NETWORK DATA

(M2 ETE Econometrics II)

Koen Jochmans

Toulouse School of Economics

Last revised on April 2, 2024

Outline

Basic terminology

Let V be a set of vertices (nodes) with $|V| = n$; can normalize $V = \{1, \ldots, n\}$.

Let E be a set of edges between nodes in V. This is a set of pairs of nodes $(i, j) \in V \times V$.

Then $G = (V, E)$ is a graph.

G is (un)directed if the pairs in E or (un)ordered.

G is a simple graph if a pair (i, j) has at most one edge between them.

G is a multigraph if the latter is not the case.

Edges can carry a weight, $w_{i,j} \geq 0$, in which case G is a weighted graph.

G can be represented by its (weighted) $n \times n$ adjacency matrix $\boldsymbol{A}$.

In the unweighted case,

$$(\boldsymbol{A})_{i,j} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ \\ 1 & \text{if } (i,j) \in E \end{cases}$$

In the weighted case,

$$(\boldsymbol{A})_{i,j} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ \\ w_{i,j} & \text{if } (i,j) \in E \end{cases}$$

We usually exclude self links, so the diagonal of $\boldsymbol{A}$ contains only zeros.

In the undirected case, $\boldsymbol{A}$ is symmetric.

The (weighted) degree of vertex i is

$$d_i = \sum_{j=1}^{n} w_{i,j} = \sum_{j:(i,j)\in E} w_{i,j}.$$

In the unweighted case d_i is the number of (if the graph is directed, outgoing) edges involving node i.

The volume of a set of nodes $V_1 \subseteq V$ is

$$\mathrm{vol}(V_1) = \sum_{i\in V_1} d_i = \sum_{i\in V_1} \sum_{j:(i,j)\in E} w_{i,j}.$$

A subset V_1 of V is connected if any two vertices in V_1 can be joined by a path of which all the intermediate points also lie in V_1.

The subset V_1 is a connected component if V_1 is connected and there are no connections between V_1 and $\bar{V}_1 = V \backslash V_1$.

The collection of sets $V_1, \ldots, V_m$ (for some m) form a partition of G if each of them is connected and

$$V_i \cap V_j = \emptyset \text{ for all } i \neq j \qquad \text{and} \qquad V_1 \cup \ldots \cup V_m = V.$$

(For an undirected graph) the Laplacian of G is

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A},$$

for

$$\boldsymbol{D} = \text{diag}(\boldsymbol{d}) = \text{diag}(d_1, \ldots, d_n)$$

Note that, for any $\boldsymbol{v} \in \mathbb{R}^n$,

$$\begin{aligned}\boldsymbol{v}' \boldsymbol{L}\, \boldsymbol{v} &= \sum_{i=1}^{n} v_i^2\, d_i - \sum_{i=1}^{n}\sum_{j=1}^{n} v_i\, v_j\, w_{i,j} = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i^2\, w_{i,j} - \sum_{i=1}^{n}\sum_{j=1}^{n} v_i\, v_j\, w_{i,j} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i\, (v_i - v_j)\, w_{i,j} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} (v_i - v_j)^2\, w_{i,j},\end{aligned}$$

where we have used the accounting identity $(v_i - v_j)^2 = v_i(v_i - v_j) + v_j(v_j - v_i)$.

So, $\boldsymbol{L}$ is positive semi-definite. It is singular because

$$\boldsymbol{L}\, \mathbf{1}_n = \boldsymbol{D}\, \mathbf{1}_n - \boldsymbol{A}\, \mathbf{1}_n = \boldsymbol{d} - \boldsymbol{d} = 0\, \mathbf{1}_n$$

and so its smallest eigenvalue is zero (with eigenvector $\mathbf{1}_n$).

The number of connected components in graph G equals the multiplicity of the eigenvalue zero of $\boldsymbol{L}$.

Consider first a connected graph. Let $\boldsymbol{v}$ be an eigenvector of $\boldsymbol{L}$ associated with eigenvalue 0. Then

$$0 = \boldsymbol{v}' \boldsymbol{L} \boldsymbol{v} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (v_i - v_j)^2 \, w_{i,j}.$$

As $w_{i,j} \geq 0$ this sum can only be zero if $(v_i - v_j)^2 \, w_{i,j} = 0$ for all $(i,j) \in E$. For such (i,j) we necessarily have $v_i = v_j$. Further, any third node k that can be reached via the path $i \to j \to k$ then also has $v_k = v_j = v_i$. Extending the argument to longer paths and invoking that the graph is connected implies that $\boldsymbol{v} = \mathbf{1}_n$.

Next take m connected components. Then G is the collection of m connected subgraphs $G_1, \ldots G_m$, each with their own proper Laplacian, $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_m$. To each of these the above argument applies. Further, we may always re-arrange the nodes such that $\boldsymbol{L} = \text{blockdiagonal}(\boldsymbol{L}_1, \ldots, \boldsymbol{L}_m)$. The eigenvectors of $\boldsymbol{L}$ are the eigenvectors of the $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_m$ (supplemented with zeros as additional entries), so that the eigenvalue 0 has m distinct eigenvectors associated with it; these are $\mathbf{1}_{V_1}, \ldots, \mathbf{1}_{V_m}$, where $\mathbf{1}_V$ has the n entries $(\mathbf{1}_V)_i = \{i \in V\}$. □

The normalized adjacency matrix is

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$$

(the notation presumes that G is connected; otherwise use the Moore-Penrose pseudo inverse).

Note that

$$(\boldsymbol{P})_{i,j} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ w_{i,j}/d_i & \text{if } (i,j) \in E \end{cases}$$

and that the rows of $\boldsymbol{P}$ all sum up to one.

Row i of $\boldsymbol{P}$ gives the probability distribution of taking a random step through the graph $\boldsymbol{G}$ when starting at node i.

$\boldsymbol{P}$ is the transition matrix of a Markov chain on G.

If G is connected (and not bipartite) then $\boldsymbol{P}$ has a steady-state distribution given by

$$p_i = \frac{d_i}{\text{vol}(V)}$$

Can define (at least) two normalized version of the Laplacian matrix $\boldsymbol{L}$.

A symmetric normalization:

$$\boldsymbol{L}_{\text{sym}} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\,\boldsymbol{D}^{-1/2} = \boldsymbol{I}_n - \boldsymbol{D}^{-1/2}\boldsymbol{A}\,\boldsymbol{D}^{-1/2}$$

and a 'random-walk' normalization:

$$\boldsymbol{L}_{\text{rw}} = \boldsymbol{D}^{-1}\boldsymbol{L} = \boldsymbol{I}_n - \boldsymbol{P}.$$

In both cases, can show similar properties as for $\boldsymbol{L}$ before.

