Homework 1
Networked Life (NETS 112)
Fall 2018
Prof Michael Kearns

Posted September 26, 2018. Due in hard-copy format at the start of lecture on Thursday, October 11. Please don't forget to write your name and staple the pages together.

Collaboration of any kind is NOT permitted on the homework.

Your Name:

Problem 1. The following two websites:

https://mathscinet.ams.org/mathscinet/freeTools.html?version=2

https://oracleofbacon.org/

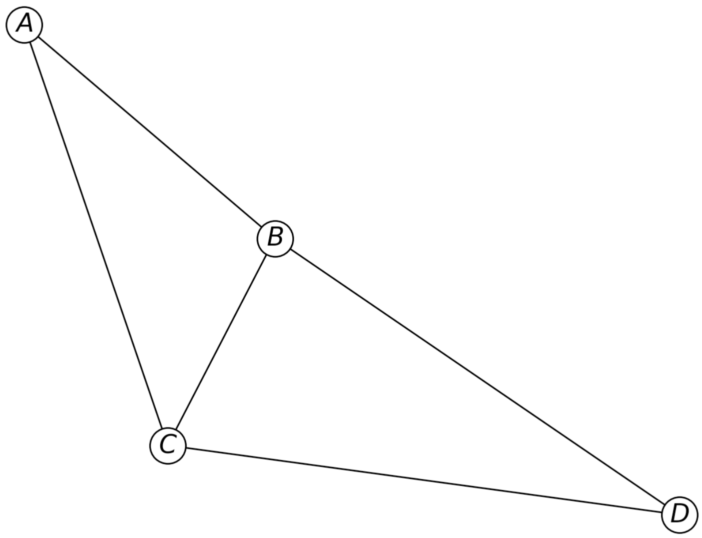
…each provide tools for computing shortest-path distances in collaboration networks (of mathematicians and actors, respectively). Use these tools to find the pair of mathematicians, and the pair of actors, with the ***largest*** shortest-path distance you can. The pairs you find must be in the same connected component --- i.e. there must be a finite distance between them.

You should provide screenshots documenting the longest distances you are able to find, and the paths found by the tools. You should provide a detailed description of the methodology and ideas you used to find your longest distances. Your methodology may include any material or research you like --- the sites themselves, information you find on the open Internet, systematic or random exploration, etc.

This problem will be graded on a combination of the actual distances you're able to find (larger is better), the creativity of your methodology, and your clear description of that methodology.

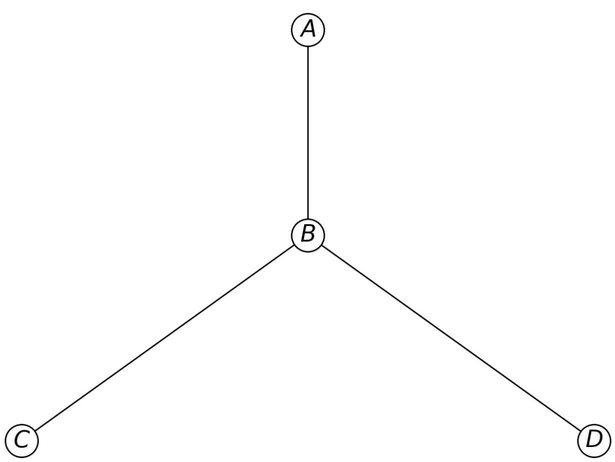
A prize of some kind (to be determined) will be awarded to the student(s) who find the largest distances.

Problem 2. Mark Zucchini, after taking NETS 112, is now interested in how content on Facebook is reshared throughout the network. Given the underlying graph of the social network and the timestamps at which the content is shared, Mark wants to see what the corresponding ***cascade tree*** looks like. For instance, here's the underlying social network of {A,B,C,D} and the timestamps at which some content was shared.



Time t	*people who shared at t*
1	A
2	B
3	C, D

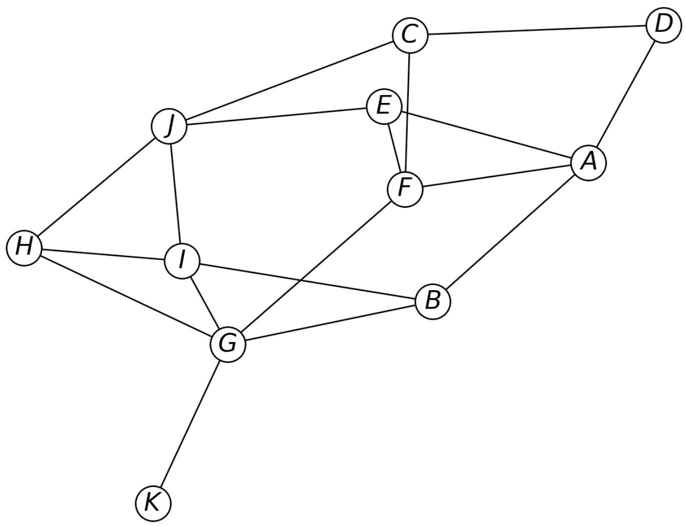
Then the corresponding cascade tree will look as follows:



