Estimating the Structure of Social Interactions Using Panel Data *

Elena Manresa
MIT Sloan, Applied Economics

10th March 2016

Abstract

I consider settings where outcomes depend on own characteristics and on the characteristics of other individuals in the population. I propose a method to estimate both the identity of individuals generating spillovers and the strength of their spillover effects using panel data on outcomes and characteristics. This is in contrast to existing approaches, which assume a priori knowledge of the structure of interactions. The method is suitable when the structure of interactions is stable over time and few individuals generate spillovers distinct in magnitude from the rest. To estimate the model, I introduce the Pooled Lasso estimator, a panel-data counterpart of the Lasso estimator and develop an iterative algorithm for computation that alternates between Lasso estimation and pooled panel regression. While spillover effects are estimated at the rate $\frac{\log N}{T}$ average marginal spillover effects may be estimated at a much faster rate of convergence. I apply this methodology to study technological spillovers in productivity in a panel of US firms. I find evidence that spillovers are asymmetric across firms, arising mostly from small, highly productive, and $R\&D$ intensive firms.

JEL codes: C23.

Keywords: Social interactions, spillovers, panel data, high-dimensional models, LASSO, model selection, technological spillovers.

*Special thanks go to Stéphane Bonhomme and Bryan Graham. Also, I would like to thank Victor Aguirregabiria, Dante Amengual, Manuel Arellano, Ivan Canay, Gary Chamberlain, Victor Chernozhukov, Lars P. Hansen, Xavier d'Haultfoeuille, Guido Imbens, Horacio Larreguy, Simon Lee, Monica Martínez, Konrad Menzel, Anna Mikusheva, Aureo de Paula, Franco Peracchi, Luigi Pistaferri, Demian Pouzo, Diego Puga, Andres Rodriguez Clare, Enrique Sentana, Petra Thiemann, Francis Vella, for useful comments and discussion. Also, audiences at numerous places including CEMFI PhD workshop, MOOD 2013 conference, the Bristol Econometric study group 2013, seminar participants at the econometrics seminar at UC Berkeley, and the Econometric Society European Winter Meeting 2013. All remaining errors are my own.

1 Introduction

Externalities or spillovers arise as a result of social interactions in a wide range of economically relevant contexts.[1] Quantifying these spillovers is often important from a policy perspective. As a result, a large empirical literature on estimation of spillover effects has grown over the past decade. In these exercises, the structure of interactions, that is, who interacts with whom, is often assumed to be known to the researcher and taken as given. In this paper I propose a methodology to estimate both the structure of interactions and the spillover effects using panel data. The methodology is useful when the structure of interactions is hard to observe or difficult to measure, or when the definition of the relevant structure of interactions is unclear.

Lack of observation of the structure of interactions has been tackled in the literature with the use of additional data. For example, collection of survey data with self-reported links according to a particular type of social interactions (e.g. friendship), has been increasing over the past years.[2] However, the extent to which collection of survey data mitigates the lack of observability problem is limited. Depending on the economic setting, individuals might not have incentives to reveal their links. For instance, competing firms might not be willing to disclose their sources of technological improvement. Collection of data on the structure of interactions can also be costly, since the number of potential links among individuals grows exponentially with the number of individuals. Studies looking at the effect of social interactions on different outcomes might end up collecting huge amounts of data since structures of interactions can differ depending on the outcome of study.[3]

A different challenge in the empirical study of spillovers arises when the definition of the relevant structure of interactions is unclear. In these cases, estimates on spillover effects can be biased due to misspecification of the structure of interactions. The risk sharing literature provides an example of this situation. In this literature, spillovers are quantified by the degree of insurance against consumption risk that households in developing countries achieve. Co-movement of household consumption with aggregate village consumption suggests that all households in the village interact in a single group. At the same time, the model of full risk sharing is typically rejected in the data (Townsend, 1994). Recently, Ambrus *et al.*, (2013) are able to shed some light on this puzzle by considering a different type of structure of interactions: geographic and family groups of households with few connections with

[1] Some settings where allowing for spillovers has been important are: education (e.g. Graham, 2008), crime practices (Lee *et al.*, 2013), consumption behavior (e.g. De Giorgi *et al.*, 2013) , technology adoption (e.g. Conley *et al.*, 2009), productivity (e.g. Griliches, 1979, Bloom *et al.*, 2013)

[2] An example of this type of data is the Add Health data. This is a longitudinal study of a US representative sample of adolescents in grades 7-12 in 1994-95 school year. The Add Health has information on two different types of social interactions: friendship and romantic relationships. Card and Giuliano (2012) and Lee *et al.*, (2013) are two recent papers that make use of this dataset.

[3] Banerjee *et al.*, (2013) collect information on 13 different types of structures of social interactions after they ask about 13 different types of favor exchange between individuals in 75 rural villages in India.

other groups. Another example of the consequences of misspecification of the structure of interaction is given, in an education context, by Carrel Sacerdote West (2013).

In this paper spillovers arise in a linear panel data regression framework when the characteristics of individuals not only affect their own outcome but also the outcome of other individuals in the sample. As a novelty, I leave unrestricted the identity of the sources of spillovers of individuals, which I allow to differ from individual to individual. Hence, for a given individual, I do not specify which other individuals affect its outcome. I do not restrict either the magnitude of the effect of each source of spillovers. Instead, spillover effects are pair-specific and not necessarily symmetric.

I focus on structures of interactions that are sparse and persistent over time. Specifically, each individual, over time, is influenced by the same small number of other individuals. The methodology can also handle settings where individuals receive a common spillover effect, but for each individual, few other individuals generate a distinct in magnitude spillover effect.

The literature on the economics of social networks provides a motivation for combining sparse structures of interactions with unrestricted heterogeneity in spillover effects.[4] This literature emphasizes differences in the intensity of spillover effects depending on the relative position of individuals in the social network.[5] These individuals are often identified with centrality measures according to a given definition of the structure of interactions. In this paper, the identity of these individuals, together with their spillover effects, are estimated from the data.

The identification of sources of spillovers comes from the co-movements of individual's outcomes with characteristics of other individuals over time. The panel dimension of the data, together with the stability of the structure of interactions, is crucial in order to identify sources of spillovers.

There are several settings where this methodology can be useful. For example, in a context of randomized treatments, where the effect of the treatment is subject to generate externalities, the methodology can help disentangle direct treatment effects from spillover effects. Moreover, it can be used to design efficient treatment rules that take spillovers into account.

Another example where the method could be useful is in the context of production functions, where productivity is subject to generate spillovers. In this context, a particular policy relevant type of spillovers are technological spillovers arising from $R\&D$ investments. This type of spillovers have been long studied in the literature both at the micro and the macro level.[6] Technological spillovers at the firm level will be the object of my empirical application.

To estimate the model I develop a pooled panel data counterpart of the Lasso estimator (Tibshirani, 1996) that I call Pooled Lasso estimator. This estimator minimizes the sum of squared errors of the model, across individuals and time periods, subject to a constraint on the sum of the absolute values of

[4] See Jackson (2008) for an overview of the literature.

[5] An example of this type of special individuals is the key player (e.g. Ballester *et. al*, 2006).

[6] For example, at the country level, technological spillovers have been related to foreign direct investment (see Coe and Helpman, 1991).

the spillover effects. The particular geometry of the constraint, in terms of the sum of absolute values, is responsible for the sparsity in the structure of interactions of the estimator. That is, many spillover effects are estimated as zero. This property is inherited from the Lasso estimator. The Pooled Lasso estimator, as Lasso, is able to deal with lots of potential structures of interactions, even if the number of individuals in the sample exceeds the time dimension of the data.

Computation of the Pooled Lasso estimator is efficient given its convex nature. I propose an iterative algorithm that combines two steps: computation of the Lasso estimator on individual time-series for each individual, and a panel OLS regression. The first step estimates the sources of spillovers and spillover effects for each individual, while the second step recovers the effect of own characteristics. The first step makes use of efficient algorithms to compute the Lasso (Efron *et al.*, 2004).

I study the rate of convergence of the Pooled Lasso estimator in a simplified model with no common parameters, where errors show moderate time-series correlation. We assume that the number of units, the number of time periods, and the number of sources of spillovers for each individual grows with the sample size. I compare the rate of convergence of the parameters of the spillover effects of an OLS estimator, in a model in which the structure of interactions is known, to the rate of convergence of the Pooled Lasso estimator, in a model where the structure of interactions is unknown. Under the assumption of mild collinearity across (subsets of) characteristics of all individuals in the sample (Restricted Eigenvalue Conditions, e.g. Bickel, Ritov, Tsybakov), the Pooled Lasso estimator suffers a loss in the convergence rate proportional to $\log N$, where N is the number of potential sources of spillovers.[7]

Average marginal effects, i.e. averages over individual parameters, can be of interest for several reasons: First, they can be interpreted as policy parameters, as will be clear in the empirical application. Second, when analyzing the results, if there are many individual parameters, looking at statistics on the cross-sectional distribution of the parameters might be more fruitful than looking at each of them separately. Finally, when the time dimension of the data is only moderately large, each parameter might not be very precisely estimated, however averages are potentially much more precisely estimated (e.g. Chamberlain, 1992). In particular, it can be seen that average marginal effects can be estimated at the rate $\log N/NT$. Hence, gains in the rate of convergence of average spillover effect computed with the Pooled Lasso, i.e. average of spillovers generated by a particular individual on the rest of individuals in the sample, are analogous to those enjoyed when the structure of interactions is known provided the noise in estimation is independent across individuals.

Finally, I use the methodology to investigate technological spillovers in a sample of 200 US firms from 1985 to 2000. Quantifying spillovers is challenging since, among other difficulties, spillovers are hard to observe. In particular is not clear which firms should generate spillovers on others (Syverson, 2011 and Jaffe, 1986). The literature has proposed several proxy measures of spillovers given the

[7] In the baseline model N is also the number of individuals in the sample.

unobservability of the structure of spillovers. In particular, technological spillovers are usually proxied with aggregates of $R\&D$ in the economy weighted by a technological distance between firms (Jaffe, 1986). In the framework of a Cobb Douglas production function, I use the methodology to recover the effect of other firm's $R\&D$ on productivity. The methodology can capture the effect of other firms' $R\&D$ even when there are many potential firms affecting productivity. Specifically, I shed light on the nature of firms generating spillovers and receiving spillovers. In particular, firms sources of spillovers tend to be small, but hihgly intensive RD firms with more highly cited patents conditional on size. These results survive an specification where we control for the state of the art spillover measures based on technological distances, constructed using historical patent data. These measures are likely to underestimate the effect of small young firms since those don't have a history of patents.

The rest of the paper is organized as follows: Section 2 presents the model, section 3 introduces the Pooled Lasso estimator, section 4 proposes a method of computation of the estimator, section 5 discusses statistical properties of the Pooled Lasso estimator and section 6 illustrates the methodology in the context of $R\&D$ spillovers between firms during the years 1985 to 2000. Section 7 presents results on a small Monte Carlo exercise and finally section 8 concludes.

2 The Model

In this section I present the baseline model and two settings in which the model can be useful.

2.1 A panel data regression model

Let y_{it} be an individual outcome, x_{it} be an individual characteristic which may generate spillovers, w_{it} a vector of p controls. I denote $i = 1, \ldots, N$ individuals in the sample, and $t = 1, \ldots, T$ time. I consider the following linear model:

$$y_{it} = \alpha_i + \beta_i x_{it} + \sum_{j \neq i} \gamma_{ij} x_{jt} + \theta' w_{it} + \delta_t + u_{it}. \tag{1}$$

In (1) α_i is an individual-specific intercept that captures persistent unobserved heterogeneity across individuals, β_i is an individual specific slope capturing the heterogeneous effect of own characteristics, and γ_{ij} are pair-specific parameters capturing the effect of the characteristic of individual j on the outcome of individual i. Parameter θ captures the common effect of some controls w_{it}. The u_{it}'s are idiosyncratic shocks that I assume to be uncorrelated with all individual characteristics, x_{it}, w_{it}, and with the characteristics of the rest of individuals in the sample, x_{jt} for $j \neq i$. Finally, the δ_t's are time-specific dummies capturing aggregate shocks.

This model is a linear panel data regression model with spillovers, where the x's of others are additional explanatory variables of the outcome. Notice that when the number of individuals is large, the number of regressors can exceed the number of periods of observation.

In this paper a zero spillover effect is interpreted as an absence of interaction from one individual to another, hence the pair-specific parameters γ_{ij} capture both the spillover effects and the structure of interactions. More precisely, the extensive margin, $\gamma_{ij} \neq 0$ or $\gamma_{ij} = 0$, is informative on the structure of interactions, while the intensive margin, the magnitude of γ_{ij} when $\gamma_{ij} \neq 0$, captures the spillover effect of individual j on individual i.

The structure of interactions is modeled using a "fixed effects" approach. There are two features that characterize this approach: first, the structure of interactions is persistent over time (i.e. γ_{ij} does not carry a t sub-index), and second, the structure of interactions is allowed to be endogenous with respect to covariates and unobservables. In other words, the γ_{ij}'s can correlate with α_i, β_i, other γ_{ij} and any observable individual characteristic in an unspecified way.[8]

Allowing for general forms of endogeneity in the structure of interactions and spillover effects allows the researcher to relate the structure of interactions and intensity in spillover effects with characteristic of individuals, either observables or unobservables, after the model has been estimated. This is in contrast with "random effects" type models (see Arellano, 2003), where the researcher parametrizes the structure of interactions and the spillover effects in terms of covariates and unobservables ex-ante, at the risk of misspecification.[9]

Stability in the structure of interactions can arise as an equilibrium outcome in some dynamic network formation processes. For instance, Watts (2001) develops a dynamic game in which, in every period, a link between two individuals is randomly identified, then according to a specified payoff function, the link is severed or built depending on whether doing so is beneficial for both individuals. Stability of the network is reached after some periods, when no pair of individuals has an incentive to deviate from their status (either linked or not), under some conditions on the payoff function.

Another situation in which the structure of interactions does not change over time is when individuals establish links on the basis of time-invariant characteristics, as for instance, gender, race, skills, or other time-invariant features (observables or unobservables). This is consistent, for instance, with the concept of homophily, where individuals tend to establish relations with individuals with similar characteristics. This feature seems to be prevalent in many observed social structures (e.g. McPherson *et al.*, 2001).

Persistence of the structure of interactions allows to recover the sources of spillovers and pair-specific spillover effects using the longitudinal dimension of the data. In particular, sources of spillovers

[8]The structure of interactions can be endogenous with respect to time-invariant unobservables, like α_i and β_i but I rule out cases when the structure of interactions depends on the idiosyncratic shocks u_{it}.

[9]This is the case of many network formation models (e.g. Imbens *et al.*, (2011), where the researcher sets up a rule of decision of individuals to form links according to parametric functions. Another example, more related with the specification of spillover effects is Arcidiacono *et al.*, (2012), where spillovers in the classroom arise linearly through time - invariant unobservables of students.

of individuals are identified when there are co-movements of the output (net of all other covariates) with characteristics of other individuals. Nonetheless, in Appendix A I show how to partially relax the time-invariance assumption on the structure of interactions while keeping the fixed effects approach.

Model (1) can be related with a model of spillovers that has been extensively used in the literature: the linear in means model (e.g. Manski, 1993). In this model outcomes of individuals within a reference group depend on the average characteristics of the individuals in that group.[10] The spillover effect is the same for all individuals in the same group, and individuals outside of the reference group are typically assumed to have a 0 spillover effect. Model (1) extends the linear in means model in at least three ways: First, it does not restrict the spillover effect to be homogeneous within groups. Instead, spillover effects can be different for every pair of individuals. Second, spillover effects are not limited to the group. Finally, individuals with the same characteristics can generate different spillover effects, since spillover effects can differ across individuals with other observables and unobservables.[11]

Finally, spillover effects in (1) are not limited to the γ_{ij}'s. In particular, the δ_t can capture, in an unspecified way, aggregate spillover effects in the economy. Alternatively, the researcher can parametrically specify aggregate spillover effects and include them as controls in w_{it}. In both cases, the γ_{ij}'s need to be reinterpreted as spillover effects in deviation to an aggregate spillover effect.

Model (1) can be useful in the two following settings:

Example 1: Treatment effect in the presence of externalities Quantifying the direct effect of a treatment when the comparison group enjoys positive externalities from the treated can be challenging, even in a randomized treatment setting.[12] In these cases the difference in outcomes between the treated group and the non-treated group can underestimate the treatment effect.

Model (1) can be useful to disentangle the direct effect from the spillover effect of the treatment. In particular, when: 1) the treatment histories are heterogeneous across individuals, and 2) the structure of interactions remains constant over time. An example of heterogeneity in treatment histories is when treatment is assigned to a different individual each time. Heterogeneity in the time of treatment assignment across individuals is necessary to recover the structure of interactions. If two individuals receive treatment on the same period, they cannot be distinguished as sources of spillovers.[13]

Recovery of the structure of interactions is useful to design policies that efficiently take into account externalities. As an example, consider the following thought experiment inspired in Miguel and Kremer (2003): Assume that drug ingestion, the treatment subject to generate spillovers, can affect the health

[10] According to the terminology of Manski (1993), here I am only considering exogenous effects.

[11] In Graham *et al.* (2013) terminology I am relaxing homogeneity and interchangeability assumptions. The authors argue that the motivation behind these assumptions is normally the lack of further information on the structure of interactions.

[12] This is the violation of the SUTVA assumption.

[13] See Example 2 on identification in the Appendix for further intuition.

status of pupils in a school class in a rural country.[14] Assume the class is small in comparison with the time period of observations. Let d_{it} be a binary variable indicating that individual i has been treated at time t. Assume that the treatment is assigned according to an arbitrary order of pupils where the first pupil is treated on the first period, the second one is treated on the second period, etc. When all pupils have been treated once, the first pupil takes treatment again and so on. Let y_{it} be a particular health outcome that drug ingestion can improve but not harm. Consider the following model nested by (1):

$$y_{it} = \alpha_i + \beta_i d_{it} + \sum_{j \neq i} \gamma_{ij} d_{jt} + \theta' w_{it} + \delta_t + u_{it},$$

where α_i captures pupil-specific unobserved heterogeneity in health status, β_i captures heterogeneity on the treatment for pupil i, γ_{ij} captures the effect of treatment of individual j on the health outcome of i.

Allowing for heterogeneity in the treatment effect, β_i, is possible given the panel dimension of the data. The effect of the treatment for each individual can be recovered comparing its outcome over time, conditional on observables, w_{it}, and controlling for aggregate shocks (δ_t). Also, the effect of the treated pupil on other pupils is recovered because in each point in time the treatment is taken by only one pupil: if at the time that pupil 1 takes the drug, pupil 2 experiences an improvement on its outcome, then pupil 2 is affected by the treatment of pupil 1.

