Assessment of Kossinets and Watts

Sanittawan Tan

10/13/2018

In their article *Origins of Homophily in an Evolving Social Network*, Kossinets and Watts examined the question of how homophily, the tendency of "like to associate with like," originates. More specifically, they set out to answer the question of how and on what grounds individuals choose to make or break certain ties over others and how the former choices might be able to explain why similar people tend to become acquainted than dissimilar counterparts. Historically, two distinct theoretical approaches explain why similar people form ties with the like, choice homophily (i.e. individual preferences) vis-à-vis induced homophily (i.e. structural proximity). To see which school of thought is correct, the authors set out to answer their research question by using a network data set from a particular research university. Their analysis shows that both choice and induced homophily play an important, but partial, role in originating homophily in their population of interests. In other words, both individual preferences and structural proximity reinforce each other in how the network of like people forms.

Kossinets and Watts constructed their data set by merging information obtained from three different databases which are (1) logs of email interactions of undergraduate and graduate students, faculty and staff within a university over one academic year, (2) individual attributes such as status, gender, age etc. and (3) records of course registration. The population of the study consists of 30,396 observations which include undergraduate students (21%), graduate and professional students (27%), faculty members (13%), administrators and staff (13.4%) and affiliates (which includes postdoctoral researchers, visiting scholars, exchange students and recent alumni; 25%) within a large U.S. research university. The data spans over a one year period, but the authors note that their email logs include 7,156,162 messages during 270 days of observation (p. 411), which corresponds to the length of the two academic semesters. The description and definitionof all variables can be found in Appendix A of the paper (p. 439-442).

Regarding their choice of data cleaning, there are three potential problems. Firstly, Kossinets and Watts included only messages that were sent to a single recipient. Although this category makes up 82 percent of thier data, cleaning data in this fashion does not take into account emails that were sent to multiple recipients which can shed light on the origins of homophily. For instance, some users may send emails to a group of acquaintances without sending them to a particular person because they rely on other mode of communication. This also leads us to question on to what extent email logs can explain origins of homophily which will be discussed in the next section. Secondly, the authors noted that email accounts provided by university departments were excluded from the data set because they cannot be matched with employee records. Exclusion of departmental emails poses a problem to the analysis because it is questionable if the data set accurately reflects how individuals with both departmental and general email accounts choose to make or break ties, especially communication among users from the same department. If departmental emails were to be included, it may increase the role of structural proximity and change the authors' conclusion. Finally, Kossinets and Watts mentioned that a set of heuristics were used to determine conflicting values such as gender. This approach also casts doubt on how accurate the authors capture "similarity" among observations, especially when the authors did not report the percentage of conflicting and missing observations.

One weakness of the match between explaining social relationships and email logs is that social relationships have more aspects than what can be captured by one means of communication. For example, younger users in the study may use other means of communication such as private messengers and telephone to communicate rather than emails. Some users may even have more than one email accounts for use in the personal and professional contexts. In addition, due to privacy concerns, email logs are limiting because the authors were not able to analyze the content of email conversation. Since the content was censored, the authors adopted the sliding window filter method to track how network of individuals evolve over time, determining how ties were established and lapsed. However, it is difficult to capture the depth of these relationships. For instance, some users may frequently communicate with each other through university emails, but they may not choose to spend time with each other outside of work or study context. As Kossinets and Watts noted, electronic communication through emails in certain organizations may systematically differ from one another and from normal everyday interaction. The authors addressed this limitation by proposing that future studies could conduct comparative analyses of network evolution of various environments such as businesses and

government agencies. Finally, the authors admitted that their email logs data lack some attributes such as race and socioeconomic status. Due to limited available information, they were not able to address this concern in the research design. The authors may have missed an important clue to the origin of homophily because subjects in the study who are not considered similar based on available categories may be bonded by races or socioeconomic status. Overall, the use of university email logs to explain origins of homophily casts doubt on how much the result of the study can be generalized because of the intrinsic complexity of social relationships.