A connection between the two is as follows:

$\lambda \geq 0$ is an eigenvalue of $\boldsymbol{L}_{\text{rw}}$ with eigenvector $\boldsymbol{v}$ if and only if λ is an eigenvalue of $\boldsymbol{L}_{\text{sym}}$ with eigenvector $\boldsymbol{D}^{1/2}\boldsymbol{v}$.

The number of zero eigenvalues continues to give the number of connected components in G.

For a connected graph G the second smallest eigenvalue $\lambda_2 > 0$ is a measure of global connectivity of G.

This spectral gap is connected to the Cheeger constant

$$C = \min_{\{V_1 \subset V : 0 < \sum_{i \in V_1} d_i \leq \sum_{i \in \bar{V}_1} d_i\}} \frac{\sum_{i \in V_1} \sum_{j \in \bar{V}_1} w_{i,j}}{\sum_{i \in V_1} d_i}$$

through the inequalities

$$1/2 C^2 \leq \lambda_2 \leq 2C.$$

$C \in [0, 1]$ measures how difficult it is to separate G into two disconnected components by removing edges from it.

The numerator in the definition of C is the total weight of the removed edges, the denominator is the total degree in the smallest of the two components.

Small positive C show that there is a bottleneck in G, and that the graph is only weakly connected.

Outline

Motivation

Aim:

Partition set of nodes into m clusters (groups) so that they are similar within a cluster and heterogeneous across clusters.

Here, consider undirected (non-bipartite) graphs.

m is taken as given.

Several types of clustering exist. Here, a flavor is given.

We will discuss spectral clustering based on the Laplacian and on the (random-walk) normalized Laplacian.

Unnormalized clustering

Consider clustering V into clusters $V_1, \ldots, V_m$.

One motivation: minimize the ratio cut of G:

$$\text{RatioCut}(V_1, \ldots, V_m) = \sum_{k=1}^{m} \frac{\text{cut}(V_k, \bar{V}_k)}{|V_k|}$$

for

$$\text{cut}(V_k, \bar{V}_k) = \frac{1}{2} \sum_{i \in V_k} \sum_{j \in \bar{V}_k} w_{i,j} + \frac{1}{2} \sum_{i \in \bar{V}_k} \sum_{j \in V_k} w_{i,j} = \sum_{i \in V_k} \sum_{j \in \bar{V}_k} w_{i,j}$$

(using undirectedness in the last transition).

This minimization problem tries to minimize the weight assigned to edges that go across different clusters.

The weighting makes larger clusters more attractive.

For $k = 1, \ldots, m$, introduce the n-vectors

$$(\boldsymbol{h}_k)_i = \begin{cases} 1/\sqrt{|V_k|} & \text{if } i \in V_k \\ 0 & \text{if } i \notin V_k \end{cases}.$$

For each $k = 1, \ldots, m$ we have (from calculations given above) that

$$\begin{aligned} \boldsymbol{h}_k' \boldsymbol{L}\, \boldsymbol{h}_k &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} ((\boldsymbol{h}_k)_i - (\boldsymbol{h}_k)_j)^2 \, w_{i,j} \\ &= \frac{1}{2} \sum_{i \in V_k} \sum_{j \notin V_k} \frac{1}{|V_k|} \, w_{i,j} + \frac{1}{2} \sum_{i \notin V_k} \sum_{j \in V_k} \frac{1}{|V_k|} \, w_{i,j} \\ &= \frac{\text{cut}(V_k, \bar{V}_k)}{|V_k|}. \end{aligned}$$

Therefore,

$$\text{RatioCut}(V_1, \ldots, V_m) = \sum_{k=1}^{m} \frac{\text{cut}(V_k, \bar{V}_k)}{|V_k|} = \text{trace}\, \boldsymbol{H}' \boldsymbol{L}\, \boldsymbol{H}$$

where $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_m)$.

Note that, for $k, \ell \in \{1, \ldots, m\}^2$

$$\boldsymbol{h}_k' \boldsymbol{h}_\ell = \sum_{i=1}^{n} (\boldsymbol{h}_k)_i \, (\boldsymbol{h}_\ell)_i = \begin{cases} 1 & \text{if } k = \ell \\ 0 & \text{if } k \neq \ell \end{cases},$$

so that $\boldsymbol{H}' \boldsymbol{H} = \boldsymbol{I}_m$ meaning that $\boldsymbol{H}$ is an $n \times m$ orthonormal matrix.

Minimizing the RatioCut can then be written as the discrete optimization problem

$$\min_{V_1, \ldots, V_m} \operatorname{trace} \boldsymbol{H}' \boldsymbol{L} \boldsymbol{H}$$

subject to $\boldsymbol{H}$ being orthonormal and being of the form given on the previous slide.

This problem is NP hard.

A relaxation of the discrete problem is to solve

$$\min_{\boldsymbol{H}} \operatorname{trace} \boldsymbol{H}' \boldsymbol{L} \boldsymbol{H}$$

subject to $\boldsymbol{H}$ being orthonormal (but letting its columns take on arbitrary values in $\mathbb{R}^n$)

This is a well-known problem.

The solution is that $\boldsymbol{H}$ are the eigenvectors associated with the m smallest eigenvalues of $\boldsymbol{L}$.

Let $\boldsymbol{q}_i$ be the ith row of $\boldsymbol{H}$. This is not a classifier.

To obtain one we solve

$$\min_{V_1,\ldots,V_m} \sum_{k=1}^{m} \frac{\sum_{i\in V_k, j\in V_k} \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2}{|V_k|} = \min_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m} \sum_{i=1}^{n} \min_g \|\boldsymbol{q}_i - \boldsymbol{\mu}_g\|^2$$

which minimizes the within cluster sum of squares. This is the k-means algorithm.

This step is ad hoc without further justification (!)

Normalized clustering

An alternative to RatioCut is

$$\mathrm{Ncut}(V_1, \ldots, V_m) = \sum_{k=1}^{m} \frac{\mathrm{cut}(V_k, \bar{V}_k)}{\mathrm{vol}(\mathrm{V_k})},$$

which features a different normalization in the summands.

Recall that we want to look for a partitioning so that (i) edges within a cluster have a high weight and (ii) edges across different clusters have a low weight.

(ii) amounts to making $\mathrm{cut}(V_k, \bar{V}_k)$ small.

(i) amounts to making

$$\sum_{(i,j)\in V_k} w_{i,j} = \sum_{i\in V_k} \sum_{j=1}^{n} w_{i,j} - \sum_{i\in V_k} \sum_{j\in \bar{V}_k} w_{i,j} = \mathrm{vol}(V_k) - \mathrm{cut}(V_k, \bar{V}_k)$$

large. This is done by both minimizing the cut and maximizing the volume.

