The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–25

netivreg: Estimation of Peer Effects in Endogenous Social Networks

Pablo Estrada
Emory University
Atlanta, USA
pestrad@emory.edu

Juan Estrada
Emory University
Atlanta, USA
jjestra@emory.edu

Kim P. Huynh
Bank of Canada
Ottawa, Canada
kim@huynh.tv

David Jacho-Chávez
Emory University
Atlanta, USA
djachocha@emory.edu

Leonardo Sánchez-Aragón
ESPOL University
Guayaquil, Ecuador
lfsanche@espol.edu.ec

Abstract. The command `netivreg` implements the Generalized Three-Stage Least Squares (G3SLS) estimator developed in Estrada et al. (2020, "*On the Identification and Estimation of Endogenous Peer Effects in Multiplex Networks*") and the Generalized Method of Moments (GMM) estimator in Chan et al. (2022, "*On the Estimation of Social Effects with Observational Network Data and Random Assignment*") for the endogenous linear-in-means model. The two procedures utilize full observability of a two-layered multiplex network data structure using Stata's new multiframes capabilities and Python integration (version 16 and above). Applications of the command include simulated data and three years' worth of data on peer-reviewed articles published in top general interest journals in Economics.

Keywords: st0001, Instrumental variables, `ivregress`, multiplex networks, Python

1 Introduction

In various settings, the decision of agents (people, firms, or countries, for example) to exert effort in some activity depends not only on their own characteristics (direct effects), but also on the efforts (spillover effect) and characteristics of their peers (contextual effects). In general, the literature has focused on the linear models of peer effects to test the existence of potential influences of connections on individual outcomes. Linear models with social interactions tend to have identification challenges widely recognized in the econometrics of networks literature. One of the outstanding identification issues in the field is how to address the endogenous network formation problem (Jackson et al. 2017). This paper presents the Stata implementation of two estimators capable of estimating social (spillover, contextual, and direct) effects in contexts where the network of interest is endogenous. The command `netivreg` implements the Generalized Three-Stage Least Squares (G3SLS) estimator developed in Estrada et al. (2020) and the Generalized Method of Moments (GMM) estimator in Chan et al. (2022) for the endogenous networks of linear models of social effects.

We consider Manski's (1993) linear peer effects specification, widely known as the *linear-in-means* model, where an outcome variable for agent $i \in \{1, \ldots, n\}$, y_i, is deter-

 st0001

mined according to

$$y_i = \alpha + \beta \sum_{j \neq i} w_{i,j} y_j + \sum_{j \neq i} w_{i,j} \mathbf{x}_j^\top \boldsymbol{\delta} + \mathbf{x}_i^\top \boldsymbol{\gamma} + v_i, \tag{1}$$

where $j \in \{1, \ldots, n\}$, $\mathbf{x}_i$ is agent i's $k \times 1$ vector of attributes; $w_{i,j} = 1$ if agent j shares a social connection with i, and is 0 otherwise; v_i represents agent i's unobservables, and n is the number of agents in the sample. The social network structure is fully characterized by the square $n \times n$ matrix, $\mathbf{W}$, with (i, j) entry given by $w_{i,j}$; i.e., the adjacency matrix. This general econometric network model can be written in matrix form as

$$\mathbf{y} = \boldsymbol{\iota}\alpha + \mathbf{W}\mathbf{y}\beta + \mathbf{W}\mathbf{X}\boldsymbol{\delta} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{v}, \tag{2}$$

where the *peer effect* (captured by β) measures how an agent's outcome may depend on those of her peers. The *contextual effect*, captured by the coefficients $\boldsymbol{\delta}$, occurs when an agent's outcome may depend on the exogenous characteristics of her peers, and the *direct effects*, captured by the coefficients $\boldsymbol{\gamma}$, occurs when an agent's outcome may depend on her characteristics.

Under the assumption that $E[\mathbf{v}|\mathbf{X}, \mathbf{W}] = \mathbf{0}$, the `netivreg` command implements Bramoullé et al.'s (2009) Generalized Two-Stage Least Squares (G2SLS) estimator of the structural parameters $\boldsymbol{\psi} \equiv [\alpha, \beta, \boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top]^\top$. From equation (2), we see that the G2SLS estimator is a special case of Estrada et al.'s (2020) Generalized Three-Stage Least Squares (G3SLS) estimator that assumes $E[\mathbf{v}|\mathbf{X}, \mathbf{W}_0] = \mathbf{0}$ instead. The square $n \times n$ matrix, $\mathbf{W}_0$, with (i, j) entry given by $w_{0;i,j}$ represents *another* adjacency matrix of exogenous connections. The `netivreg` command also implements the Generalized Method of Moments (GMM) estimator in Chan et al.'s (2022) where the social interaction network $\mathbf{W}_0$ is randomized instead.

The `netivreg`'s internal numerical implementation of the G3SLS and GMM estimators use the Python language (version 16 and above). It makes full use of Stata's new integrability with Python, as well as Stata's new data frames capabilities to handle the data sets $[\mathbf{y}, \mathbf{X}]$, $\mathbf{W}$, and $\mathbf{W}_0$; see, e.g., Ho et al. (2021). The command also exploits Python architecture to handle sparse matrices by asking the user to provide the $\mathbf{W}$, and $\mathbf{W}_0$ adjacency matrices as simple (i, j) lists for all pairs in which $w_{i,j} = 1$ and $w_{0;i,j} = 1$.

1.1 Related Literature

Estimating network effects (peer and contextual effects) under network endogeneity can be problematic and is an active area of research among social scientists. Recent developments in the literature propose methodologies that generally require augmenting the standard linear-in-means model to include specific network formation processes. Early methods using network formation models to control for network endogeneity include Goldsmith-Pinkham and Imbens (2013) and Qu and Lee (2015). More recent approaches that also use auxiliary network formation models include Johnsson and Moon (2021), who take a control function approach based on the fitted values of a network formation model *a la* Graham (2017). Auerbach (2022) also uses a method based on matching pairs of similar agents based on the columns of the adjacency matrix representing the network of interest. Cerulli (2017) presents a complete literature review of the topic from the perspective of estimating treatment effects under potential network interference.

The estimation methods that we showcase in this paper differs from previous literature in that we do not require to specify a structural network formation model. To be specific, typical network formation models involve additional assumptions such as the absence of strategic interactions on individuals' utilities of forming peers. Our agnostic approach to the network formation mechanism therefore offers an important advantage. For a complete discussion on the importance of strategic interactions to network formation models' point-identification, see Graham (2017), De Paula et al. (2018), and Graham and Pelican (2020).

1.2 Command Prerequisites

The `netivreg` command requires a working Python 3.7 or higher (Van Rossum and Drake 2009) distribution already installed – the Anaconda's (Anaconda Software Distribution 2020) distribution is strongly recommended. The user will also need to install the NetworkX (Hagberg et al. 2008), Numpy (Oliphant 2006), Pandas (McKinney et al. 2010), Scikit-learn (Pedregosa et al. 2011), and SciPy (Virtanen et al. 2020) Python packages and their dependencies. The command also makes use of the Python native `os` and `sys` modules.

The `netivreg` command works with Stata version 16.0 or higher. The Stata Function Interface (sfi) Python module shipped with the installed Stata version and flavor is must work properly as it provides a bidirectional connection between the local installation of Stata and Python. Table 1 lists the required software and needed versions.

Table 1: Required Software

Language	Version	Python Packages	Version
Stata	16.0 or higher	NetworkX	2.4.x
Python	3.7 or higher	Numpy	1.20.x
		Pandas	1.2.x
		Scikit-learn	0.24.x
		SciPy	1.6.x

The user is strongly encouraged to create a virtual environment with the required

Python packages versions listed in Table 1 in order to maintain backward compatibility as time passes. For example, on Windows, the user can open a command prompt and create a conda environment using Anaconda as follows,

```
> conda create -n env_netivreg -c conda-forge python=3.8 ipykernel=6.13.
0 networkx=2.4 numpy=1.20.1 pandas=1.2.4 scikit-learn=0.24.1 scipy=1.6.2
> pip install stata-setup==0.1.2
```

The previous code creates a conda environment named `env_netivreg` with Python 3.8 and the necessary packages that `netivreg` needs to run without errors.

1.3 Real Data Running Example

We use a running real-data example throughout this paper to illustrate the ideas in the methodology and estimation sections. We also present the results of estimating a linear-in-means model that aims at quantifying the potential existence of human capital externalities (peer effects) among scholars publishing in four of the top general-interest journals in economics. In particular, we use the 729 peer reviewed research articles published in the *American Economics Review* (`AER`), *Econometrica* (`ECA`), the *Journal of Political Economy* (`JPE`), and the *Quarterly Journal of Economics* (`QJE`) from 2000 to 2002 taken from Estrada et al. (2020). Section 5.2 below describes the data and the empirical model of interest.

