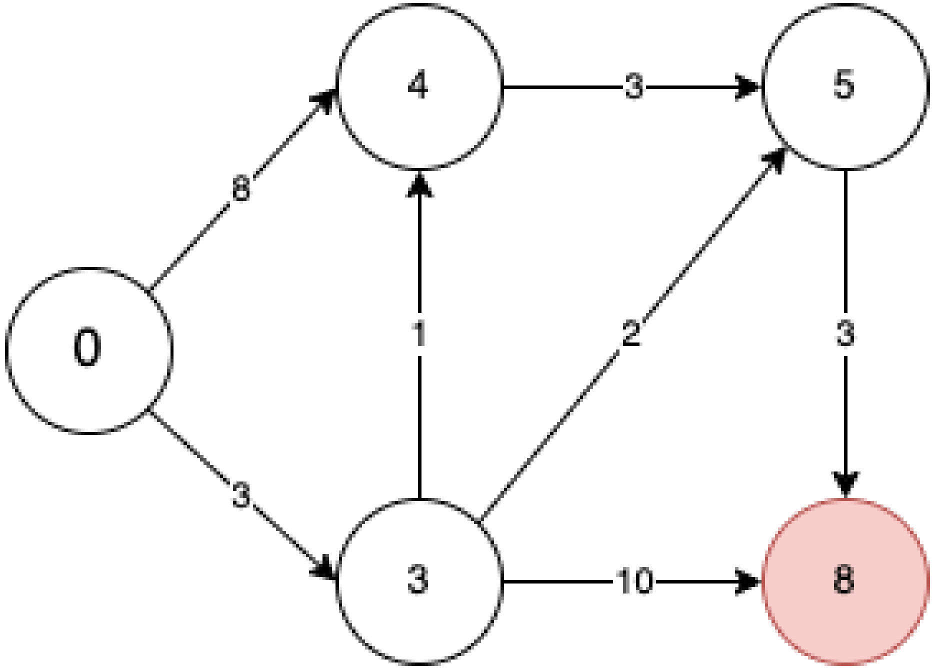


Recall

universität koblenz
weiter:denken

Dijkstra‘s algorithm

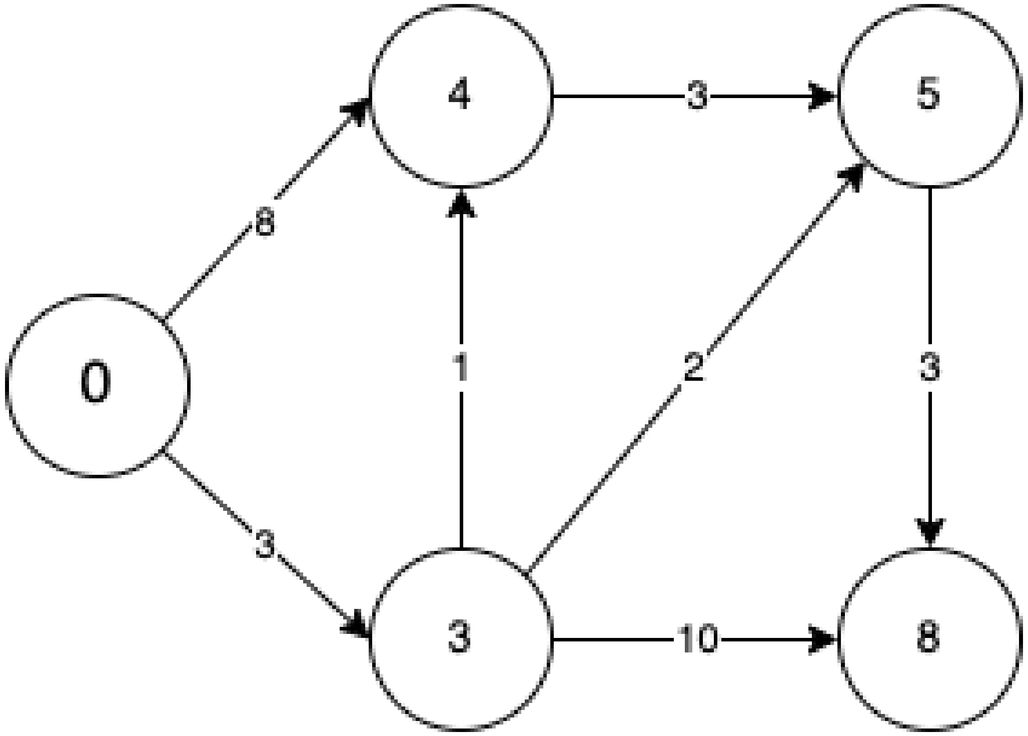




Recall

universität koblenz
weiter:denken

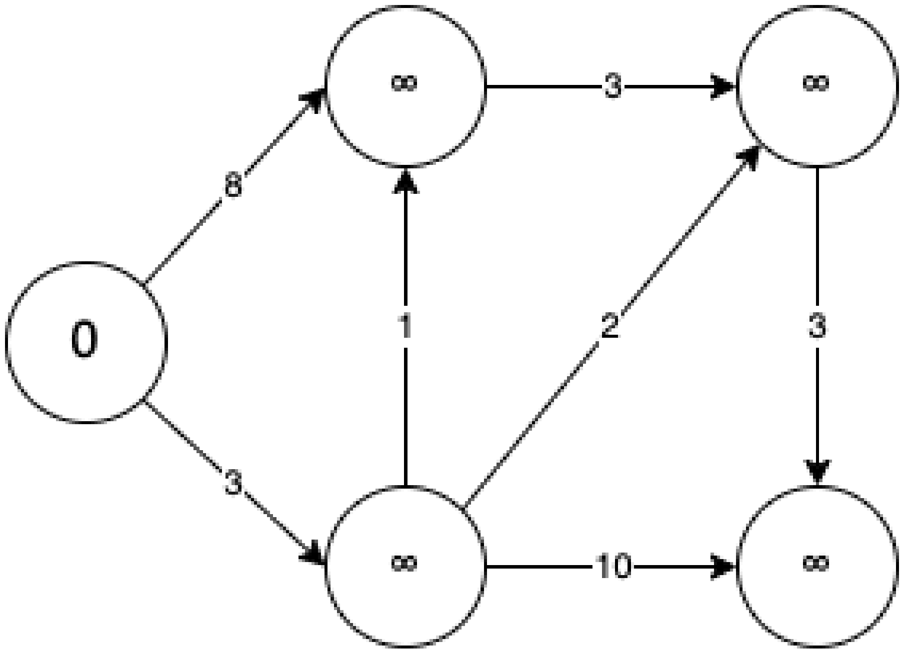
Dijkstra‘s algorithm



