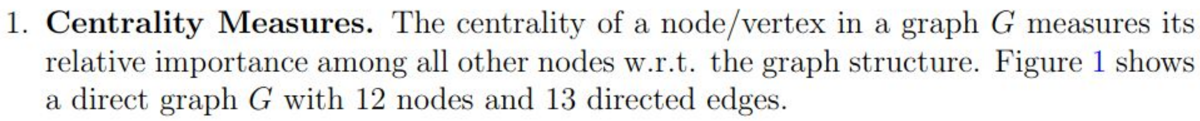
NUS | Computing

National University of Singapore

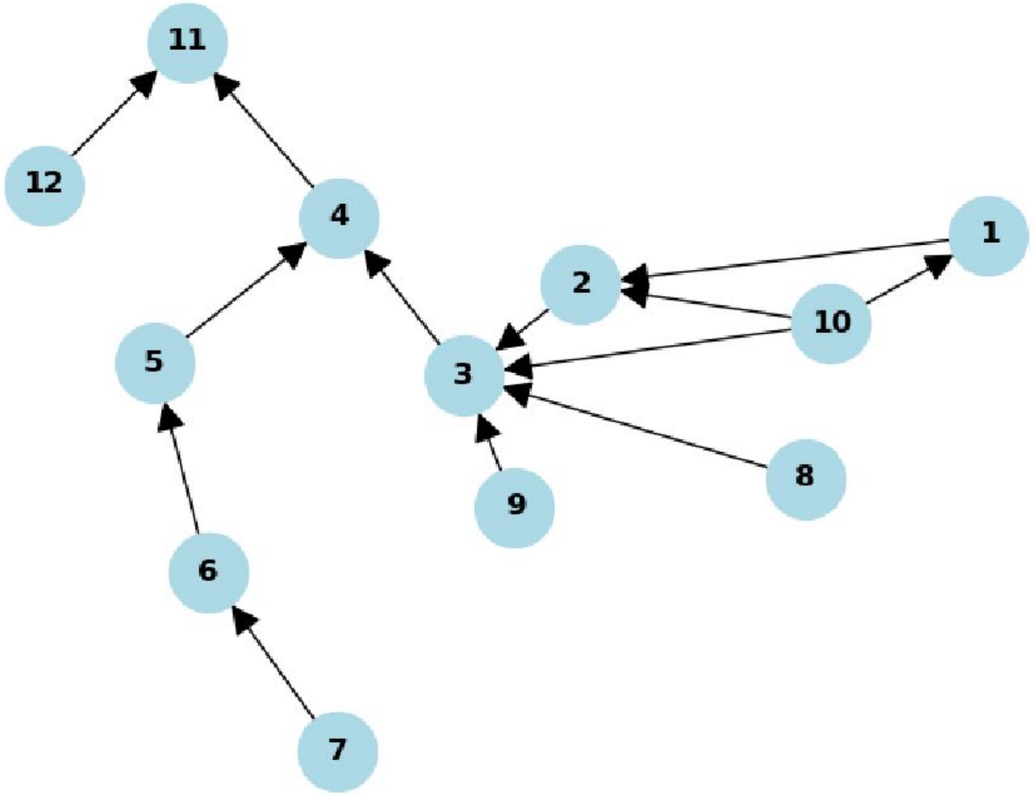
CS5228: Knowledge Discovery and Data Mining

Tutorial 9 — Graph Mining

Question 1



1. **Centrality Measures.** The centrality of a node/vertex in a graph G measures its relative importance among all other nodes w.r.t. the graph structure. Figure 1 shows a direct graph G with 12 nodes and 13 directed edges.



