Homophily and Transitivity in Dynamic Network Formation

Econometric Methods for Social Spillovers and Networks,

University of St. Gallen, October 1st to 9th, 2018

Bryan S. Graham

University of California - Berkeley

Econometrics of Network Formation in Two Slides

Network formation as a large game: Mele (2017, EM), Christakis et al. (2010, WP), de Paula et al. (2018, EM), Sheng (2013, WP), Menzel (2015, WP)

- modeling strategic behavior central

- each paper "deals with" incompleteness in different ways

- close connections with econometrics of games literatures

Econometrics of Network Formation in Two Slides (continued)

Network formation with agent heterogeneity: Graham (2017, EM), Dzemski (2014, WP), Jochmans (2018, JBES), Yan et al. (2018, JASA), Shi and Chen (2016, WP)

- focus on incorporating rich/high dimensional unobserved agent-level heterogeneity into (generally) non-strategic (dyadic) models

- close connections with panel data (and related) literatures

This Paper

Attempts to include (elements of) two main approaches into one model and study its parameter's identification, estimation and inference.

Studies a simple model of dynamic network formation where

1. agents respond to existing network structure when forming, maintaining or dissolving links;

2. model is non-dyadic: networks structure matters;

3. agents are (super) heterogenous.

Presentation Outline

1. Notation and motivation

2. Likelihood

3. Identification

4. Monte Carlo

5. Extension to directed networks / digraphs

6. Some open questions

Setup

- Large (sparse) network consisting of $i = 1, \ldots, N$ potentially connected agents.

- Observe all ties in each of $t = 0, 1, 2, 3$ periods.

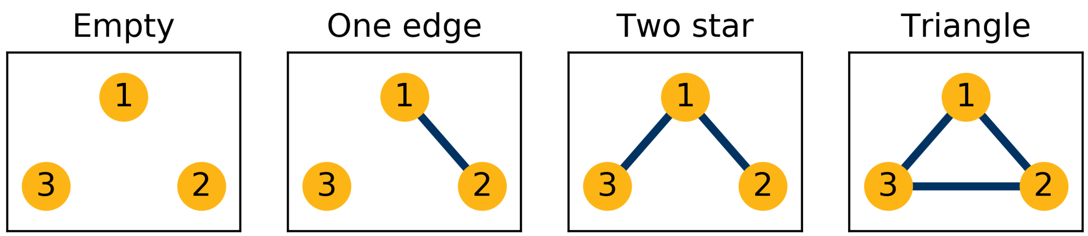
- $\mathbf{D}_t$ denotes the period t adjacency matrix:

  - $D_{ijt} = 1$ if agents i and j are connected in period t and zero otherwise

  - Ties are undirected: $D_{ijt} = D_{jit}$

  - No self-ties: $D_{iit} = 0$

Stylized Fact: Links are clustered



- Real world networks exhibit substantial clustering/transitivity in ties

- Transitivity indices often substantially exceed network densities

$$\begin{aligned}\rho_{\mathrm{CC}} &= \Pr\left(D_{ij}=1 \middle| D_{ik}=1,\, D_{jk}=1\right)\\ &> \Pr\left(D_{ij}=1\right)=\rho_{\mathrm{D}}\end{aligned}$$