The school director can be interested in maximizing the aggregate improvement on health status of the class given a constraint on the number of doses of treatment she can buy. The improvement on the aggregate health status of the class if pupil k takes the drug is:

$$\sum_{i=1}^{N} \left(y_i \left(d_k = 1 \right) - y_i \left(d_k = 0 \right) \right) = \beta_k + \sum_{i=1}^{N} \gamma_{ik}, \tag{2}$$

where $y_i(d_k = 1)$ and $y_i(d_k = 0)$ denote health status of pupil i if pupil k takes the drug and if pupil k does not take the drug, respectively. The increase in aggregate health status depends on the effect of the drug on the health status of pupil k, but also on the capacity of pupil k to generate positive externalities on the health status of other pupils in the class. Assuming that the cost of each dose is homogeneous and unitary, the maximization problem of the school director is then:

$$\max_{(D_1,\ldots,D_N)} \quad \left\{ \sum_{k=1}^{N} \sum_{i=1}^{N} \left(y_i \left(d_k = 1 \right) - y_i \left(d_k = 0 \right) \right) \cdot D_k \right\}$$
$$s.t. \sum_{k=1}^{N} D_k = C,$$
$$s.t. \left(D_1, \ldots, D_N \right) \in \{0,1\}^N,$$

[14] The framework in Miguel and Kremer (2003) is a randomized experiment at the school level. They exploit variation in the density of schools in a certain area around a treated school to estimate spillover effects of deworming treatment. They randomize at the school level and not at the class level precisely to be able to control for externalities.

where C is the total budget for buying drugs, and $D_1, \ldots, D_N$ denotes the assignment treatment variable of pupil 1 to pupil N.

If the direct effect and spillovers effects are either positive or zero, the optimal allocation of the amount of doses available is to the C pupils with the highest capacity to generate an aggregate improvement on health status in the class (2). Notice that this allocation does not necessarily coincide with an allocation of doses to pupils with the highest individual treatment effect (β_i). Instead, imagine a pupil fairly resistant to infections, not experiencing huge losses of health status if infected (low β_i), but with a high degree of interactions with several pupils that if infected suffer from huge losses of health (high γ_{ji}'s). In that situation, it might be optimal to treat the "popular" pupil given its capacity to generate positive spillover effects on other pupils, even though himself does not benefit too much from it.

Example 2: Technological Spillovers in Productivity Technological spillovers in productivity between firms, industries and countries are conceptualized as knowledge transfers. Producers are likely to attempt to adopt practices of productivity leaders in their own and related industries. At the same time, producers are unable to fully appropriate all the benefits of their discoveries. Knowledge is unobservable, and is often proxied with $R\&D$ investments under the assumption that $R\&D$ enhances knowledge, which in turn enhances productivity.[15]

Policy makers are often interested in understanding the social returns to $R\&D$ investment, that is, the aggregate increase in output after a marginal increase in $R\&D$ investment of a single or several producers. However, quantifying spillovers from $R\&D$ is hard.[16] A first challenge is the lack of observation of the channels through which $R\&D$ spillovers arise. Several proxies of the structure of interactions between producers, mainly based on distances, have been related in the literature to the correlation of productivity between producers: geographic distance (Moretti, 2004), market distance, or technological distance (Jaffe, 1986). A second challenge is to disentangle correlated productivity shocks from spillovers: closely related producers are likely to experience the same productivity shocks.[17]

The methodology is potentially useful in shedding light on the channels through which technological spillovers between producers arise, since the structure of interactions is estimated from the data. In particular, I can allow for asymmetries in spillover effects, which might be important to capture that less efficient producers might be willing to replicate industry leader's best practices. (Syverson, 2011).

I will develop further this example in section 6, where I study $R\&D$ spillovers in a production function framework in a panel of US firms.

[15] $R\&D$ might not be the only source of productivity. See Syverson (2011) for a survey on what determines productivity.

[16] See Hall *et al.* (2010) for a broad relation of challenges in the empirical literature on measuring the returns to $R\&D$.

[17] This problem, named by Manski (1993), as the reflection problem is prevalent in the literature of peer effects.

2.2 Sparse structures of interactions

I focus on sparse structures of interactions, namely, structures of interactions where individuals have relatively few sources of spillovers. Sparsity on the structure of interactions is written in terms of the parameters of model (1) as:

$$\sum_{j \neq i} \mathbb{I}\{\gamma_{ij} \neq 0\} = s_i << T \text{ for all } i.$$

In words, for each individual i, the number of sources of spillover different from zero, s_i, which is unknown, is relatively small in comparison to the time dimension T.[18]

The sparsity assumption only limits the number of sources of spillovers and leaves unrestricted its identity. Also, it leaves unrestricted the intensity of the spillover effects generated by those few individuals. In this perspective, the sparsity assumption only restricts the extensive margin of the γ_{ij}'s while leaves completely unrestricted its intensive margin.

A first motivation for sparsity can be found in the literature of economics of social networks. This literature emphasizes the existence of few individuals in the network with a differentiated capacity to impact other individuals.[19] Identifying these individuals is important in this literature, specially in policy relevant settings. The sparsity assumption, in combination with heterogeneity in spillover effects, can be a good setting to recover these individuals from the data.

Sparsity also allows to identify few important sources of spillovers even when the number of potential sources of spillovers is large in relation to the time periods of observation (i.e. $N > T$). This is the case in our empirical application, where the dataset comprises $N = 200$ firms and $T = 16$. In this perspective, sparsity is a way to reduce the dimensionality of the problem.[20]

More generally, sparse structures of interactions cover a wide range of different structures of interactions. Moreover, they fulfill realistic features of social structures of interactions such as: average number of connections per individual growing slower than the number of individuals, clustering, homophily, etc.[21] Following, some more concrete examples of sparse structures of interactions.

What type of structures can be captured under the sparsity assumption? Adjacency matrices A are used in the literature on economic networks to describe the structure of interactions. These $N \times N$ matrices contain in each position a_{ij} either a 1 if individual j is related to individual i,

[18] An alternative characterization of the sparsity condition according to our relative rates beteen N, T and s_i is $s_i^2 = o(\log N)$. This condition is typical of sparse networks.

[19] Individuals related with high index of centrality measures and/or are located in strategic positions in the network. See for example Jackson (2008) for an overview on centrality measures.

[20] A different approach to reduce the dimensionality of the model would have been to impose symmetry in the structure of interactions ($\gamma_{ij} = \gamma_{ji}$). In general, the suitability of this assumption depends on the specific application one has in mind. In the context of technological spillovers allowing for asymmetry is interesting to study the interaction of spillovers with differences in productivity between firms.

[21] Chandrasekhar and Jackson (2013) also focus on sparse networks to develop an estimator of static network formation.

Figure 1: Star network

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Note: Adjacency matrix with $N = 9$, $s_i = 1$ for all $i \neq 6$ and $s_6 = 0$. The source of spillover for all individuals is the same individual: $i = 6$.

or a 0 otherwise.[22] In our framework, $a_{ij} = 0$ indicates that the output of individual $i$does not depend on the characteristic of unit j. Conversely, $a_{ij} = 1$ indicates that the output of individual i depends on the characteristic of individual j. The elements of A can be written in terms of the parameters of model (1) as follows:

$$a_{ij} = \mathbb{I}\left(\gamma_{ij} \neq 0\right),$$

where $\mathbb{I}\left(\right)$ denotes an indicator function. In words, a_{ij} capture the extensive margin of the γ_{ij}'s.

Our notion of sparsity does not impose conditions on the identity of the units generating spillovers. In particular, sparsity imposes restrictions on the sum of the rows of the adjacency matrix, but not on the sum of the columns.

A first example of a sparse structure of interactions is one in which all individuals receive spillovers from the same individual. In the context of the empirical application this would amount to have a single firm, a sort of technological leader, generating spillovers on the productivity of the rest of the firms in the sample. Figure 1 is a representation of this type of network in terms of an adjacency matrix. This type of structure is called a star network in the literature on economic networks, and the technological leader a central player. The star network is sparse since each firm has exactly one other firm from whom it receives spillovers. A structure of interactions in which there is a central player in each industry is also a sparse structure of interactions.

Another example in which each individual has only one source of spillover is in a lattice: each individual is a source of spillover to another individual and only to one other individual. The main feature in these structures is that all individuals can be reached in the structure through "friends of friends". That is, all individuals are connected.

[22] In its most general definition, adjacency matrices do not necessarily contain only 1's or 0's.

A structure of social interactions where individuals interact in pairs, triads or small groups is also sparse. De Giorgi and Pelizzari (2012) consider this type of structure in their study of the mechanisms of interactions between pairs of individuals in the same class, in terms of their outcomes in test scores. Moreover, this structure can arise after a dynamic network formation process, as discussed before, where pair stability is reached after some periods.

A wide category of large sparse networks are the so-called small world network (Watts, 1999). These structures of interactions share some features with the lattice, in the sense that most individuals are connected indirectly through other individuals, but at the same time show high degree of clustering or lots of small groups. Small worlds are found in many real social structures of interactions, one of the most famous one being the the Co-stardom network or the film or actor collaboration graph, where two actors are linked if they have appeared in the same movie.[23]

Sparse structures of interactions cover many different adjacency matrices. However, some applications might not be suited to a small number of sources of spillovers. Also, in some dataset, the longitudinal dimension of the panel might be too short and might tighten too much the number of sources of spillovers allowed for each individual. In those cases, allowing for additional sources of spillovers can be done by imposing prior information on the structure of interactions. In the context of the empirical application, spillovers might come from two different sources: an average $R\&D$ spillover at the industry level, capturing average level of knowledge in the industry, and additionally, some firm-specific sources of spillover, either within the same industry or not. An illustration in terms of the adjacency matrix can be found in Figure 2.

2.3 Quantities of interest

There are two different types of objects of interest: the parameters related to the effect of the characteristics generating spillovers, and the common parameters.

2.3.1 Average marginal effects

Model (1) is a fixed-effects model, where the number of parameters grows with the sample size of individuals. When N is large, making sense of the estimates by looking at each of them separately might not be a good strategy. Instead, looking at distributional characteristics of the fixed effects might be more informative. In addition, if T is small with respect to N, fixed effects are imprecisely estimated.[24] Aggregate quantities are potentially more precisely estimated (e.g. Chamberlain (1992)).

What features of the distribution or marginal effects are interesting depends on the application. In what follows I define two types of quantities of interest: private effects and social effects. The

[23]The data to build this structure of interactions can be found in the Internet Movie Database (http://www.imdb.com/).

[24]This phenomenon is the incidental parameter problem (Neyman and Scott, 1948), and arises when the number of parameters grows with the sample size.

Figure 2: On top of industry spillovers

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Note: Representation of a structure of interactions with two different layers of spillover effects: Individuals receive spillovers within small pre-specified groups (e.g. firms within the same industry). Additionally, individuals receive spillovers from few other individuals inside or outside its own group. The light coloring indicates the spillover effect from the reference group while the dark coloring indicates the additional sources of spillovers of each individual.

first type of effect involves marginal effects with respect to own characteristics, while the second type involves, in addition, marginal effects with respect to characteristics of other individuals.

Private effects Increase in outcome due to an increase in own characteristics:

- Increase in i's outcome after increase in own characteristics:

$$P_i = \frac{\partial\, y_i}{\partial\, x_i} = \beta_i.$$

- Average increase in outcome due to increase in own characteristics:

$$P = \frac{1}{N}\sum_{i=1}^{N} \beta_i.$$

Social effect Increase in outcome due to increase in others' characteristics:

- Increase in i's outcome after increase of characteristics of j :

$$M_{ij} = \frac{\partial\, y_i}{\partial\, x_j} = \gamma_{ij}.$$

- Average increase in outcome after increase of characteristics of j:

$$M_j = \frac{1}{N}\sum_{i=1}^{N} \gamma_{ij} + \frac{1}{N}\beta_j.$$

- Aggregate increase in outcome after increase of characteristics of all individuals:

$$M = \sum_{j=1}^{N} \sum_{i=1}^{N} \gamma_{ij} + \sum_{j=1}^{N} \beta_j.$$

Aggregate marginal increase in outcome after a marginal increase on the characteristics of all individuals in the sample is the sum of the aggregate increases in output due to spillovers plus the aggregate increase in output due to the increase in own characteristics.

2.3.2 Common parameters

Alternatively, the object of interest might be θ, the common parameter. In this case, model (1) can be seen as a model where cross-sectional dependence is unobserved and is modeled in a flexible way through the parameters γ_{ij}'s.[25] When characteristics of others have an effect on the outcome, and additionally, characteristics are correlated among individuals, not controlling for this cross-sectional dependence can lead to omitted variable bias on the common parameters.

3 Estimation

I propose to estimate model (1) as the minimizer of the following criterion:

$$\left(\widehat{\beta}, \widehat{\Gamma}, \widehat{\theta}, \widehat{\delta}\right) = \underset{(\beta,\Gamma,\theta,\delta)}{\operatorname{argmin}}\, Q\left(\beta, \Gamma, \theta, \delta\right) + \lambda \sum_{\substack{j=1 \\ j \neq i}}^{N} \left|\gamma_{ij}\right| \phi_{ij}, \tag{3}$$

where

$$Q\left(\beta, \Gamma, \theta, \delta\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{y}_{it} - \beta_i \tilde{x}_{it} - \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma_{ij} \tilde{x}_{jt} - \theta' \tilde{w}_{it} - \tilde{\delta}_t)^2,$$

where $\tilde{z}_{it}$ denotes deviations to the mean, $\beta = (\beta_1, \ldots, \beta_N)$ are $N \times 1$ vectors containing all firm-specific parameters, $\delta = (\delta_1, \ldots, \delta_T)$ is a $T \times 1$ vector containing all time-specific parameters, and Γ is a matrix with zeros in the diagonal containing all the γ_{ij}. Finally, ϕ_{ij}, are pair-specific weights that are chosen together with λ to achieve attractive rates of convergence.

Criterion (3) has two parts. The first part, Q, is a sum across all individuals and time periods of squared residuals, where residuals are defined in terms of a transformation of model (1) in which we take deviations to the means in order to get rid of fixed effects. Q coincides with the within group criterion of a pooled panel data regression model. The second part is a penalization on the parameters of the structure of interactions, γ_{ij}'s. This penalization is an increasing function in the sum of the absolute values of the parameters of the structure of interactions as $\lambda \geq 0$.

[25] An alternative way of capturing cross-sectional correlations is by means of interactive fixed effects (Bai, 2009), or by adding grouped patterns of heterogeneity (Bonhomme and Manresa, 2012).

The two parts of the criterion are exerting opposite forces. The OLS part decreases with the number of parameters γ_{ij} that are different from zero. The penalty, in contrast, increases with the absolute value of the parameters γ_{ij}, and in particular, with the number of parameters γ_{ij} that are different from zero. The relative importance of the two terms is modulated through the value of λ. For instance, when λ is large, the minimizer of (3) is likely to deliver $\widehat{\gamma}_{ij} = 0$ for all i, j.

The Pooled Lasso estimator coincides with the Lasso estimator (Tibshirani in 1996) on each of the N time-series regressions of the outcome on the characteristics of the rest of individuals in the sample when $\delta = 0$, $\beta = 0$ and $\theta = 0$. In its original formulation the Lasso estimator minimizes the sum of the squared errors of a linear regression in a cross-section, with a penalty on the sum of the absolute values of the parameters:

$$\widehat{\gamma}_i = \underset{\gamma_i}{\operatorname{argmin}} \sum_{t=1}^{T} (\tilde{y}_{it} - \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma_{ij} \tilde{x}_{jt})^2 + \lambda \sum_{j \neq i} |\gamma_{ij}| \phi_{ij} \tag{4}$$

A related estimator is the Post-Lasso estimator: a two - step estimator in which in a first step the Lasso is used as a device for selecting the sources of spillovers, and in a second step estimates of the spillover effects are obtained by regression by OLS on the selected regressors. In the context of the Pooled Lasso estimator we will also consider performing post-lasso regressions to get rid of the shrinkage bias on the estimated parameters.[26]

The Pooled Lasso estimator delivers sparse structures of interactions, as it inherits this property from the Lasso estimator.[27] In addition, the criterion is globally convex, as Lasso. However, it differs from the Lasso and even the Cluster Lasso, (Belloni, Chernozhukov, Hansen, Kozbur, 2014), another Lasso-type estimator in panel data, as with the pooled regression framework the presence of common random coefficients that are penalized naturally arise. In addition, we also consider common parameters, as time dummies δ_t.

Finally, an estimator for average marginal effects is simply:

$$\widehat{M}_j = \frac{1}{N} \sum_{i=1}^{N} \widehat{\gamma}_{ij}^P$$

where $\widehat{\gamma}_{ij}^P$ is the Post-Pooled Lasso estimator. Another reason to perform OLS estimation after model selection to construct Average Marginal Effects estimators is to get rid of cross-section dependence in $\{\widehat{\gamma}_{iji}$ arising from dependence across weights $\{\phi_{ij}\}_i$ for fixed j.

[26]The superiority of the Post-Lasso estimator with respect to the Lasso estimator is granted when there is perfect model selection.

[27]See Belloni Chen Chernozhukov Hansen 2012 for a proof for Lasso.

4 Statistical properties

In this section I discuss statistical properties in a setting where N, the number of individuals, T, the number of time periods and s_i, the number of sources of spillovers for each individual, tends to infintiy. The relative rates of convergence of these three quantities will be made more explicit below.

I restrict the analysis to the study of rates of convergence of the parameters of a simplified model where there is no common parameters (i.e. $\delta = 0$ and $\theta = 0$), as well as the rate of convergence of average marginal effects. The analysis of a model with common parameters θ is analogous (work in progress) but incorporating time-dummies is substantially more complex and is left for future research. Similarly, inference in the Lasso-type estimators is a very active area of researcher nowadays and I don't cover it in this discussion. At the end of the paper, in the conclusions, I get back to this very important question.

The longitudinal dimension of the data allows to consistently recover the identity of the sources of spillovers for each individual. In this subsection I provide an intuition on consistency of model selection in a simplified case.

Consider the following simplified model of spillovers, as in (4):

$$\tilde{y}_{it} = \sum_{j=1}^{N} \gamma_{ij} \tilde{x}_{jt} + \tilde{u}_{it} \tag{5}$$

where for each i we have $\sum_{j \neq i} \mathbf{1}\{\gamma_{ij} \neq 0\} = s_i << T$, where the number of spillovers received by others is small in relation to the entire population. Errors $\{u_{it}\}_{t=1}^{T}$ have limited dependence on the time series and regressors are predetermined. A precise statement of the conditions can be found below.

The Pooled Lasso estimator of (5) is equal to the Lasso estimator on each individual time series regression, after getting rid of the fixed effects by doing deviations to the mean as in Belloni, Chernozukov, Hansen, Kozbur (2015):

$$\tilde{y}_{it} = y_{it} - \frac{1}{T} \sum_{s=1}^{T} y_{is}.$$

Notice that this transformation mechanically introduces dependence in the errors.

Consider the following Lasso estimator in each individual time series:

$$\widehat{\gamma}_i = \underset{(\gamma_{i1}, \ldots, \gamma_{iN})}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} (\tilde{y}_{it} - \sum_{j \neq i} \gamma_{ij} \tilde{x}_{jt})^2 + \frac{\lambda}{T} \sum_{j \neq i} |\gamma_{ij}| \, \phi_{ij} \tag{6}$$

where λ is a penalty parameter and ϕ_{ij} are pair-specific weights.

