Weak Identification and Estimation of Social Interaction Models*

Guy Tchuente†

School of Economics, University of Kent

September 2019

Abstract

The identification of the network effect is based on either group size variation, the structure of the network or the relative position in the network. I provide necessary conditions for identification of undirected network models based on the number of distinct eigenvalues of the adjacency matrix. Identification of network effects is possible; although in many empirical situations existing identification strategies may require the use of many instruments or instruments that could be strongly correlated with each other. The use of highly correlated instruments or many instruments may lead to weak identification or many instruments bias. This paper proposes regularized versions of the two-stage least squares (2SLS) estimators as a solution to these problems. The proposed estimators are consistent and asymptotically normal. An empirical application, assessing a local government tax competition model, shows the empirical relevance of using regularization methods.

Keywords: High-dimensional models, Social network, Identification, Spatial autoregressive model, 2SLS, Regularization methods.

JEL classification: C13, C31.

*Comments from Marine Carrasco, Eric Renault, James G. MacKinnon, Silvia Goncalves, Vincent Boucher, Russell Davidson, Yann Bramoullé, Lynda Khalaf, John Peirson and A. Colin Cameron are gratefully acknowledged. The author also thanks seminar participants at the University of Kent, University of Ottawa, Universidad Pontificia Javeriana, RES 2016, EEA-ESEM 2016, AFES 2016, CIREQ Econometrics Conference 2017, the University of Bristol's Econometrics Study Group 2017, SCSE 2017 and NASM 2018 for their comments. Many thanks to Xiaodong Lui for kindly providing his code for bias-corrected 2SLS methods and to Teemu Lyytikäinen for providing data on Finnish municipalities.

†E-mail: g.tchuente@kent.ac.uk. Address: Kennedy Building, Park Wood Road, Canterbury, Kent, CT2 7FS, Tel:+441227827249

1 Introduction

This paper investigates the identification and estimation of social interaction models with network structures and the presence of endogenous, contextual, correlated and group fixed effects (see Manski (1993) or Liu and Lee (2010)for a description of these models). In his seminal paper on network model estimation, Manski (1993) argues that solving the reflection problem in identifying and estimating the endogenous interaction effects is of significant interest in social interaction models. He shows that the separate identification of the network effects, in a linear-in-means model, is impossible. Following Manski (1993), the literature on identification of network effects has proposed three main identification strategies. They are based on either the variation in the size of the group of peers or on the structure through which peers interact. The present paper, after proposing easy-to-check identification conditions based on graph spectral decomposition[1], investigates a robust to weak identification estimation strategy. Weak identification can occur in limit cases for all the existing identification strategies.

The first method for identification was proposed by Lee (2007). He shows that both the endogenous and exogenous interaction effects can be identified if there is sufficient variation in group sizes. However, with large groups, identification can be weak in the sense that the estimator converges in distribution at low rates (Lee (2007)). The low rate of convergence means that we need a larger sample to have enough exogenous variation. Indeed as the group size increase, the marginal effect of an individual on its peer becomes small and more observations are needed for identification.

In a more general framework, Bramoullé, Djebbari, and Fortin (2009) investigate identification and estimation of network effect. They use the structure of the network to identify the network effect. Their identification strategy relies on the use of spatial lags of friends' (or friends of friends') characteristics as instruments. But, if the network is highly transitive (i.e. if a friend of my friend is also likely to be my friend), the identification is also weak. Weak identification can also occur if there are too many isolated individuals, the weak identification correspond to the classical weak instruments as in Staiger and Stock (1997). This paper focuses its attention on highly transitive networks that could lead to near rank deficiency and thus weak identification.

More recently, Liu and Lee (2010) have considered the estimation of a social network where the endogenous effect is given by the aggregate choices of an agent's friends. They show that

[1] The main innovation of the paper, compare to Kwok (2019) who has found the same result independently is my proof strategy.

different positions of agents in a network captured by the Bonacich (1987) centrality measure can be used as additional instrumental variables to improve estimation efficiency. The number of such instruments depends on the number of groups, and can be very large. Liu and Lee (2010) propose two-stage least squares (2SLS) and generalized method of moments (GMM) estimators. The proposed estimators have an asymptotic bias due to the presence of many instruments.

The existing papers in the literature of network model estimation use instrumental variable (IV) methods or quasi-maximum likelihood method to estimate the network effects. The present paper is interested in the use of IV when identification is weak in the sense of having near rank deficiency on the set of instruments. We will show that, in the estimation of peer effects using IV methods, highly transitive network or large group size imply the use of highly correlated instruments (where the set of instrumental variables contains the included and excluded instruments). If the Bonacich (1987) centrality measure are used the number of instruments increase with the number of groups. In both cases, the structure of the interaction generates a weak identification issue. The weak identification problem comes from the near-perfect collinearity of the first-stage regression.

Before characterizing the discussion on weak identification, this paper proposes simple-to-check necessary conditions for identification based on the spectral decomposition of the network adjacency matrix. It shows that the identification of the network effects is possible in many cases. However, given that all exogenous variations come from the same adjacency matrix, weak identification may be a concern. It proposes a regularized 2SLS estimators for network models with spatial autoregressive (SAR) representations. Regularization techniques are used to mitigate the finite sample bias of the 2SLS estimators stemming from the use of many or highly correlated instruments. The regularized 2SLS estimators are based on three ways of computing a regularized inverse of the (possibly infinite dimensional) covariance matrix of the instruments. The regularization methods come from the literature on inverse problems (see Kress (1999) and Carrasco, Florens, and Renault (2007)). The first estimator is based on Tikhonov (ridge) regularization. The Tikhonov (ridge) regularization is known in the machine learning literature for its ability to address near-perfect collinearity problems. The second estimator is based on the iterative Landweber-Fridman method. It has the same regularization properties as the ridge method, with the advantage of being appropriate for larger-scale problems. The third estimator is based on the principal components associated with the largest eigenvalues. The use of the principal components is very popular to estimate models with factors. In the presence of many instruments, the use of few principal components can help reduce the first-stage regression dimension. The regularized estimators presented in the paper depend on tuning parameters, I also proposed a data-driven method for its selection based the estimation of an approximation of the mean square error of the estimator.

The regularized 2SLS estimators are consistent and asymptotically normal and unbiased. The regularized 2SLS estimators achieve the semiparametric efficiency bound. However, the consistency and asymptotic normality conditions require more regularization than in Carrasco (2012). A Monte Carlo experiment, in supplement material, shows that the regularized estimators perform well. In general, the quality of the regularized estimators improves as the density of the network increases.

I demonstrate the empirical relevance of my estimators by estimating a model of tax competition between municipalities in Finland. The size of the tax competition parameter seems larger than what is suggested by Lyytikäinen (2012). However, the regularized estimators are not statistically different from zero. This leaves the conclusion unchanged that tax competition is absent between municipalities in Finland.

The large existing literature on network models focuses on two main issues: identification and the estimation of the network effects. In his seminal work, Manski (1993) shows that linear-in-means specifications suffer from the reflection problem, so endogenous and contextual effects cannot be separately identified. Lee (2007) and Bramoullé, Djebbari, and Fortin (2009) propose identification strategies for a local-average network model based on differences in group sizes and structures. Liu and Lee (2010) show that the Bonacich (1987) centrality measure can also be used as additional instruments to improve identification and estimation efficiency. Lee (2007) and Bramoullé, Djebbari, and Fortin (2009) use the instrumental variables method to estimate the parameter of interest. Liu and Lee (2010) propose a generalized method of moments (GMM) estimation approach, following Kelejian and Prucha (1998, 1999), who propose 2SLS and GMM approaches for estimating SAR models. The inclusion of the measure of centrality implies the use of many moment conditions (see Donald and Newey (2001), Hansen, Hausman, and Newey (2008) and Hasselt (2010) for some recent developments in this area).

In this paper, I assume that there are many instruments at hand (they are generated by the structure imposed on the data), and therefore use a framework that allows for an infinite number of instruments. Thus, this paper contributes to the literature on models for which the number of instruments exceeds the sample size. In a linear model framework without network effects, Carrasco (2012) proposes an estimation procedure that allows for the use of many instruments; the number of instruments may be smaller or larger than the sample size, or even infinite. Moreover, Carrasco and Tchuente (2016) show that these methods can be used to improve identification in weak instrumental variables estimation. Closely related papers also include Kuersteiner (2012), who considers a kernel-weighted GMM estimator; Okui (2011), who uses shrinkage with many instruments; and Bai and Ng (2010) and Kapetanios and Marcellino (2010), who also assume that the endogenous regressors depend on a small number of factors that are exogenous. Using estimated factors as instruments, they assume that the number of variables from which the factors are estimated can be larger than the sample size. Belloni, Chen, Chernozhukov, and Hansen (2012) propose an instrumental variables estimator under the first-stage sparsity assumption. Hansen and Kozbur (2014) propose a ridge-regularized jackknife instrumental variable estimator in the presence of heteroscedasticity, which does not require sparsity, and with good sizes.

Another important focus in the instrumental variables estimation literature is on weak instruments (see, for example, Chao and Swanson (2005) and Newey and Windmeijer (2009)). In this paper, I assume that the concentration parameter grows at the same rate as the sample size. However, I allow for the possibility of weak identification resulting from near-perfect collinearity in the set of instruments. My framework is similar to Caner and Yıldız (2012), with the difference that the near-singular design does not come for the proliferation of instruments, but from the structure of the social or spatial interaction.

The paper is organized as follows. Section 2 presents the network model. Section 3 discusses identification and estimation in network models. It proposes the regularized 2SLS approach to estimating the model. The selection of the regularization parameter is discussed in Section 4. An empirical application on local government tax competition is proposed in Section 5. Section 6 concludes. Supplementary material contains series of Monte Carlo evidence on the performance of the proposed estimators for small samples.

2 The Model

The following social interaction model is considered:

$$Y_r = \lambda W_r Y_r + X_{1r}\beta_1 + W_r X_{2r}\beta_2 + \iota_{m_r}\gamma_r + u_r \tag{1}$$

with $u_r = \rho M_r u_r + \varepsilon_r$ and $r = 1...\bar{r}$, where $\bar{r}$ is the total number of groups and m_r is the number of individuals in group r.

$Y_r = (y_{1r}, ..., y_{m_r r})'$ is an m_r-dimensional vector that represents the outcomes of interest. y_{ir} is the observation of individual i in group r. The total number of individuals in the sample is $n = \sum_{r=1}^{\bar{r}} m_r$.

W_r and M_r are $m_r \times m_r$ sociomatrices of known constants, and may or may not be the same.

λ is a scalar that captures endogenous network effects. I assume that this effect is the same for all individuals and groups. The outcomes of individuals influence those of their successors in the network graph (the successors are usually a friends or peers).

In such a linear model, the parameter λ is usually interpreted as the partial effect of a one-unit change in the explanatory variable on the outcome. The explanatory variable in the present case is a product of the a sociomatrix W_r and friends' outcomes Y_r. If the sociomatrix W_r is row-normalized, the endogenous network effect captured by λ represents the expected change in the outcome of an individual if all his friends' outcomes were changed by one unit. This corresponds to the "local average" endogenous effect in the terminology of Liu, Patacchini, and Zenou (2014). On the other hand, if W_r is not row-normalized, it is impossible to know which intervention is the source of the exogenous change in $W_r Y_r$ (see Goldsmith-Pinkham and Imbens (2013) and Angrist (2014) for a discussion on the causal interpretation of the network effect). The unit variation in $W_r Y_r$ could come from a change in the allocation of friends, from an intervention on friends' outcomes or from both. This should be done in a specific manner to obtain a unit change. Such a situation corresponds to the "local aggregate" endogenous effect in the terminology of Liu, Patacchini, and Zenou (2014).

