Homework 2 - Jay Alexander Trevino (jat2211)

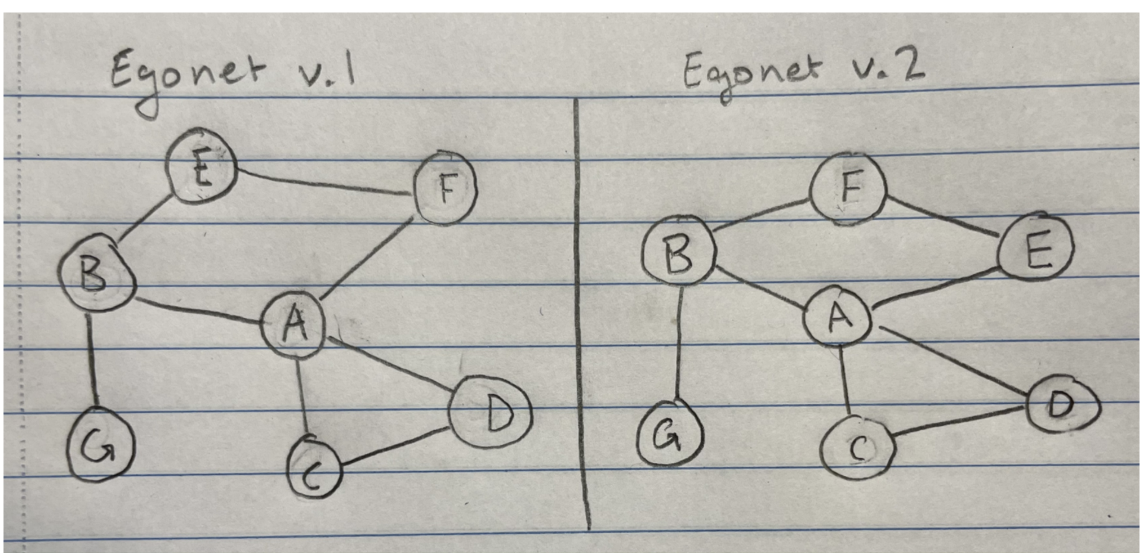
Exercise 1: Assembling a Network from Egonets

1.

Yes, because we can identify that there is one triangle, meaning this group of nodes consisting of subgraphs A, D, and C will only have 3 edges total (not including the two extra non-triangle edges in A). Amongst the rest of the edges in the egonets, since each edge not in the triangle will appear exactly once more amongst the rest of the edges in the egonets, then we can count the total of edges not in the triangle and divide by 2. This gives us 10 edges, which means the network has 5 non-triangle edges and thus a total of 8 edges.

2.

Yes, the network can be reconstructed but not uniquely because there are two ways to place the nodes E and F since they have the same egonet structure but will point to different nodes in the full network .



3.

No, because if we look at the graphs above, we can see that A connects to B in both i.e. having this information does not tell us anything new about the actual unknown unique graph. Had we known that relationship of either F or E with either B or A, then we would be able to reconstruct the unique graph.

4.

Number of drinks for each connected pair for Egonet v.1 provided above:

A — B: 6

A — C: 13

A — D: 13

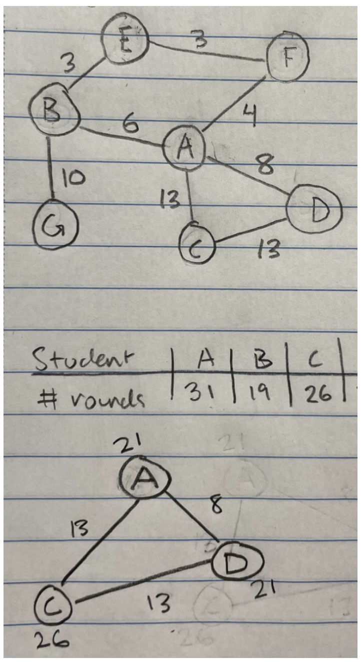
A — F: 4

B — E: 3

B — G: 10

C — D: 13

E — F: 3



Moreover, I've provided the triangle graph visually to show how I figured this part out. The only way for these nodes to satisfy the table is for C to be split evenly amongst A and D because since A and D share the same amount of drinks amongst *only* these nodes (I distributed the drinks for A to B and F beforehand since their values could easily be found), then can only be a balanced split amongst the nodes.

5.

