CSI 445/660 – Network Science – Fall 2015

Handout 3.1 – A Simple Model for Measuring Homophily

Notes:

(a) It is assumed that the characteristic for which homophily needs to be measured has only two distinct values. (The discussion below uses gender as the characteristic.)

(b) The derivation below is based on a simple random mixing model and is discussed in Chapter 4 of the text by Easley & Kleinberg.

Derivation of the Homophily Test:

1. Let N_B and N_G denote respectively the numbers of boys girls in the network.

2. So, the total number of nodes n in the network is given by $n = N_B + N_G$. Let m denote the number of edges in the network.

3. A node chosen uniformly randomly will represent a boy with probability $p = N_B/(N_B + N_G)$ and will represent a girl with probability $q = 1 - p = N_G/(N_B + N_G)$.

4. What is the probability of generating a cross-gender edge?

   Each of these events happens with probability pq. So, the probability of generating a cross-gender edge in the model is $2pq$.

5. As a consequence, the average number of cross-gender edges under the model is $2pq \times m$.

6. Thus, under the simple random mixing model, on the average, the fraction $2pq$ of the edges will be cross-gender edges.

So, the test for homophily can be summarized as follows.

Homophily Test: If the fraction of cross-gender edges in the network is significantly below $2pq$, then there is evidence for homophily.