The research question of this paper is “How does homophily form?”. According to McPherson and Smith-Lovin (1987), there are 2 mechanisms by which homophily arises—choice homophily, the factors attributed to individual, psychological preference, and induced homophily, the factors attributed to the homogeneity of structural opportunities for interaction. The research aimed to find the relative roles of those two mechanisms in forming observed homophily.

The authors used a data set of 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” (Kossinets & Watts, 2009, p.410). It was formed by combining 3 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” (Kossinets & Watts, 2009, p.410). The data span is 1 academic year (2 semesters). The detailed definition of all variables is included in appendix A.

There were some potential problems might be introduced by the data cleaning process. For example, it said in the paper that they included only messages that were sent to a single recipient other than the sender to ensure that the data represent interpersonal communication, but it was possible that some such emails were for purposes other than interpersonal communication, like a student asked questions to another student who was TA for a course. That error could

lead to misestimating some parameters and correlations.

There were also other problems. For example, the match of email logs to social relationships had some weaknesses. One was that email exchanges are discrete and often cluster in some short periods, while social relationship is persistent and continuous. The authors employed a simple but effective method known as a sliding window filter to solve that problem. They used geometric average of the number of messages exchanged by users per unit of time to describe the instantaneous strength of their relationship. For each time point t, they checked email exchanges in a period of time before that time point, i.e., (t-r, t] to judge the strength of the relationship. The checked time period slid on the time-axis to depict the curve of strength of relationship while holding the same length, just like a sliding window.

References

Kossinets, Gueorgi and Duncan J.Watts, “Origins of Homophily in an Evolving Social Network", *American Journal of Sociology*, September 2009, 115 (2), 405-450.