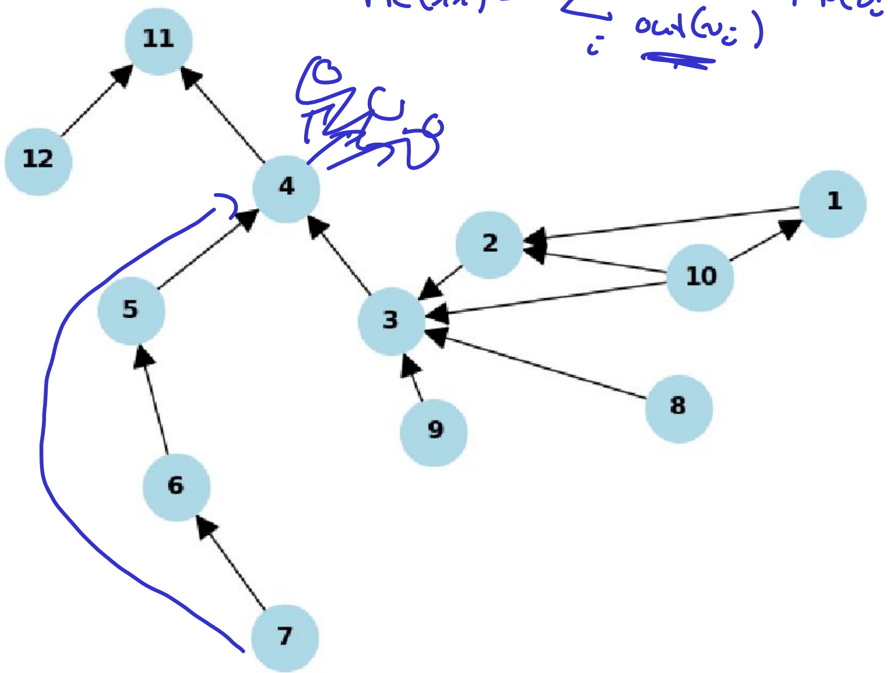
Question 1

$$PR(11) = \frac{1}{1} PR(4)$$

(a) Simply by eye-balling graph G in Figure 1 try to identify the nodes with the highest score according to the following 5 centrality measures

- OutDegree: 10
- InDegree: 3
- PageRank: 11
- Closeness: 4
- Betweenness: 3

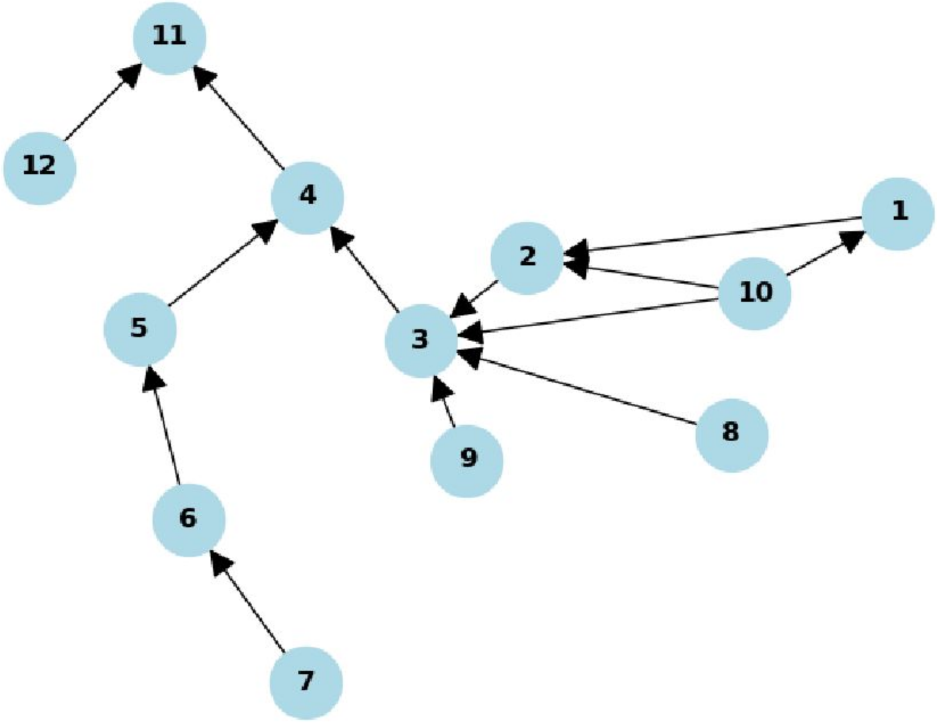
$$PR(11) = \sum_i \frac{1}{out(v_i)} PR(v_i)$$



