The issue of the discrepancy between probit coefficients and AME differences is presented as a puzzle in the paper and addressed rather extensively.

In the methods and material section, it is described in words as "Hence, when evaluating interactions between variables in a probit model, it is entirely possible that two variables (here, the deskmate indicator and dyadic similarity) interact in their effect on the latent continuous outcome (here, the latent friendship propensity) but do not interact in their effect on the probability of the manifest binary outcome (here, the probability of friendship nominations), or vice versa—and this can result simply from transforming probit coefficients into AMEs *after* a given model has been estimated on given data.".

Although this has explained the problem quite clearly, I think readers like me might find it useful to see a few formulas, maybe just as a short footnote. At least for me, I have to check the formulas every time I see these issues of limited outcomes to get a concrete idea of what's going on. Something like what follows is what I have in mind.

Extending a little the equations in the mostly harmless book (page 97 and 104), the latent friendship propensity function can be written as

$$Y_i^* = \beta_0^* + \beta_1^* D_i + \beta_2^* X_i - v_i$$

$$v_i \sim N(0, \sigma_v^2)$$

, where D_i is a binary treatment, X_i is the covariate that potentially moderates the treatment effect. Here, there is no interaction between X and D in the latent friendship propensity equation.

And the CEF for the observed binary Y_i is

$$P(Y_i = 1|D_i, X_i) = E(Y_i|D_i, X_i) = \Phi\left[\frac{\beta_0^* + \beta_1^* D_i + \beta_2^* X_i}{\sigma_v}\right]$$

First, if X_i is continuous,
write the marginal effect for Y_i as

$$\mathrm{ME}_i = \Phi\left[\frac{\beta_0^* + \beta_1^* + \beta_2^* X_i}{\sigma_v}\right] - \Phi\left[\frac{\beta_0^* + \beta_2^* X_i}{\sigma_v}\right]$$

Take derivative with regard to X_i, the potential moderator,

$$\frac{\partial \mathrm{ME}_i}{\partial X_i} = \phi\left[\frac{\beta_0^* + \beta_1^* + \beta_2^* X_i}{\sigma_v}\right] \cdot \frac{\beta_2^*}{\sigma_v} - \phi\left[\frac{\beta_0^* + \beta_2^* X_i}{\sigma_v}\right] \cdot \frac{\beta_2^*}{\sigma_v}$$

Hence, as long as $\phi\left[\frac{\beta_0^*+\beta_1^*+\beta_2^* X_i}{\sigma_v}\right] \neq \phi\left[\frac{\beta_0^*+\beta_2^* X_i}{\sigma_v}\right]$ and $\beta_2^* \neq 0$, the marginal effect of D_i on Y_i is, in general, a function of X_i. Note that $\phi\left[\frac{\beta_0^*+\beta_1^*+\beta_2^* X_i}{\sigma_v}\right] = \phi\left[\frac{\beta_0^*+\beta_2^* X_i}{\sigma_v}\right]$ only when $\beta_1^* = 0$ or $2\beta_0^* + \beta_1^* + 2\beta_2^* X_i = 0$.

Second, if X_i is binary, there is another way to formally give the intuition (again, just a small extension to p97 in the Mostly Harmless).

As before, the CEF is

$$E(Y_i|D_i, X_i) = \Phi\left[\frac{\beta_0^* + \beta_1^* D_i + \beta_2^* X_i}{\sigma_v}\right]$$

As D_i and X_i are both binary, the CEF can be written as

$$\Phi\left[\frac{\beta_0^*}{\sigma_v}\right] + \left\{\Phi\left[\frac{\beta_0^* + \beta_1^*}{\sigma_v}\right] - \Phi\left[\frac{\beta_0^*}{\sigma_v}\right]\right\} \cdot D_i + \left\{\Phi\left[\frac{\beta_0^* + \beta_2^*}{\sigma_v}\right] - \Phi\left[\frac{\beta_0^*}{\sigma_v}\right]\right\} \cdot X_i + \left\{\Phi\left[\frac{\beta_0^* + \beta_1^* + \beta_2^*}{\sigma_v}\right] - \Phi\left[\frac{\beta_0^* + \beta_1^*}{\sigma_v}\right] - \Phi\left[\frac{\beta_0^* + \beta_2^*}{\sigma_v}\right] + \Phi\left[\frac{\beta_0^*}{\sigma_v}\right]\right\} \cdot X_i D_i$$

So, as long as $\Phi\left[\frac{\beta_0^*+\beta_1^*+\beta_2^*}{\sigma_v}\right] - \Phi\left[\frac{\beta_0^*+\beta_1^*}{\sigma_v}\right] - \Phi\left[\frac{\beta_0^*+\beta_2^*}{\sigma_v}\right] + \Phi\left[\frac{\beta_0^*}{\sigma_v}\right] \neq 0$, the interaction term $X_i D_i$ will have a non-zero coefficient in the CEF, hence the population AME function as well. Note that the equality can only be achieved if $\int_{\frac{\beta_0^*+\beta_1^*}{\sigma_v}}^{\frac{\beta_0^*+\beta_1^*+\beta_2^*}{\sigma_v}} \phi dv = \int_{\frac{\beta_0^*}{\sigma_v}}^{\frac{\beta_0^*+\beta_2^*}{\sigma_v}} \phi dv$, which is generally not true.