My model specification allows for the use of the "local average" and "local aggregate" endogenous effects. Micro-foundations developed in Liu, Patacchini, and Zenou (2014) suggest that "local average" should be used in situations where the network effect comes from individuals trying to conform to the social norm and the "local aggregate" for a situation where there is leakage.

X_{1r} and X_{2r} are $m_r \times k_1$ and $m_r \times k_2$ matrices, respectively. They represent individuals' exogenous characteristics. β_1 is the parameter measuring the dependence of individuals' outcomes on their own characteristics. The outcomes of individuals may also depend on the characteristics of their predecessors via the exogenous contextual effect, β_2. ι_{m_r} is an m_r-dimensional vector of ones and γ_r represents the unobserved group-specific effect (it is treated as a vector of unknown parameters that will not be estimated).

Aside from the group fixed effect, ρ captures unobservable correlated effects between individuals and their connections in the network.

ε_r is the m_r-dimensional disturbance vector, ε_{ir} are *i.i.d.* with a mean of 0 and variance of σ^2 for all i and r. I define $X_r = (X_{1r}, W_r X_{2r})$.

For a sample with $\bar{r}$ groups, the data is stacked up by defining $V = (V_1', ..., V_{\bar{r}}')'$ for $V =$

Y, X, ε or u.

I also define $W = D(W_1, W_2, ..., W_{\bar{r}})$ and $M = D(M_1, M_2, ..., M_{\bar{r}})$, $\iota = D(\iota_{m_1}, \iota_{m_2}, ..., \iota_{m_{\bar{r}}})$, where $D(A_1, .., A_K)$ is a block diagonal matrix in which the diagonal blocks are $m_k \times n_k$ matrices, denoted as A_k, for $k = 1, ..., K$.

The full sample model is

$$Y = \lambda WY + X\beta + \iota\gamma + u \tag{2}$$

where $u = \rho Mu + \varepsilon$.

I define $R(\rho) = (I - \rho M)$. The Cochrane-Orcutt-type transformation of the model is obtained by multiplying equation (2) by $R = R(\rho_0)$, where ρ_0 is the true value of the parameter ρ:

$$RY = \lambda RWY + RX\beta + R\iota\gamma + Ru.$$

This leads to the following equation:

$$RY = \lambda RWY + RX\beta + R\iota\gamma + \varepsilon. \tag{3}$$

When the number of groups is large, there exist an incidental parameter problem (see Neyman and Scott (1948) and Lancaster (2000) for a discussion of the consequences of this problem).

To eliminate unobserved group heterogeneity, I define

$$J_r = I_{m_r} - (\iota_{m_r}, M_r\iota_{m_r})[(\iota_{m_r}, M_r\iota_{m_r})'(\iota_{m_r}, M_r\iota_{m_r})]^{-}(\iota_{m_r}, M_r\iota_{m_r})'$$

where A^{-} is the generalized inverse of a square matrix A. In general, J_r represents the projection of an m_r-dimensional vector on the space spanned by ι_{m_r} and $M_r\iota_{m_r}$ if they are linearly independent. Otherwise, $J_r = I_{m_r} - \frac{1}{m_r}\iota_{m_r}\iota'_{m_r}$, which is the deviation from the group mean projector.

The matrix $J = D(J_1, J_2, ..., J_{\bar{r}})$ is then pre-multiplied by equation (3) to create a model without the unobserved group-effect parameters:

$$JRY = \lambda JRWY + JRX\beta + J\varepsilon. \tag{4}$$

This is the structural equation, and we are interested in the estimation of λ, β_1, β_2 and ρ. The discussions on the identification and estimation of λ, β_1 and β_2 in this paper will be carried out under the assumption of a consistent estimation of ρ.

I define $S(\lambda) = I - \lambda W$. I assume that equation (2) is an equilibrium and that $S \equiv S(\lambda_0)$ is invertible at the true parameter value. The equilibrium vector Y is given by the reduced-form equation:

$$Y = S^{-1}(X\beta + \iota\gamma) + S^{-1}R^{-1}\varepsilon. \tag{5}$$

It follows that $WY = WS^{-1}(X\beta + \iota\gamma) + WS^{-1}R^{-1}\varepsilon$ and WY is correlated with ε. Hence, in general, equation (4) cannot be consistently estimated by ordinary least squares (OLS). Moreover, this model may not be considered as a self-contained system where the transformed variable JRY can be expressed as a function of the exogenous variables and disturbances. Hence, a partial-likelihood-type approach based only on equation (4) may not be feasible.

In this paper, I consider the estimation of the parameters of equation (4) using regularized 2SLS.

3 Identification and Estimation of the Network Models

This section presents the identification and estimation of the network model parameters using regularization techniques. It first proposes some new results on the identification of network effect using the number of distinct eigenvalues of the adjacency matrix. Using some specific social interaction models, I discuss the weak identification of network effects. I, then, propose a regularized 2SLS model using three regularized methods (Tikhonov, Landweber-Fridman and principal component). They are presented in a unified framework covering both a finite or infinite number of instruments. The focus is on estimating endogenous and contextual effects under the assumption of a preliminary estimator of the unobservable correlation between individuals and their connections in the network. I also derive the asymptotic properties of the models' estimated parameters.

3.1 Identification and Network Structure

The model presented in Equation(2) proposes an underlying structure assumed to have generated the data of the population from which our sample is drawn. The estimation strategy that I propose later aims at making statements about the parameters of this model. To that end, they shouldn't exist many parametrization compatible with the observed data. Discussing conditions under which a unique parametric characterisation exist is a considerable problem in the estimation of network models (see Bramoullé, Djebbari, and Fortin (2009)) and, in econometrics (see Dufour and Hsiao (2010) for a general discussion on identification).

The discussion on the identification is done under a number of assumptions.

Assumption 1. The elements of ε_{ir} are $i.i.d.$ with a mean of 0 and variance of σ^2, and a moment of order higher than the fourth exists.

Assumption 2. The sequences of matrices $\{W\}$, $\{M\}$, $\{S^{-1}\}$ and $\{R^{-1}\}$ are uniformly bounded (UB), and $Sup\|\lambda W\| < 1$.

Uniformly bounded in row (column) sums of the absolute value of a sequence of square matrices $\{A\}$ will be abbreviated as UBR (UBC), and uniformly bounded in both row and column sums in absolute value as UB. A sequence of square matrices $\{A\}$, where $A = [A_{ij}]$, is said to be UBR (UBC) if the sequence of the row-sum matrix norm of A (column-sum matrix norm of A) is bounded.

I take $\varepsilon(\rho_0, \delta) = JR(Y - Z\delta) = f(\delta_0 - \delta) + JRWS^{-1}R^{-1}\varepsilon(\lambda_0 - \lambda) + J\varepsilon$, with $f = JR[WS^{-1}(X\beta_0 + \iota\gamma_0), X]$, where λ_0, β_0 and γ_0 are true values of the parameters $\delta = (\lambda, \beta')'$ and $Z = (WY, X)$.

Under Assumption 2 (i.e. that $Sup\|\lambda W\| < 1$), f can be approximated by a linear combination of $(WX, W^2X, ...)$, $(W\iota, W^2\iota, ...)$ and X. This is a typical case where the number of potential instruments is infinite.

I define $Q = J[Q_0, MQ_0]$, where $Q_0 = [WX, W^2X, ...W\iota, W^2\iota, ..., X]$ is the infinite dimensional set of instruments.

We can also consider the case where only a finite number of instruments, such as $m_1 < n$, is used.

For this case, I define

$$Q_{m_1} = J[Q_{0m_1}, MQ_{0m_1}]$$

where $Q_{0m_1} = [WX, W^2X, ...W^{m_1}X, W\iota, W^2\iota, ..., W^{m_1}\iota, X]$.

As discussed in Liu and Lee (2010), δ is identified if $Q'_{m_1}f$ has full column rank $k+1$. This rank condition requires that f has full rank $k+1$. Note that this assumes that Q_{m_1} is full column rank (meaning no perfect collinearity between instruments). If instruments are near-perfectly or perfectly collinear, f having full rank $k+1$ does not ensure identification. If W_r does not have equal degrees in all its nodes, is different from M_r and W_r is not row-normalized, the centrality score of each individual in his group helps to identify δ. This is possible even if $\beta_0 = 0$. However, if W_r has constant row sums, then $f = JR[WS^{-1}X\beta_0, X]$ and the identification is impossible for $\beta_0 = 0$. Under Assumptions 1 and 2, δ is identified.

The identification in the general case with an infinite number of instruments is possible if the matrix with an infinite number of rows, $Q'f$, has full column rank. The identification is based on the moment condition $E(Q'\varepsilon(\rho_0, \delta)) = 0$ (i.e. $Q'f(\delta_0 - \delta) = 0$).

For any sample of size n, $rank(Q) \leq n$. If we assume that $rank(QQ') = n$, then the full column rank condition only requires that f has full rank $k+1$, this is the same identification conditions as in the finite dimensional case.

The identification of the model parameters relies on the structure of the network through the adjacency matrix W. The adjacency matrix is an $n \times n$ matrix. Let $\tau_1 \geq \tau_2 \geq ... \geq \tau_n$ be its n eigenvalues. An eigenvalue could have multiplicity one or k depending on the number of corresponding eigenvectors. Let define ϱ_w, to be the number of distinct eigenvalues of the adjacency matrix. The results propose in proposition 1 to 3 apply to symmetric spatial and adjacency matrix W. Undirected networks' adjacency matrices is an example of a network structure represented by a symmetric adjacency matrix.

Proposition 1 *Consider a network model represented by Equation 2 with $\rho = 0$. If $\varrho_w = 2$, then the network effects are not identified.*

Proposition 1 implies that the identification of the network effect can be reduced to a spectral analysis of the adjacency matrix. It provides an easy-to-verify condition for the identification of the network effects under the assumption of network exogeneity. Indeed, if $\varrho_w = 2$, using the Cayley-Hamilton theorem, I can show that there exist μ_0 and μ_1 non-null scalars such that $W^2 = \mu_0 I + \mu_1 W$. Then, using proposition 1 from Bramoullé, Djebbari, and Fortin (2009) the network effects are not identified. The result in Proposition 1 was found simultaneously and independently by Kwok (2019) using a different proof strategy. The used of the network adjacency matrix' spectral decomposition in a regression context has also been recently discussed by Jochmans and Weidner (2016).

Proposition 2 *Consider a network model represented by Equation 2 with $\rho_0 = 0$ and $\varepsilon(\delta) = J(Y - Z\delta) = f(\delta_0 - \delta) + JWS^{-1}\varepsilon(\lambda_0 - \lambda) + J\varepsilon$, with $f = J[WS^{-1}(X\beta_0 + \iota\gamma_0), X]$, where λ_0 and $\beta_0 \neq 0$ are true values of the parameters and $\delta = (\lambda, \beta')'$ and Assumptions 1 and 2 hold. Let ϱ_w be the number of distinct eigenvalues of the adjacency matrix W. If $[WX, W^2X, ..., W^{\varrho_w - 1}X, X]$ is full rank column, the network effects are identified.*

Proposition 2 gives a relationship between the identification of network effect and the spectral decomposition of the adjacency matrix. If $\varrho_w = 2$ using the definition of X and applying the Cayley-Hamilton theorem leads to the conclusion that $[WX, X]$ is not full rank column. Thus, JWX cannot be excluded from the structural equation and, therefore, can not serve as an instrumental variable for JWY. However, if the number of distinct eigenvalues is strictly greater than 2, identification may be possible. For instance if $\varrho_w = 3$, $\rho_0 = 0$ and $[WX, W^2X, X]$ is full rank column then the network effect are identified. Indeed, JWX and JW^2X serve as excluded instruments for JWY.

