Introduction to Social Networks

Assignment 1

Siyu Jiang, Roman Kracht, Zhengting He, Seokwon Choi

Task 1: Describe and plot a social network

(a) load the affective network and the gender data in R

```
affective_network_data <- read.csv("2400_affective_w1.csv", encoding="UTF-8", header=FALSE)
affective_network_data.friendship <- data.matrix(affective_network_data)[-1,-1]
gender_data <- read.csv('2400_sex.csv', encoding="UTF-8")
```

(b) recode the affective network wave 1 into a friendship network in R

```
affective_network_data.friendship[affective_network_data.friendship == 1] <- 0
affective_network_data.friendship[affective_network_data.friendship == -1] <- 0
affective_network_data.friendship[affective_network_data.friendship == -2] <- 0
affective_network_data.friendship[is.na(affective_network_data.friendship)] <- 0
affective_network_data.friendship[affective_network_data.friendship == 2] <- 1
```

(c) basic network descriptives

- network size: 27 nodes, 132 ties
- network density: 132/702 = 0.1880
- average degree: 132/27 = 4.8889
- reciprocity ratio (dyadic.nonnull): 0.3608
- gender composition : male = 11, female = 16
- the number of same-gender ties: 108 ties
- Add the measure: degree centrality (total degree)
    - max(total degree centrality) = 22
    - min(total degree centrality) = 1
    - mean(total degree centrality) 9.7778

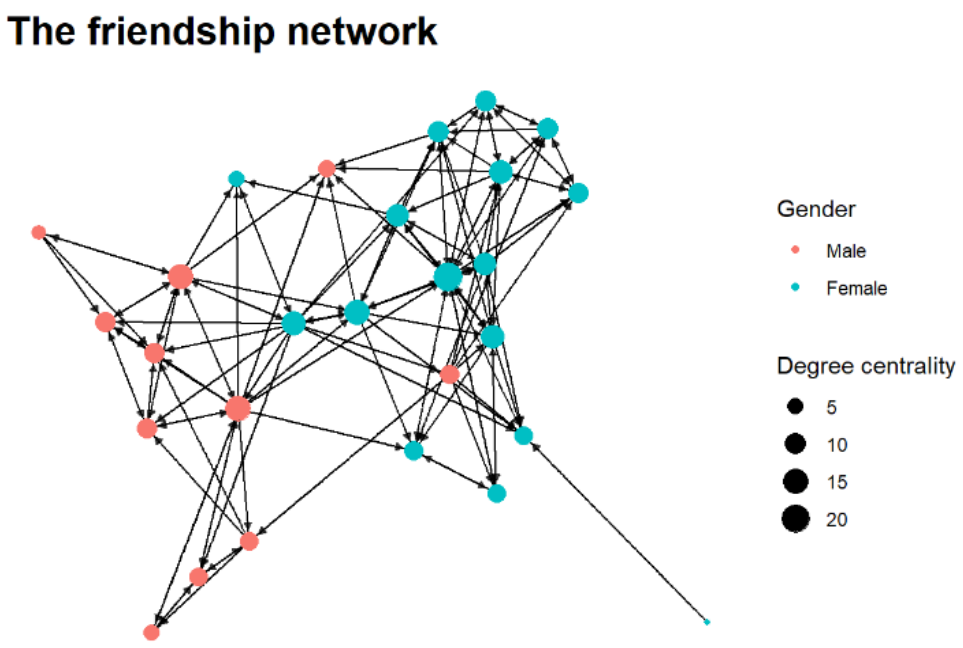
Measures interpretation:

- Network density: 18.80% of ties are present in the friendship network.
- Reciprocity ratio: the ratio of reciprocal dyads to non-null dyads is 0.3608
- Total degree centrality: On average, a node has 9.7778 (total of) incoming and outgoing ties.

