Response to "Origins of Homophily in an Evolving Social Network"

The paper by Watts and Kossinets is an attempt to answer a sociological question from a computational and mathematical perspective - and in particular, a network science perspective. One could argue that the domain of Social Network Analysis really kicked off in 2004 with the introduction of the Enron Mail Dataset [1], and since then it has not slowed down, especially with the advent of popular social media. In the context of the paper by Watts and co., we are continuing to use emails to answer our questions. So what is the question? The authors seem to state this quite explicitly early on -

"On what grounds, then, 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? "

This may seem a bit long winding, but luckily for us the authors again clarify their question more succinctly later on during the paper. This time, they state that their original question is -

"to what extent some observed pattern of homophily can be attributed to individual preferences versus structural constraints?"

Which is a more appropriate rewording of the original question. Throughout the paper the authors attempt to understand how networks of similar people form and disintegrate, while focusing on implicit and explicit foci. We can say that the previous statement the authors have given us is a good answer to the question of what the research question of the paper is.

The data for the experiment is detailed out in the **Data and Methods** section of the paper. Their analysis is based on 30,396 undergraduate and graduate students, faculty and staff in a large US University. There were three major data sources, which are described by the authors as follows:

(1) the logs of e-mail interactions within the university over one academic year,

(2) 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.

The duration for the data collection was one academic year. **Appendix A** is also an important source to describe all the data. All the features used to construct the model as well as the descriptions of these features and how they are tied to the research are explained in this section.

The data cleaning process involved only keeping mails sent to a single individual, so as to capture interpersonal relationships. The contents of the e-mail messages were not recorded. This poses two possible problems - a lot of times e-mails are sent to more than one person (indeed, 18% in this dataset), and could include a very important measure of triads or even larger groups. While bulk emails are included later to capture external foci or social foci, it may also contribute to understanding interpersonal relationships. As for the contents of the messages, it could be very important in defining *internal* foci, or interests. While it could be argued that the contents of the messages may include sensitive material, with the large number of text processing algorithms

readily available it is sure that this information could be very useful in assessing similarity and in turn in understanding homophily better. Another point to note is that they only selected users active in *both* semesters, which reduced the network size by over 12,000 users. This suggests that only users who regularly keep in touch the whole 2 semesters are considered as part of the network, which might not necessarily be true.

The underlying data structure used to analyse our question of homophily is a network of e-mails. Expecting this network to represent an actual network of social interactions is difficult to expect, even though it does go a long way in capturing some of the patterns. While we did criticise how the cleaning process may lose out some of the interactions, the discarded bulk e-mails are indeed later used to model external foci. The authors do their best to approximate a real world social network using e-mail data, but there is only so much one can do. It would be interesting to conduct the research with a more *social* media, as e-mails are largely professional in nature.

[1] B. Klimmt, Y. Yang. Introducing the Enron corpus. CEAS conference, 2004.