CSC303: A2

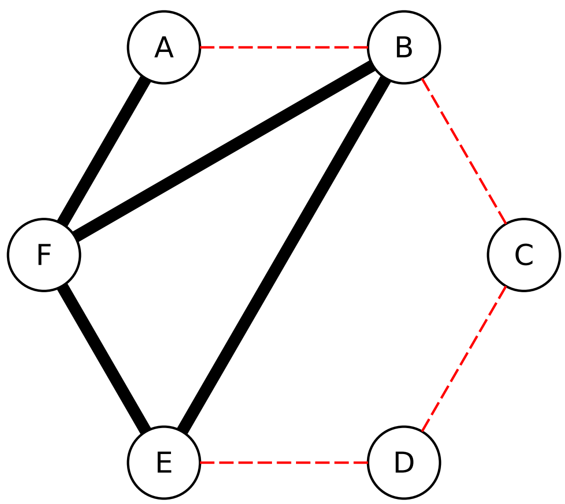
Due Mar 26 at 11:50PM, EST

Be sure to include your name and student number with your assignment. All assignments are to be submitted on Markus.

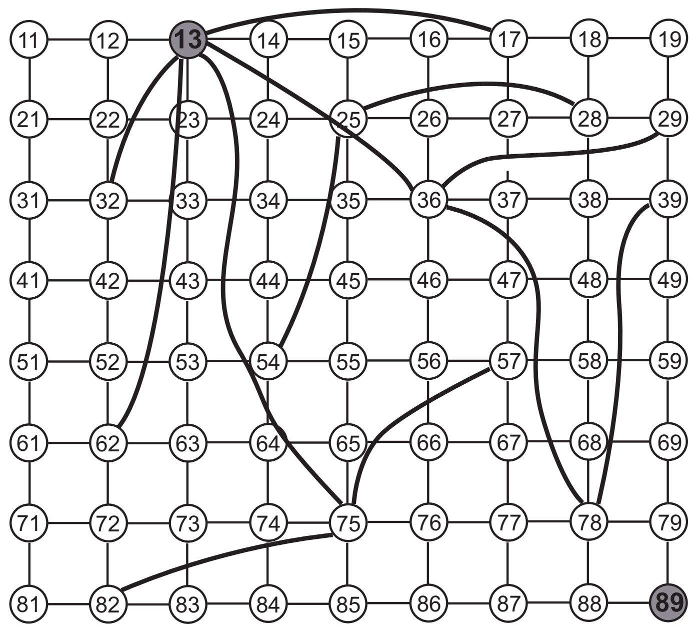
You will receive 20% of the points for any (sub)problem for which you write I do not know how to answer this question. You will receive 10% if you leave a question blank. If instead you submit irrelevant or erroneous answers you will receive 0 points. You may receive partial credit for the work that is clearly on the right track.

Question 1: (10 Points) Consider the following signed network where a solid edge denotes friendship (an edge labeled +) and a dashed edge represents enemies (an edge labeled −).

Specify the minimum number of edge signs (if any) that need to be changed so that the network can be completed to a strongly balanced network. Provide an explanation as to how the initial or modified network can be completed into a strongly balanced one.



Question 2: (10 points) Consider the following communication network



Recall from class, the process of decentralized search. In decentralized search, if a node n is asked to forward a message so that it will reach a target node t quickly, it must forward the message to one of its friends f (who will then continue the process). Node n will forward the message to the friend f that is closest to target node t, where closeness is measured by grid distance (or city block distance). The grid distance is simply the length of (smallest) path between f and t using only local edges (thin edges in the picture). If there are several friends f that are equally close to the target, n can send its message any one of these friends.

(a) [5 points] 13 is trying to get a message to node 89 using the decentralized search process. What path will the message take? (Note: There may be more than one acceptable answer but you only need to provide one path.). How many hops (links) will the message need to traverse?

(b) [5 points] What is a shortest path that the message from 13 to 89 could take (not using decentralized search)? (Note: There may be several different shortest paths; just list one). How long is it?

(c) [5 points] For each of the following scenarios, is the decentralized or shortest path more plausible? Why?

  (i) Nodes represent sorting facilities for mail. Local edges are truck routes and long-distance edges are rail routes. All edges take the same amount of time to traverse. An express package must be sent from node 13 to node 89.

  (ii) Nodes represent students in a class of high school graduates who graduated in 2010. All edges represent connections on social media, and roughly correspond to friendships that existed during highschool. Person 13 is trying to send person 89 an invitation to a high school reunion. Furthermore, person 89's social media account is under an alias, and thus they cannot be found directly through the social media's search function.

Question 3: (15 points)

A music retailer runs an online site for downloading popular international songs not well known in Canada. The retailer is charging only $.25 cents per download. The retailer will not let you listen to the music in advance, but does provide lots of information on international sales.

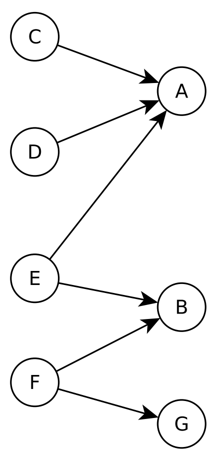
After the first month, the retailer's most popular download has sold 10,000 times. The retailer has also observed that the number of sales is satisfying a power law with exponent 1.