A first posts the content. After seeing the content shared by A, B reshares it. C could have seen the post from either A or B. However, we will ***always break ties by attributing the cause of a reshare to the friend who shared most recently.*** Hence, because by the time C shares at time t=3, B has shared the post more recently than A, we will attribute C's "infection" to B. Finally, we see that D reshares the content after seeing it from B; C and D share the post at the same time t=3, so the only friend that D could have had seen the post from is B. So the cascade tree, rooted at the person who first posts the content, illustrates who is responsible for other people resharing.

Also, Mark wants to calculate the ***virality*** of the cascade, where virality is defined as the ***average distance across all pairs of nodes in the cascade tree.*** In this case, the virality is

$\frac{d_{AB}+d_{AC}+d_{AD}+d_{BC}+d_{BD}+d_{CD}}{6} = \frac{1+2+2+1+1+2}{6} = \frac{3}{2}$, where d_{ij} is the distance between vertex i and j.



(a) Using the graph above and the following timestamps at which content was shared, draw the cascade tree, and calculate its virality.

Time t	*people who shared at t*
1	D
2	C
3	A
4	J
5	F
6	E

(b) Using the graph above and the following timestamps at which content was shared, draw the cascade tree, and calculate its virality.

Time t	*people who shared at t*
1	C
2	J
3	F
4	A
5	E
6	H

(c) Using the graph above and the following timestamps at which content was shared, draw the cascade tree. You don't need to calculate the virality.

Time t	*people who shared at t*
1	G
2	K
3	H, I, B
4	F
5	A, E
6	J
7	D
8	C

(d) Using the graph above and the following timestamps at which the conent was shared, draw the cascade tree. You don't need to calculate the virality.

Time t	*people who shared at t*
1	A
2	D
3	C
4	J
5	I
6	G
7	K
8	F
9	B
10	H
11	E

(e) Based on the structure of the cascade trees, which of the cascades from (c) and (d) was most likely to be Taylor Swift posting a photo of her breakfast, and which was most likely to be a political petition? Justify your answers from the readings, ideally with specific citations.

Problem 3. Suppose your first job out of college is as a "viral marketer". You are given the task of designing a piece of content, and a plan for disseminating it on Facebook, with the goal that it achieve the ***largest number of reshares possible***. You are able to control the following aspects of the campaign:

- The design of the content itself: It could be a photo (real or synthetic), a video an audio clip, a news article, etc. --- basically anything that it is possible for users to post on Facebook.
- The identities of the "seed" posters: Your firm has given you budget sufficient for you to convince/pay any 100 users of Facebook to post your content at any time of your choosing (not all necessarily simultaneously). So another component of your design is who these 100 users should be --- their identities, personalities, structural location in the Facebook network, etc. --- as well as the timing/schedule with which they will post your content.
- Other temporal features of the campaign --- e.g. particular dates or season of year in which you launch.
- Any other aspects of the campaign you want to design, except that in the end your only resource is the posting by the 100 seeds. In particular, you have ***no ability to do any kind of external publicity or advertising outside of Facebook.*** Your firm will judge your performance on the extent to which your content "goes viral", purely on the basis of your 100 seeds and the cleverness of your campaign design.

