CSC303: A2

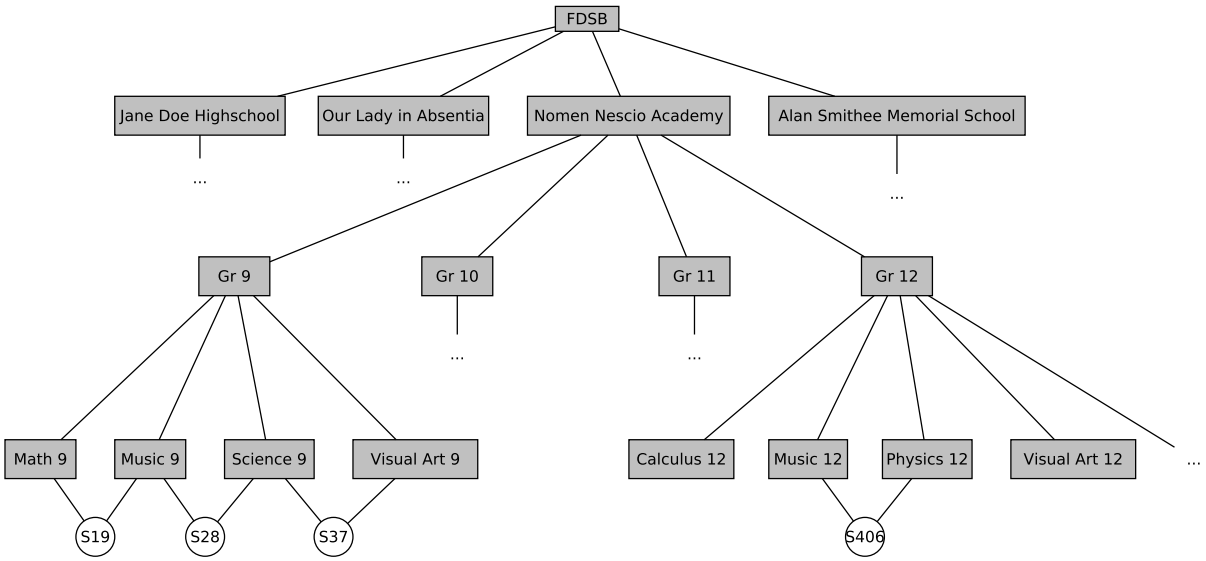
Due Mar 28 at 11:50PM, Toronto Time (EDT)

Be sure to include your name and student number with your assignment, *on the last page. The last page should contain only your name and student number (i.e., the last page should have no solutions).* This is done as part of an effort to combat unconscious bias in grading. 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." If instead you submit irrelevant, erroneous, or blank answers then you will receive 0 points. You may receive partial credit for the work that is clearly "on the right track."

Note: There is a LaTeXstarter file for this assignment with the questions and diagrams in Quercus, under the Files tab.

Question 1: (10 Points) Consider the following graph, representing the Fictitious District School Board (FDSB). The FDSB contains 4 schools. Each school is subdivided into 4 grades (Grades 9 through 12). Each grade has a given number of courses (e.g. "Math 9" for Grade 9 math). In the FDSB, each class for a grade contains only students from that grade (i.e., there are no classes that contains students from different grades). In the FDSB, for a given grade g, each grade g student is enrolled in two courses of the same grade. In the diagram we display three grade 9 students (S19, S28, and S37) as well as one grade 12 student (S406). Although not pictured, the FDSB, and Nomen Nescio Academy (N.N.) have many more students in all grades.



We are interested in the social distance between students in the FDSB, and decide to use this organizational chart to define social focii (i.e., each class, grade, school, and the FDSB itself are all considered to be separate social focii).

(a) What is:

(i) (1 point) The social distance between the students S19 and S28?

(ii) (1 point) The social distance between S28 and S406?

(iii) (1 point) The social distance between S19 and S37?

