MIDTERM EXAMINATION
Networked Life (NETS 112)
October 24, 2019
Prof. Michael Kearns

This is a closed-book exam. You should have no material on your desk other than the exam itself and a pencil or pen. If you run out of room on a page, you may use the back, but be sure to indicate you have done so. You may also make annotations directly on any diagrams given.

Name:

Problem 1: _____/10

Problem 2: _____/10

Problem 3: _____/20

Problem 4: _____/10

Problem 5: _____/15

Problem 6: _____/15

Problem 7: _____/10

Problem 8: _____/10

TOTAL: _____/100

Problem 1 (10 points). For each of the following assertions, indicate whether it is "TRUE" or "FALSE".

(a) The accuracy with which one can predict whether a Facebook cascade of size k will eventually be greater or smaller than the median for that size is slightly better than random guessing. False

(b) In order for N vertices to all be in the same connected component, the number of edges must be at least N(N-1)/2. False

(c) A tweet of Taylor Swift's breakfast will tend to have low virality. True

(d) Ramy Ashour was a famous mathematician. False

(e) The "forest fire" demo was essentially an exercise in the average or expected size of connected components. True

(f) Romantic partners on Facebook are most strongly indicated by the number of shared neighbors. False

(g) In machine learning, one may sometimes prefer a simpler model with higher training error to a more complex model with lower training error. True

(h) The purpose of the "roster" in the Travers and Milgram experiment was to prevent the letters from cycling. True

(i) If there are no triangles in a network (sets of three vertices with all three edges present), the clustering coefficient is zero. True

(j) Machine learning tends to not be a useful technique for predicting social network properties because they are so complex. False

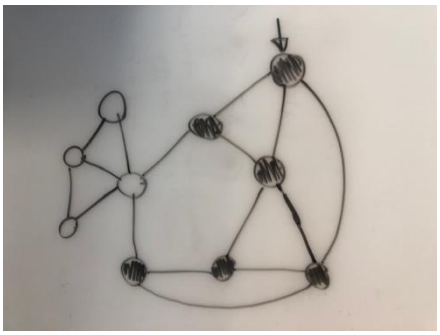
Problem 2 (10 points). Consider the network in which there is a vertex for each airport in the United States, and there is an edge between two airports if some airline provides a nonstop flight between them. For example, there would certainly be an edge between Los Angeles and Philadelphia, but there would not be an edge between Ithaca, New York and Des Moines, Iowa.

For each of the universal network properties discussed in class and readings, indicate whether you think this network will exhibit that property, and briefly but clearly justify your answer.

Two points for each property + justification
Small diameter - yes, because hub airports have flights to each other
Heavy-tailed distribution - yes, because a few hubs have extremely high degree whereas most airports are small and have low degrees
Sparsity - yes, because there are millions of possible direct flights and not all are realized - also, the necessity of layovers in many flights demonstrates sparsity
Giant component - yes, because almost every airport is connected to some hub
Clustering - both yes/no answers were accepted, based entirely on their justification.

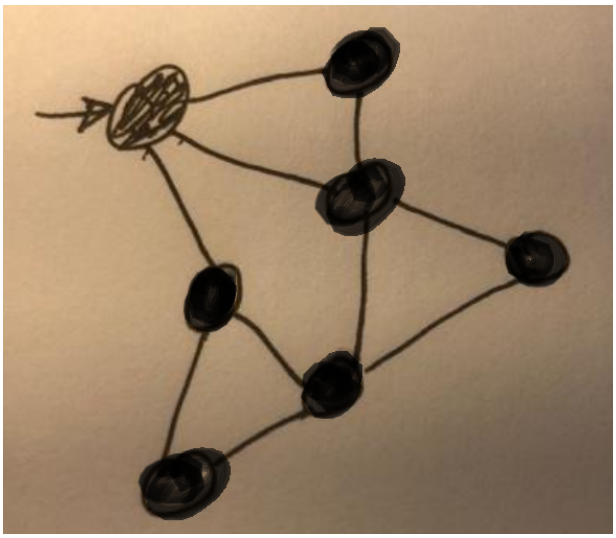
- Answers were evaluated based on whether they demonstrated a) accurate knowledge of the definition of clustering b) recognition of key facts about US airports (e.g. geographic region separation, hub and spokes structure).
- In particular, it was not sufficient to justify clustering with “big airports will all be connected to each other”. Big airports, the outliers in the heavy-tailed distribution, are outliers! Answers that justified clustering on the basis of (the more numerous) smaller airports were better.