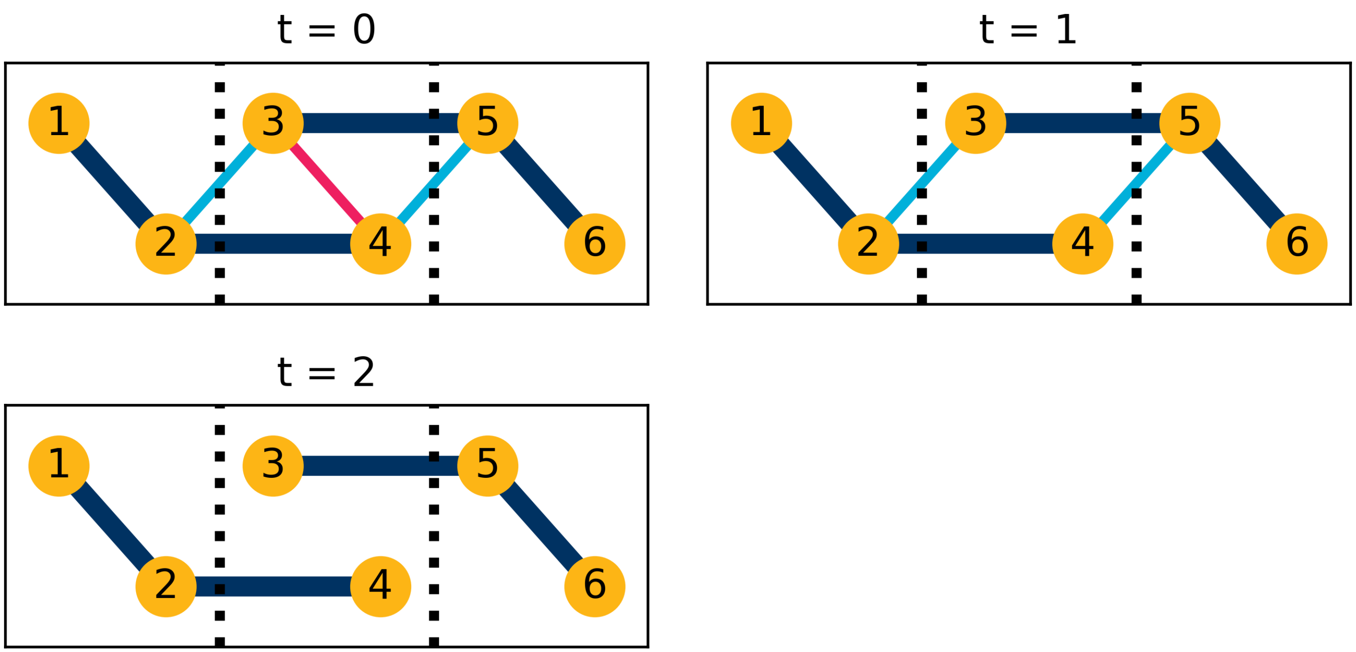
Homophily versus Transitivity

Two explanations for clustering:

- Homophily – '*birds of a feather flock together*' (assortative mixing, community structure)
  - sorting may be on both observed and, problematically, *unobserved* agent attributes

- (Structural) taste for transitivity ('triadic closure') – '*a friend of a friend is also my friend*'

Homophily versus Transitivity: Policy implications



Link formation model

- Agents i and j form a link in periods $t = 1, \ldots, 3$ according to the rule

$$D_{ijt} = \mathbf{1}\left(\beta D_{ijt-1} + \gamma R_{ijt-1} + A_{ij} - U_{ijt} > 0\right)$$

- $R_{ijt} = \sum_{k=1}^{N} D_{ikt} D_{kjt}$ equals the number of period t friends i and j have in common

- $A_{ij} = A_{ji}$ is dyad-specific unobserved heterogeneity

- U_{ijt} is iid across links and over time with distribution function $F(u)$

Comments on model

Model captures three key features of link formation (cf. Snijders, 2011)

1. State dependence – β;

2. Structural taste for transitivity or 'triadic closure' – γ;

3. (Time invariant) dyad-specific heterogeneity, A_{ij}:

   (a) Degree heterogeneity (van Dujin et al., 2004; Graham, 2017);

   (b) Homophily (Assortative Mixing on *unobservables*).

Comments on model (continued)

Dyad-specific heterogeneity, A_{ij}, admits many specifications (cf., Krivitsky, Handcock, Raftery and Hoff, 2009; Zhao, Levina, Zhu, 2012).

Example #1

$$A_{ij} = v_i + v_j - g\left(\xi_i, \xi_j\right)$$

The v_i term induces degree heterogeneity.

$g\left(\xi_i, \xi_j\right)$ measures distance in ξ_i attribute space (assortative linking on ξ_i).

Comments on model (continued)

Example #2

$$A_{ij} = \upsilon_i + \upsilon_j + C_i' P C_j$$

C_i is a $K \times 1$ vector with a 1 in k^{th} row if i belongs to community k and zeros elsewhere (and P a $K \times K$ real symmetric matrix).

In what follows $\mathbf{A} = (A_{12}, \ldots, A_{N-1N})'$ *is left unrestricted*.

Comments on model (continued)

In each period agents take initial structure of the network as fixed when deciding whether to form, maintain or dissolve links:

- (myopic) Best-reply type dynamics (e.g., Jackson & Wolinsky, 1996);

- no completeness/coherence problems;

- measurement challenges (cf. Chamberlain, 1985; Snijders, 2011).