Question 1

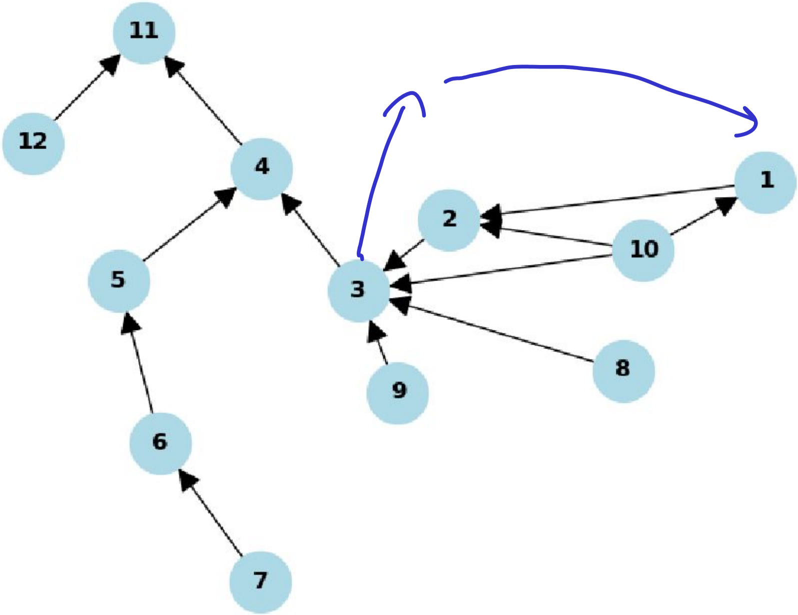
Solution

- OutDegree: 10
- InDegree: 3
- PageRank: 11
- Closeness: 4
- Betweenness: 3

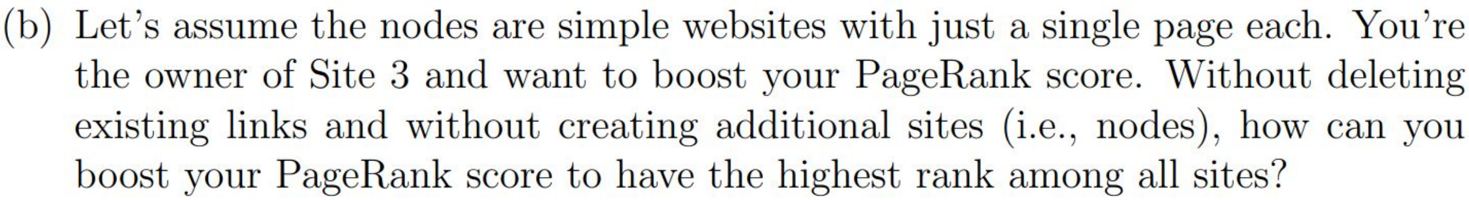


Question 1

(b) Let's assume the nodes are simple websites with just a single page each. You're the owner of Site 3 and want to boost your PageRank score. Without deleting existing links and without creating additional sites (i.e., nodes), how can you boost your PageRank score to have the highest rank among all sites?



Question 1



(b) Let's assume the nodes are simple websites with just a single page each. You're the owner of Site 3 and want to boost your PageRank score. Without deleting existing links and without creating additional sites (i.e., nodes), how can you boost your PageRank score to have the highest rank among all sites?

Solution

- Any link from Node 3 to any node that links to 3 will already do the trick (e.g., 3 → 9).
- Note: Nodes 11 and 4 have a higher PageRank because they can only be reached from Node 3.
- Any additional link from Node 3 to a node that (directly or indirectly) links back to 3 immediately emphasizes 3 and de-emphasizes 4 and 11.

Question 1

(c) PageRank has become famous for ranking websites w.r.t their relative importance compared to other sites. The underlying intuition is that a website is important (or trusted or authoritative) if many other important websites link to it. However, the idea of ranking nodes is of course not limited to websites, and there are also many other centrality measures to quantify a node's importance considering different aspects of the graph structure. For the following 5 centrality measures, for which application or data mining task would a specific measure arguably be the best choice?

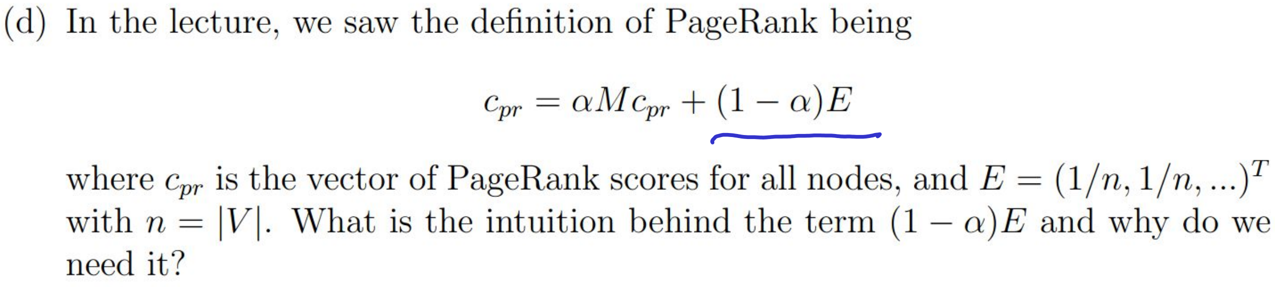
- OutDegree
- InDegree → (social media)
- PageRank → websites, social media
- Closeness
- Betweenness