We also consider the Post-Lasso estimator, where the Lasso estimator is used in a first step to estimate the identity of the sources of spillovers, and then the spillover effects are estimated with an OLS time-series regression on the selected sources of spillovers, $\widehat{T}_i \subset \{1, \ldots, N\}$:

$$\widehat{\gamma}_i^P = \underset{(\gamma_{i1},\ldots,\gamma_{iN}):\ \gamma_{ij}=0\ if\ j\notin\widehat{T}_i}{\text{argmin}} \frac{1}{T}\sum_{t=1}^{T}(\tilde{y}_{it} - \sum_{j\subset\widehat{T}_i}\gamma_{ij}\tilde{x}_{jt})^2. \tag{7}$$

The choice of λ and ϕ_{ij} determine the rate of convergence of the estimator as the rate of convergence of the Lasso estimator is related to the probability of the noise in estimation to be bounded by the penalty level uniformly. A suitable choice of ϕ_{ij} and λ is one in which, for a given $\varepsilon > 0$ small:

$$Pr\left[\max_{i,j}\frac{2\phi_{ij}^{-1}}{T}\left|\sum_{t=1}^{T}u_{it}x_{jt}\right| > \frac{\lambda}{T}\right] \leq \varepsilon.$$

Before analyzing the outlined estimators above we provide an intuition of consistent model selection in a related model.

4.1 Consistent Model Selction: An Intuition

Consider a model as in (5) where $\alpha_i = 0$ for all i. I further assume that each individual only has one source of spillover, which I denote as $j(i)$, and $\gamma_{ij(i)}$ is its spillover effect:

$$y_{it} = \gamma_{ij(i)}x_{j(i)t} + u_{it}. \tag{8}$$

Moreover, I assume that regressors are normalized: $\frac{1}{T}\sum_{t=1}^{T}x_{it} = 0$ and $\frac{1}{T}\sum_{t=1}^{T}x_{it}^2 = 1$ for all $i = 1,\ldots,N$.

In this model there are $N-1$ potential sources of spillovers but only one is the actual source of spillover, and we wish to estimate its identity and its effect. I denote R_j^2 the R-squared of the time series regression of the outcome on characteristics of individual j. An estimator of the source of spillover to i is the following:

$$\widehat{j}(i) = \underset{k\neq i}{\text{argmax}}\, R_k^2.$$

That is, the estimator of the identity of the source of spillovers is the individual whose characteristic, $x_{\widehat{j}(i)t}$, has the highest explanatory power on the outcome of i, y_{it}. Under this estimation rule, and given that regressors are normalized, maximizing the R^2 is the same as maximizing the absolute value of the sample covariance between y and x. The probability that the estimated source of spillover, $\widehat{j}(i)$, does not coincide with the true source of spillovers, $j(i)$, is:

$$\begin{aligned}\mathbb{P}\left(\widehat{j}(i)\neq j(i)\right) &= \mathbb{P}\left(\left|\widehat{Cov}\left(y_{it},x_{j(i)t}\right)\right| < \sup_{k\neq j(i)}\left|\widehat{Cov}\left(y_{it},x_{kt}\right)\right|\right)\\ &= \mathbb{P}\left(\left|\gamma_{ij(i)} + T^{-1}\sum_{t=1}^{T}x_{it}\epsilon_{it}\right| < \sup_{k\neq j(i)}\left|T^{-1}\sum_{t=1}^{T}x_{kt}\epsilon_{it}\right|\right).\end{aligned}$$

This is, the probability that the estimated source of spillovers is not the true source of spillover is equal to the probability that the maximum sample covariance between the output and all regressor is not attained using regressor $x_{j(i)t}$.

Consistency of model selection as N and T grows thus depends on the behavior of the following quantity:

$$\sup_{k \neq j(i)} \left| T^{-1} \sum_{t=1}^{T} x_{kt} \epsilon_{it} \right|.$$

In words, the noise in estimation is equal to the sup of the noise in estimation of time series regressions of outcome i on the characteristics of each individual in the sample. On the one hand, each of the $T^{-1} \sum_{t=1}^{T} x_{it}\epsilon_{kt}$ vanish at the typical $\sqrt{T}$ rate. However, since the identity of the influencing unit is unknown, the noise in estimation grows with the potential sources of spillovers. In particular there are $N-1$ potential sources of spillovers, and as a consequence $N-1$ different noise $T^{-1} \sum_{t=1}^{T} x_{it}\epsilon_{kt}$, with $k = 2\ldots, N$ to control. Given that regressors are normalized, and that they are independent of ϵ_{it}, $T^{-1} \sum_{t=1}^{T} x_{kt}\epsilon_{it}$ is distributed as $N(0, \frac{\sigma^2}{T})$. The following statistical result is informative on how the noise in estimation grows when the structure of interactions is unknown. Let ς_k be distributed as $N(0, \sigma^2)$, then:

$$\sup_{1 \leq k \leq N} |\varsigma_k| = O_p\left(\sigma \sqrt{\log N}\right)$$

Applying this result to $T^{-1} \sum_{t=1}^{T} x_{kt}\epsilon_{it}$:

$$\mathbb{P}\left(\widehat{j}(i) \neq j(i)\right) \quad \approx \quad \mathbb{P}\left(\left| \gamma_{ij(i)} \right| < O_p\left(\sigma \sqrt{\frac{\log N}{T}}\right)\right).$$

Hence, consistent model selection requires $\frac{\log N}{T} \longrightarrow 0$.

Remark 1 *The estimation rule in the above simple model, coincides with the Lasso estimator for a suitable choice of the penalization parameter λ_i (see Efron et al., 2004). More generally, the Pooled Lasso estimator can be seen as a convexification of another penalized estimator, where the penalty is in terms of the sum of the number of spillover effects parameters that are different from zero for each individual. This estimator searches among all potential structures of interactions and selects the one that minimizes the residual variance. These combinatorial-type estimators are natural estimators in a setting with sparse structures of interactions, however their computational properties are poor.*

4.2 Rates of Convergence for Time-series Lasso and Average Marginal Effects

The conditions to derive rates of convergence of the pooled lasso estimator mimick those of the simplified model using the naive estimator. In particular, the following assumption limits the cross-sectional dependence of the x's among the different potential sources of spillovers:

Assumption 1 *Condition Sparse Eigenvalues*
Let M be the Gram matrix, $M = \frac{1}{T}X'X$, where X contains all the regressors $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{iT})'$ in columns. We define the minimal and maximal m-sparse eigenvalues of M as follows:

$$\phi_{min}(Cs)(M) = \min_{\delta \in \Delta(m)} \delta' M \delta \quad \textit{and} \quad \phi_{min}(Cs)(M) = \max_{\delta \in \Delta(m)} \delta' M \delta$$

where

$$\Delta(M) = \left\{ \delta \in \mathbb{R} : \sum_{j=1}^{N} \mathbb{1}\{\delta_j \neq 0\} < m, \qquad \|\delta\|_2 = 1 \right\}.$$

For any $C > 0$, there exists constants $0 < \kappa_1 < \kappa_2 < \infty$ which do not depend on N or T but may depend on C, such that, with probability approaching 1, as $N, T, s \to \infty$, where $s/T \to 0$,

$$\kappa_1 \leq \phi_{min}(Cs)(M) \leq \phi_{max}(Cs)(M) \leq \kappa_2.$$

This assumption (see Ritov, Bickel, Tsybakov, 2009) limits the amount of cross section correlation between the x's of different individuals. As in the case of the naive estimator, based on the maximal correlation, this assumption prevents that two individuals share exactly the same history of x's so that they can be distinguished as a potential source of spillovers.[28]. This assumption also implies that there is time-series variation of each series x_{it}, given that we have taken deviations to the mean.

The behaviour of the Lasso estimator crucially depends on the behaviour fo the noise in estimation. The following set of assumptions imposes conditions on the dependence structure of the error and covariates. These conditions ensure that for a particular choice of λ and weights ϕ_{ij} the probability of the noise in estimation is uniformly bounded, for all i and all regressors $j \in \{1, \ldots, N\}$ by the penalty level.

Assumption 2 *Time Series Lasso*
Let $K > 0$, $M > 0$ constants independent of N, T, s.

a. For all i, j, $\gamma_{ij} \in \Gamma$, where $\Gamma \subset \mathbb{R}$, is a compact set.

b. $\mathbb{E}[x_{jt}^2] \leq K$, and $\mathbb{E}[\left(\frac{1}{T}\sum_{t=1}^{T} x_{jt}\right)^2] \leq K$, for all $j = 1, \ldots, N$.

c. $\mathbb{E}[u_{it}|x_{j1}, \ldots, x_{jt}] = 0$ for all i, j, t.

d. There are constants $a > 0$, $d_1 > 0$ and a sequence $\alpha[t] \leq e^{-at^{d_1}}$ such that, for all $i, j \in \{1, ..., N\}$, $\{u_{it}x_{jt}\}_t$, and $\{u_{it}\}_t$ are strongly mixing processes with mixing coefficients $\alpha[t]$.

e. There are constants $b > 0$ and $d_2 > 0$ such that $\Pr(|u_{it}| > m) \leq e^{1-\left(\frac{m}{b}\right)^{d_2}}$ for all i, j, t, and $m > 0$.

f. The following relative rates hold: (i) $\log N = o(T^{1/3})$, (ii) $s^2 \log^2 N = o(T)$.

[28] Primitive conditions in a context of cross section have been studied in Belloni, Chen, Chernozhukov, Hansen (2012), and their references

In 2.a we require the parameter space to be compact. Condition 2.b rules out non-stationary covariates and limits the time-series dependence of covariates. We allow for covariates to be predetermined in 2.c. Condition 2.d is a high level condition on the time series dependence of the process $\{u_{it}x_{jt}\}_t$. This condition is satisfied, for instance, if $\{x_{jt}\}_t$ and $\{u_{it}\}_t$ are both mixing and independent.[29] This condition, together with the mixing condition on $\{u_{it}\}_t$ ensures that the noise in estimation, once deviations to the mean have been taken, is bounded for an appropiately penalty level, when N and T go to infinity. Importantly, while limited time-series dependence in errors is required in order to recover individual spillover effects, cross-section correlation is left unrestricted. Conditions on mild cross-section correlation in errors conditional on covariates are only needed to derive rates of convergence of average marginal effects. This is the case as only the time-series dimension of the data is informative about the identity and magnitude of the sources of spillovers of each individual. In particular, this allows for correlated shocks across individuals. Finally condition 2.e requires the tail of u_{it} to decays exponentially fast.

The choices of λ and ϕ_{ij}, determine the rate of convergence of the estimator. A suitable choice of ϕ_{ij} and λ is one where:

$$Pr\left[\max_j \frac{2\phi_{ij}^{-1}}{T}\left|\sum_{t=1}^{T}\tilde{u}_{it}\tilde{x}_{jt}\right| > \frac{\lambda}{T}\right] = o(1).$$

The assumptions on exponential tails and mixing on covariates and errors, together with the relative rates of convergence, allow to prove that the nose in probability is uniformly controlled by the penalty level with high probability as N and T go to infinity.

We introduce additional assumptions in order to derive rates of convergence of average marginal effects.

Assumption 3 *Average Marginal Effects*
Let $T_i \subset \{1,\dots,N\}$ denote the subset of selected regressors, where $|T_i| < s$.

a. $|\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{jt}\tilde{x}_{kt}| \leq \frac{1}{Us}$ *for all $1 \leq j < k \leq N$ with probability 1, for each N, T and s.*

b. Let $\tilde{X}[T_i]$ denote the set of columns of $\tilde{X}$ according to the subset T_i. Then for all $1 \leq j \leq N$

$$\sup_{T_i}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}_{jt}\widetilde{u}_{it}\mathbb{1}\{j \in T_i\} = O_p\left(\frac{1}{\sqrt{NT}}\right),$$

where $\ddot{X}[T_i] = (\tilde{X}[T_i]'\tilde{X}[T_i])^{-1}\tilde{X}[T_i]$.

c. The distribution of spillover effects is such that:

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\sup_{j\subset T_i^0}\mathbb{P}_{\gamma_i}\left(\frac{K}{N} \leq |\gamma_{ij}^0| \leq \frac{\lambda}{T} \mid j \in T_i^0\right)\right) = O\left(\frac{1}{N}\right)$$

[29] Sor far the case where $x_{it} = y_{it-1}$ is not covered since AR(1) type of models are not mixing necessarily (Andrews).

as N, T and s goes to infinity, where $\mathbb{P}_\gamma$ governs the distribution of $\{\gamma_{ij}^0\}$ conditional on j generating some spillover on i and the $\mathbb{E}$ is the expectation with respect to the structure of interactions.

Assumption 3.a limits the covariance between regressors to be inversely proportional to the maximum number of regressors in the model, s. This condition implies Assumption 1.[30] In 3.b we require that the sample mean across time and individuals of the product of erros, u_{it} and standardized covariates, $\ddot{x}_j t$, vanishes at the rate $\sqrt{NT}$ irrespectively of the selection of regressors. This condition effectively limits the time-series and cross-section dependence of errors and covariates.[31] Finally, 3.c imposes conditions on the distribution of spillover effects in order to limit the impact of misselection of regressors. This is a weaker condition than the so-called beta min condition, where all spillover effects different from zero would be sufficiently far away from zero.

Proposition 1 *Rate of convergence of time-series Lasso and Average Marginal Effects*
*Suppose Assumptions 1, 2, and 4 hold. Consider the estimators (**??**) and (7) with penalty level $\lambda = K\sqrt{T\log(N/\epsilon)}$ and weights $\phi_{ij}^2 = \widehat{\mathbb{V}}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T \tilde{u}_{it}\tilde{x}_{jt}\right]$, where K is a constant independent of i. Then, we have that:*

$$\sup_i \sum_{j=1}^N \left(\widehat{\gamma}_{ij} - \gamma_{ij}^0\right)^2 = O_p\left(\frac{s\log N}{T}\right).$$

where $\widehat{\gamma}_{ij}$ denotes either the Lasso or the Post Lasso estimator. In addition, if Assumption 3 holds then:

$$\left(\frac{1}{N}\sum_{i=1}^N \widehat{\gamma}_{ij}^P - \frac{1}{N}\sum_{i=1}^N \gamma_{ij}^0\right)^2 = O_p\left(\frac{s\log N}{NT}\right).$$

where $\widehat{\gamma}_{ij}^P$ is the Post-Lasso estimator.

Proof. See the Appendix. ■

The proof on the rate of convergence of the Lasso and Post-Lasso estimators of the spillover effects draws significantly from the series of papers by Belloni, Chernozhukov, Hansen and coauthors. In particular Belloni, Chen, Chernozhukov and Hansen (2012), where they prove a similar result in the context of heteroskedastic independent errors in the cross-section. The main difference relies on the strategy to control the noise in estimation. Belloni and coauthors use Gaussian approximations of self-normalized sums to obtain feasible choices of λ and ϕ_{ij}, instead in a context of dependence in errors Fuk-Nagaev type of inequalities due to Rio (2000) are useful to control this same quantity.[32] I

[30] This assumption has implications in terms of the value of the constants κ and κ', although in our context this is unimportant. See e.g. Belloni and Chernozhukov (2012) for further details

[31] More primitive conditions are readily available in a context of exogenous regressors with mild dependence in the time-series and independence in the cross-section. Conditions in a weak dependence and predetermindness context are work in progress.

[32] An alternative strategy would be to use self-normalized sums for dependent variables (e.g. Chen et al. 2015). However this strategy would require changing the estimator and would only be useful in contexts where T is substantial.

obtain the same rate of convergence when dependence is mild.[33]

The rate of convergence of the spillover effects is slower when the structure of interactions is unknown. The cost associated to estimating the structure of interactions is proportional to $\sqrt{\log N}$. This cost reflects that there are N potential sources of spillover.[34] However, in practice, the loss in rate of convergence can be mild even when there are many potential sources of spillovers, since $\log N$ is small relative to N. In particular, the Pooled Lasso estimator is consistent when $\frac{\log N}{T}$ goes to zero, and s_i, the number of sources of spillovers, grows mildly at the rate $s_i = o_p\left(\frac{T}{\log N}\right)$. Notice that this includes the case when the number of potential spillovers, N, substantially exceeds the number of time periods of observations.

When the number of potential sources of spillovers is fixed, the rate of convergence of $\widehat{\gamma}_i$ is the same when the structure of interactions is known and when the structure of interactions is unknown.[35] An example of this situation is when N is fixed. Another example is when the set of individuals potential sources of spillovers is not the rest of individuals in the sample but a fixed subset of them. For instance, in a context of spillovers in a sample of school kids, the potential sources of spillovers are restricted to arise in the class and the size of the class is fixed. N can still grow if I think that I am adding more classes in the sample.

Estimation of the structure of interactions can be costly in terms of the rate of convergence, but this cost can be relatively mild even if N is quite large. However, the conditions on the Gram matrix under which the rate of convergence of the Pooled Lasso estimator is close to the OLS estimator with knwon structure of interactions are substantially more demanding when the structure of interactions is unknown. In particular, when the structure of interactions is known, it is enough that collinearity among characteristics of the actual sources of spillovers is limited. In contrast, when the structure of interactions is unknown the paths of the characteristics of **all** potential sources of spillovers need vary differently over time. Unless this is the case, the Pooled Lasso estimator cannot recover the structure of interactions.

When the number of periods of observations is small, each of the γ_{ij} is poorly estimated, even when the structure of interactions is known. However, average marginal effects enjoy a gain in the rate of convergence proportional to the number of individuals in the population, in spite of model selection mistakes.

The derivation of the rate of convergence of average marginal effects of the post-lasso estimator is novel in the high-dimensional literature. This result extends the results on average marginal effects of random coefficients models to the case where the number of random coefficients is high-dimensional.

[33] However, using this alternative strategy I loose the ability to obtain feasible choice of λ, although I obtain the same rates.

[34] See the next subsection for further intuition of the rate of convergence.

[35] The Lasso estimator is said to achieve the oracle rate of convergence, that is, the rate of convergence as if the true model was known.

Analogously to the random coefficients models, the speed of convergence of average marginal effects can increase substantially if the cross-section is large and there is limited cross-section dependence in errors. This results depends on the population distribution of the spillover effects, however. In particular, the result is derived under the assumption that the proportion of small but not too small spillover effects is small. This assumption is used to limit the impact of imperfect model selection on the average marginal effect estimator. Imperfect model selection impacts in two different ways in the average marginal effects estimator: first, if a particular spillover effect is not selected this creates a bias on the average marginal effect itself, similar as if we would have censored data. In addition, even if a particular regressor is selected, omitted variable bias can create a bias in the estimated effect if there is misselection of the sources of spillovers for that particular individual. Both effects are sufficiently small as to not to impact the average marginal effects rate of convergence when the distribution of marginal effects is such that . A particular case in which this assumption is fulfilled is when all spillover effects in the population are sufficiently far away from zero in relation to the level of the penalty. Then, as the size of the data increases perfect model selection is guaranteed. It is important to stress that the result does not depend on the assumption that for a given individual, its impact on the rest of the population is sparse. Sparsity is only required on the number of sources of spillovers that each individual has (by rows), but nothing is required in termes of how many individuals a particular individual can impact (by columns).

