Comment on: "Homophily and Transitivity in Dynamic Network Formation"*

Áureo de Paula †

March 2026

1 Introduction

Dano, Graham and Sbai Sassi offer an elegant econometric study on models of network formation. Their model and analysis expand on previous work by the authors in many ways and I expect it to add to the toolbox available to researchers interested in the theme. In this short comment, I highlight some of the underlying assumptions implicit in the framework adopted and present a few (admittedly underdeveloped) thoughts on potential avenues for future research.

The paper focusses on an econometric model whereby a connection in period t between two individuals i and j forms ($D_{ijt} = 1$) or not ($D_{ijt} = 0$) according to:

$$D_{ijt} = \mathbf{1}\left(\alpha_0 D_{ijt-1} + \beta_0 \left(\sum_k D_{ikt-1} D_{jkt-1}\right) + A_{ij} \geq U_{ijt}\right), t = 1, \ldots, T. \tag{1}$$

Here, as conventional, the indicator $\mathbf{1}(\cdot)$ is one whenever the expression in brackets holds and zero, otherwise. Pairwise connections between i and j in a given period t are marked by D_{ijt} as already suggested above and one observes them (or their absence) for a panel with T periods. A link between i and j in period t thus depends on whether it was present or not in the previous period (D_{ijt-1}), how many connections in common i and j have ($\sum_k D_{ikt-1} D_{jkt-1}$), and a time invariant benefit/cost constant (A_{ij}) that facilitates (when positive) or complicates (when negative) the establishment of a link between those individuals. The data generating process is finally completed with a time varying error term (U_{ijt}), which can be seen as an idiosyncratic shock to benefits and/or costs of link formation and is presumed to follow a distribution whose CDF is

*I gratefully acknowledge financial support from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Ref: EP/X02931X/1) and the Economic and Social Research Council (ESRC) through the ESRC Institute for the Microeconomic Analysis of Public Policy (Grant Ref: ES/T014434/1). I am grateful to Bo Honoré, Elie Tamer and Siqi Wei for useful exchanges in preparation of this discussion. I declare that there are no relevant financial or non-financial competing interests to report

†de Paula: University College London, CeMMAP and IFS, a.paula@ucl.ac.uk

denoted by $F_U(\cdot)$. Taking the pair (i.e., i and j) as the unit of observation, this is a natural and powerful generalisation of by now well-known dynamic, nonlinear panel models (see de Paula and Honoré (2021) and references therein).

When taking the framework above to data, the (*integrated*) likelihood function related to the model above becomes:

$$\begin{aligned} p^{\mathrm{I}}\left(\mathbf{d}^T;\theta,\eta\right) = \int \quad & \ldots \int \prod_{t=1}^{T} \prod_{i<j} \Big\{ F_U\left(\alpha d_{ijt-1} + \beta r_{ijt-1} + a_{ij}\right)^{d_{ijt}} \\ & \times \left[1 - F_U\left(\alpha d_{ijt-1} + \beta r_{ijt-1} + a_{ij}\right)\right]^{1-d_{ijt}} \Big\} \\ & \times \pi_{\mathbf{D}_0|\mathbf{A}}\left(\mathbf{d}_0 \mid \mathbf{a};\eta_1\right) \pi_{\mathbf{A}}\left(\mathbf{a};\eta_2\right) \mathrm{d}a_{21}, \ldots, \ \mathrm{d}a_{NN-1}, \end{aligned} \tag{2}$$

where $\mathbf{d}^T$ collects realisations of D_{ijt} across pairs and times in the data, which are individually denoted by d_{ijt}; r_{ijt-1} records the realised sum of common connections between i and j; and $\mathbf{d}_0$ enumerates the initial ($t = 0$) links between pairs in the data. Similarly, $\mathbf{a}$ lists the realisations fo A_{ij} for all the pairs and $\pi_{\mathbf{D}_0|\mathbf{A}}\left(\cdot;\eta_1\right), \pi_{\mathbf{A}}\left(\cdot;\eta_2\right)$ denote the distributions of $\mathbf{D}_0$ given $\mathbf{A}$ and of $\mathbf{A}$ and are presumably guided by parameters η_1 and η_2, respectively. Above, $\theta = (\alpha, \beta)$ and $\eta = (\eta_1, \eta_2)$.