Question 1

"Solution"

- OutDegree/InDegree: Any time only the direct neighborhood is important (e.g., the number of incoming or outgoing connections of an airport).
- PageRank: Given the follower network on Twitter (directed graph) a user with a high PageRank is arguably an influencer.
- Closeness: Given a traffic network of train stations, bus stops or roads (nodes are the intersections), a node with a high Closeness score would indicate a good location to build a hospital since this node can quickly be reached from anywhere else.
- Betweenness: Network of Internet router. A router with a high Betweenness score has to passed a lot of data between other routers (connected to PCs). Such routers should be particularly well maintained and checked.

Question 1

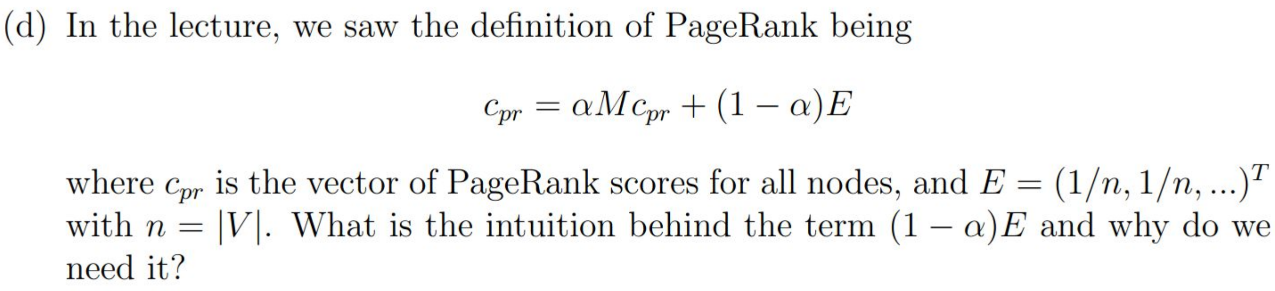


(d) In the lecture, we saw the definition of PageRank being

$$c_{pr} = \alpha M c_{pr} + \underline{(1 - \alpha)E}$$

where c_{pr} is the vector of PageRank scores for all nodes, and $E = (1/n, 1/n, ...)^T$ with $n = |V|$. What is the intuition behind the term $(1 - \alpha)E$ and why do we need it?

Question 1



(d) In the lecture, we saw the definition of PageRank being

$$c_{pr} = \alpha M c_{pr} + (1 - \alpha)E$$

where c_{pr} is the vector of PageRank scores for all nodes, and $E = (1/n, 1/n, ...)^T$ with $n = |V|$. What is the intuition behind the term $(1 - \alpha)E$ and why do we need it?

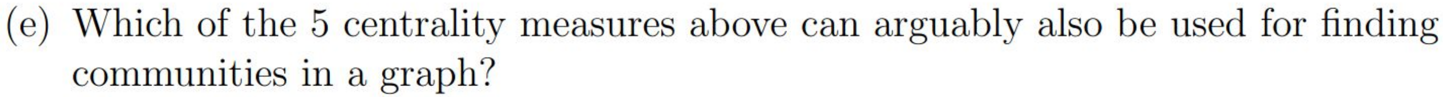
Solution

- Power Iteration method requires the directed graph to be (strongly) connected
- Problem: the Web Graph is not strongly connected (some pages have no incoming and/or outgoing links)
- *(1-α)E* introduces ”virtual links” between websites, making the graph always (strongly) connected.

Question 1

(e) Which of the 5 centrality measures above can arguably also be used for finding communities in a graph?

Question 1



(e) Which of the 5 centrality measures above can arguably also be used for finding communities in a graph?

Solution

- Recall: the Girvan-Newman algorithm for community detection relies on the notion of Edge Betweenness to find the edges that should be removed
- Edge Betweenness is naturally tightly connected to the Betweenness of a node.
- As such, nodes that connect 2 communities are likely to have high Betweenness scores.