Reaching the gap between the analysis of this simplified model and the analysis of the baseline model (1) is work in progress. Some aspects are more difficult to overcome than others. Including the effect of own characteristics (β_i) is easy. Projection arguments as in standard linear regression models can be used to separately estimate own characteristics effects, not subject to penalization, from spillover effects.[36] When including common parameters (θ) in the model, the rate of convergence of spillover effects is likely to remain unchanged but the θ's are likely to enjoy the same gain in the rate of convergence as average marginal effects. In contrast, allowing for time dummies is presumably hard.[37]

5 Practical Aspects

In this section I discuss important practical aspects of the Pooled Lasso estimator. First, I discuss choice of the ϕ_{ij} weighting parameters and the penalization parameter λ. Second, I propose a computation method.

[36] See Bulhmann van der Geer (2011) for more details.

[37] Time dummies cannot be "differenced" out as individual fixed effects because after the transformation the model is not sparse anymore.

5.1 Choice of penalization parameter λ and weights ϕ_{ij}.

The penalization parameter λ and weights ϕ_{ij} determine the number of individuals generating spillovers in the economy. When λ is large, the Pooled Lasso estimator is likely to estimate little spillovers, while when λ is small we might estimate a larger number of spillovers in the economy.

In order to implement the choice of λ consider the optimization problem (6). It can be seen (e.g. see Lemma 2.1. in Bulhmann Van der Geer, 2011) that for given λ and ϕ_{ij}, the Kuhn Tucker conditions for (6) are:

$$\begin{aligned} \frac{2}{T}\sum_{t=1}^{T} x_{jt}(y_{it} - \sum_{j=1}^{N} \gamma_{ij} x_{jt}) &= \frac{\phi_{ij}\lambda}{T} sign(\gamma_{ij}) \quad \text{if } \gamma_{ij} \neq 0, \\ \frac{2}{T}\sum_{t=1}^{T} x_{jt}(y_{it} - \sum_{j=1}^{N} \gamma_{ij} x_{jt}) &< \frac{\phi_{ij}\lambda}{T} \quad \text{if } \gamma_{ij} = 0. \end{aligned}$$

Hence, in order to ensure that $\widehat{\gamma}_{ij}$ is close to γ_{ij}^0 we want to pick λ and ϕ_{ij} such that:

$$Pr\left[\sup_{i,j} 2c\left|\frac{\phi_{ij}^{-1}}{T}\sum_{t=1}^{T} x_{jt}\epsilon_{it}\right| \leq \frac{\lambda}{T}\right] \leq 1 - \varepsilon \qquad \text{for all} i, j.$$

As pointed out first in Bickel, Ritov Tsybakov (2009) a sensible choice of λ and ϕ_{ij} is one where the score is uniformly bounded by the penalization parameter λ after weighting by ϕ_{ij}. Given the choice of $\phi_{ij}^2 = \widehat{\mathbb{V}}[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_{jt}u_{it}]$ the central limit theorem suggests picking $\lambda/\sqrt{T}2c$ slightly above the $(1-\epsilon)$ quantile of the maximum of N^2 Gaussians.[38]

Finally, in order to implement the choice of ϕ_{ij}^2, we consider estimators of variances that take into account dependence over time, such as the Newey West (1987) HAC estimator. In addition, we make use of a feasible variation of the Lasso estimator, the iterative Lasso, proposed by Belloni *et. al* (2011) which allows to obtain consistent estimators of the error, necessary to compute the weights.

In Section 7 I illustrate the finite sample behaviour of the rate of convergence of marginal effects with the above described choices of λ and ϕ_{ij} through a Monte Carlo simulation.

5.2 Computation

The Pooled Lasso estimation problem is a convex optimization problem. I propose a method of computation that boils down to two well known estimation steps: an OLS regression and N different Lasso estimations on the times series of each individual. I make use of existing very efficient routines of optimization in order to solve the Lasso problems. In particular, I use the *LARS* algorithm (Efron

[38] Note that the choice of penalization parameter in Proposition 1 is not feasible as it depends on u', defined in Assumption 2.

et. al, 2004) to compute the time-series Lasso. The *LARS* algorithm is able to calculate the Lasso estimates according to all relevant values of the penalization parameter λ_i in a very efficient way.[39]

I propose the following iterative algorithm to compute the Pooled Lasso estimator:

1. Choose θ^0, β^0, δ^0. Set $m = 1$.

2. Obtain $\alpha^{(m)}$ and $\gamma_i^{(m)}$ by solving the Lasso estimator for each i:

$$\left(\alpha_i^{(m)}, \gamma_i^{(m)}\right) = \underset{(\alpha_i, \gamma_i)}{\operatorname{argmin}} \left\{ Q_i(\alpha_i, \beta_i^{(m)}, \gamma_i, \theta^{(m)}, \delta^{(m)}) + \lambda_i \sum_{j \neq i} |\gamma_{ij}| \, \phi_{ij} \right\}$$

   where $Q_i(\alpha_i, \beta_i, \gamma_i, \theta, \delta) = \sum_{t=1}^{T} (y_{it} - \alpha_i - \beta_i x_{it} - \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma_{ij} - \theta w_{it} - \delta_t)^2$

3. Update the values of θ, β and δ by OLS

$$\left(\beta^{(m+1)}, \theta^{(m+1)}, \delta^{(m+1)}\right) = \underset{(\beta, \theta, \delta)}{\operatorname{argmin}} \left\{ Q(\alpha^{(m)}, \beta, \Gamma^{(m)}, \theta, \delta) \right\}.$$

4. Set $m = m + 1$. Go to Step 2 until convergence.

The final estimators can be defined as : $\widehat{\alpha}_i = \widehat{\alpha}_i \left(\widehat{\theta}, \widehat{\beta}_i, \widehat{\delta}\right)$, $\widehat{\gamma}_i = \widehat{\gamma}_i \left(\widehat{\theta}, \widehat{\beta}_i, \widehat{\delta}\right)$, $\widehat{\beta} = \widehat{\beta} \left(\widehat{\alpha}, \widehat{\Gamma}\right)$ $\widehat{\theta} = \widehat{\theta} \left(\widehat{\alpha}, \widehat{\Gamma}\right)$, and $\widehat{\delta} = \widehat{\delta} \left(\widehat{\alpha}, \widehat{\Gamma}\right)$.

Notice that in each of the steps the objective function is non-increasing. As a consequence, since (3) is globally convex, this iterative algorithm attains the global minimum.

[39] In the case that the number of regressors is smaller than the sample size computing the whole sequence of Lasso estimators involves the same number of operations as computing an OLS estimation.

6 Empirical Application

I use this methodology to study $R\&D$ spillovers in a production function framework.

6.1 Background and Motivation

$R\&D$ investments enhance firm's knowledge, which in turn increase firm's productivity. Knowledge spillovers are likely to arise between firms since firms might not be able to keep every aspect of their production process secret. As a consequence, firms might take advantage of each other's knowledge.

There is a large empirical literature devoted to quantifying the returns to $R\&D$ investments, both from the private perspective (i.e. returns enjoyed by the firm investing in $R\&D$) and the social perspective (i.e. returns enjoyed by the firms in the economy plus spillover effects).[40] This literature is partly motivated by the policy implications that the presence of $R\&D$ spillovers can have for the productivity of the whole economy: If social returns are high, the policy maker might be interested in promoting $R\&D$ investment among firms. On the other hand, policies meant to increase spillovers might destroy firm's incentives to innovate if it gets too hard to appropriate the benefits of their innovations.

In order to quantify $R\&D$ spillover effects one would like to measure the impact that a firm's investment in $R\&D$ has on the productivity of another firm. However, this empirical strategy might be hard to implement for several reasons. In general, theory does not provide with a clear guidance on which firms generate spillovers on others. Spillovers might need not be tied to any single geographic and or input market (Syverson, 2011). If there are many firms, searching for spillover effects among all the potential pairs of firms might be costly.[41]

Given that interactions are unobservable, the literature has used several between-producer distance definitions to account for potential spillovers between firms. The literature has documented the importance of spillovers by adding these proxies as additional inputs in a production function framework (Griliches, 1979). The Jaffe measure (Jaffe, 1986), which takes into account the overlap in historical patenting behavior of firms, is the preferred measure of technological spillovers in the literature (Bloom, *et al.* 2013).

In this empirical application I use the methodology to uncover actual spillovers between firms. For a given firm, the methodology is well suited to recover the identity of the firms whose $R\&D$ has an effect on productivity, even if the number of firms is large. In addition, the methodology captures asymmetries in spillovers, which are ruled out by construction when using distances as proxy for spillovers between firms. Allowing for asymmetric spillover effects is important since, as stated in Syverson (2011) page 349: "Firms are likely to attempt to emulate productivity leaders in their own

[40] See Hall, Mairesse and Mohnen (2011) for a review on the literature of measuring returns to $R\&D$.

[41] Other problems associated to this strategy are the presence of correlated productivity shocks that can be confounded with spillovers (Manski, 1993), and the exogeneity of $R\&D$.

and closely related industries". That is, spillovers are likely to originate in high productive firms and cascade down to less productive firms.

6.2 Data and Sample Selection

I use the NBER match of the USPTO patent database with the Compustat Accounting database (see Hall *et al.*, 2001). The first dataset contains information on patenting behavior of firms and citations, while the second one contains information on firm-level accounting data (sales, employment, capital, etc.).

I take as starting point the selected firms in Bloom, Schankerman and Van Reenen (2013) (BSV from now on): 715 firms for which there is patenting history as well as availability of data on their segments of sales. Time span from 1980 to 2000. I first consider a balanced panel of firms between the years 1985 to 2000. The $R\&D$ activity of these firms is heterogeneous both in the intensive and the extensive margin. Out of the 463firms, 168 firms (36.3%) do not invest in $R\&D$ at all throughout the whole period, or start investing in $R\&D$ at some point between the years 1980 - 2000. I keep the firms whose $R\&D$ stock is positive throughout the whole period of time. Using the entire sample of firms in estimation is work in progress.

In Table 5 I compare the mean and the median values of a list of descriptive statistics for the whole sample and the selected sample. Firms in the restricted sample are, in mean and median, larger in terms of real market value, sales, capital, labour and $R\&D$ expenditures. Moreover, firms in the restricted sample show higher ratios of $R\&D$ stock over capital stock. Overall the restricted sample contains larger firms with larger $R\&D$ stocks relative to capital stocks.

In Table 6 I show the distribution of firms in the sample across industries in terms of the SIC2 industry classification index. The most prevalent industries coincide with industries that are more intensive in $R\&D$ expenditure, as is the Chemical Industry (28), that contains the Pharmaceutical Industry, the Electronic Industry (36), or the Industrial Machinery industry (35), that contains the computer industry. Table 7 shows descriptive statistics by industry of the selected sample. The most prevalent industries also tend to have high investment and high stocks of $R\&D$. Moreover, they also have large ratios of $R\&D$ stocks to capital stocks (above 50%). However, the industry with the largest ratio (on average) is industry (73): *Business Services*. This industry contains firms in the business of Software, Computer programming, etc. Moreover this same industry is the highest in patent per years and also patent citation.

6.3 Production Function

I assume firms produce according to a Cobb Douglas production function augmented with the stock of $R\&D$ or knowledge capital (K):

$$Y_{it} = A_{it} L_{it}^{\theta_l} C_{it}^{\theta_c} K_{it}^{\beta_i} \prod_{j \neq i} K_{jt}^{\gamma_{ij}} e^{\epsilon_{it}}. \tag{9}$$

Y denotes output, L is labor, C is capital, and ϵ is an idiosyncratic error.[42] The technological progress component, A_{it}, is a firm-specific shock to productivity. It has a firm-specific time invariant part and an aggregate part at the industry level. The firm-specific productivity level captures persistent differences in productivity across firms. It can be interpreted as time-invariant characteristics of the firm that determine its productivity level (e.g. specific corporate governance practices). The common shock or time effects capture determinants of productivity at the industry level:

$$A_{it} = A_i D_{I(i)t}$$

where $I(i)$ indicates the industry to which i belongs.

The Total Factor Productivity (TFP) in this model is defined as:

$$\frac{Y_{it}}{C_{it}^{\theta_c} L_{it}^{\theta_l}} = A_i D_{I(i)t} K_{it}^{\beta_i} \prod_{j \neq i} K_{jt}^{\gamma_{ij}} e^{\epsilon_{it}}.$$

There are two sources of spillovers in the TFP. The first source of spillovers is explicitly modeled by including the rest of the firms $R\&D$ as additional inputs in the production function.[43] The second (potential) source of spillover is captured through the aggregate time-industry effects in $D_{I(i)t}$. These time-industry aggregate effects contain any aggregate shocks, including spillover effects, at the industry level. As a consequence, the γ_{ij}'s are spillover effects in deviation to an aggregate industry spillover effect.

In order to capture heterogeneity in the effect of knowledge across firms, I allow for firm-specific output elasticities to own knowledge , β_i, and pair-specific spillover effects γ_{ij}. In particular, I allow for asymmetric spillover effects: when the knowledge of firm i has an impact on the productivity of firm j, the reverse is not necessarily true.

6.4 Taking the model to the data

Taking logs in (10) I obtain the following regression model:

$$y_{it} = \alpha_i + \beta_i x_{it} + \sum_{j \neq i} \gamma_{ij} x_{jt} + w_{it}' \theta + \delta_t + \epsilon_{it}, \tag{10}$$

[42] I abstract here from intermediate inputs, such as materials or energy.

[43] An early paper with the same modeling approach is Bernstein and Nadiri (1988) where the authors sutdy inter-industry spillovers amongst five important technological industries allowing for each industry to be a potential source of spillovers to other industries.

where $\log(A_{it}) = \alpha_i + \delta_t$ and $w_{it} = (l_{it}, c_{it})$. Lower case letters denote the log of the capital letters in (10).

Estimation of returns to $R\&D$ in a production function framework is known to be challenging in several dimensions.[44] In what follows I outline each of these issues and the fixes I propose.

6.4.1 Measurement

I follow the extensive literature in measuring returns to $R\&D$, in particular BSV, in order to construct measures for output, labor, capital, and knowledge. When applicable, all variables are deflated by the CPI in 1994.

I measure output as real sales. The use of sales as output introduces measurement error since differences in revenue due to demand shocks cannot be disentangled from productivity shocks (e.g. see Foster *et al.*, 2008 for a detailed discussion on the problem).[45] In the absence of information on firm prices, and if firm-specific market power is persistent within industry, this issue can be addressed by including industry price index. Unfortunately, this approach is not successful if industry price index do not fully incorporate changes in prices due to differences in quality arising from $R\&D$, as Hall *et al.* (2011) document. The inclusion of time-industry dummies solves the problem. However, the time-industry dummies might also be capturing part of the spillover effects between firms. In particular, the common spillover effects at the industry level. I will take this fact into account when interpreting the spillover effects that I estimate.

The empirical counterpart of labour is number of employees and capital is measured as in book values.

Knowledge is proxied with a capital stock of $R\&D$ constructed using real $R\&D$ expenditures (e.g. Griliches, 1979). The conceptual framework underlying the use of this measure is that $R\&D$ creates a firm-level stock of knowledge that yields returns into the future. I use lagged stock of $R\&D$ as a measure of the knowledge of the firm today since it is unlikely that the latest addition to $R\&D$ stock becomes productive immediately. The distribution of returns to $R\&D$ over time can be motivated on the basis of the lag from expenditure to innovation.

The common approach to construct the $R\&D$ capital stock is to use a perpetual inventory method with depreciation rate 15%. The initial benchmark stock is measured assuming a a 5% growth in $R\&D$ expenditures.[46]

[44] For an exhaustive review of the challenges in this literature see Hall, Mairesse and Mohnen (2010).

[45] Hall *et. al* (2011) suggest modeling the demand side in order to be able to distinguish between the two. At this point this is out of the scope of this project

[46] See more details in the Appendix B of BSV.

6.4.2 Inputs as choices

Current choices of inputs are likely to be correlated with contemporaneous firm-specific productivity shocks. Given the panel dimension of the data I can use lags to instrument current levels of capital and labour.[47] However, our methodology so far does not cover instrumental variables. In order to preserve the regression framework, and given that our main interest is on the estimates of spillovers and the elasticity to $R\&D$, I substitute capital and labour by their respective lags in the production function (10). TFP is now proxied as the residuals of the reduced form regression:

$$y_{it} = \alpha_i + \beta_i rd_{it-1} + \sum_{j \neq i} \gamma_{ij} rd_{jt-1} + \widetilde{\theta}_c c_{it-1} + \widetilde{\theta}_l l_{it-1} + \delta_{I(i)t} + \widetilde{\epsilon}_{it}. \tag{11}$$

$\widetilde{\theta}_c$ and $\widetilde{\theta}_l$ are the coefficients of lagged log capital and lagged log labour, which can not be interpreted anymore as elasticities.

In a large T perspective, and if the structure of interactions is known, OLS estimates of (11) are consistent as long as: lagged expenditures in $R\&D$ are uncorrelated with current productivity shocks, and there is no time-series dependence on the shocks.[48] If errors are serially correlated further lags in time of capital and labour can be used to proxy TFP consistently. However, if there is correlation between past $R\&D$ investments and current productivity shocks instruments are needed. BSV makes use of tax credit combined with Federal rules in order to construct exogenous demand shifters of $R\&D$ at the firm level. At this point, since my framework does not allow for endogenous regressors, I plan to explore the extent of this issue using the reduced form, where I substitute the stock of $R\&D$ by the demand shifters.

Correlated Shocks Identification of spillover effects can be challenged by the presence of correlated shocks (Manski, 1993) even in the absence of endogenous effects.[49] I illustrate this problem with the following example provided in BSV: If firm i and j receive the same productivity shock they both enjoy an increase in output. If additionally firm j increases its $R\&D$ expenditure as a consequence of the shock, the output of firm i correlates with the increase in $R\&D$ stock of firm i. This increase generates a positive correlation between the output of firm i and the $R\&D$ stock of firm j. This correlation can then be wrongly interpreted as an spillover effect from firm j to firm i.

Nonetheless, given that I am using lagged expenditures in $R\&D$ as input in the production function, as long as lagged $R\&D$ expenditures do not correlate with current productivity shocks, spillover effects are correctly identified in (11). When lagged $R\&D$ expenditures correlate with current productivity shocks, as for instance when errors show dependence, instrumental variables are needed.

[47] This approach has been used in instrumental variables or GMM settings by e.g. Blundell and Bond (2000).

[48] See chapter 8 of Arellano (2003) for more details.

[49] In Manski's terminology, endogenous effects arise if there is strategic output decision of the firm in terms of the output of other firms. Given the production function framework this is unlikely once I allow for spillovers in $R\&D$.