The highlight several difficulties with the likelihood function above. For example, the model "provides no guidance on how to specify $\pi_{\mathbf{D}_0|\mathbf{A}}\left(\mathbf{d}_0 \mid \mathbf{a};\eta_1\right)$". In circumventing those, they instead focus on the *conditional* likelihood function (for logistically distributed U_{ijt}), which allows them to appeal to appropriate sufficient statistics (for $\mathbf{A}$ and π) and network sequences. The starting point is to operationalise this are p-star subnetworks whereby a focal individual is connected with p other ones. They then demonstrate that p-starts in stable neighbourhoods can be used to obtain a conditional likelihood that "differences away" the nuisance objects in the more complex likelihood functions above (e.g., the A_{ij} pairwise terms). A stable neighbourhood in turn is a subgraph where peers to members of the star maintain their link status (with those as well as others in the network) and members of the star (other than the focal indvidual) maintain their link status (among themselves) across periods. If such a p-star embedded in a stable neighbourhood is such that the links between focal vertex and others in the star differ across periods, those "identifying p-stars" can be used to form a conditional likelihood function that is still informative about θ. The paper presents an extensive examination of this estimation protocol and examines its finite and large sample properties.

2 A Few Underlying Assumptions

While the model in expression (1) is seemingly simple, it is already sufficiently rich and complex for many purposes. That said, it is probably important for potential users to bear in mind some of the underlying assumptions that it encodes.

One of them is that although it depends on other features of the network, it does so only with

a lag: the network formation is essentially *dyadic*. More precisely, it rules out processes like:

$$D_{ijt} = \mathbf{1}\left(\rho_0\left(\sum_k D_{ikt}D_{jkt}\right) + \alpha_0 D_{ijt-1} + \beta_0\left(\sum_k D_{ikt-1}D_{jkt-1}\right) + A_{ij} \geq U_{ij}\right), t = 1, \ldots, T,$$

where $\rho_0 \neq 0$ and contemporaneous "externalities" (who is linked to either i or j) matter. This may or may not be an important guide in the formation of connections depending on context and other aspects such as temporal aggregation of the data. For example, third-party connections to whom a potential risk-sharing partner is exposed to in the present period may matter for risk-sharing relationship to form as documented in Graham and Pelican (2020), Graham and Pelican (2026). Accommodating those contemporenously though introduces simultaneity and, in a nonlinear system like this, potential multiplicity (see, for example, de Paula and Honoré (2021)).

Another aspect to bear in mind is that the economic behaviour narrative used in the paper to render (1) relies on pairwise stability (with transfers). Pairwise stability in networks is a (rightly!) celebrated solution concept first put forward by Jackson and Wolinsky (1996). This equilibrium notion has been used extensively in the literature (see, e.g., de Paula et al. (2018) and de Paula (2020)) and requires that: (a) any observed link have positive surplus and (b) non-links have negative surplus. Among other things, this requires knowledge of the surplus across *all* potential links and *all* agents. This may be reasonable in a small to moderately sized setting, but may be epistemically demanding in larger groups (e.g. $N = 5,000$). Whereas (a) remains reasonable, it may be that links did not form for lack of "opportunity" and it is unclear whether relaxing (b) would be possible in this setting. This could in principle be done by having a truncated *meeting protocol* (e.g., Mele (2017) or Christakis et al. (2020)) or a *dyad sampling* scheme though it is unclear whether this would preserve the properties exploited by Dano, Graham and Sbai Sassi. Of course, the model (1) remains a valid statistical one even without a behavioural interpretation and it (b) may well be a plausible condition even in larger groups.