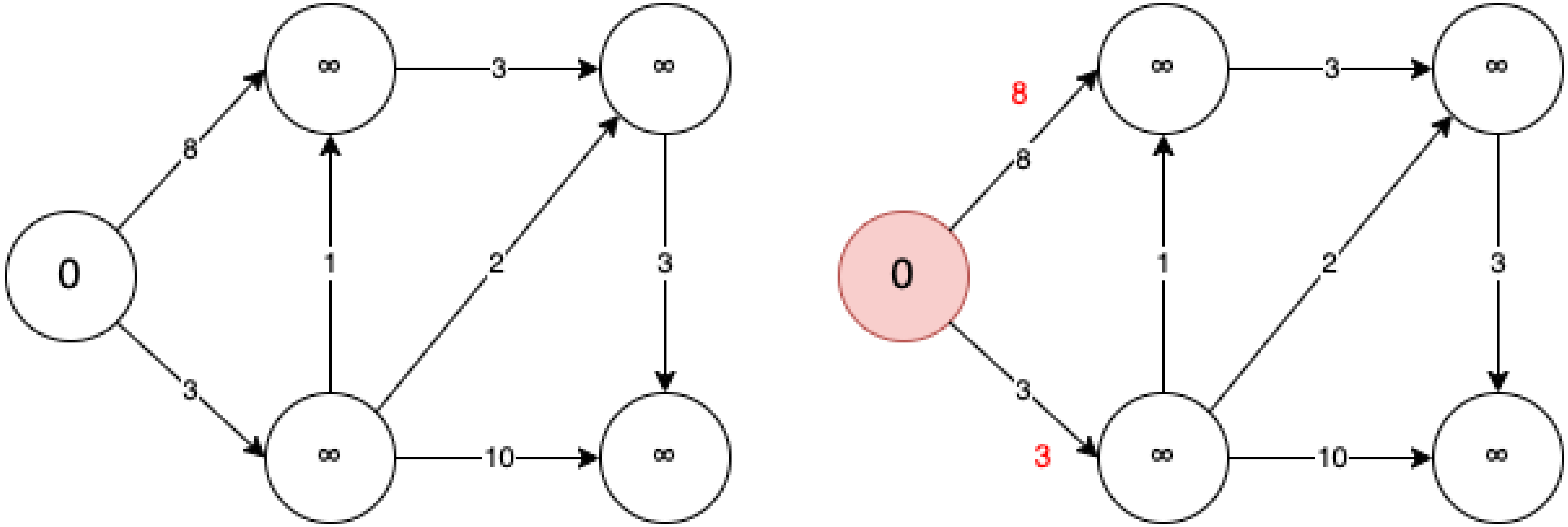
Recall

Parallel BFS in Pregel

You're given a weighted graph (Same as in previous task)
Please perform Parallel BFS in Pregel on it. Remember, that for every step, you should redraw the graph. (See the examples from the lecture)