Comments on model (continued)

A link forms if its net surplus is positive; *utility is transferrable.*

R_{ijt-1} measures opportunities to engineer 'triadic closure' or the number of triangles an agent (myopically forecasts) a period t ij link will create.

If agents have a structural taste for transitivity the network will evolve in a way that fills these so-called 'structural holes'.

Initial condition

The link rule specified above applies only to periods $t = 1, \ldots, 3$.

The *initial condition* is unspecified.

Assume

$$(\mathbf{D}_0, \mathbf{A}) \sim \Pi_0$$

with $\mathbf{A}$ denoting the $\frac{1}{2}N(N-1)$ vector of dyad-specific heterogeneity terms.

Initial condition (continued)

Π_0 is unrestricted:

- $\mathbf{D}_0$ and $\mathbf{A}$ may covary;

- elements of $\mathbf{A}$ may also be dependent.

In a single cross-section ***any*** network configuration can be generated by an appropriately chosen draw of $\mathbf{A}$ (graphon).

Likelihood

The joint probability density at $\mathbf{D}_0^T = \mathbf{d}_0^T$ and $\mathbf{A} = \mathbf{a}$ is:

$$\begin{aligned} p\left(\mathbf{d}_0^T, \mathbf{a}, \theta\right) \;&=\; \pi\left(\mathbf{d}_0, \mathbf{a}\right) \\ &\times \prod_{i<j} \prod_{t=1}^{T} F\left(\beta d_{ijt-1} + \gamma r_{ijt-1} + a_{ij}\right)^{d_{ijt}} \\ &\times \left[1 - F\left(\beta d_{ijt-1} + \gamma r_{ijt-1} + a_{ij}\right)\right]^{1-d_{ijt}}. \end{aligned}$$

$\pi\left(\mathbf{d}_0, \mathbf{a}\right)$ is the density of the 'initial network condition' (high dimensional nuisance parameter).

<u>Comments on likelihood</u>

Since $\mathbf{A}$ is unobserved, the econometrician has three options:

1. <u>**random effects**</u>: specify a distribution for $\mathbf{A}$ given $\mathbf{D}_0$ and base inference on the corresponding integrated likelihood; also specify distribution of U_{ij}.

2. <u>**joint fixed effects**</u>: treat the $\binom{N}{2}$ components of $\mathbf{A}$ as additional (incidental) parameters to be estimated; also specify distribution of U_{ij}.

3. <u>**conditional fixed effects**</u>: find an (identifying) implication of the model that is invariant to $\mathbf{A}$; distribution of U_{ij} may or may not be specified.

Comments on likelihood (continued)

First option (random effects) is difficult conceptually and computationally (cf., van Dujin et al., 2004; Goldsmith-Pinkham & Imbens, 2013).

Second option (joint fixed effects) will have poor statistical properties in the present setting (cf., Graham, 2017).

Third option (conditional fixed effects) is pursued here.

Research question

- Can we learn anything about β and γ without imposing (strong) restrictions on $\pi(\mathbf{d}_0, \mathbf{a})$ and/or $F(\bullet)$?

- Need an (identifying) implication of the model that is invariant to $\mathbf{A}$:

  - this is a high-dimensional object;

  - initial condition is also high dimensional;

  - likelihood interdependencies...

<u>Likelihood interdependencies</u>

If we change the value of a <u>single</u> link (i, j) from, say, zero to one, <u>many</u> components of the likelihood may change.

Dyad-specific decisions today may alter the incentives for link formation across many other dyads in subsequent periods.

Two networks sequences $\mathbf{D}_0^T = \mathbf{d}_0^T$ and $\mathbf{D}_0^T = \mathbf{v}_0^T$ may differ in only a small number of elements, yet have very different likelihoods.

Stable neighborhoods

Idea: we can learn about the β and γ by comparing the frequency of different link histories for a given pair (i, j) holding other (local) features of the network fixed.

Problem: Changing the link history of a single (i, j) pair has effects which cascade throughout the likelihood.

Solution: Look for pairs embedded in ‘stable neighborhoods’.

Stable neighborhoods (continued)

The pair (i, j) are embedded in a stable neighborhood if

1. all their links, except possibly those with each other, are stable across periods 1, 2, 3;

2. the links belonging to their friends are stable in periods 1, 2.