3 No man is an island ...

Once one embraces the assumptions underlying (1), the prevalence identifying p-star system is important for the procedure. Since those are key for the conditional likelihood and in informing the estimation of θ, their frequency affects convergence rates in the large sample analysis as established by the authors, for example. This is similar to the role played by "informative quadruples" in the static version of (1) explored in Graham (2017), Charbonneau (2017) and Jochmans (2018).

It is not surprising that the prevalence of stable neighbourhoods and p-star systems is connected to inference and estimation precision. I conjecture that they may also carry information on the parameters (θ) themselves and thus be useful for identification. To (perhaps absurdly) illustrate

this point, imagine two island societies, A and B, populated by two castaways each. Call the inhabitants of island A, Robinson and Friday (in a nod to Daniel Defoe's classic novel), and the castaways in island B, Chuck and Wilson (in reference to the feature film *The Castaway*). Robinson and Friday as well as Chuck and Wilson relationships are governed by (1). Suppose A_{ij} is moderate within each island and α_0 is small. Then, it is likely that sometimes we will see islanders bicker between themselves ($D_{ijt} = 0$) and otherwise, make peace ($D_{ijt} = 1$). At same time, assume that the two islands are nonetheless separated by shark-infested waters and it is rather costly for relationships across islands to form (i.e., A_{ij} is very negative, possibly $-\infty$).

No man is an island... but each of these islands is an identifying p-system (given the variation in the within-island friendships) within a stable neighborhood (since there is never a link between Robinson and Chuck or Wilson or between Friday and Chuck or Wilson)! If on the other hand the four castaways had landed on the same island, A_{ij} being moderate (and similar) for any two pairs since there are no longer shark-infested waters separating them, it is perhaps less likely that we find a stable neighborhood (giving the alternating bickering and peace making among them) or therefore an identiying p-star.

One may thus wonder: if A_{ij} is more homogeneous and α_0 is larger, do identifying p-star systems become rarer? Turning this around: would it be possible to use the prevalence of stable neighbourhoods and p-star systems to gain further information on α_0 and β_0?

4 Beyond α_0 and β_0 and Alternative Routes

As noted previously, the article points out difficulties with the (integrated) likelihood in (2). It focusses on a conditional likelihood function in a "fixed effects" paradigm, which is agnostic about $\pi_{\mathbf{D}_0|\mathbf{A}}(\mathbf{d}_0 \mid \mathbf{a}; \eta_1)\,\pi_{\mathbf{A}}(\mathbf{a}; \eta_2)$. Occasionally though, '...if, as is often the case, one wants to use the model for prediction or for calculating the effect of various 'what-if's', then a random effects model would be preferable" (Honoré (2002)).

As pointed out by the authors, one of those difficulties in taking this forward relates to the fact that the model "provides no guidance on how to specify $\pi_{\mathbf{D}_0|\mathbf{A}}(\mathbf{d}_0 \mid \mathbf{a}; \eta_1)$". It may nonetheless be more palatable to condition on $\mathbf{D}_0$ and specify $\pi_{\mathbf{A}|\mathbf{D}_0}(\mathbf{a} \mid \mathbf{d}_0; \eta_1)$ (see, e.g., Wooldridge (2005) or Todd and Wolpin (2006)).

In doing so, one reasonable starting point is to assume that supp($\mathbf{A}$) is finite (as in the MC exercise presented in the paper). The following figure, for example, plots estimated "individual effects" from a study on "favor networks" in India using the data in Banerjee et al. (2013). There Dzemski (2017) estimates a related, but different model where A_{ij} is the sum of individual effects depicted in the figure.[1] The salient point here is that estimated "types" cluster in a few groups.

[1] The figure corresponds to Figure B1 in the working paper version of Dzemski (2019).