The main goal is to estimate the potential peer effects in citations and the contextual effects from gender and editor-in-charge status while controlling for a set of direct effects, including articles' characteristics such as the number of pages, authors, and references. Assuming that articles are the unit of observation, we estimate the parameters of interest. Two articles are connected if at least two of their authors are linked in one of the two observed networks: the coauthors' network $\mathbf{W}$ and the alumni network $\mathbf{W}_0$.

The rest of the paper is organized as follows: Section 2 introduces the theoretical framework and the identification conditions of model (2) with endogenously-formed social interactions. The estimation algorithm implemented by the `netivreg` command is provided in Section 3, while Section 4 provides the command syntax information. Section 5 illustrates how to use the command with simulated data and an empirical application. Section 6 concludes.

2 Methodology

The main idea of the methodology is to propose a set of sufficient conditions to identify the parameters of interest in equation (1) when the network of interest represented by $\mathbf{W}$ is formed endogenously. We propose two approaches for identification, both founded on the idea of the existence of an additional set of connections that are exogenous with respect to $\mathbf{v}$. Let $\mathbf{S}$ be a $n \times (k+1)$ matrix given by $\mathbf{S} \equiv [\mathbf{y} \quad \mathbf{X}]$ and let $\boldsymbol{\theta} \equiv (\beta, \boldsymbol{\delta}^\top)^\top$ be a $(k+1) \times 1$ vector of parameter such that $\beta\mathbf{W}\mathbf{y} + \mathbf{W}\mathbf{X}\boldsymbol{\delta} = \mathbf{W}\mathbf{S}\boldsymbol{\theta}$. Therefore, equation

(2) can be written as

$$\mathbf{y} = \alpha\boldsymbol{\iota} + \mathbf{WS}\boldsymbol{\theta} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{v}. \tag{3}$$

Estrada et al. (2020) introduces an additional auxiliary system of equations given by

$$\mathbf{WS} = \mathbf{W}_0\mathbf{S}\boldsymbol{\Pi} + \mathbf{U}, \tag{4}$$

where $\boldsymbol{\Pi} = [\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{k+1}]^\top$ represents a full rank $(k+1) \times (k+1)$ matrix of system coefficients and the $n \times (k+1)$ matrix of system errors $\mathbf{U}$ is such that $E[\mathbf{U}|\mathbf{S}, \mathbf{W}_0] = \mathbf{O}$ (a matrix of zeros). In our running example, the outcome in equation (3) is the natural logarithm of article i's citations eight year post publication. The matrix $\mathbf{S}$ includes the outcome variable, whether the research team is all the same gender, and whether the team consists of an editor in charge of one of the four journals included in the sample. Notice that $\mathbf{WS}$ in equation (4) represents the average values of the variables in $\mathbf{S}$ for the set of connections of each article i in the coauthors network $\mathbf{W}$, while $\mathbf{W}_0\mathbf{S}$ represents the average for the same variables for the set of connections of each article i in the alumni network. Therefore, equation (4) represents a regression of average values of $\mathbf{S}$ in $\mathbf{W}$ on average values of $\mathbf{S}$ in $\mathbf{W}_0$.

We use the projection in equation (4) and the assumption that $E[\mathbf{v}|\mathbf{X}, \mathbf{W}_0] = \mathbf{0}$ to identify the parameters of interest. By substituting (4) in (3), one has

$$\mathbf{y} = \alpha\boldsymbol{\iota} + \mathbf{W}_0\mathbf{S}\boldsymbol{\Pi}\boldsymbol{\theta} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{e}, \tag{5}$$

where $\mathbf{e} \equiv \mathbf{U}\boldsymbol{\theta} + \mathbf{v}$. However, (5) cannot be estimated by simple Ordinary Least Squares (OLS) because $E[\mathbf{S}^\top\mathbf{W}_0\mathbf{e}] \neq \mathbf{0}$; i.e., the simultaneity of $\mathbf{W}_0\mathbf{y}$ still persists and an Instrumental Variable (IV) procedure is required. Estrada et al. (2020) show that $\mathbf{W}_0^p\mathbf{X}$, where $p > 1$ is a valid instrument for $\mathbf{W}_0\mathbf{y}$ in (5). In the case when $p = 2$, if agents (i, j) have a connection and (j, l) have a connection, it does not necessarily imply that (i, l) also have a connection. Therefore, following Bramoullé et al. (2009), results in Estrada et al. (2020) have shown that if the matrices $[\mathbf{I},\ \mathbf{W}_0, \ldots\ \mathbf{W}_0^p]$ are linearly-independent and $\beta(\gamma_k\pi_{1,1} + \pi_{k,1}) + \sum_{i=1}^{k} \delta_i(\gamma_i\pi_{1,i+1} + \pi_{k,i+1}) \neq 0$ for all k, the social effects $\boldsymbol{\psi}$ are identified. In our running example, the instrumental variable when $p = 2$ is the matrix $\mathbf{W}_0^2\mathbf{X}$, which contains the characteristics of articles that are indirectly connected via the alumni links.[1]

The GMM estimator proposed by Chan et al. (2022) does not involve an additional auxiliary system of equations in (4). Instead, identification follows from the moment condition that aggregates local heterogeneous identifying information for all the individuals in the population given by $\boldsymbol{m}(\boldsymbol{\psi}) := \sum_{i\in\mathcal{I}_N} \mathbf{z}_i v_i$, where $\mathbf{z}_i$ is the ith column of the matrix of instruments $\mathbf{Z} \equiv [\mathbf{W}_0^p\mathbf{X}\ \ \mathbf{W}_0^{p-1}\mathbf{X}\ \ \ldots\ \ \mathbf{W}_0\mathbf{X}\ \ \mathbf{X}\ \ \boldsymbol{\iota}]$ for some $p \geq 2$.

1. When there are not indirect links, such as when the network is partitioned into connected groups, the identifying variation comes from differences in groups sizes (Bramoullé et al. 2009).

Chan et al. (2022) shows that identification is possible in a context where the network of interest is formed endogenously by taking advantage of the exogeneity (randomization) and exclusion restrictions on the network represented by $\mathbf{W}_0$.

3 Estimation

This section describes the estimation algorithms for the equation of interest 1. We describe both the Generalized Three-Stage Least Squares (G3SLS) estimator in Estrada et al. (2020) and the Generalized Method of Moments (GMM) estimator in Chan et al. (2022).

3.1 Generalized Three-Stage Least Squares (G3SLS)

This estimator uses equation (4) to estimate the social parameters in (2) via (5). This approach can be reduced to Bramoullé et al.'s (2009) methodology for the case when $\mathbf{W} = \mathbf{W}_0$; i.e., the network is exogenous. The algorithm operates as follows.

Step 1: Regress $\mathbf{Wy}$ on $[\mathbf{W}_0\mathbf{y} \;\; \mathbf{W}_0\mathbf{X}]$ by Ordinary Least Squares (OLS) and get $\widehat{\mathbf{Wy}} = \mathbf{W}_0\mathbf{y}\widehat{\pi}_{1,1} + \mathbf{W}_0\mathbf{X}\widehat{\boldsymbol{\pi}}_{12}$, and $\widehat{\mathbf{u}}_1 = \mathbf{Wy} - \widehat{\mathbf{Wy}}$. In our running example, the regression of $\mathbf{Wy}$ on $[\mathbf{W}_0\mathbf{y} \;\; \mathbf{W}_0\mathbf{X}]$ means to run a regression of the average citations for the articles connected to i in the coauthorship network on the average values of the outcome and the regressors calculated using the alumni network.

Regress $\mathbf{WX}$ on $[\mathbf{W}_0\mathbf{y} \;\; \mathbf{W}_0\mathbf{X}]$ by OLS and get $\widehat{\mathbf{WX}} = \mathbf{W}_0\mathbf{y}\widehat{\boldsymbol{\pi}}_{21} + \mathbf{W}_0\mathbf{X}\widehat{\boldsymbol{\Pi}}_{22}$ and $\widehat{\mathbf{U}}_2 = \mathbf{WX} - \widehat{\mathbf{WX}}$. This is a system of regressions where the number of outcomes is determined by the number of variables included in $\mathbf{X}$ that generate contextual effects. For our running example, we include the articles' characteristics: editor-in-charge and different gender. Each outcome in the regression system then represents the average value of the regressor for the articles connected to i in the coauthorship network.