Ncut achieves both (i) and (ii) while Ratiocut only looks at (ii).

Like RatioCut, the Ncut problem has a connection to the Laplacian matrix.

Redefine the columns of $\boldsymbol{H}$ as

$$(\boldsymbol{h}_1)_i = \begin{cases} 1/\sqrt{\mathrm{vol}(V_k)} & \text{if } i \in V_k \\ 0 & \text{if } i \notin V_k \end{cases} .$$

Then

$$\mathrm{Ncut}(V_1, \ldots, V_m) = \mathrm{trace}\, \boldsymbol{H}' \boldsymbol{L}\, \boldsymbol{H}$$

subject to $\boldsymbol{H}' \boldsymbol{D} \boldsymbol{H} = \boldsymbol{I}_m$ and $\boldsymbol{H}$ being of the form given above.

Can substitute in $\dot{\boldsymbol{H}} = \boldsymbol{D}^{1/2} \boldsymbol{H}$ to obtain the alternative representation

$$\mathrm{Ncut}(V_1, \ldots, V_m) = \mathrm{trace}\, \dot{\boldsymbol{H}}' \boldsymbol{D}^{-1/2} \boldsymbol{L}\, \boldsymbol{D}^{-1/2} \dot{\boldsymbol{H}} = \mathrm{trace}\, \dot{\boldsymbol{H}}' \boldsymbol{L}_{\mathrm{sym}}\, \dot{\boldsymbol{H}}$$

subject to $\dot{\boldsymbol{H}}' \dot{\boldsymbol{H}} = \boldsymbol{I}_m$ and $\dot{\boldsymbol{H}}$ being of the form given above.

Hence, relaxing discreteness yields the solution that $\dot{\boldsymbol{H}}$ are the eigenvectors of $\boldsymbol{L}_{\mathrm{sym}}$ associated with the m smallest eigenvalues.

But then $\boldsymbol{H}$ are the corresponding eigenvectors of $\boldsymbol{L}_{\mathrm{rw}}$.

An alternative view on Ncut comes from the following observation.

Let $\{X_t\}$ be a stationary Markov process on G. Then

$$\mathrm{Ncut}(V_k, \bar{V}_k) = \Pr(X_t \in V_k | X_{t-1} \in \bar{V}_k) + \Pr(X_t \in \bar{V}_k | X_{t-1} \in V_k).$$

This is the probability of switching to/from a cluster.

Normalized spectral clustering aims to make the probability of such an event small.

A proof of the above comes from the observation that, for any set $A \subset V$,

$$\Pr(X_t \in A, X_{t-1} \in \bar{A}) = \sum_{i \in \bar{A}} \sum_{j \in A} \Pr(X_t = j, X_{t-1} = i) = \sum_{i \in \bar{A}} \sum_{j \in A} \pi_i \boldsymbol{P}_{i,j}$$

but

$$\pi_i = \frac{d_i}{\mathrm{vol}(V)}, \qquad \boldsymbol{P}_{i,j} = \frac{w_{i,j}}{d_i},$$

and so

$$\Pr(X_t \in A, X_{t-1} \in \bar{A}) = \frac{1}{\mathrm{vol}(V)} \sum_{i \in \bar{A}} \sum_{j \in A} w_{i,j}.$$

Therefore,

$$\Pr(X_t \in A | X_{t-1} \in \bar{A}) = \frac{\Pr(X_t \in A, X_{t-1} \in \bar{A})}{\Pr(X_{t-1} \in \bar{A})} = \frac{1}{\mathrm{vol}(\bar{A})} \sum_{i \in \bar{A}} \sum_{j \in A} w_{i,j},$$

which equals

$$\frac{\mathrm{cut}(\bar{A}, A)}{\mathrm{vol}(\bar{A})}.$$

Reversing the roles of A and $\bar{A}$ in the above and collecting terms yields the result.

Outline

Classic SBM

Consider an unweighted undirected graph G. (If weighted, can always reduce to binary case.)

Suppose that nodes are of one of m latent types.

Can think about this as (independent) latent variables α_i that take values from the set $\{1, \ldots, m\}$.

Conditional on the types of all nodes, edges are placed independently.

The probability of an edge between (i, j) depends on α_i, α_j.

Thus

$$\Pr(\boldsymbol{A}) = \sum_{(z_1,\ldots,z_n)\in\{1,\ldots,m\}^n} \Pr(\boldsymbol{A}, \alpha_1 = z_1, \ldots, \alpha_n = z_n)$$

where the summand factors as

$$\prod_{i=1}^{n} \Pr(\alpha_i = z_i) \prod_{i<j} \Pr((\boldsymbol{A})_{i,j} | \alpha_i = z_i, \alpha_j = z_j).$$

This is a simple but flexible model that allows for (dis-)assortative matching between nodes.

Parametrization of heterogeneity is low dimensional:

$m \times m$ symmetric matrix $\boldsymbol{B}$ of bernoulli success probabilities for pairs of types:

$$(\boldsymbol{B})_{z_1,z_2} = (\boldsymbol{B})_{z_2,z_1} = \Pr(w_{i,j} = 1 | \alpha_i = z_1, \alpha_j = z_2)$$

and an m-vector of type probabilities.

In this model there are m latent communities. Spectral clustering correctly classifies nodes to communities as $n \to \infty$.

Spectral clustering in SBM (fixed effects)

The intuition behind the consistency of spectral clustering has two steps:

(i) The eigenvectors of the (normalized) Laplacian converge in probability to those of the expected Laplacian.

(ii) The expected Laplacian has only m eigenvectors, which identify group membership.

Point (i) is a technical issue. We focus on Point (ii).

We do this conditional on the variables $\alpha_1, \ldots, \alpha_n$; so we treat these as fixed.

Note that the $n \times n$ matrix $\boldsymbol{A}$ has (conditional) expectation

$$\boldsymbol{A}_0 = \mathbb{E}(\boldsymbol{A}|\alpha_1, \ldots, \alpha_n) = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}'$$

for $n \times m$ matrix $\boldsymbol{Z}$ whose ith column is composed of all zeros expect at location α_i.

Will assume that all columns of $\boldsymbol{Z}$ contain at least one 1 (so all communities have at least one member, and so the matrix has full column rank) and that $\boldsymbol{B}$ has full rank.

The normalized Laplacian associated with $\boldsymbol{A}_0$ is

$$\boldsymbol{L}_0 = \boldsymbol{I}_n - \boldsymbol{D}_0^{-1/2} \boldsymbol{A}_0 \boldsymbol{D}_0^{-1/2}$$

for $\boldsymbol{D}_0$ the diagonal degree matrix.