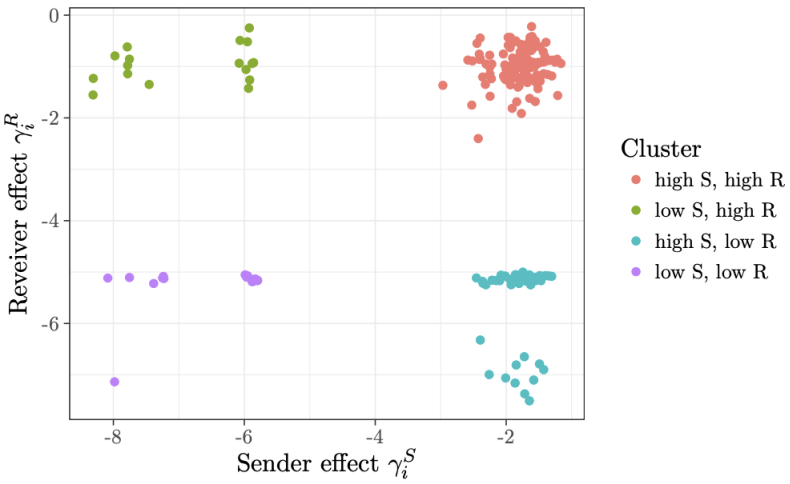

Figure B.1: Distribution of estimated agent effects in median village (village = 15).

Source: Dzemski (2017)

As Dano, Graham and Sbai Sassi duly note, it is desirable to choose a specification that "allows A_{ij} and A_{ik} to covary." Goldsmith-Pinkham and Imbens (2013), for instance, employ:

$$A_{ij} = \gamma_0 + \gamma_1 \xi^*_{ij}, \qquad \xi^*_{ij} = |\xi_i - \xi_j|, \quad \xi_i, \xi_j \sim \text{Ber}(p),$$

setting $p = 1/2$. This induces correlation (controlled by p) between A_{ij} and A_{kl} when $\{i,j\} \cap \{l,k\} \neq \emptyset$ and independence, otherwise. There is by now a sizeable literature on the estimation and simulation of (high-dimensional) multivariate (flexible) Bernoulli distributions that could be employed instead (see, for example, Dai et al. (2013), Jiang et al. (2021), Fontana and Semeraro (2023) and references therein).

This being a "mixture", one can in principle then use a *stochastic* Expectation-Maximisation (EM) algorithm (see, e.g., Arellano and Bonhomme (2017)), alternating until convergence:

> E-step. Draw from the implied posterior $\pi_{\mathbf{A}|(\mathbf{D}_t)_{t=0}^T} \propto \mathcal{L}((\mathbf{D}_t)_{t=0}^T \mid \mathbf{A}; \alpha^{(k)}, \beta^{(k)}) \pi_{\mathbf{A}|\mathbf{D}_0}(\eta^{(k)})$ using, e.g., simulation plus a Metropolis-Hastings sampler;
>
> M-step. Given the (complete data) likelihood, update the parameter estimates to $\alpha^{(k+1)}, \beta^{(k+1)}$ and $\eta^{(k+1)}$.

The simulation in the E-step avoids the high-dimensional integration which is correctly highlighed in the paper as another difficulty with the integrated likelihood (see Diebolt and Celeux (1993) and Nielsen (2000)). Because the EM algorithm can be slow, acceleration methods can also be deployed (see, e.g., Wei (2025)).

One reason why protocols such as the one above may be useful is that, while the coefficients α_0 and β_0 may be of interest on their own, one may need information about A_{ij} for other goals. One potential motivation for estimating the network formation model in (1), for instance, would

be to control for link endogeneity/selection in an outcome equation of interest. In this case, one needs more information about the "pairwise effects". Goldsmith-Pinkham and Imbens (2013) (see above) allow ξ_i to correlate with unobservables in an outcome equation (on academic achievement) in their application. Their estimates suggest endogeneity/selection is not salient. Bernard et al. (2022), on the other hand, investigate firm size (outcome) in a supply chain (network formation) setting. Here, correlation in productivity and "relationship costs" is estimated to be important. In this application, $N = 94,147$ firms (in 2014)! The EM algorithm outlined above (even with the acceleration methods alluded to) might be slow then. It would be interesting to examine whether one employ the conditional likelihood ideas in the present paper alongside these other protocols to learn about the distribution of A_{ij}.