6.5 Quantities of interest

Let

$$Y = \sum_{i=1}^{N} Y_i$$

be the aggregate output in the economy in an unspecified point in time (no t dependence). An informative quantity to assess the return to investment in $R\&D$ is the elasticity of aggregate output to a given firm's knowledge. For instance, the increase in output of the economy (in percentage terms) after an increase of 1 percentage points of firm k's $R\&D$ is:

$$\begin{aligned} \frac{\partial Y}{\partial K_k}\frac{K_k}{Y} &= \sum_{i=1}^{N} \frac{\partial Y_i}{\partial K_k}\frac{K_k}{Y} \\ &= \sum_{i=1}^{N} \gamma_{ik}\frac{Y_i}{Y} + \beta_k \frac{Y_k}{Y}. \end{aligned}$$

The elasticity has two parts. The first part takes into account the spillover effect that the knowledge of firm k produces on the rest of the firm's output. The second part reflects the increase in the output of firm k due to an increase in its own knowledge. This second part is the private return to $R\&D$, while the sum of both parts is the social return to $R\&D$.

Given that in our framework there is no interaction between firms in their decision to invest in $R\&D$, the elasticity of aggregate output to the knowledge of all firms in the economy is the sum of the elasticities of the whole output to the knowledge of each firm. I denote it as $\widetilde{M}$:

$$\begin{aligned} \widetilde{M} = \sum_{k=1}^{N} \frac{\partial Y}{\partial K_k}\frac{K_k}{Y} &= \sum_{k=1}^{N}\sum_{i=1}^{N} \frac{\partial Y_i}{\partial K_k}\frac{K_k}{Y} \\ &= \sum_{k=1}^{N}\sum_{i\neq k}^{N} \gamma_{ik}\frac{Y_i}{Y} + \sum_{k=1}^{N} \beta_k \frac{Y_k}{Y} \\ &= \sum_{i=1}^{N} \left(\sum_{k\neq i}^{N} \gamma_{ik} + \beta_i \right) \frac{Y_i}{Y}. \end{aligned}$$

6.6 Assessment of conditions for consistent model selection and Implications

The conditions for consistent model selection of the Pooled Lasso estimator in this empirical application are related to the amount of colinearity between $R\&D$ paths of firms. Figure 5 in the appendix shows the distribution of the absolute value of the pairwise correlation between paths of $R\&D$.[50] The distribution has substantial mass close to 1, which means that consistent model selection is potentially threaten.

[50] This distribution provides with informal evidence on the collinearity between $R\&D$ paths, and is not a formal test of any condition stated in section 5.

Moreover, the effect of own $R\&D$ can be contaminated by the spillover effect of firms in the sample whose $R\&D$ path is colinear to that of the firm. This implies that, in this application, the private effect and the spillover effects are not separately identified. However, the social effect, the sum of both the private and the spillover effect, is indeed identified.

I provide additional intuition on the challenges of identification of the model with colinearity on the regressors in Example 2 in Appendix B.

6.7 Results

In this section I present results of the baseline empirical model (10). I analyze differences in characteristics of firms according to their role in the structure of interactions: receivers of spillovers, sources of spillovers or none of the two. Then, I look at the estimated structure of interactions at the industry level. Finally I provide results on elasticities of output to $R\&D$ in order to give a sense of the social returns of $R\&D$ investment in this data and under these adopted assumptions. Further results are in progress.

Making sense of the structure of interactions: characteristics of receivers and sources of spillovers One of the features of the methodology is that it recovers the structure of interactions endogenously from the data. Moreover, given the fixed effects approach in estimation, I can relate the estimated structure of interactions with firms observables and unobservables. In what follows I relate the different types of the firms in the structure of interactions: sources of spillovers, receivers of spillovers, or none of the former, with firms observable characteristics and productivity levels.

In Table **??** I report average firms characteristics for the three different types of firms mentioned above. Firms sources of spillovers and receivers of spillovers are different on average according to several characteristics. First, sources of spillovers (3rd column) are on average small firms: they have low expenditures and stock on $R\&D$ and low averages of employment and capital. At the same time, they also generate smaller amounts of output. Second, firms sources of spillovers seem to be more productive than firms receivers of spillovers, where productivity is measured as the fixed effect.[51] This finding supports the idea that less productive firms adopt best practices in production from more productive firms. Finally, firms sources of spillovers have, on average, more citations in their patents than firms receiving spillovers, reflecting the idea that their innovations have a higher quality. Nonetheless, non-linked firms have even higher number of patent citation of both sources and receivers. In order to avoid mechanical effects arising from firm size Table **??** in the Appendix presents results of multinomial logit regression where the outcome is being of one of the three firm categories: receiver, non-linked and source. In Columns (5) it can be seen that both type of firms, receiving spillovers or

[51] In order to make this statement I am relying on the fact that each category has sufficient number of firms so that the average of fixed effects is sufficiently well estimated.

	Mean		
	not linked	receivers	sources
real sales	3,120.23	3,741.94	696.85
capital	1,249.63	1,159.66	216.11
labor	17.88	19.64	6.30
expenditure in *R&D*	147.57	211.14	32.41
stock of *R&D*	783.22	1151.16	210.89
stock *R&D* / stock capital	0.58	0.74	0.62
patents	21.95	11.89	10.56
patent cite	172.15	81.98	89.64
productivity	5.96	3.30	5.97
firms %	71%	19%	11%

non-linked, produce less patents than the sources of spillovers conditional on size. That is, they are less patent intesnive. Similarly, in column (6) it can be seen that patent citation is also highest among sources of spillover firms conditional on firm size. These results seem to be consistent with theoretical models where the decision of *R&D* investment by small and large firms differ due to their different purposes: smaller or younger might invest in exploitation related investments in order to develop new products versus larger firms motive for *R&D* investment might be more related to exploration in order to improve their mark-ups on their existing products (See Akcigi and Kerr, 2010). At the same time, these results also show that smaller firms, as sources of spillovers, are largely the most relevant firms generating spillovers.

Structure of interactions at the indsutry level I now present results on the estimated structure of interactions. I present results at the industry SIC2 level for two reasons: First, it allows to better interpret the results. Second, given that the conditions for identification of the structure of interactions are not fulfilled in this data, looking at an aggregate of the structure of interactions might be more robust. I define an induced estimated structure of interactions at the industry level using the estimated structure of interactions at the firm level. I consider there is an spillover effect from industry a to industry b if there is at least one firm in industry a that generates spillovers on a firm in industry b. The results in terms of an adjacency matrix are shown in the upper panel of Figure 3.

Spillovers across industries are concentrated in few industries. The two industries generating most spillovers are industries 35 and 36: *Industrial Machinery and Computers* and *Electronic and other Electrical Equipment except Computers.* These industries generate spillovers to industries such as *Transportation Equipments* (37) or *Measuring and Analyzing Instruments* (38). Consistent with this result industries (35) and (36) are among the highest industries in patent citations, as well as

Figure 3: Industry-level structure of interactions

Note: The upper panel shows the estimated structure of interactions at the industry level. The bottom panel shows the average across Monte Carlo simulations of the same matrix. The design of the Monte Carlo exercise is tailored to the data. In particular, I fix the true parameters of the DGP as the estimated ones. Then, I generate data conditioning on the true data and adding normal shocks with variance equal to the empirical individual variance.

expenditure on $R\&D$ (see Table 7 in the Appendix for descriptive statistics of the firms at the industry level). On the other hand, industry (36) is the industry receiving most spillovers from other industries. In particular it is influenced by the above mentioned industries, (37) and (38), but also industries (35) and (34), that includes companies in the business of Electronic computers, computer storage devices and others.

Another interesting source of spillovers arises from industry (73). This industry is named *Business Services* and as opposed to industries (35) and (36) is not particularly represented in the sample (the distribution of firms by industry in the sample can be seen in Table 6). Importantly, industry (73) comprises firms in the business of Computer Programming Services, Prepackaged software and others. This industry generates spillovers on industries such as (36) and (37) and receives influence of these same industries. Interestingly, industry (73) also generates spillovers on industry (28), *Chemical and Allied Products*. On the other hand the contrary is not the case. It seems likely that software development might be important in order to produce chemical products, at the o

I capture asymmetries in spillovers between industries. For instance, while indsutry (28), *Chemicals and Allied Products*, receives spillovers from industries (34), (35), (36), and (37), it only seems to influence industry (36) and (37), but not the others. The chemical industry and the electronic industry are connected due to innovation in semiconductors.[52]

The lower panel of Figure 3 shows an average across Monte Carlo simulations of the same estimated structure of interactions at the industry level. The design on this Monte Carlo is tailored to the empirical application (see more details in the following section). According to this results the estimated structure of interactions at the industry level seems to be reasonably well estimated.

Elasticities I compute the elasticity of aggregate output to the knowledge of all firms in the economy, $\widetilde{M}$. This elasticity is a measure of the social rate of return of $R\&D$ investment in the economy. Table 1 reports the results of the elasticities in the first column, while the second column shows the standard

[52] "It is not an exaggeration to view semiconductor manufacturing facilities as large chemical factories. Chemistry has played a role in the scaling of silicon circuits for the past 50 years" Extracted from *The Chemistry Innovation Process: Breakthroughs for Electronics and Photonics.*

deviation of the same quantities across monte carlo simulations.

The point estimate for the overall rate of return in the economy is 3.15%, which is low in relation to the values obtained in the literature, even after considering the uncertainty in estimation of this quantity, which is substantial. A low rate of return to $R\&D$ is consistent with the empirical fact that differences in productivity across firms are persistent over time. If spillovers would be large I would observe a convergence of the productivity levels of firms over time (Syverson, 2011). On the other hand, this result masks considerable heterogeneity in the social returns of $R\&D$. Computing the same elasticity by size quartiles (in terms of output) I find that smaller firms obtain significantly higher social returns of $R\&D$ investment than bigger firms.

In this data, smaller firms seem to be both the instigators of $R\&D$ spillovers, and the benefactors of $R\&D$ investment in the economy. According to the results on the previous section, smaller firms are more likely to generate spillovers, and at the same time, smaller firms obtain the highest rates of return of additional investments in $R\&D$. However, my results cannot contribute to the current debate of whether small firms should be supported by the government in their $R\&D$ enterprises or not. In particular, there are not enough "small" firms in my dataset (only 10%), according to the eligibility of some of these programs, to assess these type of policies. [53] Second, the estimates of spillovers in this model are "in deviations" to an aggregate industry spillover effect. It could be that bigger firms benefit more from aggregate spillovers while smaller ones do so from particular innovations of other firms.

Table 1: Social returns

	Elasticity	Monte Carlo Std
1st quartile	24.93%	5.91
2nd quartile	12.85 %	4.90
3rd quartile	8.70%	5.82
4th quartile	1.53%	12.94
All	3.15%	10.92

Note: The first column shows the elasticity of aggregate output to R&D investment. That is, the sum of the elasticities of aggregate output to each firm's R&D in the economy. When broken down by quartiles the elasticity is the sum of elasticities of output of firms within the quartile to each firm's R&D in the economy. The second column shows the Monte Carlo standard deviation of the same quantity on a Monte Carlo experiment tailored to the empirical application. See the details of the design in the next section.

[53] According to the Small Business Innovation Research program (SBIR), small firms are those with less than 500 employees. This program had a budget of 1 Bn $ in the year 2010.

6.8 Comparison with the Jaffe parametrization of technological spillovers

I conclude this section of results by comparing the obtained measure of spillovers with the measure of technological spillovers introduced by Jaffe (1986). In particular, I use the measure of spillovers constructed in BSV (2013). I start by providing some background on the Jaffe parametrization of technological spillovers.

6.8.1 Pool of Knowledge

The conceptual framework for technology spillovers was first introduced by Griliches in 1979. In his seminal work, technological spillovers arise through a common pool of technological knowledge. Firms make use of the pool of knowledge as an additional input in their production functions.

The technological pool of knowledge evolves over time: it grows when individual firms produce technological advances, and it depreciates as knowledge becomes obsolete. Technological advances affect productivity through two different channels: directly, on the productivity of the firm originating the innovation, and indirectly, on the rest of the firms in the economy through the pool of knowledge. In this way, innovations arising in one particular firm can affect the productivity of another firm.

Jaffe (1986) introduced heterogeneity in the pools of knowledge across firms to take into account that different types of innovations might have different effects on the productivity of firms. In particular, Jaffe introduced a technological distance between firms in order to capture differences in the likelihood that knowledge spillovers would arise between two firms. According to the Jaffe measure, when two firms patent their innovations on the same technological disciplines, they are technologically close, and hence more likely to experience technological spillovers one from each other. On the other hand, when two firms patent in different technological categories they are technologically far, and hence not likely to experience spillovers from each other.

Interestingly, Jaffe, in his original paper in 1986, acknowledges that this measure of spillover only provides "indirect inference [of the spillover phenomenon] which is made necessary by the extreme difficulty of observing the actual spillovers".

6.8.2 Variance decomposition

In this subsection I compare the explanatory power of two different measures of spillovers, one based on the Jaffe measure, and another one obtained using the methodology. I denote S^J the measure based on Jaffe, and S^E the estimated one. The two measures are functions of other firms $R\&D$ in the economy:

$$S_{it}^J = \sum_{j \neq i} \omega_{ij} RD_{jt}$$

$$S_{it}^E = \prod_{j \neq i} \widehat{\gamma}_{ij} RD_{jt}$$

where ω_{ij} denotes the Jaffe distance between firm i and firm j, and $\widehat{\gamma}_{ij}$ are the spillover effects obtained in the estimation of (11).

Variance decomposition pooled regression The objective of this exercise is to compare the explained variation in the data of each measure of spillovers. I first explore the variation explained in a pooled sense. To do so, I compare the variance of the residuals of the following model of spillovers:

$$y_{it} = \alpha_i + \beta rd_{it-1} + \gamma s_{it-1} + \theta' w_{it} + \delta_{I(i)t} + \epsilon_{it},$$

where s is either s^J or s^E, and as before lower case letters denote the log of capital letters, with the variance of a model with no spillovers:

$$y_{it} = \alpha_i + \beta rd_{it-1} + \theta' w_{it} + \delta_{I(i)t} + \epsilon_{it}.$$

The results, in table 2, show that the variance of the residuals when using S^E as a measure of spillovers is lowest. However, the decrease in variance of the residuals with respect to the the model with no spillovers is mild (16%), but substantially larger than the decrease in the variance using S^J (0.4%).

Table 2: Variance decomposition - Pooled Regression

All	None	S^J	S^E
Variance Residuals	0.0359	0.0357	0.0301
Decrease Variance (%)		0.47%	15.99%

Variance decomposition time series I now turn to explore the variation explained of each measure of spillovers in the time series. To do so, I work with the residuals of the following regression:

$$y_{it} = \alpha_i + \beta_i rd_{it-1} + \theta' w_{it} + \delta_{I(i)t} + \epsilon_{it}.$$

I allow now for individual-specific elasticities to $R\&D$ to capture more variability in the time series.

The results for the overall sample, shown in table 3, reveal that the estimated measure does a better job also at explaining the time-series variation of productivity. This result is not evident ex ante since there is a substantial part of firms that do not enjoy spillovers according to the estimated spillover measure. For all those firms, the additional variation explained by the estimated measure is 0. On the other hand, the additional variability explained by the Jaffe measure is always positive.

The average additional variability explained in the time series by the estimated measure of spillovers is close to 23%, while the average variability explained by the Jaffe measure of spillovers is 5%. When I restrict the analysis to those firms that enjoy spillovers according to the estimated measure, the additional variability explained by the estimated measure goes up to 71.5%, on average.

Table 3: Variance decomposition - Time series (I)

All	None	S^J	S^E
Average Variance Residuals	0.0234	0.0222	0.0181
Decrease Variance (%)		5.34%	22.74%

Table 4: Variance decomposition - Time series (II)

Receivers	None	S^J	S^E
Average Variance Residuals	0.0364	0.0337	0.0104
Decrease Variance (%)		7.44%	71.49%

7 Finite sample properties

7.1 Simple DGP

I illustrate the result on the rate of convergence of average marginal effects stated in Proposition 1 with the following small Monte Carlo experiment.

In consider a sequence of DGPs as in (2) where $N = T^2$ and $T = 3, \ldots, 12$. I consider $s_i = 1$ for all i and for all sequences of DGPs. Moreover, x_{it} is i.i.d. $\mathcal{N}(0{,}1)$ in both dimensions and ϵ_{it} is i.i.d. also in both dimensions $\mathcal{N}(0{,}0.1)$.

The thick solid line in Figure 4 shows the average difference between the estimated average marginal effect and the true average marginal effect for the first individual in the sample across DGPs. Each average is taken over 100 Monte Carlo replications. The dashed line is $\frac{\log(N)}{NT}$ and the thin line is $\frac{1}{NT}$, which approach the difference in average marginal effect as N and T grow.

Figure 4: Rate of Convergence Average Marginal Effect

Notes: The thick solid line plots the average of $\left|\frac{1}{N-1}\sum_{i\neq j}\widehat{\gamma}_{ij}, -\frac{1}{N-1}\sum_{i\neq j}\gamma^0_{ij}\right|$ *over 100 Monte Carlo replications, the dashed line is* $\frac{\log(N)}{NT}$ *while the thin solid line is* $\frac{1}{NT}$.

7.2 Monte Carlo tailored to the empirical application

I generate data according to the following DGP:

$$y_{it} = \alpha_i^0 + \alpha_k^0 k_{it} + \alpha_l^0 l_{it} + \alpha_{RD_i}^0 rd_{it-1} + \sum_{j\neq i}\gamma_{ij}^0 rd_{jt-1} + \delta^0_{I(i)t} + \epsilon_{it}.$$

and I set $N = 200$ and $T = 16$ and I consider ϵ_{it} i.i.d shocks $\mathcal{N}(0, \sigma_i^0)$ distributed. I set the parameters ^{0} equal to the obtained estimates using the original data. Finally regressors k_{it}, l_{it} and rd_{it-1} are kept fixed. I perform $S = 130$ replications.

8 Conclusions

In this paper I present a methodology to estimate both the structure of interactions and the spillover effects when the structure of interactions is not observable to the econometrician. This method is useful when the structure of interactions is sparse and persistent over time. Both of these assumptions can be partially relaxed: Sparsity can be relaxed by adding a priori information on the structure of interactions. Persistence over time can be relaxed by splitting the sample, parametrizing the spillover effects as a function of time, or by augmenting the number of regressors as explained in Appendix A.

Spillovers arise when characteristics have an impact on the outcome of other individuals in the sample. This model is useful in at least two cases: First, in the context of randomized treatment experiments, when the treatment is subject to generate externalities. Second, in production function frameworks, where productivity generates spillovers.

I propose a new estimator, the Pooled Lasso estimator, that can be seen as a panel data counterpart of the Lasso estimator. I provide an iterative computation method that combines the Lasso estimator with OLS pooled regression. Computation is fast, in relation to the large number of potential structures of interactions, given the global convex nature of the criterion.

I analyze the properties of the Pooled Lasso estimator in a simplified model with no common parameters under assumptions of Gaussian and independent errors, both in the time and cross-sectional dimension. Based on a recent paper by Lam and Souza (2013), these strong conditions on the errors are likely to be relaxed. First, gaussianity can be replaced by conditions on the tail probability of errors. Second, limited time-series dependence can also be incorporated using Nagaev-type inequalities. Finally, mild cross-sectional dependence in the errors is also likely to be incorporated.