$\boldsymbol{L}_0$ and $\bar{\boldsymbol{L}}_0 = \boldsymbol{I}_n - \boldsymbol{L}_0$ have the same eigenvectors so focus on the latter here.

Note that $\boldsymbol{D}_0 = \text{diag}(\boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}'\mathbf{1}_n)$ and, therefore, that

$$(\boldsymbol{D}_0)_{i,i} = \sum_{z=1}^{m} \boldsymbol{B}_{Z_i,z}\, n_z$$

for $n_z = \sum_{i=1}^{n} \{\alpha_i = z\}$. Hence, the entries of the $n \times n$ diagonal matrix $\boldsymbol{D}_0$ can only take on m different values. Collect these values in the $m \times m$ diagonal matrix

$$(\bar{\boldsymbol{D}}_0)_{z,z} = \sum_{z'=1}^{m} \boldsymbol{B}_{z,z'}\, n_{z'}.$$

Then

$$\bar{\boldsymbol{L}}_0 = \boldsymbol{Z}\bar{\boldsymbol{D}}_0^{-1/2}\boldsymbol{B}\bar{\boldsymbol{D}}_0^{-1/2}\boldsymbol{Z}' = \boldsymbol{Z}\bar{\boldsymbol{B}}\boldsymbol{Z}'$$

for $\bar{\boldsymbol{B}} = \bar{\boldsymbol{D}}_0^{-1/2}\boldsymbol{B}\bar{\boldsymbol{D}}_0^{-1/2}$.

Next, consider the (small) $m \times m$ matrix

$$\boldsymbol{C}\bar{\boldsymbol{B}}\boldsymbol{C}$$

for $\boldsymbol{C} = (\boldsymbol{Z}'\boldsymbol{Z})^{1/2}$. Note that $\boldsymbol{C}\bar{\boldsymbol{B}}\boldsymbol{C}$ is a full rank and symmetric. So we have the eigendecomposition

$$\boldsymbol{C}\bar{\boldsymbol{B}}\boldsymbol{C} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}'$$

where all m eigenvalues are non-zero.

With $\boldsymbol{C}$ invertible, we have $\bar{\boldsymbol{B}} = \boldsymbol{C}^{-1}\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}'\boldsymbol{C}^{-1}$ and, therefore, also that

$$\bar{\boldsymbol{L}}_0 = \boldsymbol{Z}\bar{\boldsymbol{B}}\boldsymbol{Z}' = \boldsymbol{Z}(\boldsymbol{C}^{-1}\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}'\boldsymbol{C}^{-1})\boldsymbol{Z}' = \boldsymbol{Z}\boldsymbol{Q}\,\boldsymbol{S}\,\boldsymbol{Q}'\boldsymbol{Z}'$$

for $\boldsymbol{Q} = \boldsymbol{C}^{-1}\boldsymbol{V}$.

Note that $(\boldsymbol{Z}\boldsymbol{Q})'(\boldsymbol{Z}\boldsymbol{Q}) = \boldsymbol{Q}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{Q} = \boldsymbol{V}'(\boldsymbol{Z}'\boldsymbol{Z})^{-1/2}(\boldsymbol{Z}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1/2}\boldsymbol{V} = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_m$. So the columns of the $n \times m$ orthonormal matrix $\boldsymbol{Z}\boldsymbol{Q}$ are the eigenvectors of the matrix $\bar{\boldsymbol{L}}_0$ (and so equally of $\boldsymbol{L}_0$).

Finally, letting $\boldsymbol{z}_i$ the ith row of $\boldsymbol{Z}$, we have that $\boldsymbol{z}_i\,\boldsymbol{Q} = \boldsymbol{z}_j\,\boldsymbol{Q}$ if and only if

$$\boldsymbol{z}_i = \boldsymbol{z}_j \qquad (\text{and so } \alpha_i = \alpha_j)$$

because $\boldsymbol{Q}$ is invertible. Hence, clustering on the eigenvectors of $\boldsymbol{L}_0$ yields correct assignment of nodes to communities.

Consistency of spectral clustering has been established in more general stochastic block models.

This includes the setting where m grows slowly with n, and also several forms of sparsity.

There is not a single definition of a sparse graph.

Roughly, sparse means that the number of edges does not grow proportional to n^2 (which would give rise to a dense network). This is often formulated in the (expected) degree d_i growing only slowly with n.

In the stochastic block model, usually done by including a (common) sparsity parameter that shrinks the probability of edge formation to zero as n grows.

A weighted graph can always be translated into an unweighted graph so the above equally applies to the weighted case.

In principle, with a consistent classification in hand, we can treat cluster membership as known and then estimate the model parameters by standard methods.

Identification in small networks (random effects)

We can see the stochastic block model as a (nonlinear) model with two-way heterogeneity:

$$(\boldsymbol{A})_{i,j} = w_{i,j} = \varphi(\alpha_i, \alpha_j, \varepsilon_{i,j}),$$

where the α_i are i.i.d. across i and the $\varepsilon_{i,j}$ are i.i.d. across dyads, independent of the α_i.

Similar to a finite-mixture model in panel data.

Edge weights are dependent through their joint dependence on the α_i only.

By using such conditional-independence we can nonparametrically identify

The distribution of $w_{i,j}|\alpha_i, \alpha_j$ and,

the distribution of α_i,

from the joint distribution of edge weights in small graphs.

Easiest to see under the assumption that the functions

$$G_z(w) = \mathbb{P}(w_{i,j} \leq w | \alpha_i = z) = \sum_{z'=1}^{m} \mathbb{P}(w_{i,j} \leq w | \alpha_i = z, \alpha_j = z')\, p_{z'}$$

are linearly independent.

In that case, $n \geq 4$ suffices for identification.

Consider a grid of m points $w_1, \ldots w_m$ and construct the $m \times m$ matrix

$$\boldsymbol{G} = \begin{pmatrix} G_1(w_1) & \ldots & G_m(w_1) \\ \vdots & \ldots & \vdots \\ G_1(w_m) & \ldots & G_m(w_m) \end{pmatrix}.$$

By linear independence we can always find points such that this matrix has full rank.

Next, for distinct integers $i, j, k, \ell \in \{1, \ldots, n\}^4$, let

$$\begin{aligned}(\boldsymbol{M}\ \ )_{m_1,m_2} &= \mathbb{P}(w_{i,j} \leq w_{m_1}, w_{i,k} \leq w_{m_2}),\\ (\boldsymbol{M}_w)_{m_1,m_2} &= \mathbb{P}(w_{i,j} \leq w_{m_1}, w_{i,k} \leq w_{m_2}, w_{i,\ell} \leq w),\end{aligned}$$

for $(m_1, m_2) \in \{1, \ldots, m\}^2$ and any w.