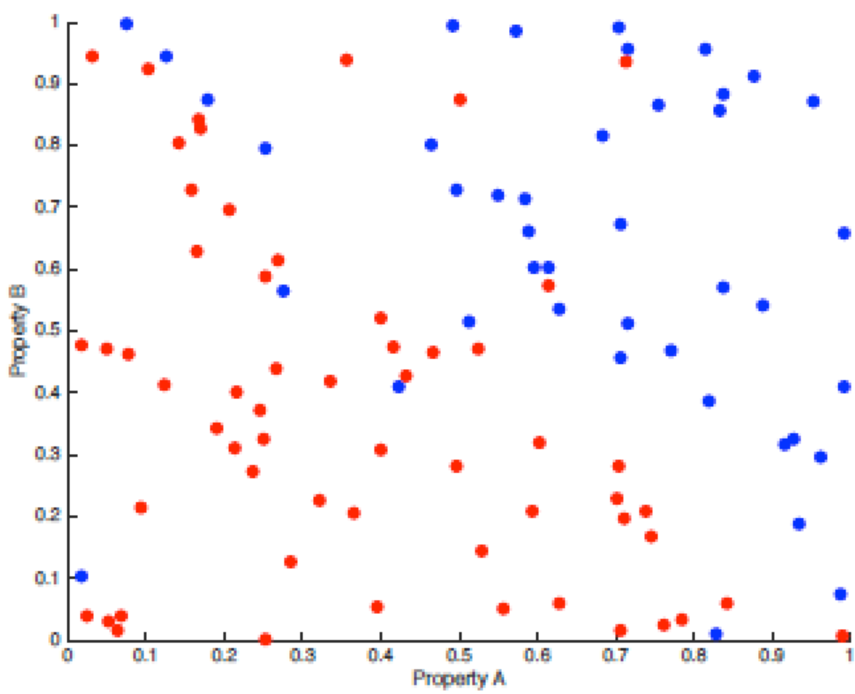
Write an essay in which you describe your proposed campaign design as clearly as you can, along the dimensions suggested above or others you'd like to discuss. ***You must justify as much of your design as possible by making reference to specific results in the contagion papers we have read.*** Be as precise as you can in these references (e.g. paper title, quotations, page or figure numbers, etc.)

Problem 4. Suppose there is a very large population of people for whom we can measure or obtain two numerical properties --- let's call them Property A and Property B --- and from these two properties, we'd like to predict some binary outcome y. For example, in class we discussed the problem of using high school GPAs (Property A) and SAT scores (Property B) to predict whether students admitted to Penn will graduate within four years or not (the outcome y). Importantly, we don't have Properties A and B and the associated y values for the entire population --- we only have a much smaller ***training sample*** from the overall population. The "machine learning" approach to this problem is to learn a "model" that does a good job of predicting y from A and B on the training sample --- and then hoping that this model will "generalize", in the sense that it will also do well at predicting y from A and B for the overall population.

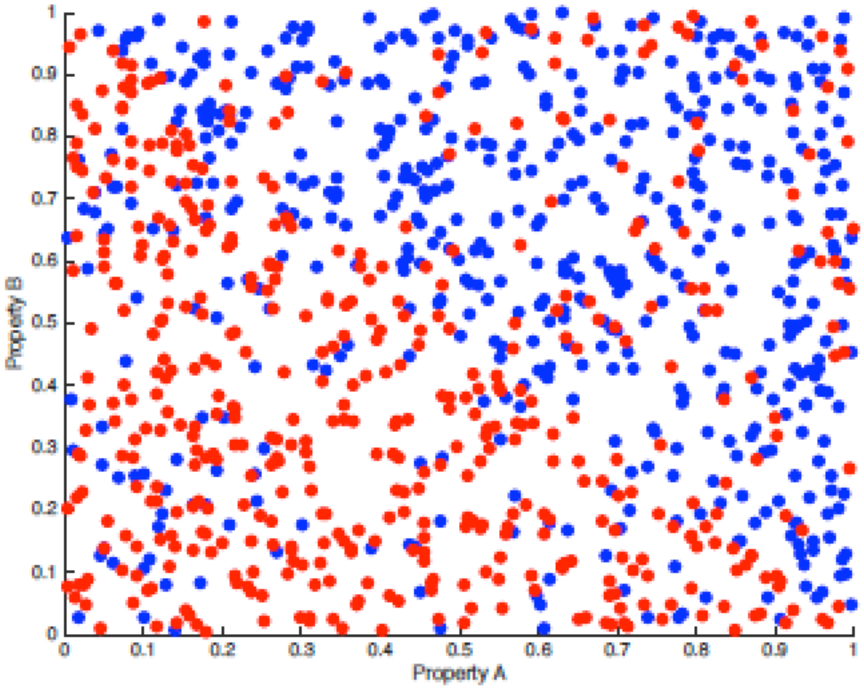
In each of the figures below, Property A is measured on the x axis, Property B on the y axis, and the dots show the (A,B) values of the training sample. There two possible outcome values are y = red and y = blue, and the color of a dot indicates its outcome.

On each figure, you should draw a "model" that does a good/reasonable (though not necessarily perfect) job of separating red points from blue points, while remembering that the goal is not perfect fit to the training sample, but generalization to the unseen overall population. Your model should take the form of a curve in the plain whose shape/simplicity/complexity is up to you. In each case, you should write a brief justification of why you chose the curve you did, relying on our in-class discussion of overfitting. The following website might also be useful: https://en.wikipedia.org/wiki/Overfitting.