I study the rate of convergence of cross-sectional spillover effects and, more generally, aggregate spillover effects. These quantities can be interpreted as relevant policy parameters depending on the application. Under conditions of cross-sectional independence on the error in estimation of the spillover effects, average spillover effects are estimated at a much better rate than individual spillover effects.

I use the methodology to study technological spillovers in productivity in a panel of US firms. In my main specification I assume that $R\&D$ is predetermined and I include time-industry effects to account for correlated productivity shocks. These industry-specific shocks might be capturing aggregate spillover effects at the industry level, and as a consequence my estimated spillovers have to be interpreted in deviation to these aggregate spillover effects. I find evidence of asymmetry in spillover effects. In particular, I find that less productive firms receive spillovers from more productive firms, consistent with the idea that less productive firms try to enhance their productivity by acquiring the knowledge of technological leaders (Syverson, 2011). Moreover, firms receiving spillovers seem to be less successful at innovating since their patents, on average, seem to be less cited. Finally, my estimated structure of interactions seems to do a better job at explaining the variation of the data than other measures of spillovers.

Inference methods in the context of the Lasso are hard to derive due to the non-differentiability of the criterion. However, recent works in econometrics and statistics show good progress in this direction. For instance, Belloni *et al.*. (2013), show how to conduct inference in a post-lasso setting. That is, after using Lasso as a model selection devise, an ex-post OLS regression is run conditional on the estimated model. In particular, after using a Lasso-type estimator twice to select relevant controls in a treatment effect framework, they derive the asymptotic distribution of the treatment effect estimator and provide a formula for confidence intervals. One of the main features of their work is that, even in spite of imperfect model selection, their results hold uniformly for a large class of DGPs. Another recent work, by Lockhart *et al.*, (2013) focus in developing a significance test of the predictor variable that enters the current lasso model, in the sequence of models visited along the lasso solution path of *LARS* (e.g. Efron *et al.*, 2004). The Lasso solution path are the different solutions that the Lasso delivers when the penalty parameter decreases.

I am currently working on an extension of the model that is particularly relevant in the study of social and economic networks. In this extension the outcome of individuals directly depends on the outcome of other individuals in the sample. In this case, technical difficulties arise due to the simultaneous determination of the outcome of (some) individuals in the sample. This problem was first described by Manski in his seminal work in 1993 as the "Reflection Problem". I am exploring recent advances on the properties of the Lasso in instrumental variable settings (e.g. Ying Zhu, 2013 or Gautier and Tsyvakov, 2013) to overcome this problem.

An interesting avenue for future research is exploring other types of penalization. For instance, a combinatorial-type estimator where the number of sources of spillovers are restricted, as outlined in the last subsection of section 5, is a natural alternative to the Pooled Lasso estimator. Moreover, in this combinatorial setting imposing cross-section restrictions on the spillover effects could be easier than in the Lasso setting. Combinatorial estimators are likely to have similar properties as the Pooled Lasso estimator. In particular, in the case where the number of sources of spillovers is fixed, the time dimension grows, and under separation conditions on the regressors, the Lasso and the combinatorial estimator attain a rate of convergence equal to the rate of convergence of an infeasible OLS estimator where the structure of interactions is known (the so-called oracle property).[54] However, computation of combinatorial estimators is in general much more intensive than the Lasso estimator, due to the non-convexity of the penalization.

Finally, a version of the Lasso, the Elastic Net (Zou and Hastie, 2005), introduces a penalty convex combination of the sum of absolute values of the parameters and the sum of the squared values of the parameters. This estimator has two interesting properties: First, it delivers sparsity, as the Lasso, but

[54] This result connects with other results in the literature on estimation of discrete heterogeneity (e.g. Bonhomme and Manresa (2013) and Hahn and Moon (2010)).

the constraint on the number of possible regressors is milder than in the Lasso. In particular, it can accommodate more sources of spillovers than periods of observations. Second, it is more robust to colinearity than the Lasso. In particular, in the presence of a group of highly colinear regressors, the Elastic Net selects the whole group. Contrarily, the Lasso tends to randomly select one among all of them. Given the substantial colinearity between paths in $R\&D$ in this data, a panel data counterpart of the Elastic Net estimator might seem a suitable alternative to the Pooled Lasso estimator. However, at this point, the economic interpretation of this alternative estimator is still unclear to me.

References

[1] Akcigit, U., W. Kerr (2010): "Growth through heterogeneous innovations" NBER Working Paper 16443

[2] Ambrus, A., M. Mobius and A. Szeidl (2010): "Consumption Risk-sharing in Social Networks," forthcoming in *The American Economic Review.*

[3] Arellano, M. (2003): "Panel Data Econometrics," Oxford University Press.

[4] Arellano, M. and S. Bonhomme (2012): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *The Review of Economic Studies*, 79.

[5] Arcidiacono, P., G. Foster, N. Goodpaster, and J. Kinsler (2012): "Estimating Spillovers using Panel Data, with an Application to the Classroom," *Quantitative Economics*, 3 (3), 421-470.

[6] Bai, J. (2009), "Panel Data Models with Interactive Fixed Effects," Econometrica, 77, 12291279.

[7] Ballester, C., A. Calvó-Armengol, and Y. Zenou (2006): "Who's Who in Networks. Wanted: The Key Player," *Econometrica*, 74(5), 1403-1417.

[8] Banerjee, A., A. Chandrasekhar, E. Dufflo, and M.O. Jackson (2012): "The diffusion of micro-nance," MIT working paper.

[9] Belloni, A., V. Chernozhukov, and C. Hansen (2011): "Inference Methods for High-Dimensional Sparse Econometric Models," *Advances in Economics Econometrics.*

[10] Belloni, A., V. Chernozhukov, and C. Hansen (2013): "Inference on treatment effects after selection amongst high-dimensional controls," forthcoming in *The Review of Economic Studies.*

[11] Bernstein, J. and M. Nadiri (1988): "Interindustry R&D Spillovers, Rates of Return, and Production in High-Tech Industries," *The American Economic Review: Papers and Proceedings*, LXXVIII, 429-34

[12] Bickel, P. J., Y. Ritov, and A. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics* 37 (4), 1705-1732.

[13] Bloom, N., M. Schankerman, J. Van Reenen (2013): "Identifying Technology Spillovers and Product Market Rivalry," forthcoming in *Econometrica*

[14] Blume, L. E., W. A. Brock, S. Durlauf, and Y. M. Ioannides (2010): "Identification of Social Interactions," *Economics Series 260, Institute for Advanced Studies.*

[15] Blundell, R. and S. R. Bond (2000): "GMM Estimation with Persistent Panel Data: An Application to Production Functions," *Econometric Reviews*, 19 (3), 32140.

[16] Bonhomme, S. and E. Manresa (2012): "Grouped Patterns of Heterogeneity in Panel Data", unpublished manuscript.

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

[18] Buhlman P., S. van der Geer (2011): "Statistics for High-Dimensional Data. Methods, Theory and Applications," Springer.

[19] Card, D. L. Giuliano (2012): "Peer Effects and Multiple Equilibria in the Risky Behavior of Friends," unpublished manuscript.

[20] Chamberlain, G.(1992). "Efciency Bounds for Semiparametric Regression," *Econometrica*, 60 (3), 567-596.

[21] Chandrasekhar, A. and M.O. Jackson (2013): "Tractable and consistent random graph models," unpublished manuscript.

[22] Chernozhukov, V. (2009): " High-Dimensional Sparse Econometric Models," (Lecture notes) Stats in the Chateau, https://studies2.hec.fr/jahia/Jahia/statsinthechateau.

[23] Coe, D. T. and E. Helpman (1995): "International RD spillovers," *European Economic Review* 39 (5), 859887.

[24] Conley, Tim and Udry, Christopher (2010): " Learning about a new technology: pineapple in Ghana," *The American Economic Review* 100 (1), 35-69.

[25] De Girogi, G., A. Frederiksen, L. Pistaferri (2013): "Consumption Network Effects," Mimeo

[26] De Girogi, G., M. Angelucci (2009): "Indirect Effects of an Aid Program: How do Cash Transfers Affect Non-Eligibles' Consumption?," *The American Economic Review*, 99 (1).

[27] De Girogi, G., M. Pellizzari (2013): "Understanding Social Interactions: Evidence from the Classroom", forthcoming at the *Economic Journal*

[28] Doraszelski, U. and J. Jaumandreu (2013): "RD and productivity: Estimating endogenous productivity," forthcoming *The Review of Economic Studies*

[29] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004): "Least angle regression," *Annals of Statistics*, 32, 407499.

[30] Gautier, E. and A. Tsybakov (2012): "High-dimensional instrumental variables regression and confidence sets", unpublished manuscript.

[31] Graham, B. (2008): "Identifying social interactions through conditional variance restrictions," *Econometrica* 76 (3), 643 - 660.

[32] Graham, B., G. Imbens, and G. Ridder (2013): "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach", forthcoming in *Quantitative Economics*.

[33] Graham, B., J. L. Powell (2012): "Identification and Estimation of Average Partial Effects in 'Irregular' Correlated Random Coefficient Panel Data Models," *Econometrica*.

[34] Griliches, Z. (1979): "Issues in assessing the contribution of RD to productivity growth," *Bell Journal of Economics*, 10 (1), 92116.

[35] Hahn, J. and R. Moon (2010): "Panel Data Model with Finite Number of Multiple Equilibria", *Econometric Theory*, 26, 863-881.

[36] Hall, B., J. Mairesse, and P. Mohnen (2010): "Measuring the returns to RD", in B. Hall N. Rosenberg, eds, Handbook of the Economics of Innovation, North-Holland, Amsterdam.

[37] Hall, B., A. B. Jaffe, and M. Trajtenberg (2001): "The nber patent citation data le: Lessons, insights and methodological tools," NBER Working Papers 8498, National Bureau of Economic Research, Inc.

[38] Harris, K.M., C.T. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J.R. Udry. 2009. The National Longitudinal Study of Adolescent Health: Research Design [WWW document]. URL: http://www.cpc.unc.edu/projects/addhealth/design.

[39] Jackson, M.O. (2008): "Social and Economic Networks", Princeton, NJ: Princeton Univ.

[40] Jaffe, A. (1986): "Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value," *The American Economic Review*, 76 (5), 984-1001.

[41] Lam, C., P. CL Souza (2013): "Regularization for Spatial Panel Time Series Using Adaptive Lasso," Mimeo

[42] Lee, L., X. Liu, E. Patachini and Y. Zenou (2012): "Crime Networks: Who is the Key Player?," CEPR Discussion paper n. 8772.

[43] Lockhart, R., J. Taylor, R. J. Tibshirani, R. Tibshirani (2013): "A signicance test for the lasso", arXiv 1301.7161

[44] Manski, C. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, Wiley Blackwell, 60 (3), 531-42.

[45] Meinshausen N. and B. Yu. (2009): "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, 37 (1), 22462270.

[46] Miguel, E. and M. Kremer, (2004): "Worms: Identifying impacts on education and health in the presence of treatment externalities". *Econometrica*, 72 (1), 159-218.

[47] McPherson, M., L. Smith-Lovin, and J. Cook, (2001): "Birds of a Feather: Homophily in Social Networks", *Annual Review of Sociology*, 27, 415-44.

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

[49] Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of Royal Statistical Society Series B*, 58, 267288.

[50] Townsend, R. (1994): "Risk and Insurance in Village India," *Econometrica*, 62 (3), 539-591

[51] Syverson, Chad (2011): "What determines productivity at the micro level?", *Journal of Economic Literature.*

[52] Watts, A. (2001): "A Dynamic Model of Network Formation," *Games and Economic Behavior*, 34, 331341.

[53] Watts, D.J. (1999): " Small Worlds: The Dynamics of Networks between Order and Randomness", Princeton University Press.

[54] Zhu, Y. (2013): "Sparse Linear Models and Two-Stage Estimation in High-Dimensional Settings with Possible Many Endogenous Regressors", Mimeo

[55] Zou, H. and T. Hastie (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of Royal Statistical Society Series B*, 67 (2), 301-320.

APPENDIX

A Extensions

In this appendix I show some extensions of the baseline model (1).

A.0.1 Spillovers occur due to more than one characteristics

So far in model (1) I have assumed that the characteristic susceptible of generating spillovers is one-dimensional. The model is well defined too when I consider $\mathbf{x}_{it}$ a vector of q characteristics susceptible of generating spillovers. A particularly interesting example is one in which $q = 2$ and $\mathbf{x}_{it} = (x_{it}, y_{it-1})$. That is, the outcome of individuals can be affected by both characteristics and lagged outcomes of other individuals in the sample:

$$y_{it} = \alpha_i + \beta_i x_{it} + \mu_i y_{it-1} + \sum_{j \in G_i} \gamma_{ij} x_{jt} + \sum_{j \in S_i} \eta_{ij} y_{jt-1} + \theta' w_{it} + \delta_t + \epsilon_{it}. \qquad (12)$$

The sparsity assumption in this model can be put in two different ways. The first way to impose sparsity is analogous to the sparsity assumption imposed on the baseline model: I restrict the number of different units that can have an effect through both x_{jt} and y_{jt-1}, where $j \neq i$. The second way to impose sparsity is directly on the number of characteristics that can affect the outcome, either x_{jt} or y_{jt-1}. Under this second sparsity framework different units can generate spillovers on other units through different characteristics. When $q = 1$ there is no distinction among these two sparsity conditions.

A.0.2 Relaxing time-invariant network structure and effects

I can extend the model to account for some dynamics in the structure of interactions as well as in the intensity of their spillovers. Let us consider a certain point in time, specific to each unit, in which there is a change in the structure of interactions. I denote it by T_i^* where $1 < T_i^* < T$. I define the following regressors after the original regressors $\{x_{jt}\}_{j \neq i}$:

$$x_{jt}^1 = \mathbb{1}\{t \leq T_i^*\} \cdot x_{jt} \quad , \quad x_{jt}^2 = \mathbb{1}\{t > T_i^*\} \cdot x_{jt}.$$

That is, the x_{jt}^1 new regressors contain, in the first T_i^* positions, the values of x_{jt} while in the bottom $T - T_i^*$ positions contains 0. Conversely, the x_{jt}^2 new regressors contain 0's on the T_i^* first positions while in the $T - T_i^*$ positions are equal to x_{jt}. I now consider the augmented model:

$$y_{it} = \alpha_i + \theta w_{it} + \beta_i x_{it} + \sum_{j \neq i} \gamma_{ij}^1 x_{jt}^1 + \sum_{j \neq i} \gamma_{ij}^2 x_{jt}^2 + \delta_t + \epsilon_{it}. \qquad (13)$$

where now the sparsity assumption comprises the vector $\left(\gamma_{ij}^1, \gamma_{ij}^2\right)$ but I allow for twice as much regressors to be different from 0.

This model is susceptible to capture a change in the influence (intensive margin) of one particular unit j on i if $\gamma_{ij}^1 \neq \gamma_{ij}^2$ but $\gamma_{ij}^1 \cdot \gamma_{ij}^2 \neq 0$. That is, there is no change in the extensive margin because the selected regressor is essentially x_{jt} but there is a change in the intensive margin when $\gamma_{ij}^1 \neq \gamma_{ij}^2$.

This model, is also susceptible to capture a change in the structure of interactions if the influencing unit changes from j_1 to j_2. This could be captured as follows: I could have $\gamma_{ij_1}^1 \neq 0$ but $\gamma_{ij_1}^2 = 0$ and $\gamma_{ij_2}^1 = 0$ but $\gamma_{ij_2}^2 \neq 0$ which would essentially mean that during the period from 1 to T_i^* the influencing unit is j_1 while from the periods $T_i^* + 1$ to T the influencing unit is j_2. This model could also capture any intermediate situation, such as for instance that unit j_1 is influencing for the whole period of observation but the unit j_2 is only influencing during the last period.

The sparsity structure in the interactions could have also been imposed by groups, restricting that if the $x_{j_1 t}$ is influencing during the first half, it has to be influencing during the second half too, although not necessarily with the same intensity. This option is convenient if influences do not disappear but rather attenuate or increase, since under this alternative specification it is unlikely that exact zero are obtained as estimates in case the influence disappears.

Finally, it is left to discuss how to choose T_i^*.

Remark 2 *A similar insight can be used in order to account for unbalanced panels*

B Examples of Failure of Identification

Example 1 ***Lack of time series variation in characteristics*** x_{it}

The model has lots of individual and pair-specific parameters the identification of which entirely relies on the time-series variation of the characteristics x_{it}*. In particular, and in the context of the empirical application, let us think that firm 1 R&D is constant over time. If that is the case, if* $x_{1t} = c_1$*, where* c_1 *is constant, then not only* α_1 *and* β_1 *are not identified, but also the spillover effect of firm 1 (*γ_{i1} *for* $i = 1, \ldots, N$*), and the* α_i *for* $i = 2, \ldots, N$ *are not identified either. This is due to the fact that the R&D of firm 1,* x_{1t}*, could (potentially) explain the output of all other firms in the sample.*

Example 2 ***Lack of independent time variation***

Another important source of lack of identification for the structure of interactions γ_{ij}*'s, and for the elasticities* β_i*'s, arises when the characteristics of several units are colinear. If this is the case I might not be able to distinguish in the data which is the influencing unit among those whose characteristics are colinear. This is intuitive when 2 firms have the same exact path of R&D. Moreover, I might not be able to identify the direct effect versus the spillover effect. To provide some intuition, let us consider the following simplified model where there are only three firm:*

$$\begin{aligned} y_{1t} &= \alpha_1 + \beta_1 x_{1t} + \gamma_{12} x_{2t} + \gamma_{13} x_{3t} + \epsilon_{1t} \\ y_{2t} &= \alpha_2 + \beta_2 x_{2t} + \gamma_{21} x_{1t} + \gamma_{23} x_{3t} + \epsilon_{2t} \\ y_{3t} &= \alpha_3 + \beta_3 x_{3t} + \gamma_{31} x_{1t} + \gamma_{32} x_{2t} + \epsilon_{3t}. \end{aligned}$$

Let us assume that : $x_{2t} = \lambda + \mu x_{1t}$*, where* λ *and* μ *are two arbitrary numbers and* x_{3t} *is independent of both* x_{1t} *and* x_{2t}*. That is, the R&D time evolution of firm 1 is perfectly colinear with the R&D time evolution of firm 2, but the third one is independent of the former two.*

Finally, let us assume that the true DGP is the following:

$$\begin{aligned} y_{1t} &= \alpha_1 + \beta_1^0 x_{1t} + \epsilon_{1t} \\ y_{2t} &= \alpha_2 + \beta_2^0 x_{2t} + \epsilon_{2t} \\ y_{3t} &= \alpha_3^0 + \beta_3^0 x_{3t} + \gamma_{31}^0 x_{1t} + \epsilon_{3t}. \end{aligned}$$

According to this DGP, the only firm receiving spillovers from another firm is firm 3, who receives spillovers from firm 1. However, since the R&D paths of firm 1 and firm 2 are colinear, the structure of interactions is not identified. Indeed, it could be that firm 1 is generating spillovers on firm 3, with influence γ_{31}^0*, but it could also be that firm 2 is generating spillovers on firm 3 with an effect of* $\mu^{-1}\gamma_{31}^0$*.*^{0}*. Now, let us focus on firm 1 in order to illustrate a different identification failure. Again, given that* x_{1t} *and* x_{2t} *are colinear, from the data I cannot distinguish whether it is the own R&D that* *is having an effect on output, with effect equal to* β_1^0, *or if firm 1 is experiencing an spillover effect from firm 2 with an effect of* $\beta_1^0 \cdot \mu^{-1}$.