Then conditional independence implies that

$$\boldsymbol{M} = \boldsymbol{G}\boldsymbol{D}\boldsymbol{G}'$$

and

$$\boldsymbol{M}_w = \boldsymbol{G}\boldsymbol{D}^{1/2}\boldsymbol{D}_w\boldsymbol{D}^{1/2}\boldsymbol{G}'$$

for $\boldsymbol{D} := \mathrm{diag}(p_1, \ldots, p_m)$ and $\boldsymbol{D}_w = \mathrm{diag}(G_1(w), \ldots, G_m(w))$ for any w.

Use an eigendecomposition of $\boldsymbol{M}$ to construct a matrix $\boldsymbol{V}$ so that

$$\boldsymbol{V}\boldsymbol{M}\boldsymbol{V}^\top = \boldsymbol{I}_m.$$

Note that

$$\boldsymbol{V}\boldsymbol{M}\boldsymbol{V}^\top = \boldsymbol{V}(\boldsymbol{G}\boldsymbol{D}\boldsymbol{G}')\boldsymbol{V}^\top = (\boldsymbol{V}\boldsymbol{G}\boldsymbol{D}^{1/2})(\boldsymbol{D}^{1/2}\boldsymbol{G}'\boldsymbol{V}^\top) = \boldsymbol{Q}\,\boldsymbol{Q}' = \boldsymbol{I}_m$$

for $\boldsymbol{Q} := \boldsymbol{V}\boldsymbol{G}\boldsymbol{D}^{1/2}$ an $m \times m$ orthonormal matrix of full rank.

Next,

$$\boldsymbol{V}\boldsymbol{M}_w\boldsymbol{V}^\top = (\boldsymbol{V}\boldsymbol{G}\boldsymbol{D}^{1/2})\boldsymbol{D}_w(\boldsymbol{D}^{1/2}\boldsymbol{G}'\boldsymbol{V}^\top) = \boldsymbol{Q}\boldsymbol{D}_w\boldsymbol{Q}'$$

for any w.

By linear independence of the columns of $\boldsymbol{G}$ we can recover the matrix $\boldsymbol{Q}$ up to sign and permutation of its columns as the joint diagonalizer of the collection of matrices $\boldsymbol{V}\boldsymbol{M}_{w_1}\boldsymbol{V}^\top, \ldots, \boldsymbol{V}\boldsymbol{M}_{w_m}\boldsymbol{V}^\top$.

Let this matrix be denoted as $\bar{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{\Delta}\boldsymbol{\Sigma}$.

Here, $\boldsymbol{\Sigma}$ is a permutation matrix and $\boldsymbol{\Delta}$ is a diagonal matrix with only 1 or -1 as diagonal entries.

Next, construct the matrix

$$(\boldsymbol{N}_w)_{m_1,m_2} = \mathbb{P}(w_{i,k} \leq w_{m_1}, w_{j,\ell} \leq w_{m_2}, w_{i,j} \leq w).$$

Then

$$\boldsymbol{N}_w = \boldsymbol{G}\boldsymbol{D}\boldsymbol{H}_w\boldsymbol{D}\boldsymbol{G}'$$

for

$$(\boldsymbol{H}_w)_{z,z'} = \mathbb{P}(w_{i,j} \leq w | \alpha_i = z, \alpha_j = z').$$

But

$$\boldsymbol{V}\boldsymbol{N}_w\boldsymbol{V}^\top = (\boldsymbol{V}\boldsymbol{G}\boldsymbol{D}^{1/2})\boldsymbol{D}^{1/2}\boldsymbol{H}_w\boldsymbol{D}^{1/2}(\boldsymbol{D}^{1/2}\boldsymbol{G}'\boldsymbol{V}^\top) = \boldsymbol{Q}\boldsymbol{D}^{1/2}\boldsymbol{H}_w\boldsymbol{D}^{1/2}\boldsymbol{Q}'$$

and so the entries of

$$\bar{\boldsymbol{Q}}'\boldsymbol{V}\boldsymbol{N}_w\boldsymbol{V}^\top\bar{\boldsymbol{Q}} = \boldsymbol{\Sigma}'\boldsymbol{\Delta}(\boldsymbol{D}^{1/2}\boldsymbol{H}_w\,\boldsymbol{D}^{1/2})\boldsymbol{\Delta}\boldsymbol{\Sigma}$$

give the

$$\mathbb{P}(w_{i,j} \leq w | \alpha_i = z, \alpha_j = z')\,\sqrt{p_z\,p_{z'}}$$

up to sign and labelling. Note that for $w = +\infty$ this gives $\sqrt{p_z\,p_{z'}}$ up to the same sign and labelling, so that the ratio yields $\mathbb{P}(w_{i,j} \leq w | \alpha_i = z, \alpha_j = z')$ while the diagonals yield the p_z.

Outline

Models on networks

If, in an (undirected) network of n agents we observe interactions between all

$$\frac{n(n-1)}{2}$$

possible pairs of nodes we can consider estimation of nonlinear fixed-effect models.

Say

$$y_{i,j}|x_{i,j}, \alpha_i, \alpha_j$$

has a parametric distribution. Can do bias-corrected maximum likelihood estimation.

This is like a panel data problem where $N, T \to \infty$ at the same rate; for each of n nodes we observe $(n-1)$ outcomes, so we get more measurements on each individual as $n \to \infty$.

(Fixed effect estimation from many small networks will not be possible.)

The case where we observe more limited interaction between nodes is much more complicated.

A linear model

Consider an undirected (multi-) graph where each edge (i, j) has associated with it an outcome $y_{i,j}$, with

$$\mathbb{E}(y_{i,j}|\alpha_i, \alpha_j, w_{i,j} > 0) = \alpha_i - \alpha_j$$

Say we observe m edges.

Stack all outcomes in m-vector $\boldsymbol{y}$.

Let $\boldsymbol{B}$ be the $m \times n$ matrix where the row for edge $e = (i, j)$ has $(\boldsymbol{B})_{e,i} = 1$ and $(\boldsymbol{B})_{e,j} = -1$. This is the (oriented) incidence matrix of the graph; we note that

$$\boldsymbol{L} = \boldsymbol{B}'\boldsymbol{B}$$

connects it to the Laplacian matrix.

We can write

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{B}) = \boldsymbol{0}.$$

The conditional-mean restriction is one of network exogeneity.

A bipartite special case

Suppose that the node set V can be partitioned into two sets V_1, V_2 and that edges are only formed between the two subsets.