The full rank condition can be generalized to a necessary and sufficient condition, under restrictive assumptions on the set in which the true model's parameters belong. This possibility is discussed in the proof of Proposition 2 in the appendix.

I now consider the case where there is spatial serial correlation. The following proposition generalizes Proposition 1 and 2.

Proposition 3 *Consider a network model represented by Equation 2, $\beta_0 \neq 0$ and Assumptions 1 and 2 hold. Let ϱ_w be the number of distinct eigenvalues of the adjacency matrix W. If $Q_{\varrho_w} = [Q_{0\varrho_w}, MQ_{0\varrho_w}]$ where $Q_{0\varrho_w} = [WX, W^2X, ...W^{\varrho_w-1}X, W\iota, W^2\iota, ..., W^{\varrho_w-1}\iota, X]$ is full rank column, the network effects are identified.*

A special case of a model with spatial serial correlation is one in which $W = M$. In such a situation, proposition 3 becomes similar to Proposition 2. Otherwise, the identification of the network effects could be achieved via the effect of unobserved shock on peers of peers via M. Having spatial correlation provides a second source of exogenous variation.

The identification of the network effects seems to rest upon the possibility of having a full rank column matrix $Q_{\varrho_w} = J[WX, W^2X, ..., W^{\varrho_w-1}X, X]$. The rank property of Q_{ϱ_w} can be measured by condition number of the matrix $Q_{\varrho_w}Q'_{\varrho_w}$. The condition number is the ratio between the largest and the smallest eigenvalue of a symmetric matrix (see Öztürk and Akdeniz (2000) for the relation between ill-conditioned and multicollinearity). Large values of the condition number correspond to situations of near-rank deficiency and near-non-identification of the model's network effects. I consider a model with near rank deficient Q_{ϱ_w} matrix as being weakly identified following the terminology of Dufour and Hsiao (2010). The following subsection provides a discussion of the empirical contexts in which existing network effects identification strategies may become weak.

3.2 Weak Identification in Network Models

Since Manski (1993), the identification problem in network models has been a major concern for econometricians. After finding that separately identifying endogenous and exogenous interaction effects in a linear-in-mean model is not possible, many subsequent studies have investigated network structures in which identification is possible. The identification of the network effect is achieved through group size variation or by exploiting the structure of the network. It is notable that in all cases, additional information is required to overcome the reflection problem.

Lee (2007) uses variations in group sizes to identify both the endogenous and exogenous interaction effects. His identification relies on having sufficient variation in group size. For example, if we assume that the we have two groups form m_1 and m_2 individual and we consider the adjacency matrix formed as follows $W_{ii} = 0$ and $W_{ij} = \dfrac{1}{m_k - 1}$ as long as i and j belong in to the same group k. W can be represented as a block diagonal matrix. Its distinct eigenvalues are $\tau_3 = -\dfrac{1}{m_1 - 1}$, $\tau_2 = -\dfrac{1}{m_2 - 1}$ and $\tau_1 = 1$. If the group sizes are equal, we have exactly two distinct eigenvalues. And, the network effect cannot be identified. Different group sizes lead to more than two distinct eigenvalues. The spectral decomposition of the adjacency matrix leads to the same conclusion as in the comments from Bramoullé, Djebbari, and Fortin (2009) on Lee's identification with two groups of different same sizes. I can show that with group large group size there is almost near-perfect collinearity between $WX, W^2X, ..., W^{\varrho_w - 1}X$ and, X. Or in other words, with large groups, the identification can be weak.

More precisely, let us consider the model presented in Section 2. To focus the discussion on the possibility of weak identification, I will consider the version of the social interaction model without spatial serial correlation.

For an individual in group r, the model above gives

$$y_{ir} = \lambda \left(\frac{1}{m_r - 1} \sum_{j \neq i}^{m_r} y_{jr} \right) + x_{1ir}\beta_1 + \left(\frac{1}{m_r - 1} \sum_{j \neq i}^{m_r} x_{2jr} \right) \beta_2 + \gamma_r + \varepsilon_{ir}. \tag{6}$$

The reduced form after a within transformation is given by:

$$y_{ir} - \bar{y}_r = (x_{1ir} - \bar{x}_{1r}) \frac{(m_r - 1)\beta_1}{m_r - 1 + \lambda} - (x_{2ir} - \bar{x}_{2r}) \frac{\beta_2}{m_r - 1 + \lambda} + \frac{m_r - 1}{m_r - 1 + \lambda} (\varepsilon_{ir} - \bar{\varepsilon}_r) \tag{7}$$

where $\bar{y}_r$, $\bar{x}_{1r}$, $\bar{x}_{2r}$, and $\bar{\varepsilon}_r$ are the group average of the variables excluding individual i (see equation 12 in Bramoullé, Djebbari, and Fortin (2009), and equation 2.5 in Lee (2007)). To simplify the discussion, without loose of generality, let us assume that $x_{1ir} = x_{2ir}$. Thus,

$$y_{ir} - \bar{y}_r = (x_{1ir} - \bar{x}_{1r}) \frac{(m_r - 1)\beta_1 - \beta_2}{m_r - 1 + \lambda} + \frac{m_r - 1}{m_r - 1 + \lambda} (\varepsilon_{ir} - \bar{\varepsilon}_r) \tag{8}$$

Each reduced form equation gives value for $\dfrac{(m_r - 1)\beta_1 - \beta_2}{m_r - 1 + \lambda}$. Identification of the parameters in this model comes from the variations in $\dfrac{(m_r - 1)\beta_1 - \beta_2}{m_r - 1 + \lambda}$. Indeed, Bramoullé, Djebbari, and Fortin (2009) show that we need at least three different group sizes to be able to identify β_1, β_2 and, ν. The parameters are obtained after solving a system of linear equations. There is a need for at least three distinct equations for a unique solution.

When the group size becomes large, $\dfrac{(m_r - 1)\beta_1 - \beta_2}{m_r - 1 + \lambda}$ converges to a constant, which means no or very small variation as m becomes large in the coefficient of the reduced form of all groups. An explanation is that with a large group, the marginal contribution of an additional member of the group is relatively small, which means that the amount of exogenous variation useful for identification vanishes as the group's size increases. This situation is a case of weak identification of the network effects.

The adjacency matrix associated with Lee's model is a block diagonal matrix. The distinct eigenvalues are given by $\tau_r = -\dfrac{1}{m_r - 1}$ $r = 1, ..., \bar{r}$ and $\tau_{\bar{r}+1} = 1$. As the group sizes increase, the difference between the eigenvalue τ_r decreases. Indeed, the number of distinct eigenvalues becomes nearly-equal to two: $\dfrac{\tau_r}{\tau_{r_1}} \to 1$ for $r \neq r_1$; and the groups sizes increase. Thus, the model's parameters are weakly identified.

If the groups are large, based on Proposition 1; X, WX and W^2X will be nearly linearly dependent, leading to weak identification.

Bramoullé, Djebbari, and Fortin (2009) use the structure of the network to identify the network effect. Their work proposes a general framework that incorporates Lee's and Manski's setups as special cases. The identification strategy proposed in their work relies on the use of spatial lags of friends' (i.e. friends of friends') characteristics as instruments. The variables WX, W^2X, and W^3X... are used as instruments for WY. The condition for identification is that I, W and W^2 (or, as noted in Proposition 1 and 4 of Bramoullé, Djebbari, and Fortin (2009), I, W, W^2 and W^3 in the presence of correlated effects) are linearly independent. Variation in group size ensures that I, W and W^2 are linearly independent.

Following Bramoullé, Djebbari, and Fortin (2009) we define the transitivity of a network as the ratio between the number of friends and the number of potential friends. If the network is highly transitive (i.e. a friend of my friend is likely to be my friend too; $W \sim W^2$), identification is also weak. The extreme case of a fully-connected graph has exactly two distinct eigenvalues. An application of Proposition 1 implies that the network effects are not identified. In practice, the near violation of the full rank condition of Proposition 2 is a potential source of weak identification. In practice, using WX, W^2X and W^3X... as instruments can lead to near-perfect collinearity, which implies weak identification (Gibbons and Overman (2012)). Because it leads to near-perfect collinearity occurring in the first-stage regression of the endogenous network effect. The use of regularization methods, such as ridge regression, has been shown to solve these problems. It should be noted that robust to weak-instrument inference method such as AR test may not be appropriate for this type of weak identification.

Liu and Lee (2010) also consider the estimation of a social interaction model. As in Bramoullé, Djebbari, and Fortin (2009), they exploit the structure of the network to identify the network effect. In addition to WX, W^2X and W^3X..., the Bonacich centrality across nodes in a network is used as an instrumental variable to identify network effects and improve estimation efficiency. The use of the Bonacich centrality measure usually leads to the use of many instruments. The 2SLS estimates obtained with these instruments are biased because of the large number of instruments used. Liu and Lee (2010) propose a bias-corrected 2SLS method to account for this.

In this paper, I use regularization techniques. These high-dimensional estimation techniques enable the use of all instruments and deliver efficiency with better finite sample properties (see Carrasco (2012) and Carrasco and Tchuente (2015)). In this case, asymptotic efficiency can be obtained by using many (or all potential) instruments. I use both the Bonacich centrality measure and WX, W^2X and W^3X... as instrumental variables and apply a high-dimensional technique to mitigate the problem of near-perfect collinearity resulting from network structure or/and the bias of many instruments.

3.3 Estimation Using Regularization Methods

The parameters of interest can be estimated using instrumental variables. We can use a finite number of instruments or all potential instruments. As the number of instruments increases, estimation becomes asymptotically more efficient. However, a large number of instruments relative to the sample size creates the many instruments problem (see, for example, Bekker (1994), Donald and Newey (2001) and Han and Phillips (2006)). The parameter of interest can also be weakly identified when a fixed number of instruments is used but the structure of the interaction does not provide sufficient exogenous variation.

The 2SLS estimator with a fixed number of instrumental variables will be consistent and asymptotically normal, but may be less efficient than using many instruments. In order to use all potential instruments (Q), I use regularization tools. In addition to addressing the many instruments bias in Carrasco (2012), my objective is to use regularization to address the weak identification problem.

Let $\varepsilon(\rho_0, \delta) = JR(Y - Z\delta)$, with $\delta = (\lambda, \beta')'$ and $Z = (WY, X)$. The estimation is based on moments corresponding to the orthogonality condition of Q and $J\varepsilon$ given by

$$E(Q'\varepsilon(\rho_0, \delta)) = 0 \tag{9}$$

with the set of instrumental variables given by $Q = J[Q_0, MQ_0]$ with $Q_0 = [WX, W^2X, ...W\iota, W^2\iota, ..., X]$. They can be normalized or standardized. My identification results are conditional on ρ_0. I should first have a preliminary estimator $\tilde{\rho}$ of ρ. I take $\tilde{R} = I - \tilde{\rho}M$ to be an estimator of R.

The regularized estimators use in this paper require the definition of some mathematical objects. My notation follows existing notation in the literature on regularization methods. The set of all potential instrumental variables (Q) is a countable infinite set. π is a positive measure on $\mathbb{N}$, and $l^2(\pi)$ is the Hilbert space of square-summable sequences with respect to π in the real space. I define the covariance operator K of the instruments as

$$K : l^2(\pi) \rightarrow l^2(\pi)$$
$$(Kg)_j = \sum_{k \in \mathbb{N}} E(Q_{ji} Q_{ki} g_k \pi_k)$$

where Q_{ji} is the j^{th} column and i^{th} line of Q. Under the assumption that $|Q_{ji}Q_{ki}|$ for all j, k and i are uniformly bounded, K is a compact operator (see Carrasco, Florens, and Renault (2007) for a definition). Indeed, under Assumption 2, the operator K is a Hilbert-Schmidt operator; I assume that it has non-zero eigenvalues. I assume that the element of X are uniformly bounded.