Step 2: Regress $\mathbf{y}$ on $[\boldsymbol{\iota} \;\; \mathbf{X} \;\; \mathbf{W}_0\mathbf{y} \;\; \mathbf{W}_0\mathbf{X}]$ by 2SLS using $\left[\boldsymbol{\iota} \;\; \mathbf{X} \;\; \mathbf{W}_0^2\mathbf{X} \;\; \mathbf{W}_0\mathbf{X}\right]$ as instruments. From 2SLS, get $\widehat{\boldsymbol{\psi}}_{2\mathrm{SLS}} \equiv [\widehat{\alpha}_{2\mathrm{SLS}}, \widehat{\boldsymbol{\gamma}}_{2\mathrm{SLS}}^\top, \widehat{\theta}_{1;2\mathrm{SLS}}, \widehat{\boldsymbol{\theta}}_{2;2\mathrm{SLS}}^\top]^\top$, where $\widehat{\boldsymbol{\theta}}_{2\mathrm{SLS}} \equiv \begin{pmatrix} \widehat{\theta}_{1;2\mathrm{SLS}} \\ \widehat{\boldsymbol{\theta}}_{2;2\mathrm{SLS}} \end{pmatrix} = \begin{pmatrix} \widehat{\pi}_{1,1} & \widehat{\boldsymbol{\pi}}_{12}^\top \\ \widehat{\boldsymbol{\pi}}_{21} & \widehat{\boldsymbol{\Pi}}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\theta}^*_{1;2\mathrm{SLS}} \\ \widehat{\boldsymbol{\theta}}^*_{2;2\mathrm{SLS}} \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\pi}}_1^\top \\ \widehat{\boldsymbol{\Pi}}_2^\top \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\theta}^*_{1;2\mathrm{SLS}} \\ \widehat{\boldsymbol{\theta}}^*_{2;2\mathrm{SLS}} \end{pmatrix}$.

In our running example, this step translates into runing an IV regression of the citation outcomes on the linear-in-means model specification using $\mathbf{W}_0$ instead of $\mathbf{W}$ and using the square of the exogenous matrix $\mathbf{W}_0$ as an instrument for the endogenous regressor $\mathbf{W}_0\mathbf{y}$. With the estimated coefficients from this step, we can calculate an initial estimator of the parameters of interest using the matrix of estimates from Step 1 given by $\boldsymbol{\Pi}$. The parameter $\widehat{\boldsymbol{\psi}}_{2\mathrm{SLS}}$ contains the peer, contextual, and direct effects on citations from equation (1).

Step 3: Regress $\mathbf{y}$ on $[\boldsymbol{\iota}\ \ \mathbf{X}\ \ \widehat{\mathbf{Wy}}\ \ \widehat{\mathbf{WX}}]$ by IV using $[\boldsymbol{\iota}\ \ \mathbf{X}\ \ \widehat{\mathbf{Z}}\widehat{\boldsymbol{\Pi}}\ \ \mathbf{W}_0\mathbf{X}\widehat{\boldsymbol{\Pi}}]$ as instruments, where $\widehat{\mathbf{Z}} = \mathbf{W}_0[\mathbf{I} - (\widehat{\boldsymbol{\pi}}_1^\top \widehat{\boldsymbol{\theta}}_{2\text{SLS}})\mathbf{W}_0]^{-1}\{\boldsymbol{\iota}\widehat{\alpha}_{2\text{SLS}} + \mathbf{X}\widehat{\boldsymbol{\gamma}}_{2\text{SLS}} + \mathbf{W}_0\mathbf{X}(\widehat{\boldsymbol{\Pi}}_2^\top \widehat{\boldsymbol{\theta}}_{2\text{SLS}})\}$. Call these IV estimates the resulting G3SLS estimator. From $\boldsymbol{\iota}\widehat{\alpha}_{\text{G3SLS}} + \mathbf{X}\widehat{\boldsymbol{\gamma}}_{\text{G3SLS}} + \widehat{\mathbf{Wy}}\widehat{\beta}_{\text{G3SLS}} + \widehat{\mathbf{WX}}\widehat{\boldsymbol{\delta}}_{\text{G3SLS}}$, we get

$$\widehat{\boldsymbol{\psi}}_{\text{G3SLS}} \equiv [\widehat{\alpha}_{\text{G3SLS}}, \widehat{\boldsymbol{\gamma}}_{\text{G3SLS}}^\top, \widehat{\beta}_{\text{G3SLS}}, \widehat{\boldsymbol{\delta}}_{\text{G3SLS}}^\top]^\top$$

and $\widehat{\mathbf{v}} \equiv \mathbf{y} - \boldsymbol{\iota}\widehat{\alpha}_{\text{G3SLS}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}_{\text{G3SLS}} - \mathbf{Wy}\widehat{\beta}_{\text{G3SLS}} - \mathbf{WX}\widehat{\boldsymbol{\delta}}_{\text{G3SLS}}$.

This final step is required in order to get the efficient estimator of the parameters of interest. The idea is to run a new IV regression where we now use the efficient instrument $\widehat{\mathbf{Z}}$ instead of $\mathbf{W}_0^2\mathbf{X}$ for the endogenous variable $\widehat{\mathbf{Wy}}$. The construction of the optimal instruments $\widehat{\mathbf{Z}}$ requires an initial estimator of the parameters of interest. We use the estimator $\widehat{\boldsymbol{\psi}}_{2\text{SLS}}$ from Step 2 to construct $\widehat{\mathbf{Z}}$.

Estrada et al. (2020) show that $\widehat{\boldsymbol{\psi}}_{\text{G3SLS}}$ is a consistent estimator of the structural parameters $\boldsymbol{\psi}$ and that $\sqrt{n}(\widehat{\boldsymbol{\psi}}_{\text{G3SLS}} - \boldsymbol{\psi})$ has an asymptotic multivariate normal distribution with a variance-covariance matrix that can be consistently estimated by

$$\widehat{\mathbf{V}}_{\boldsymbol{\psi}} = (n^{-1}\widehat{\widehat{\mathbf{Z}}}^{*\top}\widehat{\mathbf{D}})^{-1}(n^{-1}\widehat{\widehat{\mathbf{Z}}}^{*\top}\widehat{\mathbf{e}}^*\widehat{\mathbf{e}}^{*\top}\widehat{\widehat{\mathbf{Z}}}^*)(n^{-1}\widehat{\mathbf{D}}^\top\widehat{\widehat{\mathbf{Z}}}^*)^{-1},$$

where $\widehat{\mathbf{e}}^* = \mathbf{M}_{\mathbf{W}_0}\widehat{\mathbf{U}}\widehat{\boldsymbol{\theta}}_{\text{G3SLS}} + \widehat{\mathbf{v}}$ with $\mathbf{M}_{\mathbf{W}_0} \equiv \mathbf{I} - \mathbf{W_0}\mathbf{S}(\mathbf{S}^\top\mathbf{W}_0^2\mathbf{S})^{-1}\mathbf{S}^\top\mathbf{W_0}$. The residuals $\widehat{\mathbf{U}} \equiv \begin{bmatrix}\widehat{\mathbf{u}}_1 & \widehat{\mathbf{U}}_2\end{bmatrix}$ are obtained from step 1, $\widehat{\boldsymbol{\theta}}_{\text{G3SLS}} \equiv [\widehat{\beta}_{G3SLS}, \widehat{\boldsymbol{\delta}}_{\text{G3SLS}}^\top]^\top$, and the residuals $\widehat{\mathbf{v}}$ are taken from step 3. Similarly, $\widehat{\mathbf{D}} = [\iota, \mathbf{X}, \mathbf{W}_0\mathbf{y}, \mathbf{W}_0\mathbf{X}]\widehat{\boldsymbol{\Gamma}}$, where

$$\widehat{\boldsymbol{\Gamma}} = \begin{bmatrix}\mathbf{I}_{k+1} & \mathbf{O}_{k+1} \\ \mathbf{O}_{k+1} & \widehat{\boldsymbol{\Pi}}\end{bmatrix} \tag{6}$$

is a $(2k+2) \times (2k+2)$ matrix, $\mathbf{O}_{k+1}$ is a $(k+1) \times (k+1)$ matrix of zeros, and $\mathbf{I}_{k+1}$ represents the identity matrix of order $k+1$. The matrix $\widehat{\widehat{\mathbf{Z}}}^* = \widehat{\mathbf{Z}}^*\widehat{\boldsymbol{\Gamma}}$, where $\widehat{\mathbf{Z}}^* = [\boldsymbol{\iota}, \mathbf{X}, E_{\mathbf{X},\mathbf{W}_0}[\mathbf{W}_0\mathbf{y}](\widehat{\boldsymbol{\psi}}_{2\text{SLS}}, \widehat{\boldsymbol{\Pi}}), \mathbf{W}_0\mathbf{X}]$.