Can then reparametrize

$$\alpha_i = \begin{cases} \eta_i & \text{if } i \in V_1 \\ -\gamma_i & \text{if } i \in V_2 \end{cases}$$

An example is worker i works in firm j at time t; so that log-wage $y_{i,j,t}$ decomposes as

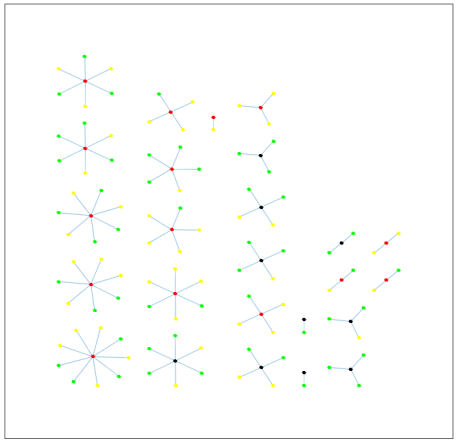
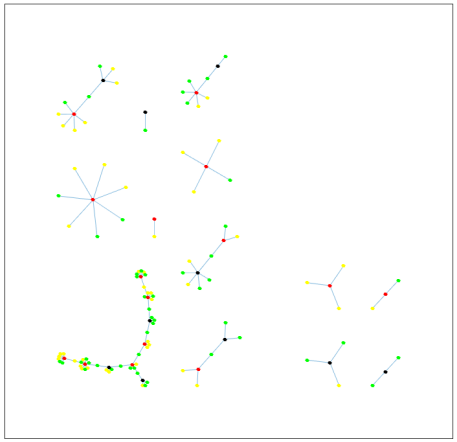
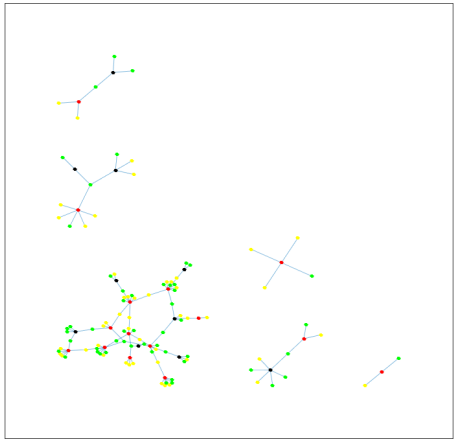
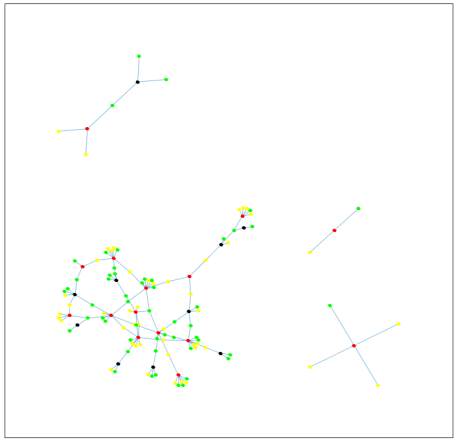
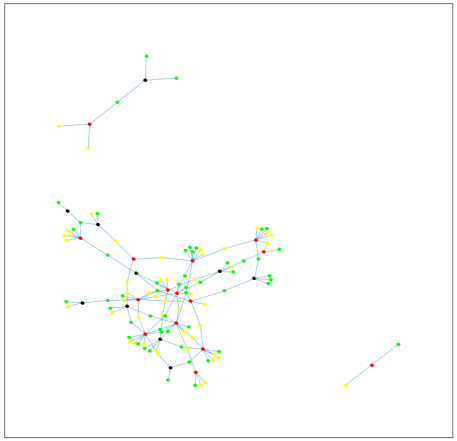
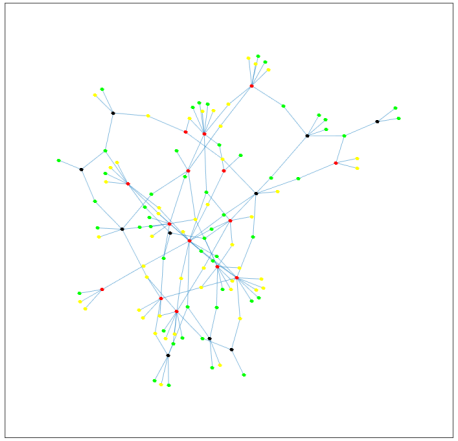
$$y_{i,j,t} = \eta_i + \gamma_{j(i,t)} + \varepsilon_{i,j,t}.$$

Here, the number of edges between any i and j is the number of periods that i is employed by j.

The connectivity structure of the induced graph drives the statistical properties of the fixed-effect estimator.

For example, if no worker ever switches employer we cannot learn anything.

(Dis-)connected graphs

The above plot is for a stationary model where workers and firms are of binary types (high and low quality), and mobility of workers is introduced through exogenous lay-offs.

A one-period network has as many components as firms.

Cannot disentangle worker effects from firm effects.

Adding time periods makes workers switch to different firms.

Makes the graph more connected.

With enough mobility, the graph becomes connected.

Connectivity allows for the least-squares estimator to be well defined (subject to one normalization on the fixed effects) but is not enough for it to have good properties.

Least-squares estimator

Each row of $\boldsymbol{B}$ sums to zero so we normalize the fixed effects so that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\boldsymbol{A})_{i,j}\,(\alpha_i+\alpha_j)=0.$$

If the graph is connected then the (constrained) least-squares estimator exists an is equal to

$$\hat{\boldsymbol{\alpha}}=\boldsymbol{D}^{-1/2}(\boldsymbol{L}_{\text{sym}})^{*}\,\boldsymbol{D}^{-1/2}\,\boldsymbol{B}'\boldsymbol{y}.$$

By strict exogeneity this estimator is unbiased.

Further note that

$$\text{var}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}|\boldsymbol{B})=\boldsymbol{D}^{-1/2}(\boldsymbol{L}_{\text{sym}})^{*}\,\boldsymbol{D}^{-1/2}\,\boldsymbol{B}'\,\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{B})\,\boldsymbol{B}\boldsymbol{D}^{-1/2}(\boldsymbol{L}_{\text{sym}})^{*}\,\boldsymbol{D}^{-1/2}.$$

Notably, with $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \, \boldsymbol{I}_m)$,

$$\mathrm{var}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} | \boldsymbol{B}) = \sigma^2 \, \boldsymbol{D}^{-1/2} (\boldsymbol{L}_{\mathrm{sym}})^* \, \boldsymbol{D}^{-1/2}.$$

So the precision of the estimator is driven by the Laplacian of the graph.