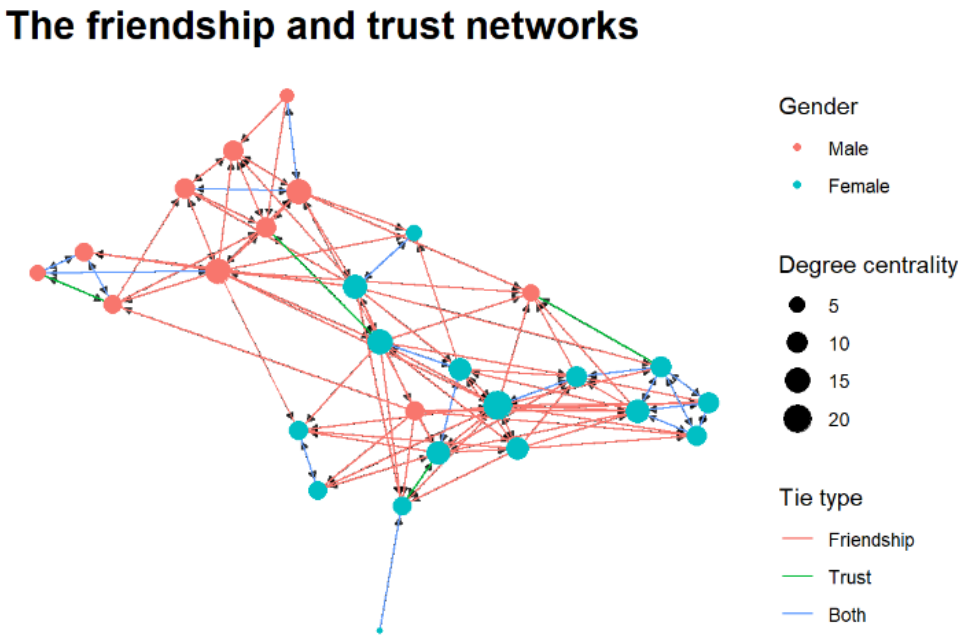
(d) plot the friendship network

Introduction to Social Networks

Assignment 1



(e) plot both friendship and trust networks into one network



(f) How large is the overlap between the two networks?

see the answer in the following (g) part.

(g) Short summary about the final plot

In the plot, females and males seem to form their own clusters. Given the previous statistics, 108 out of 132 ties are same-gender friendship ties. Besides, no individual is isolated in the plot. However, a female individual (at the middle bottom of the plot) seems to be very

peripheral in both networks, because the only connection she has is her nomination to one individual as her trustworthy friend, and she is the only individual without an in-degree. Thirdly, many people have friendship ties but not trust ties. A trustworthy friendship circle is present in a small female group (in the right corner of the plot).

Regarding the overlap between the two networks, 34 ties between two people are present in both networks.

Task 2: Data collection strategies

(a) Advantages and disadvantages of three types of network data

Complete Network Data

This data collection strategy requires us to know all the nodes in the beginning, for example, all the students in the school class. Unlike snowball-sampled network data, the network boundaries are fixed and we could not simply extend the set of nodes by other school classes.

On the other hand, collecting the complete network data minimizes systematic biases in the resulting social network. We would include isolated as well as popular participants in the analysis since the complete set of nodes is known in the first place.

Finally, collecting complete data of large networks has become very easy in modern times with social media platforms like Twitter or Facebook. One has to be careful to abide by research ethics.

Snowball-Sampled Network Data

Collecting complete network data is not possible when the set of nodes is unknown in the beginning, which is often the case for sensitive topics. Snowball sampling gives us an overview of the large-scale network without knowing the entire set of nodes initially, as discussed in the lecture for the needle sharing example.

While we do not have to worry much about systematic biases when dealing with complete network data, here we have to be much more careful though. In the context of the needle-sharing example, some highly-connected nodes in the network might actually be not only consumers but also distributors of heavy drugs. Most certainly, these distributors would not participate in the study and, therefore, not be captured in the resulting network.

Furthermore, snowball sampling only allows us to discover the cluster of the node we start with. This can be partially circumvented by increasing the number of random initial nodes, but even then, we will most likely not discover isolated nodes or additional small clusters.