(a) [5 points] Estimate the number of sales of the second most popular download. Briefly explain your estimate.

(b) [5 points] Estimate the total number of sales of the top 100 downloads. Briefly explain your estimate.

(c) [5 points] Provide an explanation as to why the number of downloads is following a power law.

Question 4: (20 points) The following is web graph with hubs C, D, E and F, and authorities A, B, and G. (This graph and question is a slight modification of Question 3 in Sec. 14.7 of the text.)



(a) [5 points] Show the (hubs and authorities) values obtained by running two rounds of the hubs and authorities algorithm on this network. Show the values both before and after the final normalization step, in which we divide each authority score by the sum of all authority scores, and divide each hub score by the sum of all hub scores. You may write the normalized scores as fractions rather than decimals.

(b) [15 points] Suppose you wish to add a new web site X to the network and want it to have as high an authority score as possible. You also have the ability to create another new web site Y which we can use as a hub to elevate the score of X. Consider the following three options:

- You add a link from Y to X (and add no other new links).
- You add a link from Y to A and X (and no other links).
- You add a link from Y to A, B, G and X (and no other new links).

For each option, show the normalized authority values that each of A, B, G and X obtain when you run 2-steps of hub-authority computation on the resulting network (as you did in part (a)). Which option gives X the highest (normalized) score?

Question 5: (10 points) Question 6 of Chapter 14 in the EK text. Reproduced below:

One of the basic ideas behind the computation of hubs and authorities is to distinguish between pages that have multiple reinforcing endorsements and those that simply have high in-degree. (Recall that the in-degree of a node is the number of links coming into it.)

Consider for example the graph shown in Figure 14.22. (Despite the fact that it has two separate pieces, keep in mind that it is a single graph.) The contrast described above can be seen by comparing node D to nodes B1, B2, and B3: whereas D has many in-links from nodes that only point to D, nodes B1, B2, and B3 have fewer in-links each, but from a mutually reinforcing set of nodes.

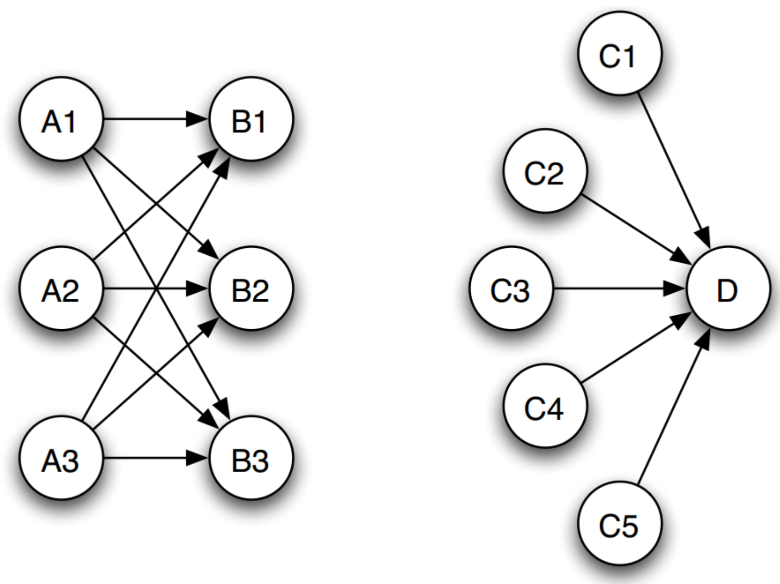


Figure 14.22:

Lets explore how this contrast plays out in the context of this stylized example.

(a) Show the values you get from running the 2-step hub-authority computation from the chapter on link analysis. (If you want, you can omit the final step in which the values are normalized; i.e., you can just leave the values as large numbers.)

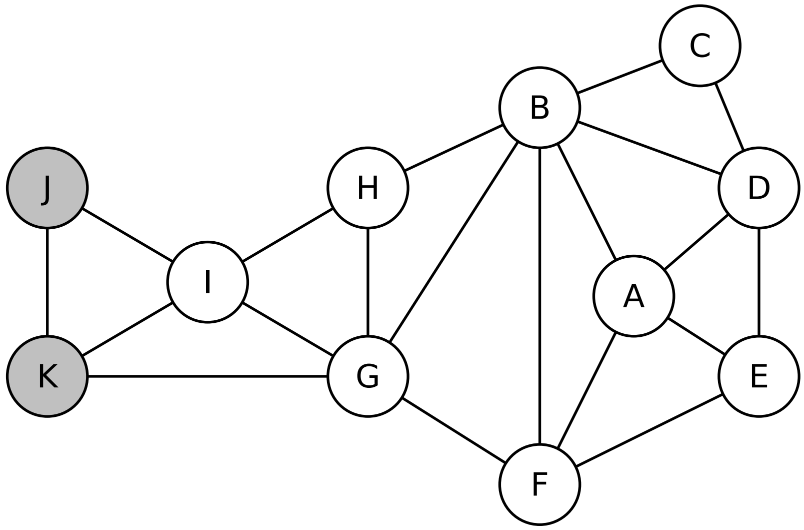
(b) Give formulas, in terms of k, for the values at each node that you get from running the k-step hub-authority computation. (Again, if you want, you can omit the final step in which the values are normalized, and give the formulas in terms of k without normalization.)