We have

$$\mathrm{var}(\hat{\alpha}_i | \boldsymbol{B}) = \sigma^2 \, \frac{(\boldsymbol{L}^*_{\mathrm{sym}})_{i,i}}{d_i}$$

The behavior of this variance depends in a very complicated way on the connectivity structure of the network.

No reason it would behave proportional to d_i^{-1}, which would correspond to the usual parametric rate.

Let λ_2 be second smallest eigenvalue of $\boldsymbol{L}_{\text{sym}}$ and

$$h_i = \left(\frac{1}{d_i} \sum_{j=1}^{n} \frac{w_{i,j}^2}{d_j} \right)^{-1}.$$

Can derive that

$$\frac{\sigma^2}{d_i} - \frac{2\,\sigma^2}{m} \leq \text{var}(\hat{\alpha}_i | \boldsymbol{B}) \leq \frac{\sigma^2}{d_i} \left(1 + \frac{1}{\lambda_2 h_i} \right) - \frac{2\,\sigma^2}{m}.$$

So,

$$\text{var}(\hat{\alpha}_i) = \frac{\sigma^2}{d_i} + o(d_i^{-1})$$

provided that $\lambda_2 h_i \to \infty$.

In this case can get conventional asymptotic normality with d_i serving as effective sample size for estimating α_i.

Example: student/teacher data

Pupils in Grades 4 and 5 of elementary school over the period 2008–2012.

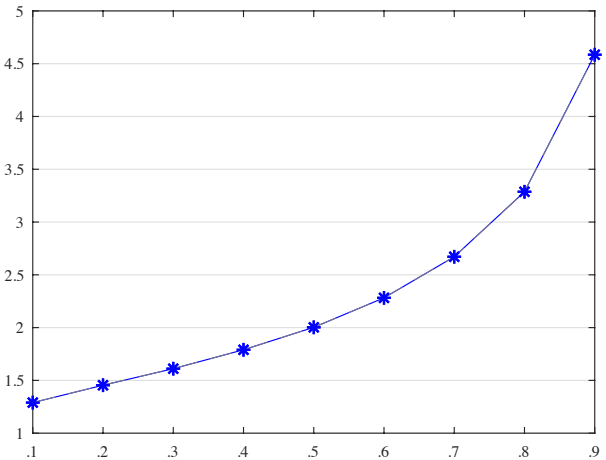
(One-mode projection has) $m = 41,612$ edges and $n = 11,945$ teachers.

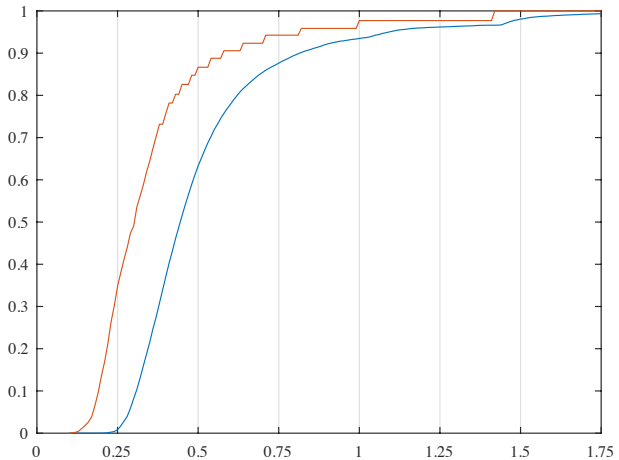
$\lambda_2 = .0039$.

	mean	stdev	$10^{\text{th}}\%$	$20^{\text{th}}\%$	$30^{\text{th}}\%$	$40^{\text{th}}\%$	$50^{\text{th}}\%$	$60^{\text{th}}\%$	$70^{\text{th}}\%$	$80^{\text{th}}\%$	$90^{\text{th}}\%$
d_i	13.87	10.76	3.00	5.50	7.50	9.00	11.00	14.00	17.50	21.50	27.50
h_i	7.15	7.13	2.43	3.30	4.01	4.72	5.48	6.36	7.44	9.12	12.56

Conventional inference procedures not applicable.

Under homokedasticity we can compute the ratio of the exact variance versus its first-order approximation.





Left plot: deciles of the distribution of this ratio.

Right plot: distribution of width of 95% confidence intervals constructed using exact (blue) and approximate (red) standard error (using $\sigma^2 = 1$).

Variance components

Interest often lies in variances and covariances between fixed effects.

Of the form

$$\boldsymbol{\alpha}' \boldsymbol{W} \boldsymbol{\alpha}$$

for some weight matrix $\boldsymbol{W}$.

Plug-in estimator is

$$\hat{\boldsymbol{\alpha}}' \boldsymbol{W} \hat{\boldsymbol{\alpha}}$$

and

$$\mathbb{E}(\hat{\boldsymbol{\alpha}}' \boldsymbol{W} \hat{\boldsymbol{\alpha}} | \boldsymbol{B}, \boldsymbol{W}) = \boldsymbol{\alpha}' \boldsymbol{W} \boldsymbol{\alpha} + \sigma^2 \operatorname{trace}(\boldsymbol{W} \operatorname{var}(\hat{\boldsymbol{\alpha}} | \boldsymbol{B})).$$

The bias again depends on graph structure. Here, can be corrected for exactly by subtracting an unbiased estimator of the bias.

Nevertheless, only in the dense-network case is the bias-corrected estimator asymptotically normal!.

For example in a graph that contains a fixed number of clusters, the estimator follows a mixture of non-central chi-squared distributions. Inference here is highly non-standard. (student-teacher/worker-firm data does not fit this setup!)

Outline

Peer effects

Observe data (y_i, x_i) for agents $i = 1, \ldots, n$.

Agents engage in social interactions.

Action and/or outcome of agent i may be influenced by action/outcome or a characteristic of agent j.

Corresponds to a graph G with adjacency matrix $\boldsymbol{A}$ and normalized $\boldsymbol{P}$.

Gives rise to peer effects:

Correlated effects:
Agents behave similarly because they are in the same environment.

Exogenous effects, contextual effects, or spillover effects:
The propensity of an agent to behave in a certain way varies with the distribution of characteristics in his/her reference group.

Endogenous effects:
The propensity of an agent to behave in a certain way varies with the prevalence of that behavior in the reference group.

A linear model

Stack observations in n-vectors $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{x} = (x_1, \ldots, x_n)'$.