I consider $\nu_j;\ j = 1, 2, ...$ to be the eigenvalues (in decreasing order) of K, and $\phi_j;\ j = 1, 2, ...$ to be the orthogonal eigenvector of K. K can be estimated by K_n, defined as:

$$K_n : l^2(\pi) \rightarrow l^2(\pi)$$
$$(K_n g)_j = \sum_{k \in \mathbb{N}} \frac{1}{n} \sum_{i=1}^{n} Q_{ji} Q_{ki} g_k \pi_k.$$

In the SAR model, the number of potential moment conditions can be infinite as in equation (9). Therefore, the inverse of K_n needs to be regularized because it is nearly singular. By definition (see Kress (1999), p. 269), a regularized inverse of an operator K is

$$R_\alpha : l^2(\pi) \rightarrow l^2(\pi)$$

such that $\lim\limits_{\alpha \to 0} R_\alpha K \varphi = \varphi$, $\forall \varphi \in l^2(\pi)$.

I consider three different types of regularization schemes: Tikhonov (T), Landweber-Fridman (LF) and principal component (PC). They are defined as follows:

- **Tikhonov (T)**

  Tikhonov regularization is also known as ridge regularization:

  $$(K^\alpha)^{-1} r = (K^2 + \alpha I)^{-1} K r$$

  or

  $$(K^\alpha)^{-1} r = \sum_{j=1}^{\infty} \frac{\nu_j}{\nu_j^2 + \alpha} \langle r, \phi_j \rangle \phi_j$$

  where $\alpha > 0$ and I is the identity operator.

- **Landweber-Fridman (LF)**

  Let $0 < c < 1/\|K\|^2$, where $\|K\|$ is the largest eigenvalue of K (which can be estimated by the largest eigenvalue of K_n). Then,

$$(K^\alpha)^{-1}r = \sum_{j=1}^{\infty} \frac{[1-(1-c\nu_j^2)^{\frac{1}{\alpha}}]}{\nu_j} \langle r, \phi_j \rangle \phi_j$$

  where $\dfrac{1}{\alpha}$ is some positive integer.

- **Principal component (PC)**

  This method consists of using the first eigenfunctions:

$$(K^\alpha)^{-1}r = \sum_{j=1}^{1/\alpha} \frac{1}{\nu_j} \langle r, \phi_j \rangle \phi_j$$

  where $\dfrac{1}{\alpha}$ is some positive integer. In general, $\langle ., . \rangle$ represents the scalar product in $l^2(\pi)$ and in $\mathbb{R}^n$ (depending on the context).

  The use of PC in the first stage is equivalent to projecting on the first principal components of the set of instrumental variables.

In the case of a finite number of moments, $P_{m_1} = Q_{m_1}(Q'_{m_1}Q_{m_1})^{-}Q'_{m_1}$ is the projection matrix on the space of instruments. The matrix $Q'_{m_1}Q_{m_1}$ may become nearly singular when m_1 gets large. Moreover, when $m_1 > n$, $Q'_{m_1}Q_{m_1}$ is singular. To address these cases, I consider a regularized version of the inverse of the matrix $Q'_{m_1}Q_{m_1}$.

I use ψ_j to represent the eigenvectors of the $n \times n$ matrix $Q_{m_1}Q'_{m_1}/n$ associated with eigenvalues, ν_j. For any vector e, the regularized version of P_{m_1}, $P^\alpha_{m_1}$ is:

$$P^\alpha_{m_1} e = \frac{1}{n} \sum_{j=1}^{n} q(\alpha, \nu_j^2) \langle e, \psi_j \rangle \psi_j$$

where for T, $q(\alpha, \nu_j^2) = \dfrac{\nu_j^2}{\nu_j^2 + \alpha}$; for LF, $q(\alpha, \nu_j^2) = [1-(1-c\nu_j^2)^{1/\alpha}]$; and for PC, $q(\alpha, \nu_j^2) = I(j \leq 1/\alpha)$.

The network models suggest the use of an infinite number of instruments or more that the number of individual, without a strong reason to discard some, which is the reason we are not using instrument selection methods.

Following Carrasco and Florens (2000), I define the counterpart of P^α for an infinite number of instruments as

$$P^{\alpha} = G(K_n^{\alpha})^{-1}G^*$$

where $G : l^2(\pi) \to \mathbb{R}^n$ with

$$Gg = \left(\langle Q_1, g\rangle', \langle Q_2, g\rangle', ..., \langle Q_n, g\rangle'\right)'$$

and $G^* : \mathbb{R}^n \to l^2(\pi)$ with

$$G^* v = \frac{1}{n}\sum_{i=1}^{n} Q_i v_i$$

such that $K_n = G^*G$ and GG^* is an $n \times n$ matrix with a typical element $\dfrac{\langle Q_i, Q_j\rangle}{n}$. Let ϕ_j, $\nu_1 \geq \nu_2 \geq ... > 0$, $j = 1, 2, ...$ be the orthonormalized eigenvectors and eigenvalues of K_n, and ψ_j be the eigenfunctions of GG^*.

$G\phi_j = \sqrt{\nu_j}\psi_j$ and $G^*\psi_j = \sqrt{\nu_j}\phi_j$. Note that in this case for $e \in \mathbf{R}^n$, $P^{\alpha}e = \sum_{j=1}^{\infty} q(\alpha, \nu_j^2)\langle e, \psi_j\rangle\psi_j$. We can also note that:

$$\begin{aligned} v'P^{\alpha}w &= v'G(K_n^{\alpha})^{-1}G^*w \\ &= \left\langle (K_n^{\alpha})^{-1/2}\sum_{i=1}^{n} Q_i\left(.\right)v_i, (K_n^{\alpha})^{-1/2}\frac{1}{n}\sum_{i=1}^{n} Q_i\left(.\right)w_i \right\rangle. \end{aligned} \tag{10}$$

Our objective is to estimate the parameters of the model.

I consider $S_n(k) = \dfrac{1}{n}\sum_{i=1}^{n}(\check{Y}_i - \check{Z}_i\delta)Q_{ik}$ with $\check{Y} = \tilde{R}Y$ and $\check{Z} = \tilde{R}Z$.

And I denote $(K_n^{\alpha})^{-1}$ as the regularized inverse of K_n and $(K_n^{\alpha})^{-1/2} = ((K_n^{\alpha})^{-1})^{1/2}$.

The regularized 2SLS estimator of δ is defined as:

$$\hat{\delta}_{R2sls} = argmin\left\langle (K_n^{\alpha})^{-1/2}S_n(.), (K_n^{\alpha})^{-1/2}S_n(.)\right\rangle. \tag{11}$$

Solving the minimization problem, we have

$$\hat{\delta}_{R2sls} = (Z'\tilde{R}'P^{\alpha}\tilde{R}Z)^{-1}Z'\tilde{R}'P^{\alpha}\tilde{R}Y. \tag{12}$$

Equation (12) defines the regularized 2SLS. The regularized 2SLS for SAR models is closely related to the regularized 2SLS of Carrasco (2012) and the 2SLS of Liu and Lee (2010). It extends Carrasco (2012) by considering SAR models and differs from Liu and Lee (2010) in that the projection matrix P is replaced by its regularized counterpart P^{α}.

The 2SLS estimators proposed in this paper are for cases with spatial serial correlation and homoscedastic errors. Extending the regularization approach to deal with heteroscedasticity is left for future research. Indeed, in a companion paper, I propose regularized GMM estimators allowing the joint estimation of all parameters of the model, and the variance-covariance estimator of the estimate is obtained using an approach similar to West and Newey (1987).

3.4 Consistency and Asymptotic Distributions of the Regularized 2SLS

The following proposition shows the consistency and asymptotic normality of the regularized 2SLS estimators. The following extra assumptions are needed.

Assumption 3. $H = \lim_{n\to\infty} \frac{1}{n} f'f$ is a finite nonsingular matrix.

Assumption 4. (i) The elements of X are uniformly bounded, X has full rank k, $E(\varepsilon|X) = 0$, and $\lim_{n\to\infty} \frac{1}{n} X'X$ exists and is nonsingular.
(ii) There is a $\omega \geq 1/2$ such that

$$\sum_{j=1}^{\infty} \frac{\left\langle E(Z(.,x_i)f_a(x_i)), \phi_j \right\rangle^2}{\nu_j^{2\omega+1}} < \infty.$$

Assumption 4 (ii) ensures that regularization allows us to obtain a good asymptotic approximation of the best instrument, f.

Proposition 4 *Under Assumptions 1-4 , $\tilde{\rho} - \rho_0 = O_p(1/\sqrt{n})$ and $\alpha \to 0$. Then, the T, LF and PC estimators satisfy:*

1. *Consistency: $\hat{\delta}_{R2sls} \to \delta_0$ in probability as n and $\alpha\sqrt{n}$ go to infinity.*
2. *Asymptotic normality: $\sqrt{n}(\hat{\delta}_{R2sls} - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 H^{-1}\right)$ as n and $\alpha^2\sqrt{n}$ go to infinity.*

The convergence rate of the regularized 2SLS estimators for SAR is different from those obtained without spatial correlation. For consistency in the SAR model, $\alpha\sqrt{n}$ must go to infinity. The Carrasco (2012) regularized 2SLS estimator is consistent with a convergence rate of $n\alpha^{\frac{1}{2}}$. Asymptotic normality is obtained if $\alpha^2\sqrt{n}$ goes to infinity, which is also different from the Carrasco (2012) asymptotic normality condition for 2SLS. The regularization parameter α is allowed to go to zero slower than in Carrasco (2012) for consistency. Compared to Carrasco (2012), more regularization is needed in order to achieve appropriate asymptotic behavior. The reinforcement of these conditions is certainly due to regularization taking into account the spatial representation of the data.
If the regularization parameter is constant, the asymptotic variance will be bigger. However, asymptotically, the use of regularization should not be needed. It is therefore reasonable to have $\alpha \to 0$.

In Liu and Lee (2010), the 2SLS estimator is biased due to the increasing number of instrumental variables. Interestingly, the regularized 2SLS estimator for SAR models is well-centered under the assumption that $\alpha\sqrt{n}$ goes to infinity.

The bias of the 2SLS estimator in Liu and Lee (2010) is of the form

$$\sqrt{n}b_{2sls} = \sigma^2 tr(P^\alpha RWS^{-1}R^{-1})(Z'RP^\alpha RZ)^{-1}e_1.$$

Using Lemma 1 and 2 in the Appendix, I show that the 2SLS bias is of order $\sqrt{n}b_{2sls} = O_p(\frac{1}{\alpha\sqrt{n}})$, which goes to zero as $\alpha\sqrt{n}$ goes to infinity. The ability to choose the regularization parameter means that we are able to control the size of $\alpha\sqrt{n}$. Therefore, selecting the appropriate regularization parameter is crucial.

The regularization methods presented involve the use of eigenvalues and eigenvectors. The eigenvalues obtained can vary greatly because of the difference in the variance of instrumental variables in the model. For example, W^2X and $W\iota$ could have different variances. To account for this difference, I use normalized instruments in the Monte Carlo simulation. We can also standardize the instruments, which means that regularization methods will be able to account for the difference in location and scale of the instruments. In addition, the regularized estimator presented in this section depend on the regularization parameter, α. The choice of this parameter is very important for the estimators' behavior in small samples. In Section 4, I discuss the selection of the regularization parameter.

4 Selection of the Regularization Parameter

This section discusses the selection of the regularized parameter for network models. I first derive an approximation of the mean-squared error (MSE) using Nagar-type expansion. I estimate the dominant term of the MSE, and select the regularization parameter that minimizes this term.