Introduction to Social Networks

Assignment 1

Ego-Centered Network Data

This approach gives a very in-depth view of a particular individual, which might not be captured by the large-scale network structure obtained by collecting complete or snowball-sampled network data.

However, the overall network structure is captured rather poorly since this technique focuses specifically on individuals while neglecting many other nodes of the network. Furthermore, it might be the most expensive strategy for data collection. In a company example, analyzing the complete network data of an internal messenger service (for the employees who agreed to the study) might be much more feasible than performing lengthy interviews with a sample of individual employees.

In conclusion, we can say that the choice of data collection strategy depends strongly on the question we want to investigate and the initial data that we start with. The best option might also be a combination of strategies. For example, we could first collect complete network data in order to identify communities and then collect ego-centered network data for randomly chosen individuals from each community.

(b) Common and unique ethical challenges of data collection strategies

Regarding complete network data

If some people decide not to participate in the study or choose to opt-out during the longitudinal study, whether they should be included in the name generator can be an ethical problem. To address this challenge, we could explicitly ask them whether they agree with their name to be nominated. If their response is “NOT to be nominated” or we cannot get their response, we should exclude their names from the name generator list.

Regarding ego-centered network data

If participants need to report the relationships between nominated people, will it invade nominated (nonparticipating) people’s privacy? To address this challenge, researchers can ask participants to assign pseudonyms to the nominated people and then report relationships among these nominated people. Even if participants use some authentic names in the responses, researchers can convert them to pseudonyms before further data analysis.

Regarding snowball-sampled data

Snowball sampling may cause privacy and confidentiality issues: Letting participants directly nominate further participants for the study to the researcher might already be a privacy violation. Firstly, the identity of the nominated person is revealed without permission. On

the other hand, contacting the nominated person might also reveal the identity of the nominating person and potentially put him/her into a trust crisis within the social group. As long as the researchers are trying to approach nominated people directly, their privacy is influenced. What researchers can do is consider how to approach those people. Since the research topics can be sensitive, researchers should respect participants' behaviors with no moral judgment when contacting them. Moreover, researchers could allow participants to send "invite links/messages" to new people who might want to join the study before researchers approach them.

Common challenge

Asking negatively connotated personal questions to participants, such as "Who do you dislike?", might cause social pressure on the potentially-nominated person. Besides, the participant needs to recall their negative perception about certain people, which may intensify their negative feelings and possibly cause a biased response. To address this challenge, researchers should be careful when reporting the research findings to avoid the identification of specific individuals from the results. Naturally, participants should be informed about these privacy-protecting countermeasures.

(c) Research questions

Previously mentioned, snowball sampling is useful when the nodes of the network are unknown at first, but has some problems with systematic biases. Using this idea, we have thought of two research questions that can and cannot be analyzed by snowball-sampled network data.

"How are social media platforms like Twitter or Facebook affecting university students' relationships?" is an example of a research question that cannot be easily addressed through snowball sampling network data. Nowadays, social media is popular among the young generation, and various types of social media platforms are kept on being created. Especially during and after the pandemic, social media works as an effective online contact method and supports many individuals in establishing relationships. Using snowball sampling, it is likely to miss a lot of individuals that are forming a small isolated community because they are not linked to a lot of nodes and have a low chance of being discovered by other samples. Since the purpose is to find a general feature about social media among university students, skipping various small communities would not be a logical approach. In addition, snowball sampling would probably lead to a bias in sampling because each initial participant recruits other participants. The existing connections' features would probably be irrelevant to the online functions, making a biased analysis in the study.

Introduction to Social Networks

Assignment 1

“The pattern of information sharing among illegal migrants” can be an example research question for being effective by using snowball-sampled network data. Illegal migrants are not completely recognized by the government, and they form their independent networks to settle in the new country. They exchange information through personal connections and give help to each other in navigating challenges. The main reason that snowball sampling would be efficient is that the researchers cannot identify the set of all illegal migrants, using the connections of each individual would lead to an approach to various subjects of the study. There might be a few isolated communities of illegal migrants that can be undetected by snowball sampling, but we think that it still has the biggest chance and is the fastest method to approach most of the nodes.