(a) State the research question of this paper.

This article investigates the origins of homophily. Do individuals selectively make or break some ties over others, and how do these choices shed light on the observation that similar people are more likely to become acquainted than dissimilar people?

(b) Describe the data that the author used.

1) How many data sources?

Our analysis is based on the population of 30,396 undergraduate and graduate students, faculty, and staff in a large U.S. university, who used their university e-mail accounts to both send and receive messages during one academic year.

(1) the logs of e-mail interactions within the university over one academic year, (2) a database of individual attributes (status, gender, age, department, number of years in the community, etc.), and (3) records of course registration, in which courses were recorded separately for each semester. How many observations (the question could have multiple dimensions)?

2) What time period did the data span?

After cleaned the data, the resulting data set comprised 7,156,162 messages exchanged by 30,396 stable e-mail users during 270 days of observation.

3) Where could you find the description and definition of all the variables?

The precise definitions of all variables are provided in app. A

(c) Highlight a potential problem that the data cleaning process might introduce in a way that diminishes the authors' ability to answer the research question.

The author states: "*To ensure that the data represent interpersonal communication, we included only messages that were sent to a single recipient (other than the sender—i.e., excluding self-addressed e-mails).*" I think this is a potential problem that could diminish the author's ability to research. The author's goal is to capture "interpersonal communication" as he said before. However, this way of data cleaning is not equivalent to his goal. We can also send emails to more than 1 people based on interpersonal communication. For example, if I want to go hiking and I prefer going with my friends. I will just send emails to all my friends who might go with me. You cannot just simply deny that this is not interpersonal communication. Also, even if there is only one recipient of this email, the email still possibly is not "interpersonal communication". The other potential problem is the author excludes the department email accounts data, like xyz@department.university.edu. As the author says in this article, "*mostly the faculty and graduate students in departments such as computer science, mathematics, and physics",* this means the social relationships built on those emails will be exclude to the analysis. In general, I think both these two potential problems will weaken the analysis to a certain extent.

(d) In this paper, the underlying theoretical construct is "social relationships" and the data are e-mail logs linked to other characteristics of the senders and receivers.

Discuss one weakness of this match of data source and theoretical construct and describe how the authors address this weakness

This article is to understand how homophily emerges over time as a function of the decisions of individuals to make and break ties, which is the process of network evolution. And the author used receiving and sending messages on the university e-mail accounts as the model to construct "social relationships". I think the weakness is that this only captures partial of their social relationships. For example, I will only use my university email account to contact the people with academic or professional reasons. And I will contact my friends on Wechat, Facebook or other social media. This pattern is also suitable for a lot of young people. Also, for those people who do not use social media, they may have other email account (like gmail or Hotmail) to contact with their friends. Capturing the data only on university email accounts will weaken the effectiveness and validity of the results to a certain extent. The author implicitly mentions this in explaining the out-degree for different groups of people. The author says: *"The average out-degree for undergraduate students (24 contacts) was somewhat lower than for faculty—a pattern that might be explained by the popularity of instant messaging among undergraduates—and for nondegree students, many of whom probably do not use the university e-mail as their primary address, it was lower still (eight contacts)."* This proves what we discussed before. For undergraduates, they might use instant messaging more frequently and for those nondegree students, they probably just use another email address as their primary address.