4.1 Approximation of the MSE

The following proposition provides an approximation of the MSE:

Proposition 5 *If Assumptions 1 to 4 hold, $\tilde{\rho} - \rho_0 = O_p(1/\sqrt{n})$ and $n\alpha \to \infty$ for LF-, PC- and T-regularized 2SLS estimators, then*

$$n(\hat{\delta}_{R2sls} - \delta_0)(\hat{\delta}_{R2sls} - \delta_0)' = Q(\alpha) + \hat{R}(\alpha),$$

$$E(Q(\alpha)|X) = \sigma_\varepsilon^2 H^{-1} + S(\alpha), \tag{13}$$

and

$$r(\alpha)/tr(S(\alpha)) = o_p(1),$$

with $r(\alpha) = E(\hat{R}(\alpha)|X)$ *and*

$$S(\alpha) = \sigma_\varepsilon^2 H^{-1}\left[\frac{f'(1-P^\alpha)^2 f}{n} + \sigma_\varepsilon^2 \frac{1}{n}\left(\sum_j q_j\right)^2 e_1\iota' D' D\iota e_1'\right] H^{-1}.$$

For LF and PC, $S(\alpha) = O_p\left(\frac{1}{n\alpha^2} + \alpha^\omega\right)$ *and for T,* $S(\alpha) = O_p\left(\frac{1}{n\alpha^2} + \alpha^{\min(\omega,2)}\right)$*, with* $D = JRWS^{-1}R^{-1}$ *and* e_1 *is the first unit (column) vector.*

For the selection of α, the relevant dominant term $S(\alpha)$ will be minimized to achieve the smallest MSE. $S(\alpha)$ accounts for a trade-off between the bias and variance. When α goes to zero, the bias term increases while the variance term decreases. The approximation of the regularized 2SLS estimator is similar to Carrasco-regularized 2SLS. However, the expression of the MSE is more complicated because of the spatial correlation.

4.2 Estimation of the MSE

The aim of this subsection is to find the regularized parameter that minimizes the conditional MSE of $\bar{\gamma}'\hat{\delta}_{2sls}$ for some arbitrary $k+1\times 1$ vector, $\bar{\gamma}$. This conditional MSE is:

$$\begin{aligned} MSE &= E[\bar{\gamma}'(\hat{\delta}_{2sls}-\delta_0)(\hat{\delta}_{2sls}-\delta_0)'\bar{\gamma}|X] \\ &\sim \bar{\gamma}' S(\alpha)\bar{\gamma} \\ &\equiv S_{\bar{\gamma}}(\alpha). \end{aligned}$$

$S_{\bar{\gamma}}(\alpha)$ involves the function f, which is unknown. We therefore need to replace $S_{\bar{\gamma}}$ with an estimate.

Stacking the observations, the reduced form equation can be rewritten as

$$RZ = f + v.$$

This expression involves $n\times(k+1)$ matrices. We can reduce the dimension by post-multiplying by $H^{-1}\bar{\gamma}$:

$$RZH^{-1}\bar{\gamma} = fH^{-1}\bar{\gamma} + vH^{-1}\bar{\gamma} \Leftrightarrow RZ_{\bar{\gamma}} = f_{\bar{\gamma}} + v_{\bar{\gamma}} \tag{14}$$

where $v_{\bar{\gamma}i} = v_i' H^{-1}\bar{\gamma}$ is a scalar.

I use $\tilde{\delta}$ to denote a preliminary estimator of δ, obtained from a finite number of instruments. I use $\tilde{\rho}$ to denote a preliminary estimator of ρ, obtained by the method of moments as follows:

$$\tilde{\rho} = armin \tilde{g}(\rho)'\tilde{g}(\rho)$$

where $\tilde{g}(\rho) = [M_1\tilde{\varepsilon}(\rho), M_2\tilde{\varepsilon}(\rho), M_3\tilde{\varepsilon}(\rho)]'\tilde{\varepsilon}(\rho)$,

$$M_1 = JWJ - tr(JWJ)I/tr(J),$$

$$M_2 = JMJ - tr(JMJ)I/tr(J),$$

$$M_3 = JMWJ - tr(JMWJ)I/tr(J),$$

and

$$\tilde{\varepsilon}(\rho) = JR(\rho)(Y - Z'\tilde{\delta}).$$

$\tilde{\delta} = [Z'Q_1(Q_1'Q_1)^{-1}Q_1'Z]^{-1}Z'Q_1(Q_1'Q_1)^{-1}Q_1'Y$, where Q_1 is a single instrument. The residual is $\hat{\varepsilon}(\rho) = JR(\tilde{\rho})(Y - Z'\tilde{\delta})$.

Let us denote $\hat{\sigma}_\varepsilon^2 = \hat{\varepsilon}(\rho)'\hat{\varepsilon}(\rho)/n$, $\hat{v}_{\bar{\gamma}} = (I - P^{\tilde{\alpha}})R(\tilde{\rho})Z\tilde{H}^{-1}\bar{\gamma}$, where $\tilde{H}$ is a consistent estimate of H and $\tilde{\alpha}$ is a preliminary value for α, $\tilde{v}_{\bar{\gamma}} = (I - P^{\tilde{\alpha}})R(\tilde{\rho})Z\tilde{H}^{-1}\bar{\gamma}$ and $\hat{\sigma}^2_{v_{\bar{\gamma}}} = \tilde{v}_{\bar{\gamma}}'\tilde{v}_{\bar{\gamma}}/n$.

I consider the following goodness-of-fit criteria:

Mallows C_p (Mallows (1973))

$$\hat{\varpi}^m(\alpha) = \frac{\hat{v}_{\bar{\gamma}}'\hat{v}_{\bar{\gamma}}}{n} + 2\hat{\sigma}^2_{v_{\bar{\gamma}}}\frac{tr(P^\alpha)}{n}.$$

Generalized cross-validation (Craven and Wahba (1979))

$$\hat{\varpi}^{cv}(\alpha) = \frac{1}{n}\frac{\hat{v}_{\bar{\gamma}}'\hat{v}_{\bar{\gamma}}}{\left(1 - \frac{tr(P^\alpha)}{n}\right)^2}.$$

Leave-one-out cross-validation (Stone (1974))

$$\hat{\varpi}^{lcv}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{RZ}_{\bar{\gamma}i} - \hat{f}^\alpha_{\bar{\gamma}-i})^2,$$

where $\tilde{RZ}_{\bar{\gamma}} = W\tilde{H}^{-1}\bar{\gamma}$, $\tilde{RZ}_{\bar{\gamma}_i}$ is the i^{th} element of $\tilde{RZ}_{\bar{\gamma}}$ and $\hat{f}^\alpha_{\bar{\gamma}-i} = P^\alpha_{-i}\tilde{RZ}_{\bar{\gamma}-i}$. The $n \times (n-1)$ matrix P^α_{-i} is such that the $P^\alpha_{-i} = G(K^\alpha_{n-i})G^*_{-i}$ are obtained by suppressing the i^{th} observation from the sample. $\tilde{RZ}_{\bar{\gamma}-i}$ is the $(n-1)\times 1$ vector constructed by suppressing the i^{th} observation of $\tilde{RZ}_{\bar{\gamma}}$.

Using (13), $S_{\bar{\gamma}}(\alpha)$ can be rewritten as

$$S_{\bar{\gamma}}(\alpha) = \sigma_\varepsilon^2\left[\frac{f_{\bar{\gamma}}'(I - P^\alpha)^2 f_{\bar{\gamma}}}{n} + \sigma_\varepsilon^2\frac{1}{n}\left(\sum_j q_j\right)^2 e_{1\bar{\gamma}}\iota' D' D\iota e_{1\bar{\gamma}}'\right].$$

Using Li (1986)'s results on C_p or cross-validation procedures, note that $\hat{\varpi}(\alpha)$ approximates to

$$\varpi(\alpha) = \frac{f'_{\bar{\gamma}} (I - P^{\alpha})^2 f_{\bar{\gamma}}}{n} + \sigma^2_{v\bar{\gamma}} \frac{tr\left((P^{\alpha})^2\right)}{n}.$$

Therefore, $S_{\gamma}(\alpha)$ is estimated using the following equation:

$$\hat{S}_{\bar{\gamma}}(\alpha) = \hat{\sigma}^2_{\varepsilon} \left[\hat{\varpi}(\alpha) - \hat{\sigma}^2_{v\bar{\gamma}} \frac{tr\left((P^{\alpha})^2\right)}{n} + \hat{\sigma}^2_{\varepsilon} \frac{1}{n} \left(tr(P^{\alpha})\right)^2 e_{1\bar{\gamma}} \iota' \tilde{D}' \tilde{D} \iota e'_{1\bar{\gamma}} \right]$$

where $\tilde{D}$ is a consistent estimator of D. The optimal regularization parameter is obtained by minimising $\hat{S}_{\bar{\gamma}}(\alpha)$ with respect to α. My selection procedure is very similar to Carrasco (2012), and its optimality can be established using the results of Li (1986) and Li (1987).

The regularized 2SLS process and the selection of the regularization parameters are based on a preliminary estimator of ρ. This means that if ρ is not correctly estimated, the estimation of δ could be biased in an unpredictable direction. Also, the use of a cross-validation-type method to choose the regularization parameter usually influences the quality of inference. This is similar to the inference problem in non-parametric estimation (see Newey, Hsieh, and Robins (1998) and Guerre and Lavergne (2005)). This paper focuses on the point estimation of the parameter; post-regularization inference is left for future research.

5 Empirical Application: Local Tax Competition in Finland

The large theoretical literature on local government tax competition can be divided in two groups: efficient local taxation (Tiebout (1956)) and tax competition models departing from Tiebout's model (Lyytikäinen (2012)). The departure from Tiebout's model leads to three types of fiscal consequences: benefit spillovers, distorting taxes on a mobile tax base, political economy considerations and information asymmetries (Lyytikäinen (2012)). While the causes of local government tax interaction are certainly present in most legislation, the empirical literature has long been divided on how to identify a causal local tax competition (interaction) effect.

The identification problem here is a special case of Manski's reflection problem. In the case of municipalities in the same legislation, the network matrix can be represented by the spatial matrix of neighbors. This neighborhood structure of the municipality can be considered as exogenous with respect to tax level. I propose a model to test the hypothesis of tax competition between municipalities:

$$T_{itr} = \lambda W_r T_{itr} + \beta_0 X_{itr} + \beta_1 W_r X_{itr} + \alpha_r + \varepsilon_{itr}$$

The identification and estimation of the tax competition parameter (λ) is achieved, in a large part of the empirical literature, via two strategies. The first strategy uses spatial lags as instruments (friends of friends' characteristics) in an instrumental variables approach, while the second uses maximum likelihood estimation, where identification is achieved via model specification. As pointed out by Gibbons and Overman (2012), the causality of the parameters obtained in these cases is not easy to defend. The validity of the exclusion restriction is not obvious and the correct specification of the model is not fully testable. As an alternative, Gibbons and Overman (2012) propose using differencing coupled with instrumental variables coming from exogenous policy variations.

Lyytikäinen (2012) estimates a tax competition parameter among Finnish local governments. He uses changes in statutory lower limits to property tax rates as a source of exogenous variation to estimate the tax competition parameter (λ) on first difference model. He estimates the following model:

$$T_{i2000} - T_{i1999} = \lambda \sum_{j \neq i} w_{ij}(T_{j2000} - T_{j1999}) + \beta_0 (X_{i2000} - X_{i1999}) + \beta_1 \sum_{j \neq i} w_{ij}(X_{j2000} - X_{j1999}) + v_i.$$

where $w_{ij} = 1/n_i$ with n_i the number of neighbor of the individual i.
The second column in Table 1 replicates the estimate using the instrument from Lyytikäinen (2012). He assumes that $\beta_1 = 0$ and use only one excluded instrument. Other estimations are carried out using spatial lag of the second-, third- and fourth-order and regularized estimators. The instrument used in Table 3 of Lyytikäinen (2012) is one of the instruments used with the spatial lag of other exogenous variables. I have augmented his model to account for an exogenous network model.