(c) As k goes to infinity, what do the normalized values at each node converge to? Give an explanation for your answer; this explanation does not have to constitute a formal proof, but it should argue at least informally why the process is converging to the values you claim. In addition to your explanation of whats happening in the computation, briefly discuss (in 1-2 sentences) how this relates to the intuition suggested in the opening paragraph of this problem, about the difference between pages that have multiple reinforcing endorsements and those that simply have high in-degree.

Question 6: (10 points)

Recall the threshold influence model from class. In this model, nodes adopt a new idea if the fraction of their neighbours having adopted the idea meets or exceeds some threshold, q.

Assume that the following network defines such a threshold influence model, where the initial adopters are J & K, and $q = \frac{1}{3}$ (i.e. a node adopts the new idea if one third or more of their neighbours have adopted the idea).



(a) [5 points] What edge could be removed to the network to prevent the node A from adopting the new idea? Why does removing this edge prevent A from adopting?

(b) [5 points] What edge could be added to the network to prevent the node A from adopting the new idea? Why does adding this edge prevent A from adopting?

Question 7: (25 points)

This final question revolves around a strange and terrible tale that strikes fear into the hearts of computer scientists and engineers everywhere – the dread tale of the hardware virus! Or more accurately, the hardware prion.

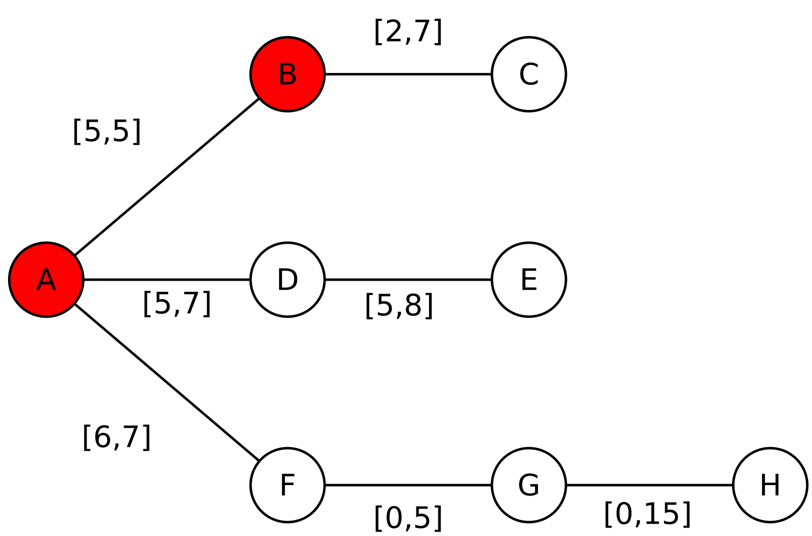
Once upon a time, in a faraway land, unsuspecting engineers daily used DVI-to-VGA adapters to connect their laptops (which only had DVI connections) to various VGA projectors. Little did they know of the horrors that would be unleashed upon them! One day, a single pin was bent in an adapter. This bent pin, when forced into a laptop's DVI port, caused the corresponding hole of the port to break. When a new adaptor was inserted into the laptop's DVI port, the broken port would bend the pin of the adapter in the same way. Thus, the infection would spread from laptop, to adapter, to laptop, and so on.

In this question we will model these events with a transient contact spread model. In the following network, nodes are pieces of hardware (either laptops or DVI adapters), and edges indicate where one piece of hardware has been connected to another (i.e. an adapter has been inserted into a laptop). The times on the edges indicate the days on which these interactions have occurred.

On day 5, the unimaginable happens. Some careless user forces laptop B to connect with DVI adapter A in such a way that both the pins of the adapter, and the DVI port of the laptop, break. Both are now contagious in the manner described above. We will now model the spread of the hardware prion throughout the company's hardware.

In this simulation we will be pessimistic: assume that any infection occurs at the start of the day (e.g., if laptop B is infected on day 5, then it can potentially infect C on day 5). The "duration" of the infection is 3 days, after which the node cannot be reinfected. In other words, it takes 3 days for someone to diagnose the problem and to send the hardware for repair. Ergo, if a piece of hardware is infected on day 5, then it can potentially infect other pieces of hardware on day 5, day 6, or day 7. After this, the node is permanently removed from the network (i.e. it is no longer in the network from day 8 onward).

Additionally, assume that on each day that an uninfected node is exposed to an infected node, there is an independent 60% chance that the infection will spread.



(a) [20 points] For every node other than A and B, calculate the probability that it was infected some time prior to day 16.

(b) [5 points] Why is it impossible for the edge (C, G) to be added to this network? Is there a corresponding concept in graph theory?