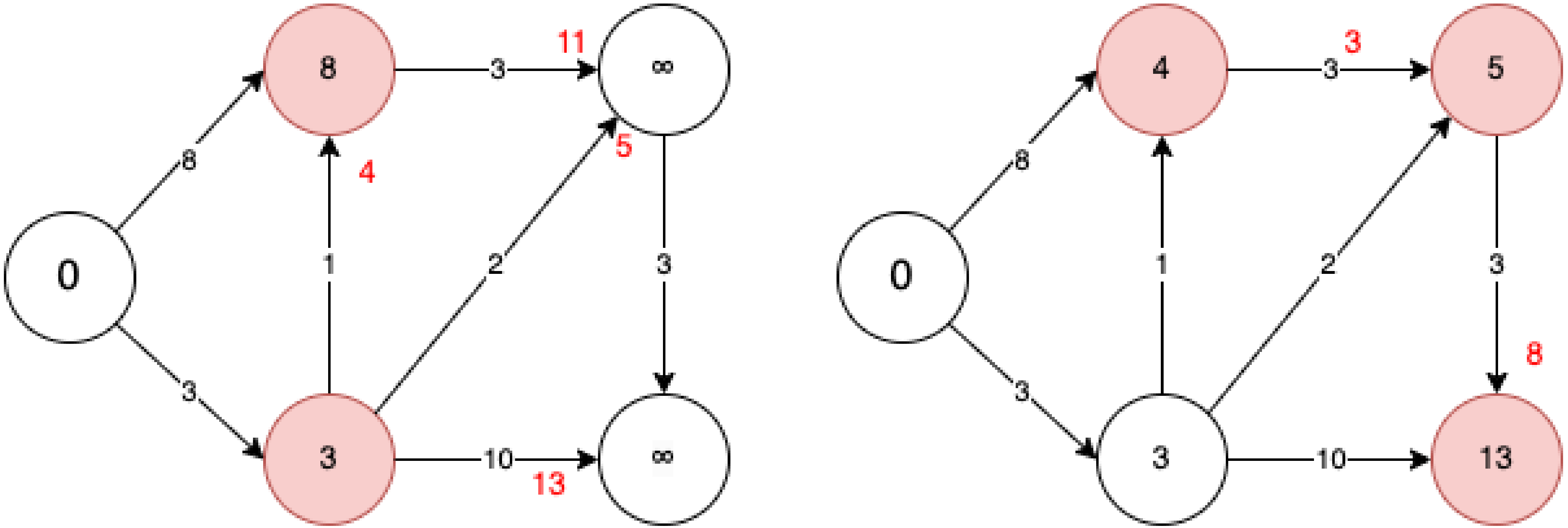


Parallel BFS in Pregel



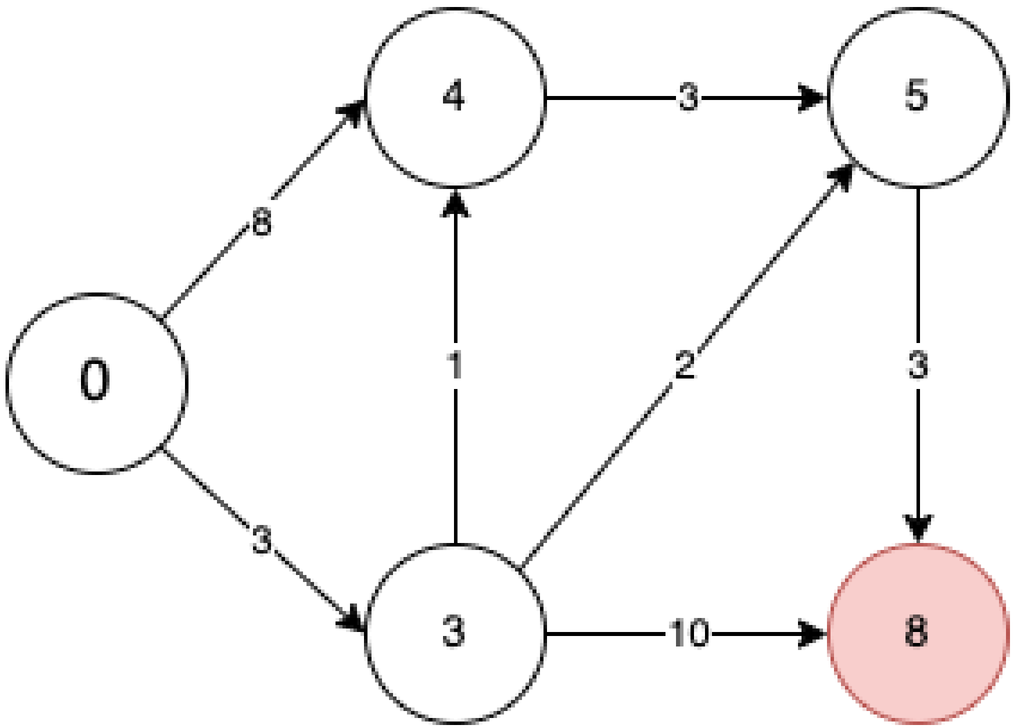
Recall

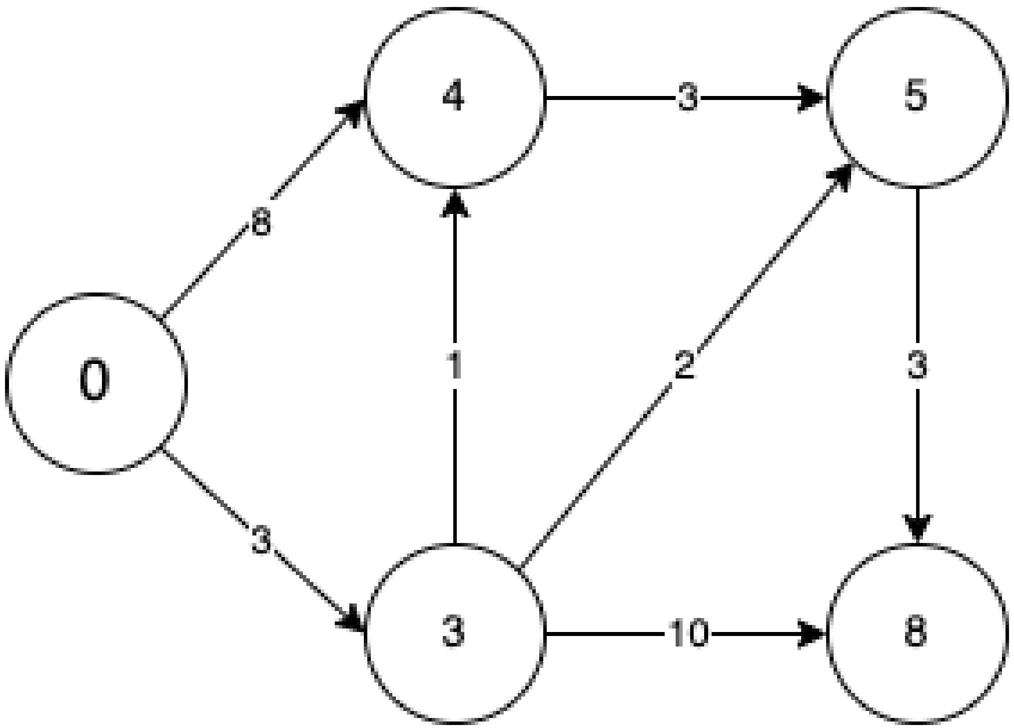
Parallel BFS in Pregel



