A note on measuring gender homophily among scholarly authors

Ted and Carl Bergstrom with Molly King, Jennifer Jacquet, Jevin West, and Shelley Correll

An *instance of authorship* consists of a person and a paper for which the person is designated as a co-author. A coauthor-pair consists of two distinct persons who are designated as coauthors of a single paper. We observe all of the co-authorships in a population of articles and persons. We know the gender of each person in the population and are interested in measures of the propensity of authors to co-author with someone of the same gender.

Let M be the set of all authorships where the author is a man and F the set of all authorships where the author is a woman. Let m and f be the numbers of instances of authorship where the authors are, respectively, men and women. (Note that a single person who is a coauthor of k papers will be counted as belonging to k distinct authorships.) Let $\lambda_m = m/(m+f)$ be the fraction of authorships in which the author is a man.

Select a co-author pair by the following random process. First select an authorship at random, where any authorship is as likely to be selected as any other. Then select one of the coauthors of this paper at random, where any coauthor other than the originally selected person is as likely to be chosen as any other.

Define the coefficient of homophily $\alpha = p - q$ where p be the probability that a randomly chosen co-author of a randomly chosen man author is also a man and q be the probability that a randomly chosen co-author of a randomly chosen woman author is a man.

For co-author pairs selected by this process, define an indicator random variable I_a to be 1 if the first person selected is a man and 0 otherwise. Define I_b to be 1 if the second person selected is a woman and 0 otherwise. Define ϕ to be the Pearson correlation coefficient

$$\phi = \frac{E(I_aI_b) - E(I_a)E(I_b)}{\sigma_a\sigma_b}$$

between I_a and I_b. This quantity is sometimes known as the phi coefficient (Warrens, M., 2008, *Psychometrika* 73: 777).

Proposition *The coefficient of homophily α is equal to the coauthor gender correlation ϕ.*

Proof. Let p_{ij} be the probability that the coauthor pair selected consists of person i and person j. This event can happen in two possible ways, each of which are equally likely. Person i can be chosen first from the entire population of authorship instances and then person j is chosen from among i's coauthors of this paper or person j can be chosen first and then i can be chosen from among j's coauthors. Where k is the total number of co-authors of the selected paper in which i and j are coauthors, each of these events has probability

$$\left(\frac{1}{m+f}\right)\left(\frac{1}{k-1}\right)$$

and hence

$$p_{ij} = \frac{2}{m+f}\frac{1}{k-1}.$$

The probability that one of the two persons selected is a man and the other is a woman is

$$\sum_{i\in M}\sum_{j\in F} p_{ij}.$$

This event can happen in two possible ways, each of which are equally likely. (i) The first person chosen is a man and the second a woman. This happens with probability $\lambda_m(1-p)$. (ii) The first person chosen is a woman and the second person a man. This happens with probability $(1-\lambda_m)q$. Therefore we have the following parity relation

$$\lambda_m(1-p) = (1-\lambda_m)q \tag{1}$$

Recall that $\alpha = p - q$. Therefore, rearranging expression (1), we find

$$\alpha = \frac{p-\lambda_m}{1-\lambda_m}. \tag{2}$$

Now let us calculate ϕ. Since each person is equally likely to be selected first as second, it must be that $E(I_a) = E(I_b) = \lambda_m$ and that $\sigma_a = \sigma_b = \sqrt{\lambda_m(1-\lambda_m)}$. Finally, $E(I_aI_b)$ is the probability that both members of the selected co-author pair are men. This is the probability λ_m that the first person selected is a man times the conditional probability p that the other member of the selected pair is a man, given that the first member selected is a man. Therefore we have $E(I_aI_b) = \lambda_m p$. It follows that

$$\phi = \frac{p\lambda_m - \lambda_m^2}{\lambda_m(1-\lambda_m)} = \frac{p-\lambda_m}{1-\lambda_m} \tag{3}$$

Comparing Equations (2) and (3), we see that the coefficient of homophily α is equal to the Pearson correlation coefficient ϕ.

Corollary *In the case of two-author papers, the coefficient of homophily α is equal to Sewell Wright's coefficient of inbreeding F.*

Proof. Take a set of two-author papers. Let x be the fraction of these that have two men authors, y be the fraction of mixed gender papers, and z be the number of papers with two women authors. Wright's coefficient of inbreeding F is given by

$$\begin{aligned} F &= \frac{y_{\text{observed}}}{y_{\text{expected}}} \\ &= 1 - \frac{y}{2\left(x + \frac{y}{2}\right)\left(\frac{y}{2} + z\right)} \\ &= \frac{(4x^2 - 4x + 4xy + y^2)}{((2x + y - 2)(2x + y))} \\ &= \frac{2x}{2x + y} - \frac{y}{y + 2z} \\ &= p - q \\ &= \alpha \end{aligned}$$