3.2 Generalized Method of Moments (GMM)

We can also estimate the social parameters in (2) by directly constructing moment conditions. First, we set up the moment conditions using the matrix of instruments $\mathbf{Z} \equiv [\mathbf{W}_0^p\mathbf{X}\ \ \mathbf{W}_0^{p-1}\mathbf{X}\ \ \ldots\ \ \mathbf{W}_0\mathbf{X}\ \ \mathbf{X}\ \ \boldsymbol{\iota}]$ for some $p > 1$, for the matrix $\mathbf{D} \equiv [\mathbf{Wy}\ \ \mathbf{WX}\ \ \mathbf{X}\ \ \boldsymbol{\iota}]$. In our running example, the construction of $\mathbf{Z}$ and $\mathbf{D}$ only requires defining the set of regressors in $\mathbf{X}$ and considering whether or not the contextual effects include all the regressors specified for the direct effects. In our empirical estimation, we use only a subset of the direct effects (editor in charge, same gender, number of pages, number of authors, number of references, and isolated) to specify the contextual effects (editor in charge and same gender). We then calculate a two-step GMM estimation as follows:

Step 1: Pick an initial weighting matrix $\mathbf{A}$, such as the identity matrix $\mathbf{I}$ or $\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}$, to calculate $\widetilde{\boldsymbol{\psi}}_{\text{GMM}} = \left[\mathbf{D}^\top\mathbf{Z}\mathbf{A}\mathbf{Z}^\top\mathbf{D}\right]^{-1}\left[\mathbf{D}^\top\mathbf{Z}\mathbf{A}\mathbf{Z}^\top\mathbf{y}\right]$.

Step 2: Calculate the efficient GMM estimator using a consistent estimator of the variance-covariance matrix $\mathbf{V}_{\boldsymbol{\psi}}$. Chan et al. (2022) propose the network Heteroskedasticity and Autocorrelation Consistent (HAC) variance estimator:

$$\widetilde{\mathbf{V}}_{\boldsymbol{\psi}} = \left[\mathbf{D}^\top\mathbf{Z}\widetilde{\boldsymbol{\Omega}}^{\star-1}\mathbf{Z}^\top\mathbf{D}\right]^{-1} \quad (7)$$

$$\widetilde{\boldsymbol{\Omega}}^{\star} = \sum_{d\geq 0} K\left(d/D\right)\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{P}_n(i,d)} \mathbf{z}_i\ \widetilde{e}_i\ \widetilde{e}_j\ \mathbf{z}_j^\top, \quad (8)$$

where $K(\cdot)$ is a kernel (weighting) function, such that $K(0) = 1$ and $K(u) = 0$ for $u > 1$, and $D = C \times [\log(\text{ average degree } \vee (1 + 0.05))]^{-1} \times \log n$ comes from the rule-of-thumb in Kojevnikov et al. (2021), $\mathcal{P}_n(i, d)$ is a set that contains the nodes at distance d of node i, and $\widetilde{e}_i = y_i - \mathbf{d}_i^\top\widetilde{\boldsymbol{\psi}}_{\text{GMM}}$ are the residuals using the GMM estimator of the first step. In the second step, the feasible efficient GMM estimator uses $\widetilde{\mathbf{V}}_{\boldsymbol{\psi}}^{-1}$ as a weighting matrix, so that

$$\widehat{\boldsymbol{\psi}}^{\star}_{\text{GMM}} = \left[\mathbf{D}^\top\mathbf{Z}\widetilde{\mathbf{V}}_{\boldsymbol{\psi}}^{-1}\mathbf{Z}^\top\mathbf{D}\right]^{-1}\left[\mathbf{D}^\top\mathbf{Z}\widetilde{\mathbf{V}}_{\boldsymbol{\psi}}^{-1}\mathbf{Z}^\top\mathbf{y}\right].$$

Again, to conduct Steps 1 and 2 empirically, we only need to define the set of regressors in the sets of variables generating both contextual and direct effects. The user also has to determine values for the kernel and some hyperparameters determining the HAC estimator to calculate the efficient GMM estimator. In the implementation, we set those values based on the rule-of-thumb in Kojevnikov et al. (2021).

Chan et al. (2022) show that, under a weak-dependence assumption in the network of agents, $\widehat{\boldsymbol{\psi}}^{\star}_{\text{GMM}}$ is a consistent estimator of the structural parameters $\boldsymbol{\psi}$ and $\sqrt{n}(\widehat{\boldsymbol{\psi}}^{\star}_{\text{GMM}} - \boldsymbol{\psi})$ has an asymptotic multivariate normal distribution with a variance-covariance matrix that can be consistently estimated by $[\mathbf{D}^\top\mathbf{Z}\widehat{\boldsymbol{\Omega}}^{\star-1}\mathbf{Z}^\top\mathbf{D}]^{-1}$, where $\widehat{\boldsymbol{\Omega}}^{\star}$ is calculated as in (8) but instead uses $\widehat{\boldsymbol{\psi}}^{*}_{\text{GMM}}$.

4 The netivreg Command

This section describes the full syntax of the new `netivreg` command. Stata 16.0 is the earliest version that can run `netivreg` and a working Python 3.0 or higher installation with the required packages listed in 1.2 are also needed. The `netivreg` does not use a Stata matrix or `spmat` object to store the adjacency matrices $\mathbf{W}$ and $\mathbf{W}_0$, but does use the Python package Numpy's sparse matrices architecture inside the NetworkX package to handle them in the numerical implementations. Therefore, once the primary node-specific dataset is loaded into memory, both adjacency matrices must be provided as adjacency *lists* instead of as Stata frames; see Section 5.1.

By default, the `netivreg` command expects these adjacency matrices to describe directed graphs. Therefore, the user must remember to list *both* entries (i, j) and (j, i) when working with undirected graphs.

4.1 Syntax

The syntax of `netivreg` is as follows:

`netivreg` *estimator depvar varlist1* `(W = W0)` [`,` *options*].

`netivreg` estimates a linear-in-means regression of *depvar* on *varlist1* and the social interaction network `W` using the exogenous network `W0` as the instrument of the endogenous network `W`. The social networks `W` and `W0` are defined by two adjacency lists stored as Stata frames.

estimator specifies the estimation procedure. There are two options: `g3sls`, which estimates via Generalized Three-Stage Least Squares, and `gmm`, which estimates a Generalized Method of Moments.

4.2 Model Options

`wx(`*varlist2*`)` indicates the variables from *varlist1* to be included as contextual effects. By default, it includes all the variables from *varlist1*.

`id(`*varname*`)` identifies the variable to match covariates with the network data. The default *varname* is *id*.

4.3 G3SLS Options

`first` reports the first-stage results of the linear projection of $\mathbf{WS}$ on $\mathbf{WS}_0$.

`second` reports the second-stage results of the 2SLS estimation of the linear-in-means model.

`transformed` estimates the linear-in-means model with the transformed variables multiplied by $(\mathbf{I} - \mathbf{W}_0)$.

`cluster(`*varname*`)` produces standard errors and statistics that are robust to both arbitrary heteroskedasticity and intragroup correlation, where *varname* identifies the group. The default is non-clustered standard errors.

4.4 GMM Options

`wz(`*varlist3*`)` is the list of variables used as instruments for *varlist2*.

`maxp(#)` is the max p-exponent of the exogenous matrix $\mathbf{W}_0^p$ to include in the set of instruments. The default is $p = 2$.

`wmatrix(`*wmtype*`)` specifies the type of weighting matrix in the GMM estimation. For the one-step GMM estimation, use `identity` or `instrument`. For a two-step efficient GMM estimation use `optimal` by default.

`kernel(`*type*`)` is the kernel function used to calculate network HAC variance estimator. There are three options: `th` for Tukey-Hanning, `truncated`, or `parzen` (this is the default).

`cons(#)` is the constant used to calculate rule of thumb of bandwidth. The default is $C = 1.8$.

4.5 Stored Results

`netivreg` stores the following in `e()`:

Scalars			
`e(N)`	number of observations	`e(mss)`	model sum of squares
`e(df_m)`	model degrees of freedom	`e(r2)`	R-squared
`e(df_r)`	residual degrees of freedom	`e(r2_a)`	adjusted R-squared
`e(rank)`	rank of `e(V)`	`e(rmse)`	root mean squared error
`e(chi2)`	chi-squared	`e(N_clust)`	number of clusters
`e(rss)`	residual sum of squares		
Macros			
`e(cmd)`	`netivreg`	`e(exogr)`	exogenous regressor
`e(wx)`	contextual effects	`e(depvar)`	name of dependent variable
`e(clustvar)`	name of cluster variable	`e(properties)`	`b V`
Matrices			
`e(b)`	coefficient vector	`e(V)`	variance–covariance matrix of the estimators
`e(first)`	first-stage regression results	`e(second)`	second-stage regression results

5 Examples

In this section, we illustrate the `netivreg` command's estimation capabilities simulated data and three years' worth of data on peer-reviewed articles in Estrada et al. (2020). The command requires two types of data files. The first one contains the outcome variable and covariates in the traditional format; i.e., a unit record per row (nodes data file). The second contains all the pair-wise associations per network among units (edges' data files). Note that at least one edge data file is needed apart from the nodes data file.