Recall

Parallel BFS in Pregel

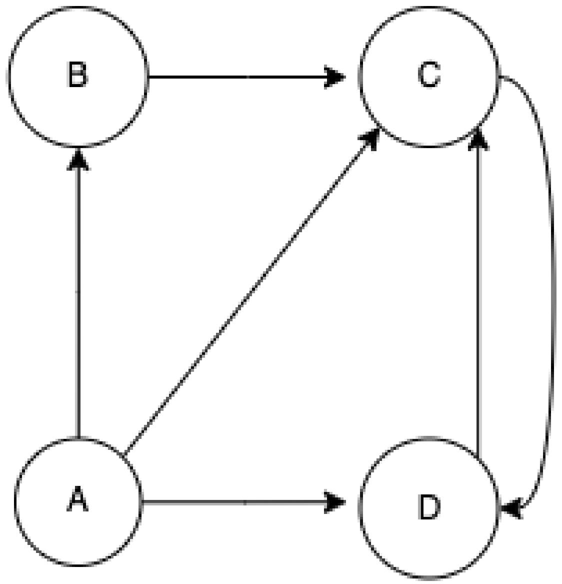




Recall

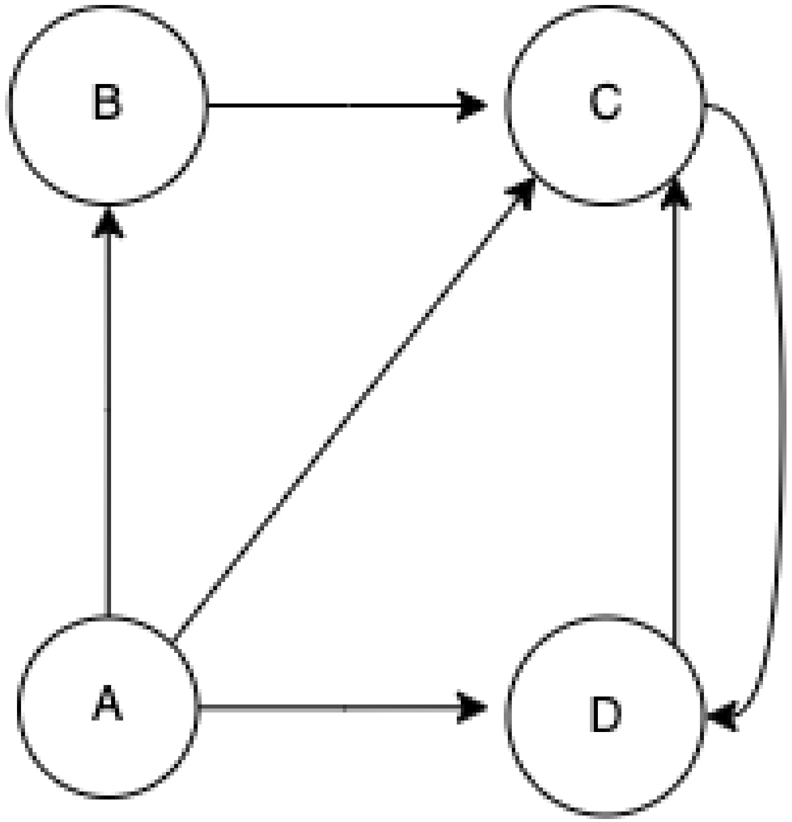
Page Rank

You are given a small network of 4 web pages - A, B, C and D. The network is modeled as a graph, where pages are represented as modes and links as edges.
Your task is to calculate Page Ranks for A, B, C and D in 2 iterations.