The results in Table 1 suggest that the use of many instruments, by adding more spatial lags, affects the results of the 2SLS estimator. Indeed, the condition number of the matrix of instrument increases as a result of the introduction of these new instrumental variables. The use of regularization seems to reduce the bias of the estimation. The simulation results indicate that T-2SLS and L-2SLS are the best methods in terms of bias correction. The point estimates obtained by both estimation methods are very similar, which suggests a bias correction relative to the 2SLS. As the number of instruments increases, the standard errors decrease for the 2SLS as well as for the regularized 2SLS. However, the standard errors are still very large,

Table 1: Estimates of the Tax Competition Parameter for Municipal Property Tax($n = 411$)

Estimators/IVs	Lyytikäinen (2012)	Spatial lags 2	Spatial lags 2 and 3	Spatial lags 2, 3 and 4
2SLS	0.06 (0.07)	0.26(0.28)	-0.02 (0.22)	-0.03(0.17)
T-2SLS	0.01(0.004)	0.19(0.30)	0.18(0.31)	0.18(0.26)
L-2SLS	0.01(0.0005)	0.20(0.22)	0.18(0.33)	0.18(0.31)
PC-2SLS	0.01(0.004)	0.26(0.28)	-0.02(0.22)	-0.03(0.17)
Cond. number ($\frac{\nu_1}{\nu_{min}}$)	2123.97	2001	17800	1.3983e+05

Standard errors are in parenthesis. The change in general property taxation between 1999 and 2000 is the dependent variable. The independent variables are changes in neighboring municipalities' tax rates, the municipality's own imposed increase, non-zero own imposed increase and changes in municipal attributes, such as grants from the central government, disposable income per capita, the unemployment rate and age structure (see Table 3 of Lyytikäinen (2012)) for more details). Tables of estimates of other coefficients are in the supplementary material. The last line of the table shows the condition numbers of QQ' matrices for different instrument sets. The values are relatively large, suggesting a near-perfect collinearity problem in small samples.

which means that the tax competition effect may not be statistically significantly different from zero. Note that inference using the standard errors of regularized estimators does not account for regularization and should be interpreted with caution, given the relatively small sample ($n = 411$ municipalities).[2]

This empirical example shows how regularized estimators can be used to improve the estimation of network models. The size of the tax competition parameter appears to be larger than is suggested by Lyytikäinen (2012). The estimators are not statistically different from zero. However, the regularized estimators (T-2SLS and L-2SLS) appear to be more stable as the number of instruments increases, which suggests that the weak identification problem may have been solved. The intensity of the network effect varies for the regularized PC-2SLS, a potential reason for the this is the absence of factor structure in the set of instruments combine with a near-singularity of the instruments set.

6 Conclusion

This paper uses regularization methods to estimate network models. It proposes easy-to-check identification conditions based on the network adjacency matrix number of distinct eigenvalues. Regularization is proposed as a solution to the weak identification problem in network models. Identification of the network effect can be achieved by using individuals' Bonacich

[2] The regularization parameter, selected in this empirical example, is based on MSE minimization, thus is not optimal for testing. Given that I have not proposed a theory for test and I am in a large sample, I use cautiously the normal asymptotic distribution. In a companion paper on regularized GMM estimator for SAR models, a test theory is developed.

(1987) centrality as instrumental variables. However, the number of instruments increases with the number of groups, leading to the many instruments problem. Identification can also be achieved using the friend-of-a-friend's exogenous characteristics. However, if the network is very dense or group size is very large, the identification is weakened.

The proposed regularized 2SLS estimators based on three regularization methods help address the weak identification and many moments problems. These estimators are consistent and asymptotically normal. The regularized 2SLS estimators achieve the asymptotic efficiency bound. I derive an optimal data-driven selection method for the regularization parameter. An application to the estimation of tax competition in Finnish municipalities shows the empirical relevance of my methods.

A Monte Carlo experiment, in supplementary material, shows that the regularized estimator performs well. The regularized 2SLS procedures substantially reduce the bias from the 2SLS estimators, specifically in a large sample. Moreover, the regularized estimator becomes more precise and less biased with increases in the network density and in the number of groups. These results show that regularization is a valuable solution to the potential weak identification problem existing in the estimation of network models.

7 Acknowledgments

I would like to thank the editors and two anonymous referees for helpful suggestions and comments that have led to substantial improvement of the paper. Comments from Marine Carrasco, Eric Renault, James G. MacKinnon, Silvia Goncalves, Russell Davidson, Yann Bramoullé, Lynda Khalaf, John Peirson and A. Colin Cameron are gratefully acknowledged. The author also thanks seminar participants at the University of Kent, University of Ottawa, Universidad Pontificia Javeriana, RES 2016, EEA-ESEM 2016, AFES 2016, CIREQ Econometrics Conference 2017, the University of Bristol's Econometrics Study Group 2017, SCSE 2017 and NASM 2018 for their comments. Many thanks to Xiaodong Lui for kindly providing his code for bias-corrected 2SLS methods and to Teemu Lyytikäinen for providing data on Finnish municipalities.

A Appendix: Summary of notation

To simplify notation, I use the following:

$P = P^{\alpha}$, $q_j = q(\nu_j^2, \alpha)$

$tr(A)$ is the trace of matrix A

e_j is the j^{th} unit (column) vector $j = 1, ..., n$

$e_f = \frac{1}{n} f'(I-P)f$

$e_{2f} = \frac{1}{n} f'(I-P)^2 f$,

$\Delta_f = tr(e_f)$ and $\Delta_{2f} = tr(e_{2f})$

B Appendix: Lemmas

Lemma 0: (Lemma 4 and Lemma 5 of Carrasco (2012))

(i) $tr(P) = \sum_j q_j = O(1/\alpha)$ and $tr(P^2) = \sum_j q_j^2 = o((\sum_j q_j)^2)$, Lemma 4 (i) of Carrasco (2012),

(ii) $\Delta_{2f} = \begin{cases} O_p(\alpha^{\omega}) \ for \ LF \ and \ SC \\ O_p(\alpha^{min(\omega,2)}) \ for \ T \end{cases}$ and $f'(I-P)\varepsilon/\sqrt{n} = O_p(\sqrt{\Delta_{2f}})$, Lemma 5 (i) and (ii) of Carrasco (2012),

(iii) $u'P\varepsilon = O_p(1/\alpha)$, Lemma 5 (iii) of Carrasco (2012),

(iv) $E[u'P\varepsilon\varepsilon'Pu|X] = (\sum_j q_j)^2 \sigma_{u\varepsilon}\sigma'_{u\varepsilon} + (\sum_j q_j^2)(\sigma_{u\varepsilon}\sigma'_{u\varepsilon} + \sigma_\varepsilon^2 \Sigma_u)$, Lemma 5 (iv) of Carrasco (2012),

(v) $E[f'(I-P)\varepsilon\varepsilon'Pu/n|X] = O_p(\Delta_{2f}/\sqrt{\alpha n})$, Lemma 5 (viii) of Carrasco (2012).

Lemma 1:

(i) $tr(P) = \sum_j q_j = O(1/\alpha)$ and $tr(P^2) = \sum_j q_j^2 = o((\sum_j q_j)^2)$.

(ii) Suppose that $\{A\}$ is a sequence of $n \times n$ UB matrices. For $B = PA$, $tr(B) = o((\sum_j q_j)^2)$, $tr(B^2) = o((\sum_j q_j)^2)$, and $\sum_i B_{ii}^2 = o((\sum_j q_j)^2)$, where B_{ii} are diagonal elements of B.

Proof of Lemma 1:

(i) Proof is in Carrasco (2012) Lemma 4 (i).

(ii) By eigenvalue decomposition, $AA' = \Pi\Delta\Pi'$, where Π is an orthonormal matrix and Δ is the eigenvalue matrix. It follows that $PAA'P \leq \nu_{max}P^2$ with ν_{max} being the largest eigenvalue. It follows that $tr(PAA'P) \leq \nu_{max} tr(P^2) = o_p((\sum_j q_j)^2)$. By the Cauchy-Schwarz inequality, $tr(B) \leq [tr(P^2)]^{1/2}[tr(PAA'P)]^{1/2} = o_p((\sum_j q_j)^2)$. Also by the Cauchy-Schwarz inequality, $tr(B) \leq tr(BB') = tr(PAA'P) = o((\sum_j q_j)^2)$.

Lemma 2: Let C and D be two UB $n \times n$ matrix sequences.

(i) $C'PD = O_p(n/\alpha)$

(ii) $\varepsilon' C' P D \varepsilon = O_p(1/\alpha^2)$ and $C'PD\varepsilon = O_p(\sqrt{n}/\alpha)$

Proof of Lemma 2:

(i) By the Cauchy-Schwarz inequality, $|e_i' C' P^{\alpha} D e_j| \leq \sqrt{e_i' C' C e_i}\sqrt{e_j' D' P^2 D e_j} = O(n/\alpha)$, which implies that $C'PD = O(n/\alpha)$.

(ii) $E|\varepsilon' C' P D \varepsilon| \leq \sqrt{E(\varepsilon' C' P^2 C \varepsilon)}\sqrt{E(\varepsilon' D' P^2 D \varepsilon)} = \sigma^2 \sqrt{tr(C'P^2C)}\sqrt{tr(D'P^2D)} = O(\frac{1}{\alpha^2})$.

By the Markov inequality, $\varepsilon' C' P D \varepsilon = O_p(\frac{1}{\alpha^2})$.

By the Cauchy-Schwarz inequality, $|e_j' C' P D \varepsilon| \leq \sqrt{e_j' C' C e_j}\sqrt{\varepsilon' D' P^2 D \varepsilon} = O_p(\sqrt{n}/\alpha)$, thus $C'PD\varepsilon = O_p(\sqrt{n}/\alpha)$.

Lemma 3: Suppose $\tilde{\rho}$ is a consistent estimator of ρ_0 and $\tilde{R} = R(\tilde{\rho})$.

Then, $\frac{1}{n} Z' \tilde{R}' P \tilde{R} Z = \frac{1}{n} Z' R' P R Z + O_p[(\tilde{\rho} - \rho_0)/\alpha]$ and

$\frac{1}{n} Z' \tilde{R}' P \tilde{R} R^{-1} \varepsilon = \frac{1}{n} Z' R P \varepsilon + O_p[(\tilde{\rho} - \rho_0)/(\alpha\sqrt{n}(1 + \alpha\sqrt{n}))]$.

Proof of Lemma 3:

$\tilde{R} = R - (\tilde{\rho} - \rho_0) M$. Thus,

$$\begin{aligned} Z' \tilde{R}' P \tilde{R} Z/n &= Z' R' P R Z/n \\ &- (\tilde{\rho} - \rho_0) Z' M' P R Z/n - (\tilde{\rho} - \rho_0) Z' R' P M Z/n \\ &+ (\tilde{\rho} - \rho_0)^2 Z' M' P M Z/n \end{aligned}$$

Let us show that $Z'R'PMZ/n = O_p(1/\alpha)$ and $Z'M'PMZ/n = O_p(1/\alpha)$.

Note that $Z = [WS^{-1}(X\beta_0 + \iota\gamma_0), X] + WS^{-1}R^{-1}\varepsilon e_1'$.