C Proofs

Proof.

We start by introducing the following regularity conditions to use feasible weights ϕ_{ij}.

Assumption 4 *Feasible weights*

a. *For any set of weights $\phi_{ij} = \widehat{\phi}_{ij}$, and denoting ${\phi^0_{ij}}^2 = \mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T \tilde{u}_{it}\tilde{x}_{jt}\right]$ we have that:*

$$\ell\phi^0_{ij} \leq \widehat{\phi}_{ij} \leq u\phi^0_{ij}.$$

with probability $1-o(1)$ uniformly on i and j with $0 < \ell \leq 1 \leq u$, such that $\ell \to 1$ and $u \to u_0$ with $u_0 \geq 1$ in probability.

b. *For all $i \in \{1, \dots, N\}$ there exists $u' > 1$, $\ell' > 0$ and $\eta > 0$ independent of i, j such that:*

$$0 < \eta < \ell' {\phi^0_{ij}}^2 < V_{ij} < u' {\phi^0_{ij}}^2$$

with $V_{ij} = V_{u_i x_j}$ and $V_{ij} = V_{u_i|x_j}$, where $V_{u_i x_j} = \sup_t \left(\mathbb{E}\left[u^2_{it}x^2_{jt}\right] + 2\sum_{s>t} |\mathbb{E}\left[u_{it}x_{jt}u_{is}x_{js}\right]|\right)$ and $V_{u_i|x_j} = \sup_t \left(\mathbb{E}\left[u^2_{it} \mid \{x_{jt}\}_{t=1}^T\right] + 2\sum_{s>t} |\mathbb{E}\left[u_{it}u_{is} \mid \{x_{jt}\}_{t=1}^T\right]|\right)$.

c. *If $a = \min_j \phi^0_{ij}$ and $b = \max_j \phi^0_{ij}$ we have that $0 < a \leq b < +\infty$ uniformly on i.*

Assumptions 4.a - 4.c collect high-level assumptions on properties that feasible weights $\widehat{\phi}_{ij}$ and ideal weights ϕ^0_{ij} need to satisfy so that the noise in estimation is uniformly bounded by the penalty level. In particular 4.a is analogous to the definition of asymptotic feasible weights from Belloni Chen Chernozhukov Hansen (2012). Assumption 4.b is partly implied by assumption 2.d as under the strongly mixing conditions V_{ij} is finite for all i, j.[55] Condition 4.c is a high level assumptions that ensures that variance of noise in estimation are bounded from above and below uniformly in i and j.

We will construct this proof in 5 steps.

Step 1 We start by defining the score, which determines the rate of convergence of the estimator:

$$S_{ij} = 2{\phi^0_{ij}}^{-1}\frac{1}{T}\sum_{t=1}^T \tilde{u}_{it}\tilde{x}_{jt}.$$

The restricted eigenvalues are useful in order to relate the prediction norm, $\delta' M \delta$ and the penalty $\|\delta_I\|_1$:

$$\kappa^2_C(M) = \min_{\delta \in D} s\frac{\delta' M \delta}{\|\delta_I\|_1^2}$$

[55]See for example Section 1.4 in in Rio (2000), Corollaries 1.1 for strongly mixing processes and 1.2 for strictly stationary processes.

where $D = \{\delta : \|\delta_I^c\|_1 \le C\|\delta_I\|_1, \delta \ne 0, |I| \le s\}$. For convenience, we also define the weighted restricted eigenvalues:

$$\kappa_C^i{}^2(M) = \min_{\delta \in D^i} s \frac{\delta' M \delta}{\|\phi\delta_I\|_1^2}$$

where $D^i = \{\delta : \|\phi_i^0 \delta_I^c\|_1 \le C\|\phi_i^0\delta_I\|_1, \delta \ne 0, |I| \le s\}$, and $\|\phi_i^0\delta\|_1 = \sum_{j=1}^N |\phi_{ij}^0 \delta_j|$.

We invoke Lemma 6 from Belloni, Chen, Chernozhukov, and Hansen (2012). The assumptions for this lemma are surprisingly mild, as they don't involve conditons on the dependence structure of variables nor their distribution. Only assumptions 4.a with $u \ge 1 \ge \ell > 1/c$, and $c \max_{j \in \{1,\ldots,N\}} |S_{ij}| \le \lambda/T$ are needed. We have that:

$$\frac{1}{T}\sum_{t=1}^T \left(\sum_{j=1}^N \widehat{\gamma}_{ij} x_{jt} - \sum_{j=1}^N \gamma_{ij}^0 x_{jt} \right)^2 \le \left(\left(u + \frac{1}{c} \right) \frac{\lambda\sqrt{s_i}}{T\kappa_{c_0}^i} \right)^2,$$

where $c_0 = \frac{uc+1}{\ell c - 1}$.

This bound depends on i in two places: the number of spillover effects different from zero, s_i, and the weighted restricted eigenvalue $\kappa_{c_0}^i$. In order to obtain results for average marginal effects we will derive uniform convergence rates for γ_i over i.

Step 2 In this step we use a Fuk-Nagaev type inequality to ensure that $c \cdot \max_{i,j \in \{1,\ldots,N\}} |\tilde{S}_{ij}| \le \lambda/T$ holds with high probability. The following result, consequence of Theorem 6.2 in Rio (2000), is useful for our purpose:

Lemma 1 *Let $\{z_t\}_{t=1}^T$ be a strongly mixing sequence of real-valued and centered random variables. Assume that there exists γ_1 and a positive c such that the strong mixing coefficients of the sequence satisfy $\alpha(t) \le exp(-at^{\gamma_1})$, and there is a constant $\gamma_2 \in (0, +\infty)$ such that $\sup_t \mathbb{P}(|z_t| > x) \le exp(1 - x^{\gamma_2})$ for any positive x. Then, defining $1/\gamma = 1/\gamma_1 + 1/\gamma_2$ we have that for any $\tilde{\lambda} \ge (TV)^{1/2}$:*

$$\mathbb{P}\left[\left| \sum_{t=1}^T z_t \right| \ge 4\tilde{\lambda} \right] \le 4exp\left(-\frac{\tilde{\lambda}^2 \log 2}{2TV} \right) + 4CT\tilde{\lambda}^{-1} exp\left(-c^2 (TV/\tilde{\lambda})^\gamma \right)$$

where $V = \sup_t \left(\mathbb{E}\left(z_t^2\right) + 2\sum_{s>t} |\mathbb{E}\left[z_t z_s\right]| \right)$.

Proof. It is enough to evaluate (1.7) in Merlevede et al. (2011) with $r = \lambda^2/(TV)$ to obtain the result. ■

Let $\varepsilon_P > 0$ be an arbitrary small number. We have that:

$$\begin{aligned} \mathbb{P}\left[c \cdot \max_{i,j} |\tilde{S}_{ij}| \ge \lambda/T \right] &\le^{(1)} N^2 \cdot \max_{i,j} \quad \mathbb{P}\left[c \cdot |\tilde{S}_{ij}| \ge \lambda/T \right] \\ &=^{(2)} N^2 \cdot \max_{i,j} \quad \mathbb{P}\left[\left| \sum_{t=1}^T \tilde{x}_{jt}\tilde{u}_{it} \right| \ge \frac{\lambda\phi_{ij}^0}{2c} \right] \end{aligned}$$

where (1) holds after the union bound and (2) holds by definition of $\tilde{S}_{ij}$.

Since $\tilde{S}_{ij} = \frac{1}{T}\sum_{t=1}^{T} x_{jt}u_{it} - \frac{1}{T^2}\sum_{t=1}^{T} x_{jt}\sum_{t=1}^{T} u_{it}$ using the traingular inequality we have that:

$$\mathbb{P}\left[|\sum_{t=1}^{T}\tilde{x}_{jt}\tilde{u}_{it}| \geq \frac{\lambda\phi_{ij}^0}{2c}\right] \quad \leq \quad \mathbb{P}\left[|\sum_{t=1}^{T} x_{jt}u_{it}| \geq \frac{\lambda\phi_{ij}^0}{2c}\right] + \mathbb{P}\left[|T^{-1}\sum_{t=1}^{T}\tilde{x}_{jt}||\sum_{t=1}^{T} u_{it}| \geq \frac{\lambda\phi_{ij}^0}{2c}\right]. \quad (14)$$

Making use of Lemma C with $z_t = x_{jt}u_{it}$ and $z_t = u_{it}$ respectively will allow us to pick a λ such that N^2 times each of the two terms above will be arbitrarily small. We now check that the assumptions of Lemma hold for $\{x_{jt}u_{it}\}_t$. First, by assumption 2.d $\{x_{jt}u_{it}\}_t$ is mixing with exponential mixing coefficients. Second, using the law of total probabilities we have:

$$\begin{aligned}\mathbb{P}(|x_{jt}u_{it}| > m) \quad &\leq \quad \mathbb{P}(|u_{it}| > m/b) + \mathbb{P}(|x_{jt}| > b) \leq^{(1)} \mathbb{P}(|u_{it}| > m/b) + \frac{\mathbb{E}[x_{jt}^2]}{b^2} \\ &\leq^{(2)} \quad exp(1 - (m/b)^{\gamma_2}) + \varepsilon_b \leq^{(3)} exp(1 - m^{\gamma_2})\end{aligned}$$

where in $^{(1)}$ we use Markov's inequality, in $^{(2)}$ we use assumption 2.e in combination with 2.b to bound the tail probability of $|u_{it}|$, and finally $^{(3)}$ holds as long as ε_b is sufficiently small so thta $b > 1$. Denoting $\tilde{\lambda} = \frac{\lambda\phi_{ij}^0}{8c}$ we have that:

$$\mathbb{P}\left[|\sum_{t=1}^{T} x_{jt}u_{it}| \geq 4\tilde{\lambda}\right] \quad \leq \quad 4exp\left(-\frac{\tilde{\lambda}^2\log 2}{2TV_{u_ix_j}}\right) + 4CT\tilde{\lambda}^{-1}exp\left(-a(TV_{u_ix_j}/\tilde{\lambda})^{\gamma}\right).$$

In addition, combining with Assumption 4.b we have that:

$$\mathbb{P}\left[|\sum_{t=1}^{T} x_{jt}u_{it}| \geq 4\tilde{\lambda}\right] \leq 4exp\left(-\frac{\tilde{\lambda}^2\log 2}{2Tu'\phi_{ij}^{0\,2}}\right) + 4CT\tilde{\lambda}^{-1}exp\left(-a(T\ell'\phi_{ij}^0/\tilde{\lambda})^{\gamma}\right).$$

Picking $\tilde{\lambda} = \phi_{ij}^0\sqrt{\frac{4u}{\log 2}T\log(N/\varepsilon_P)}$ and using the fact that $\ell'\phi_{ij} > \eta > 0$ for all j we have that:

$$\begin{aligned}N^2 \cdot \max_{i,j}\mathbb{P}\left[|\sum_{t=1}^{T} x_{jt}u_{it}| \geq \frac{\lambda\phi_{ij}^0}{2c}\right] \quad &\leq^{(1)} \quad 4\varepsilon_P^2 + O\left(N^2 \cdot \sqrt{\frac{T}{\log(N/\varepsilon_P)}}exp\left(-C_1\left(\frac{T}{\log(N/\varepsilon_P)}\right)^{\frac{\gamma}{2}}\right)\right) \\ &\leq^{(2)} \quad 4\varepsilon_P^2 + o(1).\end{aligned}$$

with $C_1 = \frac{a\ell'\sqrt{\log 2}}{4u'} > 0$. In $^{(2)}$ we make use of Assumption 4.f part (ii), $\log(N/\varepsilon_P) = O(\log(N))$, and $\gamma > 1$. A sufficient condition for $\gamma > 1$ is $\gamma_1 \geq 2$ and $\gamma_2 \geq 2$.[56]

Let us denote by $\bar{x}_j = \frac{1}{T}\sum_{t=1}^{T} x_{jt}$. Using the law of total probability we bound the second term in (14) as follows:

$$\begin{aligned}\mathbb{P}[|\bar{x}_j||\sum_{t=1}^{T} u_{it}| \geq \frac{\lambda\phi_{ij}^0}{2c}] \quad &\leq \quad \mathbb{P}\left[|\sum_{t=1}^{T} u_{it}| \geq \frac{\lambda\phi_{ij}^0}{2cM_x} \cap |\bar{x}_j| \leq M_x\right] + \mathbb{P}\left[|\bar{x}_j| > M_x\right] \\ &\leq \quad \mathbb{P}\left[|\sum_{t=1}^{T} u_{it}| \geq \frac{\lambda\phi_{ij}^0}{2cM_x}\right] + \mathbb{P}\left[|\bar{x}_j| > M_x\right].\end{aligned}$$

[56] There is a trade off between how fast the exponential decay of the mixing coefficient and the tails, captured by γ_1 and γ_2 respectively, and how fast the number of potential sources of spillovers can grow relative to the time dimension. For instance, if we relax 4.f part (ii) to $\log^2 N = o(T)$ then we need that γ is at least greater than 2.

We bound these two terms in turn. The second term can be bounded using the Markov inequality and Assumption 2.b. Hence, for an arbitrarily small $\varepsilon_x > 0$ there exists an $M_x = M_x(\varepsilon_x)$, such that $\mathbb{P}\left[|\bar{x}_j| > M_x\right] \leq \varepsilon_x$.

In order to bound $\mathbb{P}\left[\left|\sum_{t=1}^T u_{it}\right| \geq \frac{\lambda\phi_{ij}^0}{2cM_x}\right]$ we use an analogous argument as with the first term in (14). After assumtpions 2.d and 2.e Lemma C allows to pick $\ddot{\lambda} = \frac{\lambda\phi_{ij}^0}{8cM_x}$ such that the tail of the absolute value of the sum of u_{it} can be made arbitrarily small.

Finally, by choosing $\lambda = K \cdot \sqrt{T\log(N/\varepsilon_P)}$, where $K = \max\left(8c\sqrt{\frac{4u'}{\log(2)}}, 8cM_x\sqrt{\frac{4u'}{\log(2)}}\right)$, with $\varepsilon_P \to 0$ such that $\log(N/\varepsilon_P) = O(\log(max(N,T))$, and as N and T go to infinitiy

$$\mathbb{P}\left[c \cdot \max_{i,j} |\tilde{S}_{ij}| \leq \frac{\lambda}{T}\right] \to 1.$$

Step 3 $\widehat{s}_i \leq Ks$ for all i.
This result follows directly from the results derived in Lemma 8, 9 and 10 in Belloni, Chen, Chernozhukov, Hansen (2012).

Step 4 Now, we connect the rate of convergence ot the prediction norm with the rate of convergence of γ_i. Assume that $\max_{i,j}|S_{ij}| \leq \lambda/T$. We rewrite the prediction norm as:

$$\frac{1}{T}\sum_{t=1}^T\left(\sum_{j=1}^N \widehat{\gamma}_{ij}x_{jt} - \sum_{j=1}^N \gamma_{ij}^0 x_{jt}\right)^2 \quad = \quad \left(\widehat{\gamma}_i - \gamma_i^0\right)' M \left(\widehat{\gamma}_i - \gamma_i^0\right).$$

After Step 3 and making use of Assumption 1 we have that:

$$\left(\widehat{\gamma}_i - \gamma_i^0\right)' M \left(\widehat{\gamma}_i - \gamma_i^0\right) \geq \phi_{min}(2Ks)(M) \cdot \|\widehat{\gamma}_i - \gamma_i^0\|_2^2 \geq \kappa_1 \cdot \|\widehat{\gamma}_i - \gamma_i^0\|_2^2.$$

In addition, we have that, $\inf_i \kappa_{c_0}^i \geq \frac{1}{b}\kappa_{bc_0/a} = \frac{1}{b}\bar{\kappa}_C$. As a result,

$$\|\widehat{\gamma}_i - \gamma_i^0\|_2^2 \leq \frac{1}{\kappa_1}\left(\left(u + \frac{1}{c}\right)\frac{\lambda\sqrt{s_i}}{T\kappa_{c_0}^i}\right)^2 \leq \frac{1}{\kappa_1}\left(\left(u + \frac{1}{c}\right)\frac{b\lambda\sqrt{s_i}}{T\bar{\kappa}_C}\right)^2 = \frac{Ks}{\kappa_1}\left(\left(u + \frac{1}{c}\right)\frac{b\lambda}{T\bar{\kappa}_C}\right)^2 = A\frac{s\lambda^2}{T^2}$$

where $A = \left(u + \frac{1}{c}\right)^2 \frac{Kb^2}{\kappa_1\bar{\kappa}_C^2}$ is a constant that does not depend on i.

Step 5 We now put together all previous steps to derive the rate of convergence of γ_i and marginal effects. Take $\lambda = K\sqrt{T\log N}$, where $K = c8\sqrt{\frac{4u}{\log(2)}}$ is a constant indepdent of i, and let $\varepsilon > 0$ be a small number. We want to show that:

$$\sup_i \|\widehat{\gamma}_i - \gamma_i^0\|_2^2 = O_p\left(\frac{s\log N}{T}\right).$$

We study the following probability:

$$
\begin{aligned}
&\mathbb{P}\left[\|\widehat{\gamma}_i - \gamma_i^0\|_2 \geq C_\varepsilon \sqrt{\frac{s_i \log N}{T}}\right] \\
&\leq^{(1)} \mathbb{P}\left[\|\widehat{\gamma}_i - \gamma_i^0\|_2 \geq C_\varepsilon \sqrt{\frac{s_i \log N}{T}} \cap c \cdot \max_{i,j}|S_{ij}| \leq \lambda/T\right] + \mathbb{P}\left[c \cdot \max_{i,j}|S_{ij}| > \lambda/T\right] \\
&\leq^{(2)} \mathbb{P}\left[A \geq C_\varepsilon\right] + \mathbb{P}\left[c \cdot \max_{i,j}|S_{ij}| > \lambda/T\right] \\
&=^{(3)} \varepsilon_A + \varepsilon_P
\end{aligned}
$$

where the first inequality $^{(1)}$ holds after using the law of total probabilities, $^{(2)}$ uses Step 4 to bound the probability of the first term, $^{(3)}$ uses that A is a constant independent of N, T, and s and Step 2, respectively, to bound the first and second term and the first result follows.