Let $Z_{ij} = 1$ if (i, j) is a *stable dyad* – embedded in a stable neighborhood *and* $D_{ij1} \neq D_{ij2}$ – and zero otherwise.

Let $\mathcal{D}_s = \left\{ \mathbf{i} \,|\, Z_{i_1 i_2} = 1 \right\}$ denote the set of all stable dyads.

Conditioning Set

Consider the set of network sequences

$$\begin{aligned}\mathbb{V}^s \;=\; \Big\{\mathbf{v}_0^3 = (\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) \,\Big|\, & \mathbf{v}_t \in \mathbb{D} \text{ for } t = 0, \ldots, 3, \\ & \mathbf{v}_0 = \mathbf{d}_0,\ \mathbf{v}_1 + \mathbf{v}_2 = \mathbf{d}_1 + \mathbf{d}_2,\ \mathbf{v}_3 = \mathbf{d}_3, \\ & v_{ij1} = d_{ij1} \,\&\, v_{ij2} = d_{ij2} \\ & \text{if } z_{ij} = 0, \text{ for } i, j = 1, \ldots, N\Big\}.\end{aligned}$$

$\mathbb{V}^s$ contains all network sequences constructed by permutations of the period 1 and 2 link decisions of the $\mathbf{m}_N \overset{def}{\equiv} |\mathcal{D}_s|$ stable dyads.

All other link decisions are held fixed at their observed values.

The set $\mathbb{V}^s$ contains $2^{|\mathcal{D}_s|} = 2^{\mathbf{m}_N}$ elements.

Permutation lemma

For all $l \neq i, j$ let $\left(R^*_{il1}, R^*_{il2}\right)$ denote the values of (R_{il1}, R_{il2}) after permuting D_{ij1} and D_{ij2}. If the pair (i, j) is a stable dyad, then $\left(R^*_{il1}, R^*_{il2}\right) = (R_{il2}, R_{il1})$.

- Permuting D_{ij1} and D_{ij2} *does* alter period 2 and 3 link incentives for other agents to which i and j are linked, but in a *controlled* way.

- Neighborhood stability implies that $D_{il1} = D_{il2}$, so the change of incentives is entirely via transitivity effects.

Permutation lemma (continued)

Consider the period 2 and 3 likelihood contributions of an (i, l) pair that is linked in both periods.

After permutation:

$$
\begin{aligned}
& F\left(\beta d_{il1} + \gamma r^*_{il1} + a_{il}\right) F\left(\beta d_{il2} + \gamma r^*_{il2} + a_{il}\right) \\
= \; & F\left(\beta d_{il1} + \gamma r_{il2} + a_{il}\right) F\left(\beta d_{il2} + \gamma r_{il1} + a_{il}\right) \\
= \; & F\left(\beta d_{il2} + \gamma r_{il2} + a_{il}\right) F\left(\beta d_{il1} + \gamma r_{il1} + a_{il}\right) \\
= \; & F\left(\beta d_{il1} + \gamma r_{il1} + a_{il}\right) F\left(\beta d_{il2} + \gamma r_{il2} + a_{il}\right).
\end{aligned}
$$

This coincides with the pre-permutation contribution!

Permutation lemma (continued)

If i and j are embedded in a stable neighborhood, then permuting D_{ij1} and D_{ij2} leaves

1. initial condition unaffected;

2. all period 1 likelihood contributions, except those associated with (i, j), are unaffected;

Permutation lemma (continued)

3. (net) period 2 and 3 contributions from (i, l) and (j, l) dyads are unaffected (use permutation lemma);

4. period 2 and 3 contributions from all (k, l) dyads are unaffected (D_{ij1} and D_{ij2} do not enter the likelihood contributions of these pairs).