Problem 3 (20 points). Consider the following contagion process in networks. A single initial vertex is infected. Suppose that a vertex A has become infected, and let B be any neighbor of A. Then as long as A and B share a common neighbor C, B also becomes infected. This process then repeats iteratively until no further infections are possible. For example, in the network shown below, if the initial infection is at the vertex indicated by an arrow, then the process will infect all and only the shaded vertices.

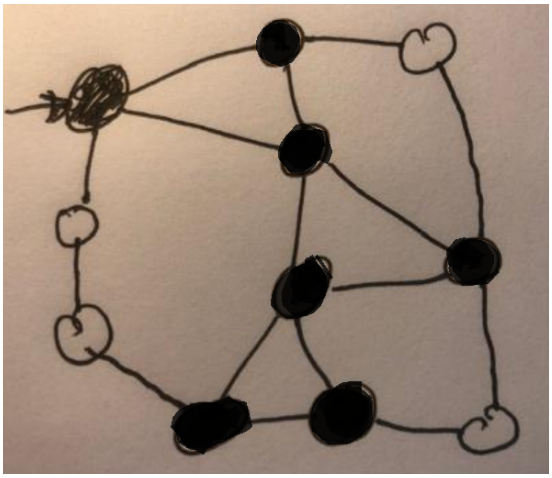


For each of the networks below, if the initial infection is indicated by the arrow on the shaded vertex, carefully shade in all and only the vertices that eventually become infected.

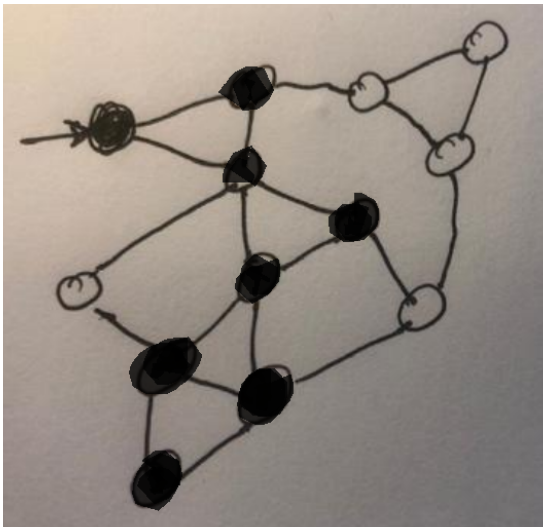
(a)



(b)



(c)



(d) Do you think that this contagion process would result in widespread infection in real social networks? Why or why not? Support your answer by referring to whichever of the universal properties of large-scale, real-world networks you think are relevant.

Yes. The contagion process spreads throughout triangles. Real social networks have large clustering → many triangles, causing a lot of contagion.

Answers that said “no, because real networks are sparse” were incorrect because sparsity doesn’t preclude clusters being connected to each other with triangles, which then facilitates the spread.

Problem 4 (10 points). Below is the opening paragraph of one of the assigned readings:

> *A growing proportion of human activities, such as social interactions, entertainment, shopping, and gathering information, are now mediated by digital services and devices. Such digitally mediated behaviors can easily be recorded and analyzed, fueling the emergence of computational social science and new services such as personalized search engines, recommender systems, and targeted online marketing. However, the widespread availability of extensive records of individual behavior, together with the desire to learn more about customers and citizens, presents serious challenges related to privacy and data ownership.*

Briefly but thoroughly describe both the methodology and main findings of this article. You don't need to provide technical detail on the methodology, but you should summarize the broad approach and the data involved, and some of the predictions made from that data. What are the authors' conclusions on the implications of their work for privacy in the digital era?

https://www.pnas.org/content/pnas/110/15/5802.full.pdf
methodology (broad approach, data involved):

- data: "58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychomet- ric tests."
    - "individuals for which between one and 700 Likes were available "
- method: "The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/ linear regression to predict individual psychodemographic profiles from Likes. "

main findings (& predictions made from the data):
"sexual orienta- tion, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental sepa- ration, age, and gender"
conclusion:

- help marketing & good tool for assessment
- threat to freedom & privacy

- facebook likes (3 pts)
- prediction of xxx (3 pts)
- conclusion (positive: help marketing 2 pts, negative: threats of freedom, privacy 2pts)

Problem 5 (15 points). For each part below, carefully draw a network exhibiting all of the specified properties.