5.1 Simulated Data

We use the following version of the linear-in-means model in (1):

$$\begin{aligned} y_i =& 1 + 0.7\sum_{j=1}^{n}\overline{w}_{ij}y_j + 0.33\sum_{j=1}^{n}\overline{w}_{ij}x_{1i} + 0.33\sum_{j=1}^{n}\overline{w}_{ij}x_{2i} + 0.33\sum_{j=1}^{n}\overline{w}_{ij}x_{3i} \\ &+ 0.33x_{1i} + 0.33x_{2i} + 0.33x_{3i} + v_i, \end{aligned} \tag{9}$$

where x_{ki} are drawn from an independent and identically (i.i.d.) normal random variable with a mean of zero and a variance of 3 for $k = 1, 2, 3$, which are independent of each other. The weights $\overline{w}_{ij}$ are row-normalized versions of the adjacency matrix $\mathbf{W} = [w_{ij}]$, i.e., $\overline{w}_{ij} = w_{ij}/\sum_{j=1}^{n} w_{ij}$. The $\mathbf{W}$ adjacency matrix is generated from $\mathbf{W}_0 = [w_{0;ij}]$ which in turn is generated from a Erdös and Rényi's (1959) random graph with a density of 0.01. Two sets of i.i.d. variables, ε^*_{1i} and ε_{2i}, are drawn from standard normal distributions and

$$w_{ij} = \begin{cases} \mathbb{I}[|\varepsilon^*_{1i} - \varepsilon^*_{1j}| < \widehat{F}^{-1}_{\varepsilon^*_1}(0.95)] \times (1 - w_{0;ij}) + w_{0;ij} & ;\text{ if } \varepsilon^*_{1i} > \Phi^{-1}(0.95), \\ \mathbb{I}[|\varepsilon^*_{1i} - \varepsilon^*_{1j}| < \widehat{F}^{-1}_{\varepsilon^*_1}(0.95)] \times w_{0;ij} & ;\text{ if } \varepsilon^*_{1i} < \Phi^{-1}(0.05), \\ w_{0;ij} & ;\text{ otherwise}, \end{cases}$$

where $\widehat{F}^{-1}_{\varepsilon^*_1}(0.95)$ represents the 95% empirical quantile of the ε^*_{1i} sample; $\Phi^{-1}(\cdot)$ represents the inverse of the cumulative distribution function of a standard normal random variable; and $I(\cdot)$ is the indicator function that equals one if its argument is true, and is zero otherwise.

The structural error in (9) is then defined as $v_i = m \times \varepsilon_{1i} + \varepsilon_{2i}$, where

$$\varepsilon_{1i} = \begin{cases} \varepsilon^*_{1i} & ;\text{ if } \varepsilon^*_{1i} < \Phi^{-1}(0.05) \text{ or } \varepsilon^*_{1i} > \Phi^{-1}(0.95), \\ 0 & ;\text{ otherwise}. \end{cases}$$

The design parameter $m \in \{0, 1\}$ acts as a switch to generate either an exogenous $\mathbf{W}$ adjacency matrix ($m = 0$) or an endogenous one ($m = 1$). The sample size is set to $n = 400$.

One first needs to import the nodes data file that contains the nodes' outcome variable and covariates.

```
. use data_sim.dta
. format y_endo y_exo x1 x2 x3 x4 %9.3f
. list in 1/5, table
```

	id	y_exo	y_endo	x1	x2	x3	x4
1.	1	4.072	4.555	-0.523	0.926	2.136	-0.546
2.	2	4.584	4.665	2.611	1.455	-0.926	0.759

3.	3	3.887	3.671	3.125	0.513	-2.718	-2.132
4.	4	3.736	3.962	-2.674	1.504	1.769	0.091
5.	5	6.360	7.002	-0.993	0.345	1.126	1.120

The data_sim.dta identifies each node by the id variable and two outcomes; i.e., y_exo when there is no endogeneity ($m = 0$), and y_endo when there is ($m = 1$). The nodes dataset also includes the covariate x4, which was generated from a standard normal distribution independent of the outcome variables and covariates x1, x2, and x3.

The edges' datasets have the following structure:

```
. use W_sim.dta
. list in 113/117, table
```

	source	target
113.	28	259
114.	28	361
115.	29	67
116.	29	79
117.	29	196

and

```
. use W0_sim.dta
. list in 113/117, table
```

	source	target
113.	30	167
114.	30	325
115.	31	38
116.	31	83
117.	31	132

where each row records the connection between the node listed as id in the data_sim.dta as either source or target. This structure allows for directed or undirected network data. When the netivreg command is invoked, all the unique identifiers under source and target are a subset of those listed as id in the data_sim.dta will be checked. The command generates an error otherwise. Nodes that are not listed in either column of these edges' datasets are assumed to be isolated and their corresponding row/column in the adjacency matrices will be zero.

Exogenous Network

If the adjacency matrix **W** is exogenous, it can then be used as an instrument for itself and Estrada et al.'s (2020) G3SLS collapses to Bramoullé et al.'s (2009) G2SLS and the `netivreg` command produces,

```
. use data_sim.dta
. frame create edges
. frame edges: use data/W_sim.dta
. netivreg g3sls y_exo x1 x2 x3 x4 (edges = edges)
Network IV (G3SLS) Regression                          Number of obs =      400
                                                       Wald chi2(10)  =  2021.41
                                                       Prob > chi2   =   0.0000
                                                       R-squared     =   0.8571
                                                       Root MSE      =     .966

        y_exo   Coefficient  Std. err.      t    P>|t|     [95% conf. interval]

W_y
        y_exo     .7187389   .0459317    15.65   0.000     .6284341    .8090436

W_x
           x1     .3628661   .0614882     5.90   0.000     .2419763    .4837559
           x2     .3197861    .051635     6.19   0.000     .2182683     .421304
           x3     .3968142   .0558112     7.11   0.000     .2870856    .5065427
           x4     .0671387   .0886706     0.76   0.449    -.1071935    .2414709

X
           x1     .3749567   .0303577    12.35   0.000     .3152716    .4346419
           x2     .3971781   .0302632    13.12   0.000     .3376787    .4566774
           x3     .3118756   .0282055    11.06   0.000     .2564218    .3673295
           x4     .0614943   .0490757     1.25   0.211    -.0349917    .1579803
        _cons     .9875521   .1453766     6.79   0.000     .7017322    1.273372
```

As expected, the G2SLS estimates are numerically close to the actual parameters in (9) and the irrelevance of `x4` is picked up by the default heteroskedastic-robust t statistics. However, the G3SLS remains a valid and consistent estimator and it can be computed as follows:

```
. netivreg g3sls y_exo x1 x2 x3 x4 (edges = edges0)
Network IV (G3SLS) Regression                          Number of obs =      400
                                                       Wald chi2(10)  =  1495.20
                                                       Prob > chi2   =   0.0000
                                                       R-squared     =   0.8562
                                                       Root MSE      =    .9712

        y_exo   Coefficient  Std. err.      t    P>|t|     [95% conf. interval]

W_y
        y_exo     .7066647   .0675054    10.47   0.000     .5739446    .8393847

W_x
           x1     .3804765   .0824964     4.61   0.000     .2182832    .5426697
           x2      .328785   .0662719     4.96   0.000     .1984902    .4590798
```

```
          x3 |   .4072332   .0686797     5.93   0.000     .2722045    .5422619
          x4 |   .0585146   .1069111     0.55   0.584    -.1516796    .2687087
-------------+----------------------------------------------------------------
X            |
          x1 |    .371649   .0390688     9.51   0.000     .2948372    .4484607
          x2 |   .3661416   .0339244    10.79   0.000      .299444    .4328392
          x3 |   .3176699   .0329671     9.64   0.000     .2528543    .3824855
          x4 |   .0702635   .0545833     1.29   0.199    -.0370508    .1775777
       _cons |   1.014104   .2062919     4.92   0.000     .6085203    1.419687
------------------------------------------------------------------------------
```

The GMM estimator also presents similar results:

```
. frame create edges0
. frame edges0: use data/W0_sim.dta
. netivreg gmm y_exo x1 x2 x3 x4 (edges = edges0)
Network IV (GMM) Regression                         Number of obs =      400
                                                    Wald chi2(10)  =  2595.42
                                                    Prob > chi2   =   0.0000
                                                    R-squared     =   0.8528
                                                    Root MSE      =    .9813
------------------------------------------------------------------------------
       y_exo | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
W_y          |
       y_exo |   .6724429   .0641775    10.48   0.000     .5462656    .7986201
-------------+----------------------------------------------------------------
W_x          |
          x1 |   .4260427   .0713216     5.97   0.000     .2858197    .5662657
          x2 |   .3527208   .0503766     7.00   0.000     .2536772    .4517644
          x3 |   .4349574    .057245     7.60   0.000     .3224099    .5475048
          x4 |   .0389376   .0578986     0.67   0.502    -.0748949      .15277
-------------+----------------------------------------------------------------
          x1 |   .3812431   .0281434    13.55   0.000     .3259113    .4365748
          x2 |   .4107493   .0194691    21.10   0.000     .3724718    .4490268
          x3 |   .3255457   .0189618    17.17   0.000     .2882657    .3628258
          x4 |   .0422976   .0460273     0.92   0.359    -.0481951    .1327904
       _cons |   1.090851    .203117     5.37   0.000     .6915098    1.490192
------------------------------------------------------------------------------
```

One observes that the estimates are again numerically close to the true values of the parameters and the regressor `x4` is insignificant in both G3SLS and GMM.