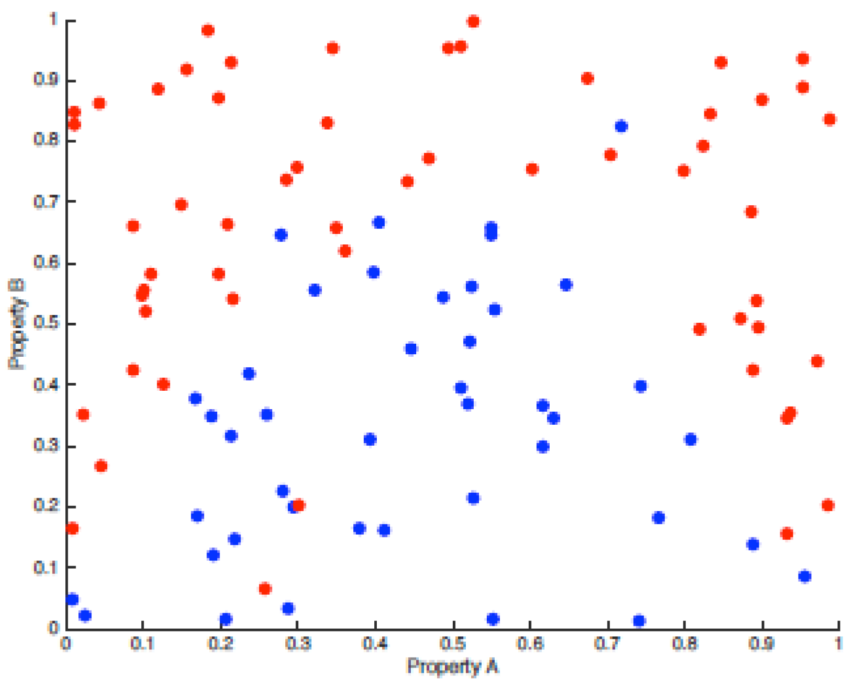
(a)



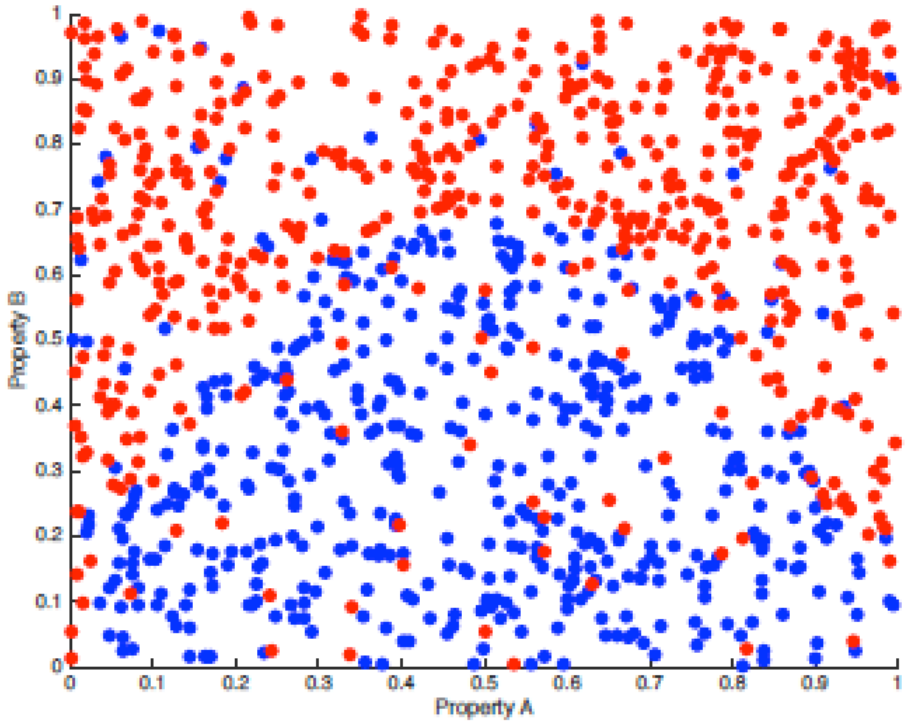
(b)


Property B
Property A
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

(c)


Property B
Property A
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

(d)