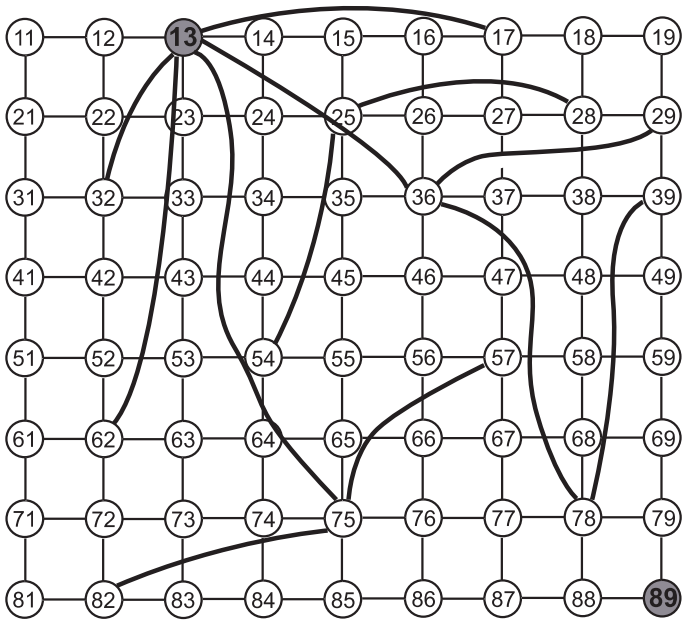
(iv) (1 point) The largest possible social distance?

(v) (1 point) The smallest possible social distance?

In all cases, be as precise as you can, and state any assumptions.

(b) (5 points) Assume you use this organizational chart (and the implied social focii) to collect data on the social distance between all pairs of students in the FDSB student body. Assume we also collect information about which pairs of students are friends. You plot this combined data as a log-log plot; the x-axis is a social distance, s, and the y-axis is the proportion of pairs of students at social distance s who are also friends (i.e., on the x-axis you have log social distance, and on the y-axis you have the log proportion of pairs of students at this social distance who are friends). What pattern in the plot would suggest that this social network supports efficient decentralized search? Briefly justify your answer.

Question 2: (15 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] Provide 5 networks and scenarios, and briefly explain for each whether the decentralized or shortest path more plausible. For example, the Milgram 6-degrees-of-separation experiment would be a contrived scenario (trying to send a letter without the recipient's address) for a social network of friendships. You are not allowed to use this example.

You must include at least 2 networks for which decentralized is more likely, and at least 2 for which shortest is more likely. You must provide 5 different networks in total (e.g., if one of your five examples involves a social media network, then your other examples cannot also be a social media network). In this example, decentralized search is more plausible than the shortest path. You can have examples with similar networks (e.g., digital vs. in-person, variants of transportation networks), but they cannot be exactly the same, and the scenarios should be different.

Question 3: (20 points)

Consider the following partial table of values, summarizing a piece of English text containing N unique words. First, the frequency of each unique word, w, was calculated (i.e., the proportion of all words, including repeats, in the text that were equal to w). For example, if we were to consider "the plane in the sky", then "the" occurs with frequency $2/5$, and all other words with frequency $1/5$.

After computing the frequencies of all unique words in our corpus, all the unique words were ranked from most frequent (rank 1), to least frequent (rank N). The frequencies of the 3 highest ranked words are reported in the table below.

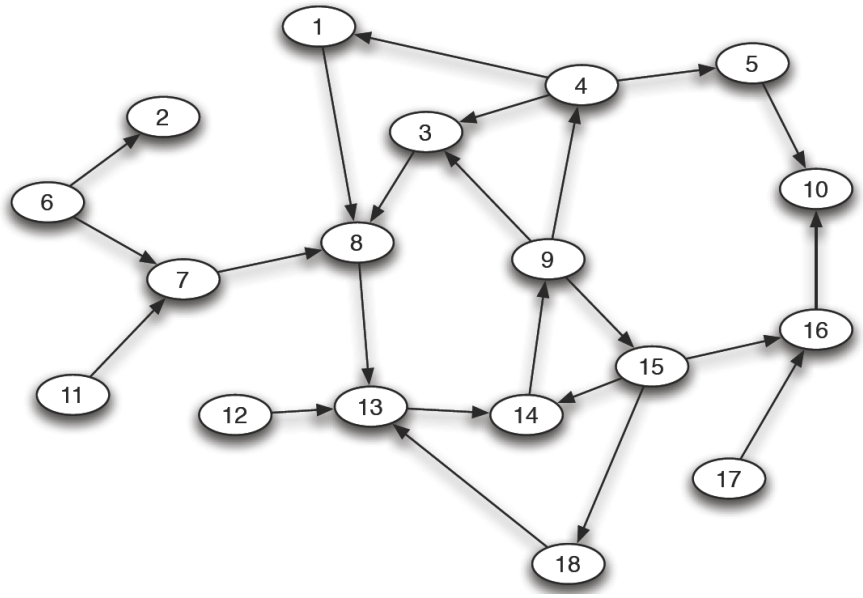
Word Rank	Word Frequency
1	0.1336
2	0.0668
3	0.0445
...	...
N	?