Endogenous Network

In the endogenous network formation case, the G2SLS becomes inconsistent; however the G3SLS and GMM both remain consistent. The basic implementation using `netivreg` for G3SLS is

```
. netivreg g3sls y_endo x1 x2 x3 x4 (edges = edges0)
Network IV (G3SLS) Regression                       Number of obs =      400
                                                    Wald chi2(10)  =   822.26
                                                    Prob > chi2   =   0.0000
```

```
                                                       R-squared     =   0.8176
                                                       Root MSE      =    1.194

------------------------------------------------------------------------------
      y_endo | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
W_y          |
      y_endo |   .7059194   .0934719     7.55   0.000     .5221476    .8896911
-------------+----------------------------------------------------------------
W_x          |
          x1 |   .3464024   .1277675     2.71   0.007     .0952031    .5976017
          x2 |   .3280795   .0870187     3.77   0.000     .1569951    .4991639
          x3 |   .3615469   .0926147     3.90   0.000     .1794604    .5436334
          x4 |   .0500988   .1476019     0.34   0.734    -.2400962    .3402939
-------------+----------------------------------------------------------------
X            |
          x1 |   .3782985   .0560235     6.75   0.000     .2681526    .4884443
          x2 |   .3287283   .0426851     7.70   0.000     .2448066    .4126499
          x3 |   .3442047   .0483468     7.12   0.000     .2491518    .4392576
          x4 |   .0895948   .0745045     1.20   0.230    -.0568859    .2360756
       _cons |   1.035534   .3189017     3.25   0.001     .4085523    1.662515
------------------------------------------------------------------------------
```

For GMM, we observe

```
. netivreg gmm y_endo x1 x2 x3 x4 (edges = edges0)
Network IV (GMM) Regression                            Number of obs =      400
                                                       Wald chi2(10)  =  5414.09
                                                       Prob > chi2   =   0.0000
                                                       R-squared     =   0.8166
                                                       Root MSE      =    1.194

------------------------------------------------------------------------------
      y_endo | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
W_y          |
      y_endo |   .7039281    .076149     9.24   0.000     .5542141     .853642
-------------+----------------------------------------------------------------
W_x          |
          x1 |   .3493545   .0781754     4.47   0.000     .1956565    .5030524
          x2 |   .3245619   .0617699     5.25   0.000     .2031183    .4460055
          x3 |   .3614537   .0623149     5.80   0.000     .2389385    .4839689
          x4 |   .0621016   .0714393     0.87   0.385    -.0783528    .2025561
-------------+----------------------------------------------------------------
          x1 |   .3877009   .0324968    11.93   0.000       .32381    .4515918
          x2 |   .3859661   .0204085    18.91   0.000     .3458416    .4260905
          x3 |   .3290334   .0246893    13.33   0.000     .2804925    .3775742
          x4 |   .0275173   .0521742     0.53   0.598    -.0750606    .1300951
       _cons |   1.025602   .2586684     3.96   0.000     .5170425    1.534161
------------------------------------------------------------------------------
```

In the case of the G3SLS, the option `first` prints the point estimates $\widehat{\mathbf{\Pi}}$, as well as the heteroskedastic-robust standard errors in Step 1 of Section 3.1.

```
. netivreg g3sls y_endo x1 x2 x3 x4 (edges = edges0), first
Projection of W on W0

------------------------------------------------------------------------------
             | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
```

W_y_endo						
W0_y_endo	.9927161	.0110994	89.44	0.000	.9708939	1.014538
W0_x1	.0038403	.0419735	0.09	0.927	-.0786823	.086363
W0_x2	-.0030277	.0372046	-0.08	0.935	-.0761744	.070119
W0_x3	.0010825	.0378073	0.03	0.977	-.0732492	.0754142
W0_x4	.002885	.0664465	0.04	0.965	-.1277531	.133523
W_x1						
W0_y_endo	-.0924276	.0041002	-22.54	0.000	-.1004889	-.0843663
W0_x1	.8743972	.0155054	56.39	0.000	.8439126	.9048817
W0_x2	.0074888	.0137437	0.54	0.586	-.0195322	.0345098
W0_x3	.0466023	.0139664	3.34	0.001	.0191436	.0740611
W0_x4	.0143088	.0245459	0.58	0.560	-.03395	.0625676
W_x2						
W0_y_endo	-.0142514	.0034378	-4.15	0.000	-.0210103	-.0074925
W0_x1	-.010898	.0130003	-0.84	0.402	-.0364575	.0146614
W0_x2	.9506398	.0115233	82.50	0.000	.9279843	.9732953
W0_x3	.0197406	.0117099	1.69	0.093	-.0032819	.0427631
W0_x4	.0103946	.0205802	0.51	0.614	-.0300675	.0508567
W_x3						
W0_y_endo	-.0311842	.0037837	-8.24	0.000	-.0386233	-.0237451
W0_x1	.0378469	.0143085	2.65	0.008	.0097153	.0659784
W0_x2	.0052678	.0126829	0.42	0.678	-.0196675	.0302031
W0_x3	.9225525	.0128883	71.58	0.000	.8972132	.9478918
W0_x4	-.0186094	.0226513	-0.82	0.412	-.0631432	.0259244
W_x4						
W0_y_endo	.0140097	.0025567	5.48	0.000	.0089831	.0190363
W0_x1	.0419752	.0096683	4.34	0.000	.0229667	.0609838
W0_x2	.021705	.0085698	2.53	0.012	.0048561	.0385539
W0_x3	-.0496623	.0087087	-5.70	0.000	-.0667841	-.0325404
W0_x4	.8994643	.0153055	58.77	0.000	.8693726	.9295559

```
Network IV (G3SLS) Regression                      Number of obs =      400
                                                   Wald chi2(10)  =   822.26
                                                   Prob > chi2    =   0.0000
                                                   R-squared      =   0.8176
                                                   Root MSE       =    1.194
```

y_endo	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
W_y						
y_endo	.7059194	.0934719	7.55	0.000	.5221476	.8896911
W_x						
x1	.3464024	.1277675	2.71	0.007	.0952031	.5976017
x2	.3280795	.0870187	3.77	0.000	.1569951	.4991639
x3	.3615469	.0926147	3.90	0.000	.1794604	.5436334
x4	.0500988	.1476019	0.34	0.734	-.2400962	.3402939
X						
x1	.3782985	.0560235	6.75	0.000	.2681526	.4884443
x2	.3287283	.0426851	7.70	0.000	.2448066	.4126499
x3	.3442047	.0483468	7.12	0.000	.2491518	.4392576
x4	.0895948	.0745045	1.20	0.230	-.0568859	.2360756
_cons	1.035534	.3189017	3.25	0.001	.4085523	1.662515

The option `second` prints the point estimates, $\widehat{\psi}_{2SLS}$, and the heteroskedastic-robust standard errors in Step 2 of Section 3.1.

```
. netivreg g3sls y_endo x1 x2 x3 x4 (edges = edges0), second

2SLS Regression                                         Number of obs =       400
                                                        Wald chi2(10)  =   839.20
                                                        Prob > chi2    =   0.0000
                                                        R-squared      =   0.8095
                                                        Root MSE       =    1.224

------------------------------------------------------------------------------
      y_endo | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
W_y          |
      y_endo |   .6568384   .1030623     6.37   0.000     .4542113    .8594655
-------------+----------------------------------------------------------------
W_x          |
          x1 |   .3798965   .1172551     3.24   0.001     .1493653    .6104278
          x2 |   .3556006   .0846906     4.20   0.000     .1890934    .5221079
          x3 |   .3883685   .0915501     4.24   0.000      .208375     .568362
          x4 |   .0558494   .1324023     0.42   0.673    -.2044623     .316161
-------------+----------------------------------------------------------------
X            |
          x1 |    .386902   .0569697     6.79   0.000     .2748958    .4989081
          x2 |   .3387842   .0433598     7.81   0.000     .2535361    .4240324
          x3 |     .35358   .0498306     7.10   0.000     .2556099    .4515501
          x4 |   .0933147   .0753217     1.24   0.216    -.0547726    .2414021
       _cons |   1.194332   .3562142     3.35   0.001     .4939918    1.894673
------------------------------------------------------------------------------

Network IV (G3SLS) Regression                           Number of obs =       400
                                                        Wald chi2(10)  =   822.26
                                                        Prob > chi2    =   0.0000
                                                        R-squared      =   0.8176
                                                        Root MSE       =    1.194

------------------------------------------------------------------------------
      y_endo | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
W_y          |
      y_endo |   .7059194   .0934719     7.55   0.000     .5221476    .8896911
-------------+----------------------------------------------------------------
W_x          |
          x1 |   .3464024   .1277675     2.71   0.007     .0952031    .5976017
          x2 |   .3280795   .0870187     3.77   0.000     .1569951    .4991639
          x3 |   .3615469   .0926147     3.90   0.000     .1794604    .5436334
          x4 |   .0500988   .1476019     0.34   0.734    -.2400962    .3402939
-------------+----------------------------------------------------------------
X            |
          x1 |   .3782985   .0560235     6.75   0.000     .2681526    .4884443
          x2 |   .3287283   .0426851     7.70   0.000     .2448066    .4126499
          x3 |   .3442047   .0483468     7.12   0.000     .2491518    .4392576
          x4 |   .0895948   .0745045     1.20   0.230    -.0568859    .2360756
       _cons |   1.035534   .3189017     3.25   0.001     .4085523    1.662515
------------------------------------------------------------------------------
```