A–C–D	0		10	12	14	16
A – C	13	...	8	7	6	5
A – D	8		3	2	1	0
C – D	13		8	7	6	5

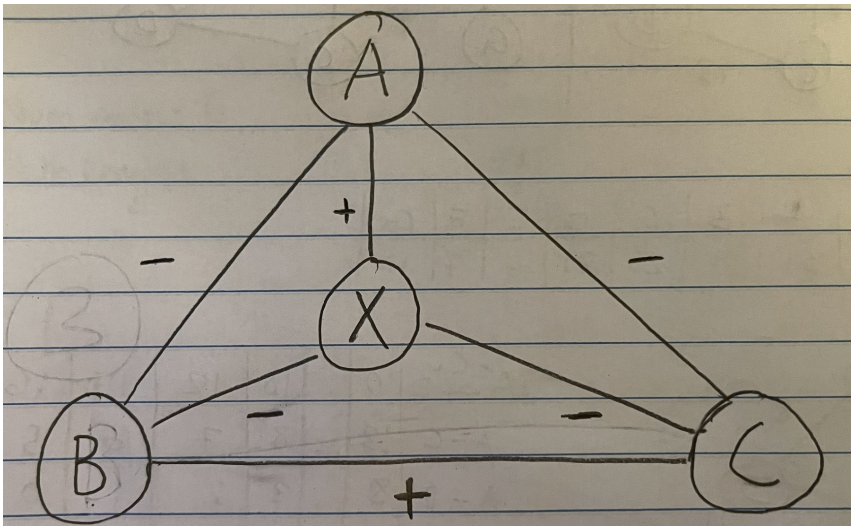
Since we only have one potential group of more than 2 drinking buddies, we can identify the number of drinks that were shared amongst the group and the number of drinks shared amongst each as a pair.

What we can see is that since the rest of the graph has the same structure (since there are no other triangles that can form a drinking *group*), then we have the same number of drinks to distribute amongst this triangle as pictured in the bottom graph in number 4. In fact, this graph shows us the visual representation of the distribution of drinks where the group of 3 has shared no drinks together (all drinks were in pairs). Since we already know that the triangle is symmetric about C, then we can find the number of drinks shared by the group, subtract this number from the total drinks that each node can share amongst the rest in pairs, then use the same method from the previous question to show that the number of drinks can be satisfied.

My results showed that this is only possible where the number of drinks shared amongst the group is an even number between 0 to 16, inclusive, where 16 is the upper bound on the number of drinks that can be shared amongst the group while still having enough drinks to allocate to the pair A—D. However, since there are multiple possible ways to distribute the number of drinks, it is not possible to find the unique number of drinks shared by this group of drinking buddies.

Exercise 2: Joining an unbalanced world

1.



2.

Using our proof from class, we have the following theorem: A complete positive/negative graph G=(V,E^-,E^+) is said to be balanced if for any triangle in the graph, either they are all friends or there is exactly one + edge between them. Thus, if we introduce node X into the graph where X connects to every other node, and if X has the freedom to choose whether or not they are friends or enemies with every other node, then we can individually consider every new triangle formed by X and two other nodes. In order to ensure that the graph is still balanced afterwards:

1. If a triangle formed by X and two other nodes has a positive edge between the other two nodes, then X can either choose to be friends with both or enemies with both and still keep the graph balanced.
2. If a triangle formed by X and two other nodes has a negative edge between the other two nodes, then X can choose to be friends with only one of the edges and still keep the graph balanced.
3. If a triangle formed by X and two other nodes has a negative edge between the other two nodes, and there is an edge already between X and one of the other edges (because this edge was created when considering scenarios 1 or 2) then:

    1. If X has a positive edge between one of the other nodes, then it will choose to have a negative edge with the other node.
    2. If X has a negative edge between one of the other nodes, then it will choose to have a positive edge with the other node.
4. If a triangle formed by X and two other nodes has a positive edge between the other two nodes, and there is an edge already between X and one of the other edges (because this edge was created when considering scenarios 1 or 2) then:
    1. If X has a positive edge between one of the other nodes, then it will choose to have a positive edge with the other node
    2. If X has a negative edge between one of the other nodes, then it will choose to have a negative edge with the other node.

These scenarios will cover all local decisions made by X when deciding to be friends or enemies with another edge when we consider all new triangles formed.

3.

There are two cases of G where the graph is not balanced: (1) there are 3 negative edges, or (2) there are 2 positive edges. We can prove that:

1. Case 1 is not possible because for every new triangle formed, X can only form one positive edge and one negative edge with the other two nodes. This means that X must have an equal amount of positive and negative edges, but this is not possible because X has 3 edges when it enters graph G and hence cannot avoid entering into an imbalanced triangle since the number of positive and negative edges will be unequal.
2. Case 2 is also not possible because if we consider the subcase 2.a where a new triangle formed with X and the two nodes that share a negative edge (the other edges are positive), then X can only form one positive and one negative edge with these two nodes. Thus, X now forms two other triangles where the other two nodes in that triangle share a positive edge and X shares both a positive edge with one of the nodes in one of these triangles (say case 2.b) and a negative edge

with one of the nodes in the other triangle (say case 2.c). Thus, X can only be in a balanced graph with 2.b if it shares a positive edge with the other node in this triangle, but would be imbalanced in case 2.c since it would now have 2 positive edges. Similarly, if it satisfies 2.c by having a negative edge for this last edge, then it would violate 2.b since it would have 2 positive edges.

4.

X can successfully join a graph G without being a part of an imbalanced triangle only if G is already balanced. This result is proven above showing that it can join a graph successfully if G is balanced and cannot join a graph successfully if G is not balanced, where the results of considering a graph G consisting of 3 nodes expands to a graph of any size.

