CS5228 – Tutorial 9

Graph Mining

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.

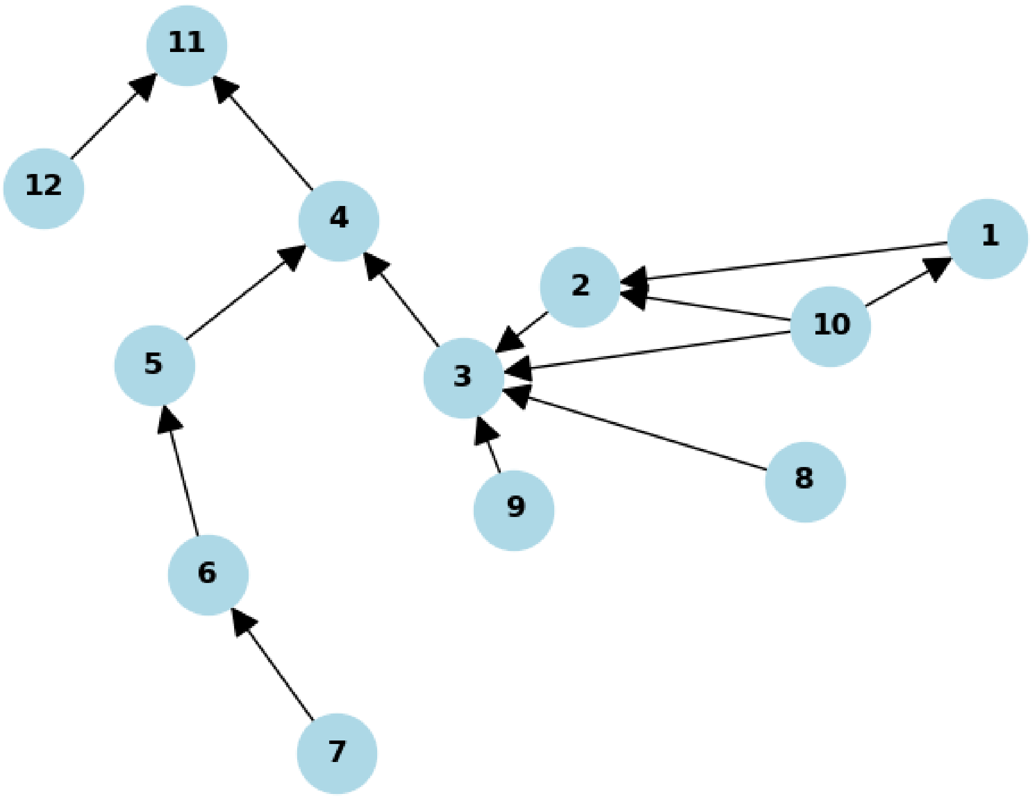

Figure 1: Example of a directed graph G

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

- OutDegree:
- InDegree:
- PageRank:
- Closeness:
- Betweenness:

Solution:

- OutDegree: 10 (the only node with more than 1 outgoing edge)
- InDegree: 3 (clearly the node with the most incoming edges)
- PageRank: 11 (other important nodes link directly or indirectly to 11; a random surfer is likely to visit Node 11)
- Closeness: 4 (Node 4 is kind of the merging point of 2 subgraphs: [4, 5, 6, 6] and [1, 2, 3 ,4, 8, 9, 10], making the easiest to reach on average)
- Betweenness: 3 (Node 3 benefits from its high InDegree, only slightly higher than score for Node 4)

(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: In this toy example, any link from Node 3 to any node that links to 3 will already do the trick (e.g., a link/edge 3 → 9). One of the main reasons why Nodes 11 and also 4 have a higher PageRank is because this is the only path from Node 3. So 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.

(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
- PageRank
- Closeness
- Betweenness

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 pass a lot of data between other routers (connected to PCs). Such routers should be particularly well maintained and checked.

(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: Calculating the PageRank using the Power Iteration method requires the directed graph to be (strongly) connected, i.e., each node can be reached from each node via some path. However, there are many websites without outgoing links and many websites that are not linked to. The term $(1 - \alpha)E$ specifies that there is a small probability to jump from one node to any other node in the graph. In some sense, this term introduces ”virtual links” between websites, making the graph always (strongly) connected.

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

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