(a) [10 points] Does it seem more likely that N is equal to: 10 unique words, 100 unique words, 1000 unique words, $10,000$ unique words, $100,000$ unique words, or $1,000,000$ unique words? Justify your answer. (Note: for this question, you may need to use a computer or look up an approximation to complete this question – be sure to state any such resources you use).

(b) [5 points] Estimate the frequency of the least common word in the corpus.

(c) [5 points] Briefly, explain the similarities and differences between the preferential attachment model, and the distribution of word frequencies in a text.

Question 4: (20 points) Consider the following tiny example of a directed web graph G = (V; E) (this is a reproduction of Fig. 13.8 in the text):



(a) [10 points] Suppose we assign all 18 nodes equal initial page rank values of $\frac{1}{18}$. Now suppose we apply the unscaled Page Rank algorithm to this graph. When the algorithm converges (i.e., reaches equilibrium), which nodes in the graph will have non-zero page rank values? What will their equilibrium page rank values be? Justify your answer.

(b) [5 points] Consider a variation of the graph in part (a) in which a single node (19) is added. This node has in-degree and out-degree 1, each going to a different node. How could you add this node such that the PageRank of this node never goes to zero? State a lower bound on the PageRank of the node, and briefly justify. Note that PageRank need not converge to an equilibrium on the altered graph.

(c) [5 points] Suppose we use scaled Page Rank on the original graph in part (a) with scaling factor s = 0.9. Which nodes will have non-zero (equilibrium) page rank values? Qualitatively, briefly describe which node will now have the highest page rank value.

Question 5: (10 points) Modified from Question 6 of Chapter 14 in the EK text:

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.

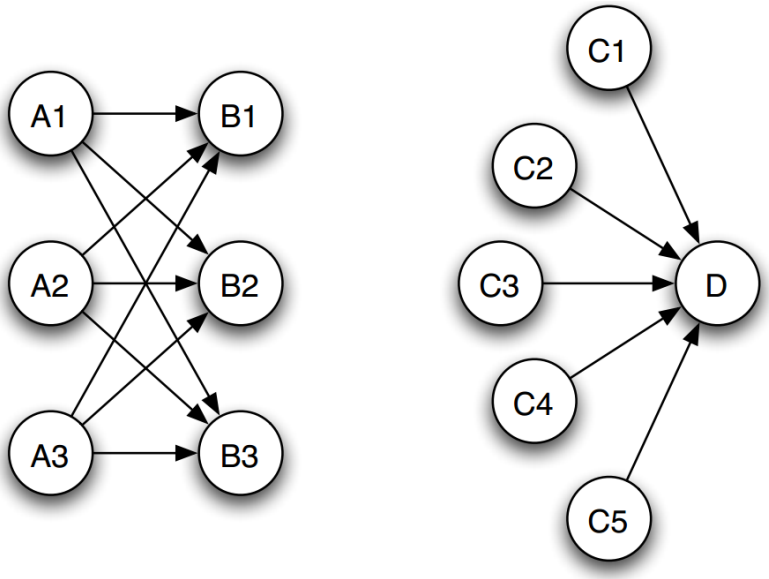


Figure 14.22:

Let's explore how this contrast plays out in the context of this stylized example.

(a) [2 points] Show the values you get from running the hub-authority update for two rounds. (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.)