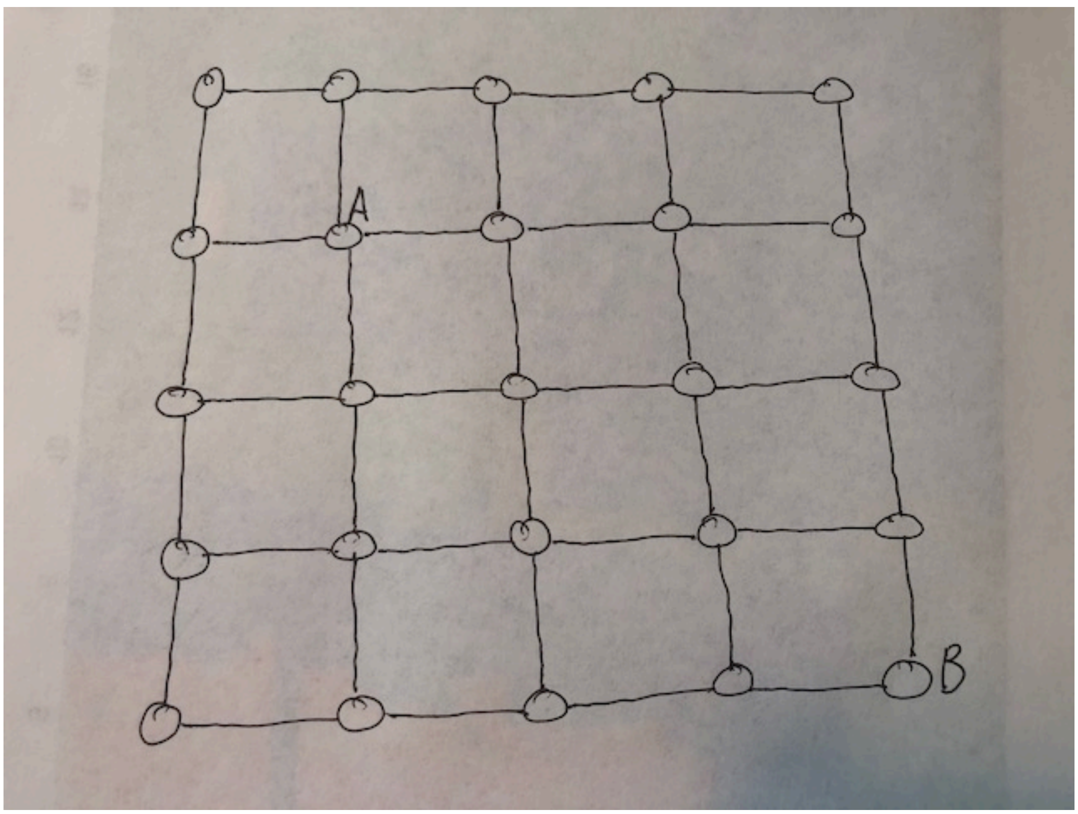
Problem 5. Consider the assigned paper "Navigation in a Small World" in the context of the network shown below.

(a) Add an edge to this network such that the shortest-path distance from A to B becomes 3, and the navigation algorithm from the paper also *could* take 3 hops from A to B.

(b) Add an edge to this network such that the shortest-path distance from A to B becomes 2, but the navigation algorithm from the paper definitely takes 6 hops from A to B.

(c) Add two edges to this network such that the shortest-path distance from A to B becomes 3, and the navigation algorithm from the paper *could* take 4 hops from A to B.

By "could take" above we mean that whenever there is a tie for which vertex the algorithm could forward to, you are allowed to specify which one is chosen. Remember that regardless of tie-breaking decisions, the algorithm will *never* take a step that increases the distance to the target.



Problem 6. For each of the following parts, carefully draw an undirected network with the specified properties.

(a) Draw a network with three connected components of sizes 8, 3 and 2 vertices, in which the total number of edges is 11 and the worst-case diameter (longest shortest-path distance) of the largest component is 2.

(b) Draw a network with three connected components of sizes 6, 4 and 1 vertices, in which the total number of edges is 9 and the worst-case diameter of the largest component is 5.

(c) Draw a network with three connected components of sizes 5, 4 and 2 vertices, in which the total number of edges is 17 and the worst-case diameter of the largest component is 1.

Problem 7. In this problem you are asked to compare and contrast the assigned readings "Navigation in a Small World" (Kleinberg) and "Identity and Search in Social Networks" (Watts,Dodd, Newman). For each of the following parts write a clear and concise response, supporting your answers with quotations or findings from the papers where possible.

(a) Both papers are trying to address a phenomenon that was implicit in the original Travers and Milgram paper, but was not explicitly examined there. What phenomenon is that?

(b) Discuss the similarities and differences between the approaches taken in the two papers. Your answer should address the models, methodologies and results of the papers.

(c) Which of the two papers do you find more compelling and convincing, and why?