The researcher must not ignore potential endogenous network formation issues; otherwise, the resulting estimators of the peer and contextual effects can be severely biased. Our estimation results suggest that if we ignore endogenous network formation and use the G2SLS estimator instead, the peer effects result in a value of 0.97, which represents a bias of 0.27 (or 38%) for the true value of 0.7. The contextual effects will also be affected with bias levels of 0.18, 0.2, and 0.09 (or 54%, 60%, and 27%), respectively. Finally, the direct effects include some bias when the researcher ignores network endogeneity.

```
. netivreg g3sls y_endo x1 x2 x3 x4 (edges = edges)
Network IV (G3SLS) Regression                      Number of obs =      400
                                                   Wald chi2(10)  =  1797.04
                                                   Prob > chi2   =   0.0000
                                                   R-squared     =   0.8379
                                                   Root MSE      =    1.104

------------------------------------------------------------------------------
      y_endo | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
W_y          |
      y_endo |   .9775033   .0571161    17.11   0.000     .8652093    1.089797
-------------+----------------------------------------------------------------
W_x          |
          x1 |   .1499146   .0678751     2.21   0.028     .0164676    .2833615
          x2 |   .1307811    .058853     2.22   0.027     .0150721      .24649
          x3 |    .247199   .0626421     3.95   0.000     .1240406    .3703573
          x4 |    .067066   .0943664     0.71   0.478    -.1184645    .2525964
-------------+----------------------------------------------------------------
X            |
          x1 |   .3444555   .0350054     9.84   0.000     .2756326    .4132784
          x2 |   .3254792   .0347623     9.36   0.000     .2571344    .3938241
          x3 |   .2759635   .0316605     8.72   0.000     .2137169      .33821
          x4 |   .0595883   .0571877     1.04   0.298    -.0528464    .1720231
       _cons |   .1586211   .2008155     0.79   0.430    -.2361952    .5534375
------------------------------------------------------------------------------
```

5.2 Real Data

We use all 729 peer reviewed research articles published in the *American Economics Review* (`AER`), *Econometrica* (`ECA`), the *Journal of Political Economy* (`JPE`), and the *Quarterly Journal of Economics* (`QJE`) from 2000 to 2002 taken from Estrada et al. (2020) as an example of using the linear-in-means model with real data. The article specific information is as follows:

```
. use articles.dta
(Data on articles published in the aer, eca, jpe, & qje between 2000-2002)
. describe
Contains data from articles.dta
  obs:           729                          Data on articles published in the aer,
                                                eca, jpe, & qje between 2000-2002
 vars:            12                          12 Sep 2020 14:09
------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------------------
id              int     %9.0g                 Article unique identifier
```

```
lcitations      float   %9.0g                Log of total citations 8 years post
                                               publication
editor          int     %8.0g                1 if at least one of the article´s authors
                                               was an editor of a T4 journal
diff_gender     int     %8.0g                1 if article´s co-authors are of different
                                               gender
isolated        int     %8.0g                1 if an article does not share a
                                               co-authorship relationship with others
n_pages         byte    %8.0g                Article´s number of pages
n_authors       int     %8.0g                Article´s number of authors
n_references    int     %8.0g                Article´s number of bibliographic references
journal         int     %9.0g     journallab
                                             Journal=aer,eca,jpe,qje
year            int     %9.0g     yearlab    Year=2000,2001,2002
c_alumni        int     %9.0g                Alumni network components unique identifiers
c_coauthor      int     %8.0g                Co-author network components unique
                                               identifiers
-----------------------------------------------------------------------------------------
Sorted by: year  journal  id
```

The AER publishes the largest number of research articles, but on average papers published in the QJE received the most citations eight years post publication. The total number of articles published in these journals increased from 2000 to 2002.

```
. gen citations = exp(lcitations)
. tabulate journal year, summarize(citations)
```

Means, Standard Deviations and Frequencies of citations

Journal=aer,eca,jpe,qje	Year=2000,2001,2002 2000	2001	2002	Total
aer	52.417721 73.653308 79	54.931821 90.712233 88	48.652174 49.893607 92	51.934364 72.800192 259
eca	49.627451 52.045351 51	43.328125 51.688336 64	37.177779 43.565161 90	42.195122 48.397466 205
jpe	34.530612 26.257619 49	32.863637 30.229811 44	46.666667 35.717172 48	38.141844 31.362075 141
qje	59.380952 74.70297 42	72.285714 47.771446 42	102.475 100.33303 40	77.653226 78.338854 124
Total	49.131221 61.651186 221	50.794119 66.741361 238	52.448149 60.112435 270	50.902607 62.735929 729

In this timeframe, papers in these four journals are on average 25 pages long, written by two co-authors, and have roughly 31 bibliographic references. Co-authors of different genders wrote about 13% of the articles and only 4.5% of them were listed as co-author and editors-in-charge of any of these journals. Finally, around 53% of the articles do not share a coauthorship relationship with others (see below).

```
. summarize editor diff_gender isolated n_pages n_authors n_references
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      editor |        729    .0452675     .208033          0          1
 diff_gender |        729    .1303155    .3368814          0          1
    isolated |        729    .5281207    .4995513          0          1
     n_pages |        729    25.15775    11.53631          3         76
   n_authors |        729    1.888889    .7486251          1          5
n_references |        729    31.40329    17.84755          0        177
```

Networks

As explained in Estrada et al. (2020), we can construct two types of connections among these 729 research articles. Since the names of each article's authors are known, a co-authors relationship can be formed among them; i.e.,

```
. frame create edges
. frame edges: use edges.dta
(Co-authorship network among articles published in the aer, eca, jpe, & qje betwe)
. frame edges: list in 1/5, table

     +-----------------+
     | source   target |
     |-----------------|
  1. |      4      472 |
  2. |      5      221 |
  3. |      5      463 |
  4. |      5      478 |
  5. |      5      665 |
     +-----------------+

. frame create edges0
. frame edges0: use edges0.dta
(Alumni network among articles published in the aer, eca, jpe, & qje between 2000)
. frame edges0: list in 1/5, table

     +-----------------+
     | source   target |
     |-----------------|
  1. |      2      482 |
  2. |      2      534 |
  3. |      4      129 |
  4. |      4      136 |
  5. |      4      407 |
     +-----------------+
```

The article with an id number of 4 is connected to the article with an id number of 472 in the `edges.dta` frame because at least one of these articles' authors is the same. We refer to these connections as the co-authors' network among articles. Similarly, since authors' information was web-scrapped or text mined from online profiles, Estrada et al. (2020) also provides alumni connections among articles. For example, the article with an id number of 4 is connected with articles 129, 136, and 407 in the `edges0.dta` frame because at least one of these articles' authors overlapped at least three years of graduate school at the same institution.

Table 2 displays network descriptive statistics for the co-authors and alumni networks. We use the Python package NetworkX to calculate the statistics of the network data.

Table 2: Network Descriptive Statistics

Statistics	Co-authors ($\mathbf{W}$)	Alumni ($\mathbf{W_0}$)
Number of nodes	729	729
Number of edges	674	8,838
Average degree	1.85	24.25
Density	0.00	0.03
Average clustering	0.71	0.55
Isolated nodes	385	41

Note: The degree of a node in a network is the number of connections (edges) it has to other nodes. The density of a network is the portion of the potential connections in a network that are actual connections. The average clustering of a network is the average of the local clustering coefficients of all the nodes, where the local clustering coefficient of a node is the proportion of edges between the nodes within its neighborhood divided by the number of edges that could possibly exist between them.

The typical article has only two connections (edges) in the co-authors' network, but we see about 24 in the alumni network; i.e., there are considerably more connections in the latter network (number of edges). Both networks have very low density and there are about ten times more articles that do not have a co-author connection than those that do not share an alumni connection.