References

Arellano, M. and S. Bonhomme (2017): "Nonlinear Panel Data Methods for Dynamic Heterogeneous Agent Models," *Annual Review of Economics*, 9, 471–496.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The Diffusion of Microfinance," *Science*, 341, 1236498.

Bernard, A. B., E. Dhyne, G. Magerman, K. Manova, and A. Moxnes (2022): "The Origins of Firm Heterogeneity: A Production Network Approach," *Journal of Political Economy*, 130, 1765–1804.

Charbonneau, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20, S1–S13.

Christakis, N., J. Fowler, G. W. Imbens, and K. Kalyanaraman (2020): "Chapter 6 - An empirical model for strategic network formation," in *The Econometric Analysis of Network Data*, ed. by B. Graham and Á. de Paula, Academic Press, 123–148.

Dai, B., S. Ding, and G. Wahba (2013): "Multivariate Bernoulli distribution," *Bernoulli*, 19, 1465 – 1483.

de Paula, A. (2020): "Strategic Network Formation," in *The Econometric Analysis of Network Data*, ed. by B. Graham and A. de Paula, Academic Press/Elsevier, 42–64.

de Paula, A. and B. Honoré (2021): "Identification in Simple Binary Outcome Panel Data Models," *Econometrics Journal*, 24, C78–C93.

de Paula, A., S. Richards-Shubik, and E. Tamer (2018): "Identifying Preferences in Networks with Bounded Degree," *Econometrica*, 86, 263–288.

Diebolt, J. and G. Celeux (1993): "Asymptotic properties of a stochastic EM Algorithm for estimating mixing proportions," *Communications in Statistics. Stochastic Models*, 9, 599–613.

Dzemski, A. (2017): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," University of Gothenburg Working Paper No. 698.

——— (2019): "An Empirical Model of Dyadic Link Formation in a Network with Unobserved Heterogeneity," *The Review of Economics and Statistics*, 101, 763–776.

Fontana, R. and P. Semeraro (2023): "Exchangeable Bernoulli distributions: High dimensional simulation, estimation, and testing," *Journal of Statistical Planning and Inference*, 225, 52–70.

Goldsmith-Pinkham, P. and G. W. Imbens (2013): "Social Networks and the Identification of Peer Effects," *Journal of Business & Economic Statistics*, 31, 253–264.

Graham, B. and A. Pelican (2026): "An optimal test for strategic interaction in network formation games," Working Paper.

Graham, B. S. (2017): "An Econometric Model of Network Formation With Degree Heterogeneity," *Econometrica*, 85, 1033–1063.

Graham, B. S. and A. Pelican (2020): "Chapter 4 - Testing for externalities in network formation using simulation," in *The Econometric Analysis of Network Data*, ed. by B. Graham and Á. de Paula, Academic Press, 63–82.

Honoré, B. E. (2002): "Nonlinear models with panel data," *Portuguese Economic Journal*, 1, 163–179.

Jackson, M. O. and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, 71, 44–74.

Jiang, W., S. Song, L. Hou, and H. Zhao (2021): "A Set of Efficient Methods to Generate High-Dimensional Binary Data With Specified Correlation Structures," *The American Statistician*, 75, 310–322.

Jochmans, K. (2018): "Semiparametric Analysis of Network Formation," *Journal of Business & Economic Statistics*, 36, 705–713.

Mele, A. (2017): "A Structural Model of Dense Network Formation," *Econometrica*, 85, 825–850.

Nielsen, S. F. (2000): "The stochastic EM algorithm: estimation and asymptotic results," *Bernoulli*, 6, 457 – 489.

Todd, P. E. and K. I. Wolpin (2006): "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *American Economic Review*, 96, 1384–1417.

Wei, S. (2025): "Estimating Latent-Variable Panel Data Models Using Parameter-Expanded SEM Methods," *Journal of Business & Economic Statistics*, 43, 324–337.

Wooldridge, J. M. (2005): "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity," *Journal of Applied Econometrics*, 20, 39–54.