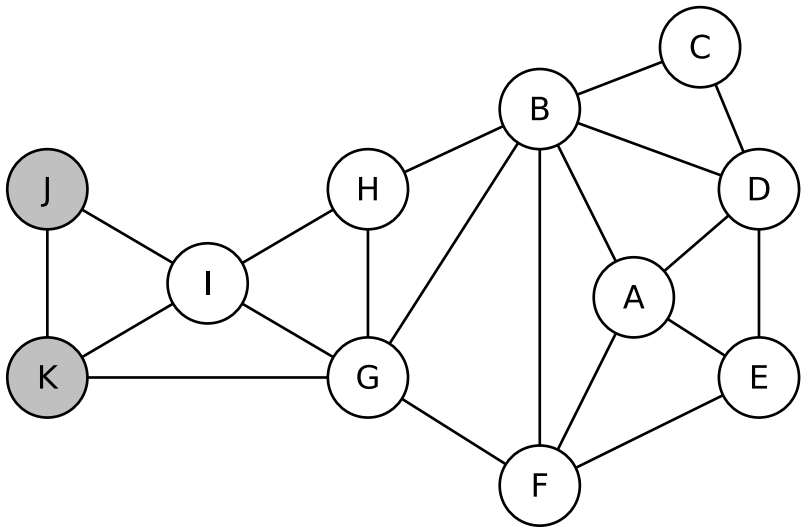
(b) [3 points] Give formulas, in terms of k, for the values at each node that you get from running the k-step hub-authority computation. (Again, if you want, you can omit the final step in which the values are normalized, and give the formulas in terms of k without normalization.)

(c) [5 points] 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 what's 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: (15 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 $q = \frac{1}{2}$ (i.e. a node adopts the new idea if one half or more of their neighbours have adopted the idea).



(a) [5 points] What is the density of the blocking cluster $\{A, B, C, D, E, F\}$? Briefly justify.

(b) [5 points] What is a minimal set of nodes that can be removed to ensure that A adopts the new idea? For the modified set of vertices V', state the density of the cluster of nodes $\{A, B, C, D, E, F\} \cap V'$ under your modified graph.

(c) [5 points] What is a minimal set of edges that could be added to the network to cause the node A to adopting the new idea? State the density of the cluster of nodes $\{A, B, C, D, E, F\}$ under your modified graph.

Question 7: (20 points)

This final question revolves around a strange and terrible legend 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.

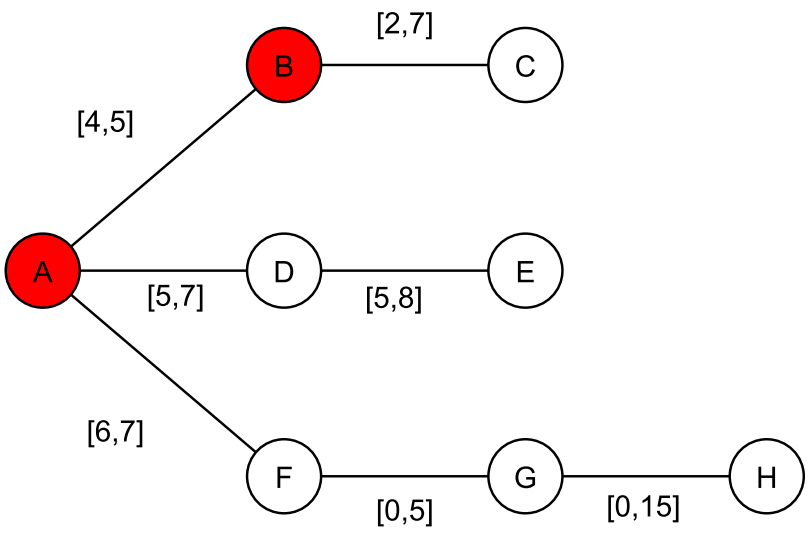
Once upon a time, in a faraway land, unsuspecting engineers daily used DVI-to-VGA adapters to connect their laptops (which only had DVI connections) to various VGA projectors. Little did they know of the horrors that would be unleashed upon them! One day, a single pin was bent in an adapter. This bent 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 the end of day 4, 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 – note that they broke at the end of day 4, and thus cannot start infecting other hardward until the start of day 5. We will now model the spread of the hardware prion throughout the company's hardware.

In this simulation we will be optimistic: assume that any infection occurs the *last* time the device is used for the day (e.g., if laptop B is infected on day 5, then it cannot potentially infect C until day 6). 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 6, day 7, or day 8. After this, the node is permanently removed from the network (i.e. it is no longer in the network from day 9 onward).

Additionally, assume that on each day that an uninfected node is exposed to an infected node, there is an independent 90% chance that the infection will spread.



(a) [15 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] Note that the graph is bipartite. Why is this? What is another situation in which a contact network could be expected to be bipartite?

END OF ASSIGNMENT 2

If you are typsetting the assignment using the provided LaTeX, then please write your name and student number below.

NAME: Your name should go here, on the last page.
STUDENT NUMBER: You student number should go here, on the last page.