Recall

Page Rank



$$PR'_n = c \cdot \sum_{m \in S_n} \frac{PR_m}{outdegree_m}$$

OR

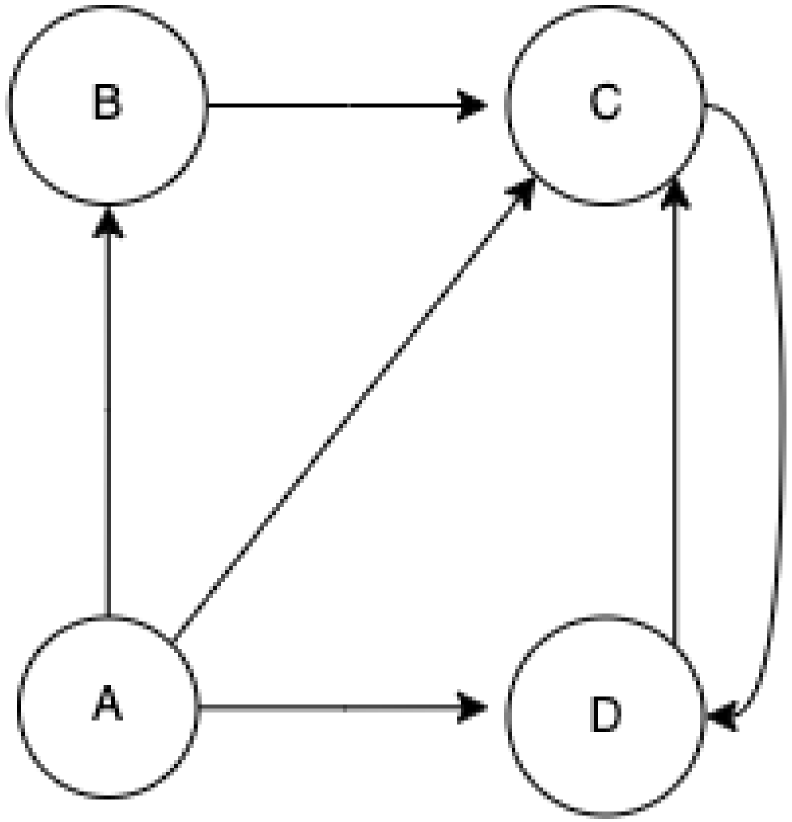
$$PR(A) = 1 - d + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

Damping factor (*d*) - The probability, at any step, that the person will continue following links is a damping factor d.