Under Assumption 3,

$Z'R'PMZ/n = O(1/\alpha) + O_p(1/\sqrt{n}\alpha) + O_p(1/n\alpha^2) = O_p(1/\alpha)$ and $Z'M'PMZ/n = O_p(1/\alpha)$ by Lemma 2 (i).

$$\begin{aligned} Z' \tilde{R}' P \tilde{R} \varepsilon/n &= Z' R P \varepsilon/n \\ &- (\tilde{\rho} - \rho_0) Z' M' P \varepsilon/n - (\tilde{\rho} - \rho_0) Z' R' P M R^{-1} \varepsilon/n \\ &+ (\tilde{\rho} - \rho_0)^2 Z' M' P M R^{-1} \varepsilon/n \end{aligned}$$

Using the same argument as in the previous case under Assumption 3,

$Z'R'PMR^{-1}\varepsilon/n = O_p(1/\sqrt{n}\alpha + 1/n\alpha^2) = O_p[1/\alpha\sqrt{n}(1 + 1/\alpha\sqrt{n})]$, $Z'M'P\varepsilon/n = O_p[1/\alpha\sqrt{n}(1+$

$1/\alpha\sqrt{n})]$ and $Z'M'P\varepsilon/n = O_p[1/\alpha\sqrt{n}(1+1/\alpha\sqrt{n})]$ by Lemma 2 (ii).

Lemma 4: If Assumptions 1-4 are satisfied and $\alpha \to 0$, then
(i)$Z'RPRZ/n = H + o_p(1)$ if $\alpha\sqrt{n} \to \infty$, and
(ii)$Z'RP\varepsilon/\sqrt{n} = f'\varepsilon/\sqrt{n} + o_p(1)$ if $\alpha^2\sqrt{n} \to \infty$.

Proof of Lemma 4:

Let $v = JRWS^{-1}R^{-1}\varepsilon$ and $JRZ = f + ve_1'$

(i) $\frac{1}{n}Z'RPRZ = \frac{1}{n}f'f - \frac{1}{n}f'(I-P)f + \frac{1}{n}e_1v'Pve_1' + \frac{1}{n}f'Pve_1' + \frac{1}{n}e_1v'Pf$

Let $e_f = \frac{1}{n}f'(I-P)f$, $e_{2f} = \frac{1}{n}f'(I-P)^2f$, $\Delta_f = tr(e_f)$ and $\Delta_{2f} = tr(e_{2f})$. By the Cauchy-Schwarz inequality, $\frac{1}{n}|e_i'f'(I-P)fe_j| \leq \frac{1}{n}\sqrt{e_i'f'fe_i}\sqrt{e_j'f'(I-P)^2fe_j} = O(\sqrt{\Delta_{2f}})$.

From Carrasco (2012) Lemma 5 (i), $\Delta_{2f} = \begin{cases} O_p(\alpha^{\omega})\ for\ LF\ and\ SC \\ O_p(\alpha^{min(\omega,2)})\ for\ T \end{cases}$. Thus, $\Delta_{2f} = o_p(1)$.

By Lemma 2 (ii), $\frac{1}{n}e_1v'Pve_1' + \frac{1}{n}f'Pve_1' + \frac{1}{n}e_1v'Pf = O_p(\frac{1}{n\alpha^2} + \frac{1}{\alpha\sqrt{n}}) = o_p(1)$.

(ii) $Z'RP\varepsilon/\sqrt{n} = f'\varepsilon/\sqrt{n} - f'(I-P)\varepsilon/\sqrt{n} + e_1v'P\varepsilon/\sqrt{n}$

By Lemma 5 (ii) of Carrasco (2012), $f'(I-P)\varepsilon/\sqrt{n} = O_p(\sqrt{\Delta_{2f}})$ and by Lemma 2 (ii), $e_1v'P\varepsilon/\sqrt{n} = O_p(1/\alpha^2\sqrt{n})$.

C Appendix: Proofs of propositions

Proof of Proposition 1:

The Cayley-Hamilton theorem in linear algebra state that each square matrix is solution to it characteristic polynomial. The adjacency matrix of the network in our case is given by W, which is an $n \times n$ matrix. If it has two distinct eigenvalues, therefore, the characteristic polynomial, $p(\tau) = det(\tau I_n - W)$, is a degree two polynomial. Thus, there exist a_0, a_1 and, a_2 with $a_2 \neq 0$ such that $a_0I_n + a_1W + a_2W^2 = 0$. I_n, W and W^2 are linearly dependant and from Proposition 1 of Bramoullé, Djebbari, and Fortin (2009) the network effects are not identified.

Proof of Proposition 2:

Under Assumption 2 (i.e. that $Sup\|\lambda W\| < 1$), f can be approximated by a linear combination of $(JWX, JW^2X, ..., JW^{\varrho_w - 1}X)$ and JX. Indeed, using Caley-Hamilton theorem and the fact that the characteristic polynomial has ϱ_w distinct eigenvalues, For any natural number $q > \varrho_w$, W^q can be written as a linear combination of $I_n, W, ..., W^{\varrho_w - 1}$. Thus, WS^{-1} can be written a linear combination of $I_n, W, ..., W^{\varrho_w - 1}$. Therefore, f can be approximated by a linear combination of $(JWX, JW^2X, ..., JW^{\varrho_w - 1}X)$ and JX.

Let assume that $[WX, W^2X, ..., W^{\varrho_w-1}X, X]$ is full rank column.

Let $Q = J[WX, W^2X, ..., W^{\varrho_w-1}X, X]$ be the set of instrumental variables. The identification of the network effects is based on the moment conditions $E(Q'\varepsilon(\rho_0, \delta)) = 0$ (i.e. $Q'f(\delta_0 - \delta) = 0$). The parameters are point identified if the solution to this equation is unique. A necessary and sufficient condition is that Q and f are full rank column. $[WX, W^2X, ..., W^{\varrho_w-1}X, X]$ is full rank column if and only if Q is full rank column. Moreover, if $[WX, W^2X, ..., W^{\varrho_w-1}X, X]$ is full rank column the f is of rank $1 + k$.

Let assume that $[WX, W^2X, ..., W^{\varrho_w-1}X, X]$ is not full rank column. Consider

$$\mathfrak{B} = \{b \in \mathbb{R}^{k\times\varrho_w}, Xb_0 + WXb_2 + ... + W^{\varrho_w-1}Xb_{\varrho_w-1} = 0\}$$

It can be observed that $f = [JWS^{-1}(X\beta_0), JX]$ is equivalent to $f = J[\sum_{k=1}^{\varrho_w-1} \varsigma_k W^k X\beta_0, X]$. Consider

$$\mathfrak{A} = \{a = (a_0, a_1) \in \mathbb{R}^k \times \mathbb{R}, Xa_0 + a_1 \sum_{k=1}^{\varrho_w-1} \varsigma_k W^k X\beta_0 = 0\}$$

f is not full rank if and only if $\mathfrak{A} \neq \{0\}$.

In other word, f is not full rank column if and only if there exist $b \in \mathfrak{B}$ such that $b_0 = a_0$, $b_k = a_1\varsigma_k\beta_0$ with β_0, ς_k known constant for all $k = 1, ..., \varrho_w - 1$ and $b \neq 0$. The condition for f not being full rank column of very restrictive. However, if we assume that there exist such a sub set in $\mathfrak{A}$, then f is not full rank.

Note that in general, it is possible to have $JWS^{-1}(X\beta_0 + \iota\gamma_0)$ linearly independent from JX without $[WX, W^2X, ..., W^{\varrho_w-1}X, X]$ being full rank column. This happen if β_0, λ and γ_0 are not in the space parameter compatible with the null space of $[WX, W^2X, ..., W^{\varrho_w-1}X, X]$.

The condition $[WX, W^2X, ..., W^{\varrho_w-1}X, X]$ is full rank column is therefore a necessary but not, in general, a sufficient condition for identification. But if we restrict the true value of the parameter to be in the compatible set as in Bramoullé, Djebbari, and Fortin (2009) Result 1 (2) Page 54 the condition is necessary and sufficient.

Proof of Proposition 3:

The proof of proposition 3 is similar to that of proposition 2 with $[WX, W^2X, ..., W^{\varrho_w-1}X, X]$ replaced by $Q_{0\varrho_w} = [WX, W^2X, ...W^{\varrho_w-1}X, W\iota, W^2\iota, ..., W^{\varrho_w-1}\iota, X]$. The identification result in this case are conditional on a consistent preliminary estimation of ρ as in Liu and Lee (2010).

Proof of Proposition 4:

The regularized 2SLS estimator satisfies $\hat{\delta}_{R2sls} - \delta_0 = (Z'\tilde{R}'P\tilde{R}Z)^{-1}Z'\tilde{R}'P\tilde{R}R^{-1}\varepsilon$.

$Z'\tilde{R}'P\tilde{R}Z/n = O_p(1) + O_n(1/\alpha\sqrt{n})$ by Lemmas 3 and 4.

$\tilde{R}'P\tilde{R}R^{-1}\varepsilon/n = O_p(1/\sqrt{n}) + O_p[(1/(n\alpha(1 + \alpha\sqrt{n}))$ by Lemmas 3 and 4.

Then, $\hat{\delta}_{R2sls} - \delta_0 = o_p(1)$ as $\alpha\sqrt{n} \to \infty$ and $\alpha \to 0$. This proves the consistency of the regularized 2SLS for SAR with many instruments:

$$\sqrt{n}(\hat{\delta}_{R2sls} - \delta_0) = (Z'\tilde{R}'P\tilde{R}Z/n)^{-1}[Z'\tilde{R}'P\tilde{R}R^{-1}\varepsilon/\sqrt{n}].$$

Using Lemmas 3 and 4, as well as the Slutzky theorem:

$$\sqrt{n}(\hat{\delta}_{R2sls} - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 H^{-1}\right)$$

if $\alpha^2\sqrt{n} \to \infty$ and $\alpha \to 0$.

Proof of Proposition 5

Let us consider the MSE of the estimated parameters:

$$n(\hat{\delta}_{R2sls} - \delta_0)(\hat{\delta}_{R2sls} - \delta_0) = \hat{H}^{-1}\hat{h}\hat{h}'\hat{H}^{-1}$$

with $\hat{H} = \dfrac{Z'\tilde{R}'P\tilde{R}Z}{n}$ and $\hat{h} = \dfrac{Z'\tilde{R}'P\tilde{R}Y}{\sqrt{n}}$. Our objective is to approximate the MSE. To achieve this, I use a Nagar-type approximation in order to concentrate on the largest part of the MSE. By Lemma 3,

$$\begin{aligned}
\hat{H} &= Z'RPRZ/n \\
&- (\tilde{\rho} - \rho_0)Z'M'PRZ/n - (\tilde{\rho} - \rho_0)Z'R'PMZ/n \\
&+ (\tilde{\rho} - \rho_0)^2 Z'M'PMZ/n.
\end{aligned}$$

And $\hat{H} = Z'RPRZ/n + O_p((\tilde{\rho} - \rho_0)/\alpha)$. By Lemma 4, we have that

$$\hat{H} = \frac{1}{n}f'f - \frac{1}{n}f'(I-P)f + \frac{1}{n}e_1v'Pve_1' + \frac{1}{n}f'Pve_1' + \frac{1}{n}e_1v'Pf + O_p((\tilde{\rho} - \rho_0)/\alpha).$$

Let us define $T^H = T_1^H + T_2^H + T_3^H$, with $T_1^H = -\frac{1}{n}f'(I-P)f$, $T_2^H = \frac{1}{n}e_1v'Pve_1'$ and $T_3^H = \frac{1}{n}f'Pve_1' + \frac{1}{n}e_1v'Pf + O_p((\tilde{\rho} - \rho_0)/\alpha)$, such that

$$\begin{aligned}
\hat{H} &= \frac{1}{n}f'f + T_1^H + T_2^H + T_3^H \\
&= H + T_1^H + T_2^H + T_3^H + o_p(1) \\
&= H + T^H + o_p(1).
\end{aligned}$$

Following similar arguments, we have

$$\hat{h} = f'\varepsilon/\sqrt{n} - f'(I-P)\varepsilon/\sqrt{n} + e_1v'P\varepsilon/\sqrt{n} + O_p[(\tilde{\rho} - \rho_0)/(\alpha(1 + \alpha\sqrt{n}))].$$

Let us also define $T^h = T_1^h + T_2^h$ with
$T_1^h = -f'(I-P)\varepsilon/\sqrt{n}$ and $T_2^h = e_1 v' P\varepsilon/\sqrt{n} + O_p[(\tilde{\rho}-\rho_0)/(\alpha(1+\alpha\sqrt{n}))]$.