Many answers were possible, each was evaluated on its own.

(a) A network with 15 vertices, in which every vertex has degree at most 3, and the worst-case diameter is at most 6.

wrong num vertices: -1, degree > 3: -2, worst case diameter >6: -2

(b) A network with 8 vertices, a single connected component, 10 edges, a clustering coefficient of 0, and no vertices of degree 1.

wrong num vertices = -1, edges != 10: -2, cc != 0: -2, vertices with degree 1 : -2

(c) A network with 6 vertices, edge density 2/3, and clustering coefficient 2/3.

wrong edge density, wrong cc each -2

Problem 6 (15 points). Give three different examples from the course readings in which the underlying structure of a social network is reflective or predictive of some activity in the network or some property of the people in the network. Be as precise as you can: identify the reading in question, and clearly describe the underlying structural property that is relevant, and the activity or property that it predicts or reflects.

5 pts for each example (must be a reading.)

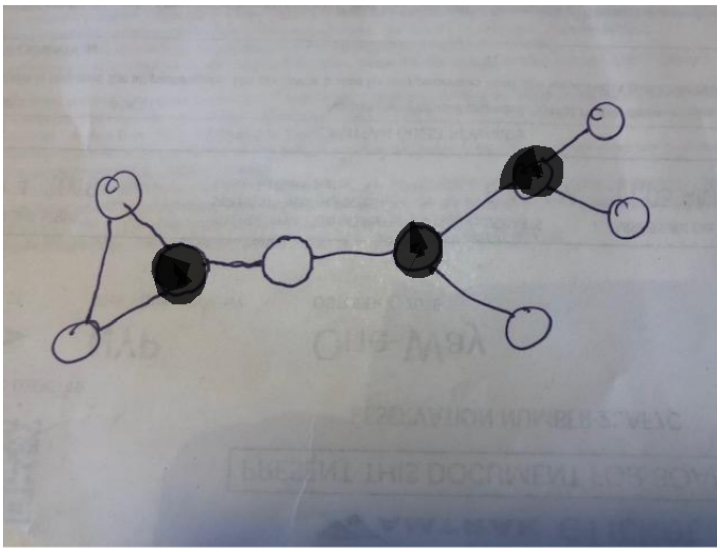
- (-3) if description of property not clear or wrong
- (-2) if did not connect back to the relevant activity/property it predicts/reflects
- (-5) no pt if it is an example mentioned in class but not a reading

Examples (not exhaustive) -
romantic partnerships - dispersion
squash - homophily
erdos - how prolific a writer is
predict cascade - the spread of information
travers and milgram - many of the chains passed through a small group of penultimate individuals with high centrality. Connectors have high degree.
Backstrom FB graph - low diameter and large component

Problem 7 (10 points). For each item on the left, write the number of the item on the right that is the best match.

(a) San Antonio component 3	1. child of divorced parents
(b) penultimate step 5	2. universal structural properties
(c) rare events 8	3. squash network
(d) five 2	4. Prof Kearns' Erdos number
(e) Facebook likes 1	5. Travers and Milgram
(f) dispersion 9	6. heavy-tailed
(g) not bell 6	7. Columbia Small Worlds
(h) clustering 10	8. big data
(i) 18 targets 7	9. romantic partners
(j) Three 4	10. triangles

Problem 8 (10 points). Recall the problem on the homework in which you were asked to simulate random walks on a network using a die --- i.e. pick a random starting vertex and then repeatedly moving to a random neighbor for many steps. Consider the network below, and clearly indicate on the diagram which vertex or vertices you think would be the most frequent final location of a long random walk. Justify your answer as precisely as you can.



The reason is, like on the homework, the vertices with the highest degrees have the highest probabilities of being the endpoint. Other factors - centrality, connection to a cycle, etc - were **not relevant**.

All three identified with justification - 10 points
Two identified - 8 points
Two identified, with justification for why the third wasn't most likely - 9 points
One identified - 6 points
One identified, with justification for why (one or two) others weren't most likely - 7 or 8 points

-1 was also applied for *incorrect* answers, i.e. vertices identified that weren't one of those 3.