Turning now to the rate of marginal effect, we introduce the following Lemma on model selection properties of the Lasso:

Lemma 2 *If* $\max_{j\in\{1,\ldots,N\}} |S_{ij}| \leq \lambda/T$ *and there is a constant* $U > 0$ *such that* $|M_{kj}| \leq 1/(Us)$ *for all* $1 \leq j < k \leq p$*, where* M *is the Gram matrix and* M_{kj} *is the* k*-th row and* j*-th column of* M*, then there is a constant* $K > 0$*, independent of* $N, Tands$*, such that:*

$$\max_{j\in\{1,\ldots,N\}} |\widehat{\gamma}_{ij} - \gamma_{ij}^0| \leq K\frac{\lambda}{T}.$$

Proof. The proof is analogous to the second part of Theorem 2 in Belloni and Chernozhukov (2013), in spite of the fact that we do not assume Gaussian errors nor independence of the observations. ■

A direct consequence of Lemma 2 is that, for a given j, if $\widehat{\gamma}_{ij} = 0$ then $|\gamma_{ij}^0| \leq K\frac{\lambda}{T}$. We will make use of this property to study the rate of convergence of average marginal effects. We introduce now some additional notation.

Let $\widehat{T}_i$ be the support of spillovers selected by the Lasso for a given individual i. Let $X[\widehat{T}_i]$ be the selection of columns of X corresponding to regressors $\widehat{T}_i$.

For a fixed j we decompose the difference of the estimated and the true average marginal effects as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\gamma}_{ij}^P - \gamma_{ij}^0\right) = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\gamma}_{ij}^P - \gamma_{ij}^0\right)\mathbb{1}\{j \in \widehat{T}_i\}}_{(b)} - \underbrace{\frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^0\mathbb{1}\{j \notin \widehat{T}_i\}}_{(a)} \tag{15}$$

We study each of these summands in turn. Starting with (a), we have that, after Lemma 2, if $c \max_{j\in\{1,\ldots,N\}} |S_{ij}| \leq \lambda/T$ then $\mathbb{1}\{j \notin \widehat{T}_i\} = \mathbb{1}\{|\gamma_{ij}^0| \leq \frac{K\lambda}{T}\}$. Hence:

$$\frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^0\mathbb{1}\{j \notin \widehat{T}_i\} = \frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^0\mathbb{1}\{|\gamma_{ij}^0| \leq K\lambda/T\}.$$

Then, using Cauchy - Schwartz inequality we have that:

$$\left(\frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^{0}\mathbb{1}\{|\gamma_{ij}^{0}|\leq K\lambda/T\}\right)^{2} \leq \frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^{0\,2}\mathbb{1}\{|\gamma_{ij}^{0}|\leq K\lambda/T\}$$

Now, we can further decompose the sum in two different subsets:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^{0\,2}\mathbb{1}\{|\gamma_{ij}^{0}|\leq K\lambda/T\} &= \frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^{0\,2}\mathbb{1}\{|\gamma_{ij}^{0}|\leq K_2/N\}+\frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^{0\,2}\mathbb{1}\{K_2/N\leq|\gamma_{ij}^{0}|\leq K\lambda/T\}\\ &\leq \left(\frac{K_2}{N}\right)^{2}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{|\gamma_{ij}^{0}|\leq K_2/N\}+\left(\frac{K\lambda}{T}\right)^{2}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{K_2/N\leq|\gamma_{ij}^{0}|\leq K\lambda/T\}.\end{aligned}$$

Taking expectations with respect to the distribution of spillover effects we have that:

$$\begin{aligned}\mathbb{E}\left(\frac{1}{N^2}\left(\sum_{i=1}^{N}\gamma_{ij}^{0}\mathbb{1}\{j\notin\widehat{T}_i\}\right)^{2}\right) &\leq \left(\frac{K_2}{N}\right)^{2}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}\{|\gamma_{ij}^{0}|\leq K_2/N|j\subset T_i^{0}\}\right)\\ &\quad+\left(\frac{K\lambda}{T}\right)^{2}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}\left(K_2/N\leq|\gamma_{ij}^{0}|\leq K\lambda/T|j\subset T_i^{0}\right)\right)\\ &\leq O\left(\frac{1}{N^2}\right)+O\left(\frac{\lambda^2}{T^2N}\right).\end{aligned}$$

where the last inequality holds since $\mathbb{P}\{|\gamma_{ij}^{0}|\leq K_2/N|j\subset T_i^{0}\}\leq 1$, and by Assumption 3.c the second summand decays as $\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_{\gamma_i}\left(K_2/N\leq|\gamma_{ij}^{0}|\leq K\lambda/T|j\subset T_i^{0}\right)\right)=O(\frac{1}{N})$. This assumpion effectively limits the impact of model selection mistakes in the average marginal effect when the size of the spillover effects is smaller. However, when coefficients are smaller than $1/N$ the average marginal effects rate of convergence is unaffected. If Assumption 3.c would not hold, the average marginal effect estimator would be consistent, albeit at a slower rate of convergence.

Now, choosing $\lambda = K\sqrt{T\log N}$ Step 2 above and the previous derivations ensure that:

$$(a)=\frac{1}{N}\sum_{i=1}^{N}\gamma_{ij}^{0}\mathbb{1}\{j\notin\widehat{T}_i\}=O_p\left(\frac{1}{N}\right)+O_p\left(\sqrt{\frac{\log N}{NT}}\right)+o_p(1)$$

where the $o_p(1)$ comes from the fact that for the particular choice of λ the regularity event holds with high probability.

We now turn to study part (b) in (15). Substituting $\left(\widehat{\gamma}_{ij}^{P}-\gamma_{ij}^{0}\right)$ by its regression expression we have:

$$(b) = \underbrace{\frac{1}{N}\sum_{i=1}^{N}e_j^{i}(X[\widehat{T}_i]'X[\widehat{T}_i])^{-1}X[\widehat{T}_i]'u_i\mathbb{1}\{j\in\widehat{T}_i\}}_{(b_1)}+\underbrace{\frac{1}{N}\sum_{i=1}^{N}e_j^{i}(X[\widehat{T}_i]'X[\widehat{T}_i])^{-1}X[\widehat{T}_i]'X[\widehat{T}_i^{c}]\gamma_i^{0}[\widehat{T}_i^{c}]\mathbb{1}\{j\in\widehat{T}_i\}}_{(b_2)}$$

where $\widehat{T}_i^c$ is the complementary of $\widehat{T}_i$ and where e_j^i denotes the vector that selects the jth regressor.[57] Part (b_1) is $O_p\left(1/\sqrt{NT}\right)$ after Assumption 3.b as by definition we have that $e_j^i(X[\widehat{T}_i]'X[\widehat{T}_i])^{-1}X[\widehat{T}_i]' = \ddot{x}_j'$.

Part (b_2) is the bias due to omitted variables arising from model selection mistakes. Eventhough we might have correctly estimated individual j as a source of spillovers for i, unless all others sources of spillovers for i have been correctly selected there will be a bias due to omitted variables in the estimate of the effect of j on i, $\widehat{\gamma}_{ij}$. We now derive the rate of convergence of the average bias. By the Cauchy - Schwartz inequality we have:

$$(b_2)^2 \leq \left(\frac{1}{N}\sum_{i=1}^N \|e_j^i(\frac{1}{T}X[\widehat{T}_i]'X[\widehat{T}_i])^{-1}\|^2\right)\left(\frac{1}{N}\sum_{i=1}^N \|\frac{1}{T}X[\widehat{T}_i]X[\widehat{T}_i^c]\gamma_{ij}^0[\widehat{T}_i^c]\|^2\right)$$

After Step 3 we have that $\sup_i \widehat{s}_i \leq Ks$. Hence,

$$\begin{aligned}\|e_j^i(X[\widehat{T}_i]'X[\widehat{T}_i])^{-1}\|^2 \quad &= \quad e_j^{i\prime}(X[\widehat{T}_i]'X[\widehat{T}_i])^{-2}e_j^i \leq^{(1)} \max_{\delta:\ \ \|\delta\|_2=1} \delta'(X[\widehat{T}_i]'X[\widehat{T}_i])^{-2}\delta \\ &=^{(2)} \quad \phi_{min}\left(\widehat{s}_i\right)(M^{-2}) =^{(3)} \phi_{max}\left(\widehat{s}_i\right)(M)^{-2} \leq^{(4)} \frac{1}{\kappa_1^2}\end{aligned}$$

which by Assumption 1 implies that $\left(\frac{1}{N}\sum_{i=1}^N \|e_j^i(X[\widehat{T}_i]'X[\widehat{T}_i])^{-1}\|^2\right) = O_p(1)$. Now, since the model is sparse, we have the following equality:

$$\|\frac{1}{T}X[\widehat{T}_i]X[\widehat{T}_i^c]\gamma_{ij}^0[\widehat{T}_i^c]\|^2 = \|\frac{1}{T}X[\widehat{T}_i]X[\widehat{T}_i^c \cap T_i^0]\gamma_{ij}^0[\widehat{T}_i^c \cap T_i^0]\|^2$$

where the cardinal of $\widehat{T}_i^c \cap T_i^0$ is, by Step 3, bounded by Ks. Now, by the Cauchy - Schwartz inequality, and by assumption 3.a we have:

$$\|\frac{1}{T}X[\widehat{T}_i]X[\widehat{T}_i^c]\gamma_{ij}^0[\widehat{T}_i^c]\|^2 \quad \leq \quad K^2\left(s^2 \cdot \sup_{j\neq k} M_{jk}^2\right)\|\gamma_i^0[\widehat{T}_i^c \cap T_i^0]\|^2 \leq \frac{K^2}{U^2}\|\gamma_i^0[\widehat{T}_i^c \cap T_i^0]\|^2.$$

If $c \max_{j\in\{1,\dots,N\}} |S_{ij}| \leq \lambda/T$ we have, by Lemma 2 that:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \|\gamma_i^0[\widehat{T}_i^c \cap T_i^0]\|^2 &= \frac{1}{N}\sum_{i=1}^N\sum_{k=1}^N {\gamma_{ik}^0}^2 \mathbb{1}\{k \notin \widehat{T}_i\}\mathbb{1}\{k \in T_i^0\} \\ &= \frac{1}{N}\sum_{i=1}^N\sum_{k=1}^N {\gamma_{ik}^0}^2 \mathbb{1}\{{\gamma_{ik}^0}^2 \leq \frac{\lambda^2}{T^2}\}\mathbb{1}\{k \in T_i^0\} \\ &= \frac{1}{N}\sum_{i=1}^N\sum_{k=1}^N {\gamma_{ik}^0}^2 \mathbb{1}\{{\gamma_{ik}^0}^2 \leq \frac{K_1}{N^2}\}\mathbb{1}\{k \in T_i^0\} + \frac{1}{N}\sum_{i=1}^N\sum_{k=1}^N {\gamma_{ik}^0}^2 \mathbb{1}\{\frac{K_1}{N^2} \leq {\gamma_{ik}^0}^2 \leq \frac{\lambda^2}{T^2}\}\mathbb{1}\{k \in T_i^0\} \\ &\leq \underbrace{\frac{K_1}{N^3}\sum_{i=1}^N\sum_{k=1}^N \mathbb{1}\{k \in T_i^0\}}_{(b_{21})} + \underbrace{\frac{\lambda^2}{NT^2}\sum_{i=1}^N\sum_{k=1}^N \mathbb{1}\{\frac{K_1}{N^2} \leq {\gamma_{ik}^0}^2 \leq \frac{\lambda^2}{T^2}\}\mathbb{1}\{k \in T_i^0\}}_{(b_{22})}\end{aligned}$$

[57]Vector e_j^i is indexed by i to account for the fact that the jth regressor might be in different position of the vector depending on $\widehat{T}_i$.

By sparsity we have $(b_{21}) \leq \frac{sK_1}{N^2}$. Taking expectations on (b_{22}) and using the law of iterated expectations we have:

$$\begin{aligned}
\mathbb{E}((b_{22})) &= \mathbb{E}\left(\frac{\lambda^2}{NT^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\mathbb{P}\{\frac{K_1}{N^2} \leq \gamma_{ik}^{0\;2} \leq \frac{\lambda^2}{T^2} \mid k \in T_i^0\}\mathbb{1}\{k \in T_i^0\}\right) \\
&\leq \mathbb{E}\left(\frac{\lambda^2}{NT^2}\sum_{i=1}^{N}\sup_k \mathbb{P}(\frac{K_1}{N^2} \leq \gamma_{ik}^{0\;2} \leq \frac{\lambda^2}{T^2} \mid k \in T_i^0)\sum_{k=1}^{N}\mathbb{1}\{k \in T_i^0\}\right) \\
&\leq K\frac{s\lambda^2}{T^2}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\sup_k \mathbb{P}(\frac{K_1}{N^2} \leq \gamma_{ik}^{0\;2} \leq \frac{\lambda^2}{T^2} \mid k \in T_i^0)\right) = O\left(\frac{s\lambda^2}{NT^2}\right).
\end{aligned}$$

Finally, for a choice of $\lambda = K\sqrt{T\log N}$ we have, using Step 2, that:

$$(b_2^2) = O_p\left(\frac{s}{N^2}\right) + O_p\left(\frac{s\log N}{NT}\right) + o_p(1)$$

where, as before, $o_p(1)$ comes from the fact that if $\lambda = K\sqrt{T\log N}$ then $c \max_{j\in\{1,\ldots,N\}} |S_{ij}| \leq \lambda/T$ is true with high probability. This concludes the proof. ■

D Tables and Figures

Table 5: Descriptive Statistics

	Final sample		Full sample	
Name variable	Mean	Median	Mean	Median
market value	5071.014	787.555	3833.266	406.283
R&D stock	1009.919	110.726	605.008	28.140
R&D stock/ capital stock	.603	.347	.468	.175
expenditure R&D	182.294	655.103	102.169	4.122
patent count	25.349	2	18.480	1
patent cite	161.455	12	143.511	9
real sales	3720.508	739.083	2848.636	449.498
capital stock	1432.829	207.546	1348.402	120.761
labour	20.120	5.621	18.437	3.821
num of firms	295		715	

Figure 5: Pairwise absolute value of correlation between $R\&D$ paths

Notes: Hisogram of the absolute values of the pairwise correlation between R&D paths in the sample.

Table 6: Distribution of firms by industry

SIC2	Descripton	Freq
13	Oil and Gas Extraction	1.0%
14	Miniming and Quarrying of Nonmetallic Minearls, Except Fuels	0.3%
20	Food and Kindred Products	3.1%
22	Textile Mill Products	0.7%
23	Apparel and Other Finished Products made from fabrics and similar materials	0.3%
24	Lumber and Wood Products, Except Furniture	0.7%
25	Furniture and Fixtures	3.4%
26	Paper and Allied Products	3.4%
27	Printing, Publishing, and Allied Industries	1.0%
28	Chemical and Allied Products	14.2%
29	Petroleum, Refining and Related Industries	1.4%
30	Rubber and Miscellaneous Plastic Products	2.7%
31	Leather and Leather products	0.7%
32	Stone, Clay, Glass, and Concrete Products	1.7%
33	Primary Metal Industries	4.1%
34	Fabricated Metal Products, Except Machinery and Transportation Equipment	4.4%
35	Industrial and Commercial Machinery and Computer Equipment	16.9%
36	Electronic and Other Electrical Equipment and Components, Except Computer	14.6%
37	Transportation Equipment	5.8%
38	Measuring, Analyzing and Controlling Instruments	11.5%
39	Miscellaneous Manufacturing Industries	1.7%
50	Wholesale Trade and Durable Goods	1.4%
51	Wholesale Trade and Nondurable Goods	0.7%
52	Building Materials, Hardware, Garden Supply, and Mobile Home Dealers	0.3%
59	Miscellaneous Retail	0.3%
73	Business Services	2.7%
99	Nonclassifiable Establishments	1.0%

Table 7: Descriptives by Industry (average)

SIC2	capital	labour	rd exp	rd stock	ratio	patents	cites	sales	m value
13	2564.6	28.6	130.6	884.5	19%	18.3	108.1	4075.9	6973.7
14	535.3	4.3	3.8	24.3	2%	0.6	1.3	692.9	1235.2
20	1468.6	22.5	14.8	92.6	4%	3.2	16.2	3878.0	8685.0
22	251.0	7.4	2.2	10.0	6%	1.1	4.8	679.8	505.1
23	150.0	13.2	0.0	0.5	0%	0.1	0.4	949.2	968.3
24	1730.8	16.6	9.4	87.6	1%	3.4	23.1	2396.6	2767.2
25	249.6	10.1	19.1	83.0	20%	3.1	17.8	1018.4	832.6
26	2188.9	19.1	82.9	496.6	17%	29.0	234.6	3025.4	4265.0
27	563.9	12.6	5.7	30.3	4%	1.4	8.6	1401.1	2894.4
28	1943.4	19.6	298.0	1519.1	50%	41.3	232.2	3824.9	11750.6
29	6332.4	11.5	54.2	449.6	16%	8.7	48.0	6578.9	7727.8
30	684.8	13.9	46.7	322.4	25%	8.3	37.6	1788.6	1513.5
31	24.5	2.8	0.7	4.5	15%	0.4	3.2	212.8	204.5
32	823.6	11.0	15.5	121.1	8%	5.4	27.4	1520.2	904.0
33	1349.4	11.5	27.0	188.4	11%	6.5	35.0	1735.5	1749.1
34	307.2	7.8	14.1	79.1	13%	3.0	13.2	924.4	1526.2
35	532.0	13.5	121.9	678.6	66%	26.1	162.5	3061.5	2446.9
36	564.1	11.4	128.7	581.8	54%	28.4	191.5	1847.9	3030.4
37	3981.7	65.4	583.8	3536.8	33%	35.7	192.2	11351.0	5009.0
38	464.0	10.5	90.7	559.7	104%	28.9	174.6	1303.4	1984.3
39	156.0	6.1	22.9	120.3	29%	5.2	23.0	668.7	906.1
50	148.6	4.4	8.0	47.6	8%	2.9	15.0	1432.3	832.3
51	1189.1	12.2	4.0	31.0	2%	0.5	2.1	2345.4	741.0
52	475.6	19.3	20.2	127.4	11%	6.4	28.5	2764.6	2903.2
59	134.5	8.4	0.0	6.0	3%	0.3	1.5	1003.2	844.7
73	2291.2	37.6	527.9	3263.2	115%	103.0	798.9	6479.7	9028.8
99	5144.9	72.0	344.3	2347.4	12%	8.6	49.1	13586.3	31924.8

??	(1)	(2)	(3)	(4)	(5)	(6)
			receivers (vs sources)			
real sales (millions)/ 1000	.358* (.188)	-	-	-	.661** (.299)	.698** (.313)
capital (millions) / 1000	-	.930* (.532)	-	-	-	-
employemnt (thousands)	-	-	.037* (.021)	-	-	-
patent count (annual)	-	-	-	-	$-.035$** (.016)	-
patent citation (annual) /100	-	-	-	$-.029$ (.124)	-	$-.479$** (.203)
			non- linked (vs sources)			
real sales (millions) / 1000	.347* (.188)	-	-	-	.613** (.298)	.648** (.311)
capital (millions) / 1000	-	.930* (.530)	-	-	-	-
employemnt (thousands)	-	-	.035* (.021)	-	-	-
patent count (annual)	-	-	-	-	$-.018$ (.012)	-
patent citation (annual) /100	-	-	-	0.071 (.095)	-	-0.228* (.134)