We assume here *d=1*

Recall

Page Rank

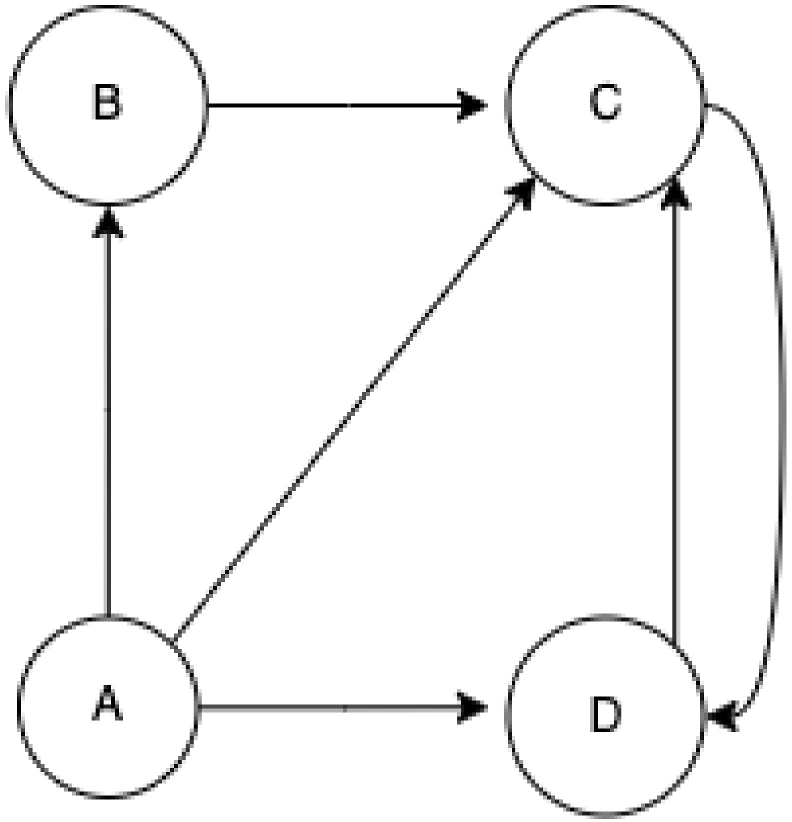


$$PR'_n = c \cdot \sum_{m \in S_n} \frac{PR_m}{outdegree_m}$$

Node	Initial	Iteration 1	Iteration 2
A	1/4		
B	1/4		
C	1/4		
D	1/4		

Recall

Page Rank



$$PR'_n = c \cdot \sum_{m \in S_n} \frac{PR_m}{outdegree_m}$$

Node	Initial	Iteration 1	Iteration 2
A	1/4	0	0
B	1/4	1/12	0
C	1/4	7/12	5/12
D	1/4	4/12	7/12

Recall

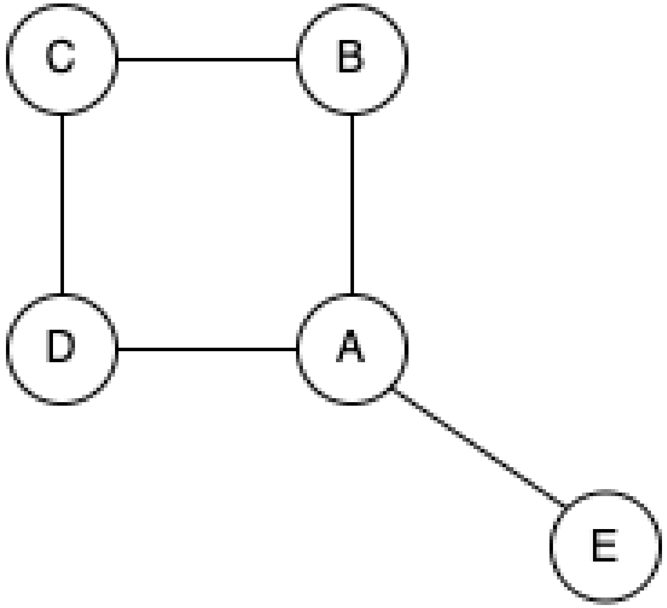
Betweenness centrality

You are given a graph. Calculate the Betweenness centrality of the node C. Don't forget to provide the formula of Betweenness centrality as well.