Estimation

The empirical model of interest is

$$\begin{aligned} y_{i,r,t} = \alpha + \beta \sum_{j \neq i} w_{i,j} y_{j,r,t} + \sum_{j \neq i} w_{i,j} \mathbf{x}_{j,r,t}^\top \boldsymbol{\delta} + \mathbf{x}_{i,r,t}^\top \boldsymbol{\gamma} \\ + \lambda_r + \lambda_t + v_{i,r,t}, \end{aligned} \tag{10}$$

where $y_{i,r,t}$ represents the natural logarithm of article i's citations eight years post publication (`lcitations`) in journal r in year t; $\mathbf{x}_{j,r,t}$ includes `diff_gender` and `editor` of article j in journal r in year t; and $\mathbf{x}_{i,r,t}$ include the same characteristics for article i plus its number of pages (`n_pages`), authors (`n_authors`), bibliographic references (`n_references`), and whether or not it shares a co-author relationship with other articles (`isolated`). Fixed effects include journal (λ_r) and year (λ_t). We assume that the co-authors' network ($\mathbf{W}$) is endogenous and that the alumni network ($\mathbf{W_0}$) is predetermined and is therefore assumed to be exogenous. The estimation of model (10) yields

```
. netivreg g3sls lcitations editor diff_gender n_pages n_authors n_references
> isolated journal2-journal4 year2-year3 (edges = edges0), wx(editor diff_gender)
> cluster(c_coauthor)
Network IV (G3SLS) Regression                          Number of obs =      729
Number of clusters (c_coauthor) =    575               Wald chi2(15)  =   253.67
                                                       Prob > chi2    =   0.0000
                                                       R-squared      =   0.1312
                                                       Root MSE       =    1.308
                          (Std. err. adjusted for 575 clusters in c_coauthor)
```

lcitations	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
W_y						
lcitations	.2694309	.1342212	2.01	0.045	.0059155	.5329462
W_x						
editor	4.309042	4.700963	0.92	0.360	-4.920321	13.53841
diff_gender	-2.588718	2.648238	-0.98	0.329	-7.787981	2.610546
X						
editor	.2174085	.1042592	2.09	0.037	.0127173	.4220996
diff_gender	.2156686	.1196222	1.80	0.072	-.0191848	.450522
n_pages	.0288559	.004115	7.01	0.000	.020777	.0369349
n_authors	.0551324	.0557331	0.99	0.323	-.0542879	.1645528
n_references	.0116497	.0023924	4.87	0.000	.0069526	.0163468
isolated	-.2149587	.0830938	-2.59	0.010	-.3780962	-.0518213
(output omitted)						
_cons	1.974159	.2279187	8.66	0.000	1.526688	2.42163

Using GMM, we obtain

```
. netivreg gmm lcitations editor diff_gender n_pages n_authors n_references
> isolated journal2-journal4 year2-year3 (edges = edges0), wx(editor diff_gender)
> wz(editor diff_gender n_pages n_authors n_references isolated) maxp(4)
Network IV (GMM) Regression                            Number of obs =      729
                                                       Wald chi2(15)  =   268.30
                                                       Prob > chi2    =   0.0000
                                                       R-squared      =   0.1637
                                                       Root MSE       =     1.08
```

lcitations	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
W_y						
lcitations	.6824698	.2845921	2.40	0.017	.1237325	1.241207

W_x						
editor	.296526	.5227753	0.57	0.571	-.7298345	1.322887
diff_gender	-2.766341	1.042483	-2.65	0.008	-4.81304	-.719642
X						
editor	-.2539077	.2636978	-0.96	0.336	-.7716236	.2638082
diff_gender	.559131	.1486833	3.76	0.000	.2672223	.8510397
n_pages	.029408	.0047013	6.26	0.000	.0201779	.038638
n_authors	.0119978	.0541688	0.22	0.825	-.0943514	.118347
n_references	.0071013	.0027243	2.61	0.009	.0017528	.0124499
isolated	1.745843	1.026178	1.70	0.089	-.2688442	3.760531

(*output omitted*)

_cons	.4873031	.9724358	0.50	0.616	-1.421872	2.396479

For the GMM estimation, we reject the hypothesis of a null peer effect at the 5% level of significance against a positive peer-effect hypothesis; i.e., a 10% increase in the number of citations of connected articles increases a paper's citations by 6.8%. Using G3SLS, we also reject the null hypothesis at a 5% level of significance with a lower estimated coefficient of peer effects. Holding everything else constant, articles with a larger number of pages and bibliographic references get cited more often, as do articles written by authors of different genders at the 1% level of significance.

6 Conclusion

This article shows how the new `netivreg` command fits a linear-in-means model with network data. Both exogenous and endogenous network formations are supported. The `netivreg` main estimation routine is written entirely in Python using Stata's integration with Python (version 16 or later). The command utilizes Stata's new multiframe capabilities to handle the required network data structure in the form of adjacency lists. These, in turn, are converted to sparse matrices within Python for numerical implementation. The basic capabilities of the `netivreg` command are illustrated with simulated data and an empirical application based on peer-reviewed articles published in four top general interest journals in economics. The empirical application uncovers positive spillover effects in terms of articles' citations eight years post publication.

7 Acknowledgments

We are grateful to the editor and referees for their very useful comments and corrections. We also thank David Drukker for his constant encouragement to make our estimation routines available in Stata and for answering various clarifying questions on Stata's new Python integration (version 16 and above). We thank Venkataraman Balasubramanian for providing invaluable expertise in optimizing the Python code that the `netivreg` command uses and helping with the extensive testing performed using the Bank of Canada's EDITH2 High-performance Cluster. Pablo Estrada and Sánchez-Aragón acknowledge financial support from ESPOL University. The views expressed in this article are those of the authors and no responsibility for them should be attributed to the Bank of Canada. All remaining errors are the responsibility of the authors.

8 References

Anaconda Software Distribution. 2020. Anaconda Documentation. https://docs.anaconda.com/.

Auerbach, E. 2022. Identification and estimation of a partially linear regression model using network data. *Econometrica* 90(1): 347–365.

Bramoullé, Y., H. Djebbari, and B. Fortin. 2009. Identification of Peer Effects through Social Networks. *Journal of Econometrics* 150(1): 41–55.

Cerulli, G. 2017. Identification and Estimation of Treatment Effects in the Presence of (Correlated) Neighborhood Interactions: Model and Stata Implementation via Ntreatreg. *The Stata Journal* 17(4): 803–833.

Chan, T. J., J. Estrada, K. P. Huynh, D. T. Jacho-Chávez, C. T. Lam, and L. Sánchez-Aragón. 2022. On the Estimation of Social Effects with Observational Network Data and Random Assignment. Unpublished Manuscript.

De Paula, Á., S. Richards-Shubik, and E. Tamer. 2018. Identifying preferences in networks with bounded degree. *Econometrica* 86(1): 263–288.

Erdös, P., and A. Rényi. 1959. On Random Graphs. *Publicationes Mathematicae (Debrecen)* 6: 290–297.

Estrada, J., K. P. Huynh, D. T. Jacho-Chávez, and L. Sánchez-Aragón. 2020. On the Identification and Estimation of Endogenous Peer Effects in Multiplex Networks. Unpublished manuscript.

Goldsmith-Pinkham, P., and G. W. Imbens. 2013. Social Networks and the Identification of Peer Effects. *Journal of Business and Economic Statistics* 31(3): 253–264.

Graham, B. S. 2017. An Econometric Model of Network Formation With Degree Heterogeneity. *Econometrica* 85(4): 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. B. Graham and Áureo de Paula, 63–82. Academic Press.

Hagberg, A., P. Swart, and D. S Chult. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Ho, A. T. Y., K. P. Huynh, D. T. Jacho-Chavez, and D. Rojas. 2021. Data Science in Stata 16: Frames, Lasso, and Python Integration. *Journal of Statistical Software* 98(1): 1–9.

Jackson, M. O., B. W. Rogers, and Y. Zenou. 2017. The Economic Consequences of Social-Network Structure. *Journal of Economic Literature* 55(1): 49–95.

Johnsson, I., and H. R. Moon. 2021. Estimation of Peer Effects in Endogenous Social Networks: Control Function Approach. *The Review of Economics and Statistics* 103(2): 328–345.

Kojevnikov, D., V. Marmer, and K. Song. 2021. Limit theorems for network dependent random variables. *Journal of Econometrics* 222(2): 882–908.

Manski, C. F. 1993. Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies* 60(3): 531–542.

McKinney, W., et al.. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*. Vol. 445, 51–56. Austin, TX.

Oliphant, T. E. 2006. *A Guide to NumPy*. Vol. 1. Trelgol Publishing USA.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al.. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12(Oct): 2825–2830.

Qu, X., and L. F. Lee. 2015. Estimating a Spatial Autoregressive Model with an Endogenous Spatial Weight Matrix. *Journal of Econometrics* 184(2): 209–232.

Van Rossum, G., and F. L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* .

About the authors

Pablo Estrada is a Ph.D. candidate in Economics at Emory University, Atlanta, United States of America.

Juan Estrada is a Post-Doctoral Fellow at Emory University, Atlanta, United States of America.

Kim P. Huynh is a Director in the Currency Department at the Bank of Canada, Ottawa, Canada.

David Jacho-Chávez is a Professor of Economics at Emory University, Atlanta, United States of America.

Leonardo Sánchez-Aragón is a Professor of Economics at ESPOL University, Guayaquil, Ecuador.