A linear model for social interactions is

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \rho \boldsymbol{P}\boldsymbol{y} + \beta \boldsymbol{x} + \delta \boldsymbol{P}\boldsymbol{x} + \boldsymbol{\varepsilon}$$

with

$$\mathbb{E}(\boldsymbol{\varepsilon} | \boldsymbol{A}, \boldsymbol{x}) = \mathbf{0}.$$

Assume that $|\rho| < 1$ so that the model has a reduced form:

$$\begin{aligned}
\boldsymbol{y} &= (\boldsymbol{I}_n - \rho \boldsymbol{P})^{-1} (\alpha \mathbf{1}_n + \beta \boldsymbol{x} + \delta \boldsymbol{P}\boldsymbol{x} + \boldsymbol{\varepsilon}) \\
&= \left(\sum_{k=0}^{\infty} \rho^k \boldsymbol{P}^k \right) (\alpha \mathbf{1}_n + \beta \boldsymbol{x} + \delta \boldsymbol{P}\boldsymbol{x} + \boldsymbol{\varepsilon}) \\
&= \frac{\alpha}{1-\rho} \mathbf{1}_n + \beta \boldsymbol{x} + \lambda \sum_{k=1}^{\infty} \rho^{k-1} \boldsymbol{P}^k \boldsymbol{x} + \sum_{k=0}^{\infty} \rho^k \boldsymbol{P}^k \boldsymbol{\varepsilon}
\end{aligned}$$

for $\lambda = \rho\beta + \delta$.

Multiplying through by $\boldsymbol{P}$ and taking conditional expectations yields

$$\mathbb{E}(\boldsymbol{P}\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{A}) = \frac{\alpha}{1-\rho}\mathbf{1}_n + \beta\boldsymbol{P}\boldsymbol{x} + \lambda(\boldsymbol{P}^2\boldsymbol{x} + \rho\boldsymbol{P}^3\boldsymbol{x} + \rho^2\boldsymbol{P}^4\boldsymbol{x} + \cdots).$$

This equation reveals that $\boldsymbol{P}^2\boldsymbol{x}$ is a relevant (and valid) instrumental variable for $\boldsymbol{P}\boldsymbol{y}$ provided that

$\lambda = \rho\beta + \delta \neq 0$ and

$\boldsymbol{I}_n, \boldsymbol{P}, \boldsymbol{P}^2$ are not colinear.

When $\rho \neq 0$ we have overidentifying restrictions coming from $\boldsymbol{P}^3\boldsymbol{x}, \boldsymbol{P}^4\boldsymbol{x}, \ldots$

Linear-in-means model

Recalling that $\boldsymbol{P}$ is a transition matrix on the network, $\boldsymbol{P}^2\boldsymbol{x}$ is an average of characteristics of peers of peers.

No-multicolinearity then requires variation in peer-group composition.

In the complete network,

$$\boldsymbol{A} = \mathbf{1}_n\mathbf{1}_n' - \boldsymbol{I}_n, \qquad \boldsymbol{D} = (n-1)\,\boldsymbol{I}_n$$

and so

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A} = \frac{1}{n-1}(\mathbf{1}_n\mathbf{1}_n' - \boldsymbol{I}_n).$$

But also

$$\boldsymbol{P}^2 = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{A} = \frac{(n-2)}{(n-1)^2}\boldsymbol{P} + \frac{1}{(n-1)}\boldsymbol{I}_n,$$

yielding colinearity.

Variation in peer-group size

Note that

$$\sum_{j=1}^{n}(\boldsymbol{P})_{i,j}\,x_j = \frac{1}{n-1}\sum_{j\neq i}x_j = \frac{1}{n-1}(n\,\bar{x} - x_i),$$

and

$$\sum_{j=1}^{n}(\boldsymbol{P})_{i,j}\,y_j = \frac{1}{n-1}\sum_{j\neq i}y_j = \frac{1}{n-1}(n\,\bar{y} - y_i),$$

So

$$y_i = \alpha + \rho\sum_{j=1}^{n}(\boldsymbol{P})_{i,j}\,y_j + \beta x_i + \delta\sum_{j=1}^{n}(\boldsymbol{P})_{i,j}\,x_j + \frac{1}{n-1}\sum_{j\neq i}x_j + \varepsilon_i$$

is equal to

$$y_i = \alpha + \rho\frac{1}{n-1}(n\,\bar{y} - y_i) + \beta x_i + \delta\frac{1}{n-1}(n\,\bar{x} - x_i) + \frac{1}{n-1}\sum_{j\neq i}x_j + \varepsilon_i.$$

This can be re-arranged to get

$$\left(1+\rho\frac{1}{n-1}\right)y_i = \alpha + \rho\frac{n}{n-1}\bar{y} + \left(\beta+\delta\frac{1}{n-1}\right)x_i + \delta\frac{n}{n-1}\bar{x} + \varepsilon_i.$$

Averaging yields

$$\left(1+\rho\frac{1}{n-1}\right)\bar{y} = \alpha + \rho\frac{n}{n-1}\bar{y} + \left(\beta+\delta\frac{1}{n-1}\right)\bar{x} + \delta\frac{n}{n-1}\bar{x} + \bar{\varepsilon}.$$

Differencing gives

$$\left(1+\rho\frac{1}{n-1}\right)(y_i-\bar{y}) = \alpha + \left(\beta+\delta\frac{1}{n-1}\right)(x_i-\bar{x}) + (\varepsilon_i-\bar{\varepsilon}).$$

The coefficients in the (population) regression

$$(y_i-\bar{y}) = a_n + b_n(x_i-\bar{x}) + (\varepsilon_i-\bar{\varepsilon})$$

are identified an equal

$$a_n = \frac{\alpha}{1+\rho\frac{1}{n-1}}, \qquad b_n = \frac{\beta+\delta\frac{1}{n-1}}{1+\rho\frac{1}{n-1}}.$$

Because b_n is known and depends on n and involves three different structural parameters, we can recover β, δ, ρ from $b_{n_1}, b_{n_2}, b_{n_3}$ if we observe three (or more) different network sizes n_1, n_2, n_3.

They are uniquely recoverable provided that $\lambda \neq 0$, where λ is the same parameter as before.

Network effects

The above model does little to accommodate correlated effects.

When at the network level, such effects are network fixed effects.

Again take

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \rho \boldsymbol{P}\boldsymbol{y} + \beta \boldsymbol{x} + \delta \boldsymbol{P}\boldsymbol{x} + \boldsymbol{\varepsilon}$$

but see α as a fixed effect for the network in question (se we are sampling many small networks here).

Can difference within the network:

$$(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \rho(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P}\boldsymbol{y} + \beta(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} + \delta(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon}.$$

Reduced form then comes from inverting

$$(\boldsymbol{I} - \rho\boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \beta(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} + \delta(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon}.$$

Now need that $\boldsymbol{I}_n, \boldsymbol{P}, \boldsymbol{P}^2, \boldsymbol{P}^3$ are linearly independent.