# Problem Set #[2]

MACS 30000, Dr. Evans
Ruixi Li

In this paper, the authors shortly present a literature review on homophily and then, bring out the research question which focuses on the origin of homophily: on what grounds 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?

To investigate more on this question, they using network data of a large university community which interactions, attributes, and affiliations are recorded. To construct the dataset, the authors merged three different databases: (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. Dataset comprised 7,156,162 messages exchanged by 30,396 stable e-mail users during 270 days of observation. The available variables could be categorized into four groups: personal characteristics (age, gender, home state, formal status, years in school); organizational affiliations (primary department, school, campus, dormitory, academic field); course-related variables (courses taken, courses taught); and e-mail-related variables (days active, messages sent, messages received, in-degree, out-degree, reciprocated degree). A precise definitions of all variables are provided in app. A.

In the data cleaning process, from 43,553 individuals, the authors identified 34,574 users who were active throughout both semesters by the principle of sending and receiving e-mail in both the first and the last months of the academic year. However, 43,553 individuals sent and received messages during the academic year. It is highly possible that they just didn't send or receive e-mail in both the first and the last months of the academic year. Simply dropped 8,979 individuals may result in a biased model and diminish the authors' ability to answer the research question.

When match the data source to theoretical construct, the authors arbitrarily suppose that the university e-mail logs fully represents the social relationships of an individual. Nevertheless, it is not necessarily the case. Let's say a student seldom use e-mail to reach his friends. Instead, apps like Whatsup and Lines replaced e-mail with their convenience. This student is likely to be removed from the dataset in the context. Here is another example about the implicit closure. If there is a fraternity which never emails its member but has a fixed time and location for gathering, e-mail logs can't reflect the existence of this implicit foci. After two people from one class made friends in the gathering before they become friends through class, they may communicate with each other through e-mail. Then, the e-mail logs may indicates that they are friends with the explicit foci, which is that they are in the same class.

Therefore, the dataset may fail to capture the real case.

In the chapter of origins of homophily, the authors proposed some problem on observed homophily.One might concerned that our measure of individual similarity acts, in effect, as an indicator variable for sharing a class, and that controlling for shared classes would effectively eliminate the potential for similarity to have any additional impact on tie formation, thereby artificially increasing the apparent importance of induced homophily vis-a-vis choice homophily. In order to address this potential systematic bias, the author consider in figure 7 (top row) the distribution of similarity for student pairs who shared classes with that for student pairs who did not. As expected from figure 6, students who shared classes (fig. 7, pt. B) are, on average, much more similar than students who did not (pt. A). However, its higher average notwithstanding, the distribution in part B of figure 7 also exhibits higher variance (1.8) than that in part A (1.3); thus, the potential for differences in similarity to impact tie formation is not in fact diminished for pairs who share classes versus those who do not. As a further check the authors compare distributions of similarity for pairs who share implicit foci (fig. 7, pt. D) with those who do not (pt. C). In this way, the concern can be properly addressed.