We therefore have

$$\begin{aligned}\hat{h} &= f'\varepsilon/\sqrt{n} + T_1^h + T_2^h \\ &= h + T_1^h + T_2^h + o_p(1) \\ &= h + T^h + o_p(1).\end{aligned}$$

Using a Nagar-type expansion on $\hat{H}^{-1}$,

$$n(\hat{\delta}_{R2sls}-\delta_0)(\hat{\delta}_{R2sls}-\delta_0) = H^{-1}[I - T^H H^{-1}][hh' + hT^h + T^h h' + T^h T^{h'}][I - H^{-1}T^H]H^{-1} + o_p(1).$$

Let us define $A(\alpha) = [I - T^H H^{-1}]\Im(\alpha)[I - H^{-1}T^H]$ with $\Im(\alpha) = [hh' + hT^h + T^h h' + T^h T^{h'}]$.

Therefore, $A(\alpha) = \Im(\alpha) + T^H H^{-1}\Im(\alpha)H^{-1}T^H - T^H H^{-1}\Im(\alpha) - \Im(\alpha)H^{-1}T^H$.

$$\begin{aligned}E[\Im(\alpha)|X] &= \sigma^2[H - 2e_f + \frac{1}{n}f'Pve_1' + \frac{1}{n}e_1 v'Pf + e_{2f}] \\ &- E[\frac{1}{n}f'(I-P)\varepsilon\varepsilon' Pve_1' + \frac{1}{n}e_1 v' P\varepsilon\varepsilon'(I-P)f|X] \\ &+ E[\frac{1}{n}e_1 v' P\varepsilon\varepsilon' Pve_1'|X].\end{aligned}$$

$E(T^H H^{-1}\Im(\alpha)|X) = -\sigma^2 e_f + o_p(1)$ and $E(\Im(\alpha)H^{-1}T^H|X) = -\sigma^2 e_f + o_p(1)$.

$$\begin{aligned}E(T^H H^{-1}\Im(\alpha)H^{-1}T^H|X) &= \sigma^2 H O_p([\frac{1}{n\alpha^2} + \frac{1}{\alpha\sqrt{n}} + \Delta_f]^2) \\ &= O_p([\frac{1}{n\alpha^2} + \frac{1}{\alpha\sqrt{n}} + \Delta_f]^2).\end{aligned}$$

We have

$$\begin{aligned}E(A(\alpha)|X) &= \sigma^2 H + \sigma^2 e_{2f} + E[\frac{1}{n}e_1 v' P\varepsilon\varepsilon' Pve_1'|X] \\ &- E[\frac{1}{n}f'(I-P)\varepsilon\varepsilon' Pve_1' + \frac{1}{n}e_1 v' P\varepsilon\varepsilon'(I-P)f|X] \\ &+ \frac{1}{n}f'Pve_1' + \frac{1}{n}e_1 v'Pf + O_p([\frac{1}{n\alpha^2} + \frac{1}{\alpha\sqrt{n}} + \Delta_f]^2).\end{aligned}$$

From Lemma 5 (viii) of Carrasco (2012), we have

$$E[\frac{1}{n}f'(I-P)\varepsilon\varepsilon' Pve_1' + \frac{1}{n}e_1 v' P\varepsilon\varepsilon'(I-P)f|X] = O_p(\sqrt{\Delta_{2f}}/\sqrt{\alpha n})$$

and $\frac{1}{n}e_1 v'(P-P^2)f = O_p(\sqrt{\Delta_{2f}}/\sqrt{\alpha n})$.

From Lemma 5 (iii) of Carrasco (2012), $\frac{1}{n}f'Pve_1' + \frac{1}{n}e_1 v'Pf = O_p(\frac{1}{n\alpha})$.

And, from Lemma 5 (iv) of Carrasco (2012),

$$E[\frac{1}{n}e_1 v' P\varepsilon\varepsilon' P v e_1'/n|X] = \frac{1}{n}(\sum_j q_j)^2 \sigma^4 e_1 \iota' D' D \iota e_1' + o_p((\sum_j q_j)^2/n)$$

with $D = JRWS^{-1}R^{-1}$.

We can conclude that

$$n(\hat{\delta}_{R2sls} - \delta_0)(\hat{\delta}_{R2sls} - \delta_0) = Q(\alpha) + \hat{R}(\alpha)$$

with $E[Q(\alpha)|X] = H^{-1}\sigma^2 + H^{-1}\left[\sigma^2 e_{2f} + \frac{1}{n}(\sum_j q_j)^2 \sigma^4 e_1 \iota' D' D \iota e_1'\right] H^{-1}$ and

$$r(\alpha) = E(\hat{R}(\alpha)|X) = o_p((\sum_j q_j)^2/n) + O_p([\frac{1}{n\alpha^2} + \frac{1}{\alpha\sqrt{n}} + \Delta_f]^2 + \frac{1}{n\alpha} + \frac{\Delta_{2f}}{\sqrt{\alpha n}}).$$

$$S(\alpha) = H^{-1}\left[\sigma^2 e_{2f} + \frac{1}{n}(\sum_j q_j)^2 \sigma^4 e_1 \iota' D' D \iota e_1'\right] H^{-1}.$$

Note that $r(\alpha)/tr(S(\alpha)) = o_p(1)$; my argument is similar to that used in Carrasco (2012). This means that $S(\alpha)$ is the dominant part of the MSE of the estimation of the model using regularized 2SLS.

References

Angrist, J. D. (2014): "The perils of peer effects," *Labour Economics*, 30, 98–108.

Bai, J., and S. Ng (2010): "Instrumental Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26(6), 1577–1606.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657–81.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80(6), 2369–2429.

Bonacich, P. (1987): "Power and centrality: A family of measures," *American Journal of Sociology*, 92(5), 1170–1182.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009): "Identification of peer effects through social networks," *Journal of Econometrics*, 150(1), 41–55.

Caner, M., and N. Yildiz (2012): “CUE with many weak instruments and nearly singular design,” *Journal of Econometrics*, 170(2), 422–441.

Carrasco, M. (2012): “A regularization approach to the many instruments problem,” *Journal of Econometrics*, 170(2), 383–398.

Carrasco, M., and J.-P. Florens (2000): “Generalization Of Gmm To A Continuum Of Moment Conditions,” *Econometric Theory*, 16(06), 797–834.

Carrasco, M., J.-P. Florens, and E. Renault (2007): “Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization,” in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 77. Elsevier.

Carrasco, M., and G. Tchuente (2015): “Regularized LIML for many instruments,” *Journal of Econometrics*, 186(2), 427–442.

——— (2016): “Efficient Estimation with Many Weak Instruments Using Regularization Techniques,” *Econometric Reviews*, 35(8-10), 1609–1637.

Chao, J. C., and N. R. Swanson (2005): “Consistent Estimation with a Large Number of Weak Instruments,” *Econometrica*, 73(5), 1673–1692.

Craven, P., and G. Wahba (1979): “Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of the generalized cross-validation,” *Numer. Math.*, 31, 377–404.

Donald, S. G., and W. K. Newey (2001): “Choosing the Number of Instruments,” *Econometrica*, 69(5), 1161–91.

Dufour, J.-M., and C. Hsiao (2010): “Identification,” in *Microeconometrics*, pp. 65–77. Springer.

Gibbons, S., and H. G. Overman (2012): “Mostly pointless spatial econometrics?,” *Journal of Regional Science*, 52(2), 172–191.

Goldsmith-Pinkham, P., and G. W. Imbens (2013): “Social networks and the identification of peer effects,” *Journal of Business & Economic Statistics*, 31(3), 253–264.

Guerre, E., and P. Lavergne (2005): “Data-driven rate-optimal specification testing in regression models,” *Annals of Statistics*, pp. 840–870.

Han, C., and P. C. B. Phillips (2006): "GMM with Many Moment Conditions," *Econometrica*, 74(1), 147–192.

Hansen, C., J. Hausman, and W. Newey (2008): "Estimation With Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26(4), 398–422.

Hansen, C., and D. Kozbur (2014): "Instrumental variables estimation with many weak instruments using regularized JIVE," *Journal of Econometrics*, 182(2), 290–308.

Hasselt, M. v. (2010): "Many instruments asymptotic approximations under nonnormal error distributions," *Econometric Theory*, 26(02), 633–645.

Jochmans, K., and M. Weidner (2016): "Fixed-effect regressions on network data," *arXiv preprint arXiv:1608.01532*.

Kapetanios, G., and M. Marcellino (2010): "Factor-GMM estimation with large sets of possibly weak instruments," *Computational Statistics and Data Analysis*, 54(11), 2655–2675.

Kress, R. (1999): *Linear Integral Equations*. Springer.

Kuersteiner, G. (2012): "Kernel-weighted GMM estimators for linear time series models," *Journal of Econometrics*, 170(2), 399–421.

Kwok, H. H. (2019): "Identification and estimation of linear social interaction models," *Journal of Econometrics*.

Lancaster, T. (2000): "The incidental parameter problem since 1948," *Journal of Econometrics*, 95(2), 391–413.

Lee, L.-f. (2007): "Identification and estimation of econometric models with group interactions, contextual factors and fixed effects," *Journal of Econometrics*, 140(2), 333–374.

Li, K.-C. (1986): "Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing," *The Annals of Statistics*, 14(3), 1101–1112.

——— (1987): "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete Index Set," *The Annals of Statistics*, 15(3), 958–975.

Liu, X., and L.-f. Lee (2010): "GMM estimation of social interaction models with centrality," *Journal of Econometrics*, 159(1), 99–115.

Liu, X., E. Patacchini, and Y. Zenou (2014): "Endogenous peer effects: local aggregate or local average?," *Journal of Economic Behavior & Organization*, 103, 39–59.

Lyytikäinen, T. (2012): "Tax competition among local governments: Evidence from a property tax reform in Finland," *Journal of Public Economics*, 96(7), 584–595.

Mallows, C. L. (1973): "Some Comments on Cp," *Technometrics*, 15(4), 661–675.

Manski, C. F. (1993): "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 60(3), 531–542.

Newey, W. K., F. Hsieh, and J. Robins (1998): "Undersmoothing and bias corrected functional estimation," .

Newey, W. K., and F. Windmeijer (2009): "Generalized method of moments with many weak moment conditions," *Econometrica*, 77(3), 687–719.

Neyman, J., and L. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica*, 16(1), 1–32.

Okui, R. (2011): "Instrumental variable estimation in the presence of many moment conditions," *Journal of Econometrics*, 165(1), 70–86.

Öztürk, F., and F. Akdeniz (2000): "Ill-conditioning and multicollinearity," *Linear Algebra and Its Applications*, 321(1-3), 295–305.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

Stone, C. J. (1974): "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 111–147.

Tiebout, C. M. (1956): "A pure theory of local expenditures," *Journal of political economy*, 64(5), 416–424.

West, K. D., and W. K. Newey (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.