This result also generalizes to a weakly balanced graph because node X can still make the same decision for every edge formed in a triangle as if it was strongly balanced. However, it can now successfully enter an imbalanced graph without entering into an imbalanced triangle simply by choosing all its connecting nodes to be negative (it can now only have 0 or 1 positive edges), which satisfies the condition for being weakly balanced.

Exercise 3: Identifying Users with Degree and Edge Information

1.

Number of potentially compromised users in meeting with one user: 17

2.

Number of potentially compromised users in meeting with two users: 29

3.

Number of potentially compromised users in meeting with two connected users: 225

4.

Number of potentially compromised users in meeting with three connected users: 1098

5.

I observed that the more users whose information is leaked, the higher the number of users whose privacy is compromised. I also noted that given the structure of the graph as it is now, any leak of data can theoretically put any user at risk given the circumstances lining up. For example, in the first part of exercise 3, only one users degree data was leaked, however, there were still 17 users (according to my calculations) who had unique degree values in comparison to all other users, meaning that it only takes a single feature to be able to uniquely identify at least one individual in this graph. Leaking two users' data increases the number of compromised users very quickly as well (perhaps exponentially but this is just a guess).

Exercise 5: Paizzi and Classera walk into a bar

1.

Dear Classera Head of Growth,

I urge you to consider a partnership with my company Paizzi. We have a strong market share in the formal academic market and our internal research shows that many ex-students we serve at these formal institutions are taking online classes *after* they use our service extensively. However, I'm sure you're aware of this. Let's talk networks. As the head of growth, I'm certain that your position deals heavily in identifying *who* would be most interested in your service. Once you've identified this person, the marketing team can do their thing and you have a new customer at your disposal. Well, here at Piazzi, we can solve that first step for you. I will introduce to you the reconciliation problem. Let's say we have two networks, one for the network of current users here at Piazzi and one for the network of both current users and potential users to reach out at Classera. What we want to do here is identify the overlapping users in both of our networks. To start, we have a list of linked users, which represent the users of *both* of our services already. Now, from these linked users, we can use this information to find "friends" of this user i.e. people that are connected to this user by means of taking a course together. Now, we can go through and find those potential users who have the most classmates that are both users of Paizzi and Classera. These are the new users you want. The research has already shown that users are more likely to use online classes *after* using Paizzi, of which we've identified a way to find those users simply by using our networks. Next, we've defined the similarity witness metric that allows you to see which potential users are already well connected to our current network of users. Lastly, we can play on the idea that these users are more likely to adopt the service if they already have friends or classmates that are using the service. Now, you just have to reach out to these users and let the effect of strong ties work its magic. One last thing I should mention, all this plays on the idea that we're trying to essentially reconcile one larger network of everyone we would consider to be a potential user. Using our network at Paizzi only makes that easier by providing additional data points for you to cross-reference with your own network to accomplish this goal. Naturally, this information should be kept strictly confidential. In identifying these individuals in the network, we're essentially putting our customers at risk because the combination of our two networks makes it all the much easier to identify these individuals, especially individuals who only agreed to share their information with each of our companies separately. Of course, they also agreed to let us share and sell this information somewhere in the fine print, so let's make sure that we are doing so morally and respectfully and not allowing this information to fall into the hands of a group who may use it for less moral reasons.

Sincerely, Paizzi Head of Growth

2.

We can consider a trajectory through class based on student enrollment in courses. Every node is a student in a course. We will consider two cases for why nodes are connected. The first is that two nodes are connected if they are enrolled in the same class. This allows us to cluster classes into dense subnetworks. This is important later. Next, we can connect two nodes if they are the same student in different classes that are taken exactly one semester apart. What this gives us is a way to visualize a student's trajectory through classes by allowing a way to (sort of) track a student's coursework over time. In additional, it allows us an easy way to deduce who is the same student since two nodes that are the same student will share very few common neighbors (on the scale of the number of classes taken in a semester) while two different students in the same class will share a lot of common neighbors (assuming classes have 20+ students). Likewise, the use of this graph will show us visual trends on a large scale of coursework taken by students. For example, assuming we can single out students in a computer science or any major, we can see roughly what classes students take in what order. In a degree path, they will mostly follow what's recommended by a university or Classera, however, students share information a lot and will recommend courses they like to other students, as well as courses that may not be required but feel necessary in the scope of the major. This will allow Classera to recommend both different course trajectories to students or even update their major requirements, moving some classes from the core to an elective, moving others from electives to the core, or even introducing new elective and core classes. One issue that may pop up are how or if to connect students across the same semester, which affects the ability to track courses over time, but also proves useful in seeing what courses are taken concurrently. Another issue is the graph is limited in the information it can provide. For example, it can show courses as a cluster, but this doesn't say anything about the variability of different professors that teach it nor does it provide a way to compensate for improving courses. Good courses may show up often, but what about bad or overly difficult courses that students feel are still necessary in their coursework trajectory? It also can't stand the test of time as it will need to be updated frequently to provide the best information because courses change over time, as do fields as a whole, and professors come and go, leaving the quality of a course very flexible.