Recall

Betweenness centrality

$$C_b(i) = \sum_{\substack{j<k \\ j \neq i \neq k}} \frac{d_{j,k}(i)}{d_{j,k}}$$



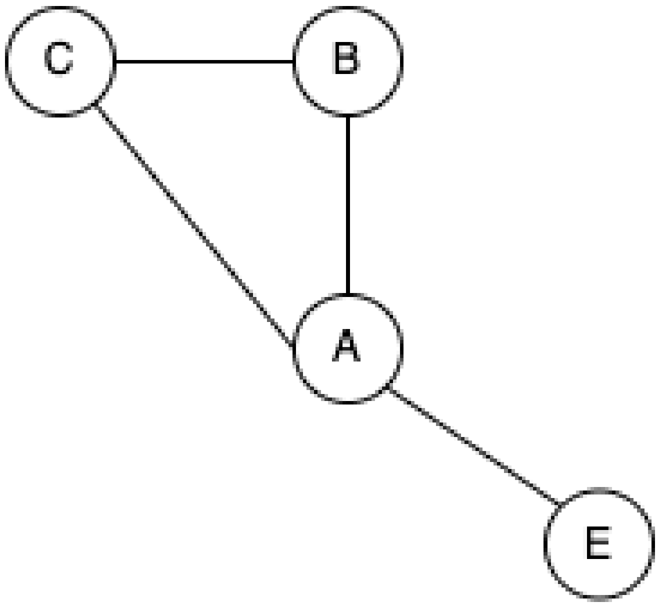
$$C_b(c) = \frac{d_{b,d}(c)}{d_{b,d}} = \frac{1}{2}$$

Based on slides from Prof. Steffen Staab

Recall

Betweenness centrality

$$C_b(i) = \sum_{\substack{j<k \\ j \neq i \neq k}} \frac{d_{j,k}(i)}{d_{j,k}}$$



$$C_b(a) = \frac{d_{b,d}(a)}{d_{b,d}} + \frac{d_{b,c}(a)}{d_{b,c}} + \frac{d_{c,d}(a)}{d_{c,d}} = \frac{1}{1} + \frac{0}{1} + \frac{1}{1} = 2$$

$$C_b(d) = 0$$

Based on slides from Prof. Steffen Staab

Knowledge Questions

Dijkstra's algorithm limitations

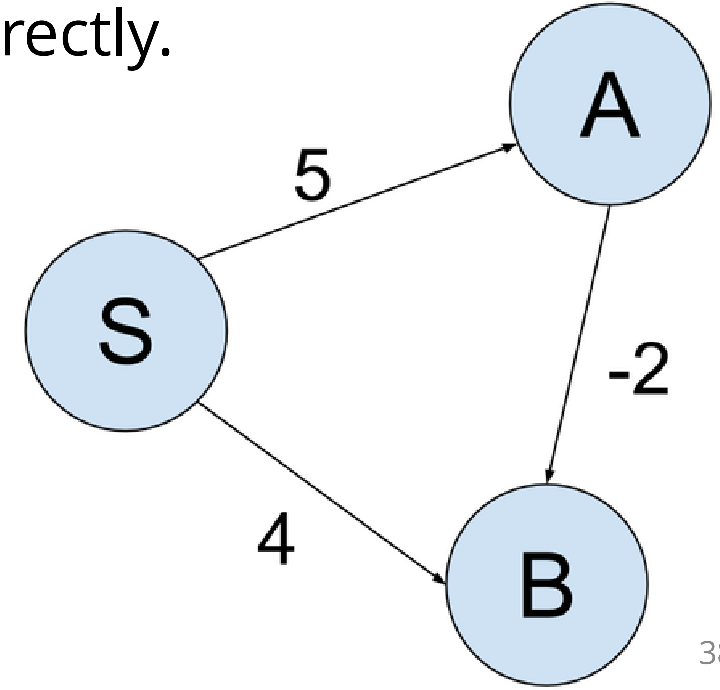
What are the limitations of Dijkstra's algorithm? When is it impossible to use it?

Knowledge Questions

Dijkstra‘s algorithm limitations

What are the limitations of Dijkstra‘s algorithm? When is it impossible to use it?

When working with graphs that have negative weights, Dijkstra’s algorithm fails to calculate the shortest paths correctly.



Knowledge Questions

Graph Partition

Based on further reading materials explain what is Graph Partition and how it works

Graph Partition is used to separate a graph into several subsets or partitions depending on preset criteria.

The process of graph partitioning involves the following steps:

1. Graph representation: Start with a graph consisting of nodes and edges.
2. Partitioning objectives: Define the objectives for partitioning, such as minimizing cut size or maximizing modularity, based on the specific requirements of the application.
3. Partitioning algorithms: Apply partitioning algorithms, such as spectral methods, Kernighan-Lin algorithm, or multilevel algorithms.
4. Partition optimization: Evaluate the resulting partitions based on the defined objectives and refine them if necessary.

Q&A

> That's all, folks!