Main result: Notation

Let $S_{ij} \overset{def}{\equiv} D_{ij2} - D_{ij1}$, $Q_{ij} \overset{def}{\equiv} \left(D_{ij0}, D_{ij3}, R_{ij0}, R_{ij1}\right)'$ and

$$
\begin{aligned}
b_{ij}^{01}\left(q_{ij}, a_{ij}, \theta\right) &= \frac{1 - F\left(\beta d_{ij0} + \gamma r_{ij0} + a_{ij}\right)}{F\left(\beta d_{ij0} + \gamma r_{ij0} + a_{ij}\right)} \frac{F\left(\beta d_{ij3} + \gamma r_{ij1} + a_{ij}\right)}{1 - F\left(\beta d_{ij3} + \gamma r_{ij1} + a_{ij}\right)} \\
b_{ij}^{10}\left(q_{ij}, a_{ij}, \theta\right) &= \frac{F\left(\beta d_{ij0} + \gamma r_{ij0} + a_{ij}\right)}{1 - F\left(\beta d_{ij0} + \gamma r_{ij0} + a_{ij}\right)} \frac{1 - F\left(\beta d_{ij3} + \gamma r_{ij1} + a_{ij}\right)}{F\left(\beta d_{ij3} + \gamma r_{ij1} + a_{ij}\right)}.
\end{aligned}
$$

c.f. Honore and Kyriazidou (2000).

Main Result (continued)

The conditional likelihood of $D_0^3 = \mathbf{d}_0^3$ given $\mathbf{d}_0^3 \in \mathbb{V}^s$,

$$l^c\left(\mathbf{d}_0^3, \mathbf{a}, \theta\right) = \frac{p\left(\mathbf{d}_0^3, \mathbf{a}, \theta\right)}{\sum_{\mathbf{v} \in \mathbb{V}^s} p\left(\mathbf{v}_0^3, \mathbf{a}, \theta\right)}, \tag{1}$$

equals

$$\begin{aligned} l^c\left(\mathbf{d}_0^3, \mathbf{a}, \theta\right) &= \prod_{\mathbf{i} \in \mathcal{D}_s} \left[\frac{1}{1 + b_{i_1 i_2}^{01}\left(q_{ij}, a_{ij}, \theta\right)}\right]^{\mathbf{1}\left(s_{i_1 i_2} = 1\right)} \\ &\quad \times \left[\frac{1}{1 + b_{i_1 i_2}^{10}\left(q_{ij}, a_{ij}, \theta\right)}\right]^{\mathbf{1}\left(s_{i_1 i_2} = -1\right)} . \end{aligned}$$

Denominator in (1) is a summation over $2^{\mathbf{m}_N}$ elements.

Main Result (continued)

...surprisingly this sum is not intractable (“binomial theorem”).

The ratio (1) can be expressed as a product of just $\mathbf{m}_N$ terms!

Main Result (comments)

An unexpected byproduct of conditioning is (conditional) independence.

Link histories of stable dyads are conditionally independent!

Distribution of U_{ijt} unspecified $\Rightarrow$ maximum score approach to estimation (Manski, 1975, 1987; Honore and Kyriazidou, 2000).

If U_{ijt} is logistically distributed, then $\mathbf{A}$ doesn't enter the conditional likelihood; criterion function takes familiar logit form.

Nonparametric case

Under the data generating process specified above

$$\Pr\left(D_{ij1}=0, D_{ij2}=1 \middle| Q_{ij}=q, Z_{ij}=1\right)$$
$$-\Pr\left(D_{ij1}=1, D_{ij2}=0 \middle| Q=q, Z_{ij}=1\right) \lesseqgtr 0$$

according to whether

$$\beta\left(d_3-d_0\right)+\gamma\left(r_1-r_0\right) \lesseqgtr 0.$$

cf. Manski (1987); suggests the following estimator:

$$\sup_{\theta:\|\theta'\theta\|=1} \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} Z_{ij}\left(D_{ij2}-D_{ij1}\right) \operatorname{sgn}\left\{X_{ij}'\theta\right\} \tag{2}$$

for $x=\left(d_3-d_0, r_1-r_0\right)'$.

Logit case

When the idiosyncratic component of surplus U_{ijt} is logistic

$$\Pr\left(S_{ij} = s \middle| Q_{ij} = q, Z_{ij} = 1\right) = \left(\frac{\exp\left(x'\theta\right)}{1 + \exp\left(x'\theta\right)}\right)^{\mathbf{1}(s=1)} \times \left(\frac{1}{1 + \exp\left(x'\theta\right)}\right)^{\mathbf{1}(s=-1)}.$$

Note: A_{ij} does not enter to the right of the equality ($\Rightarrow$ point identification up to scale).

Logit case (continued)

The *stable neighborhood logit* estimate of θ_0 is the maximizer of

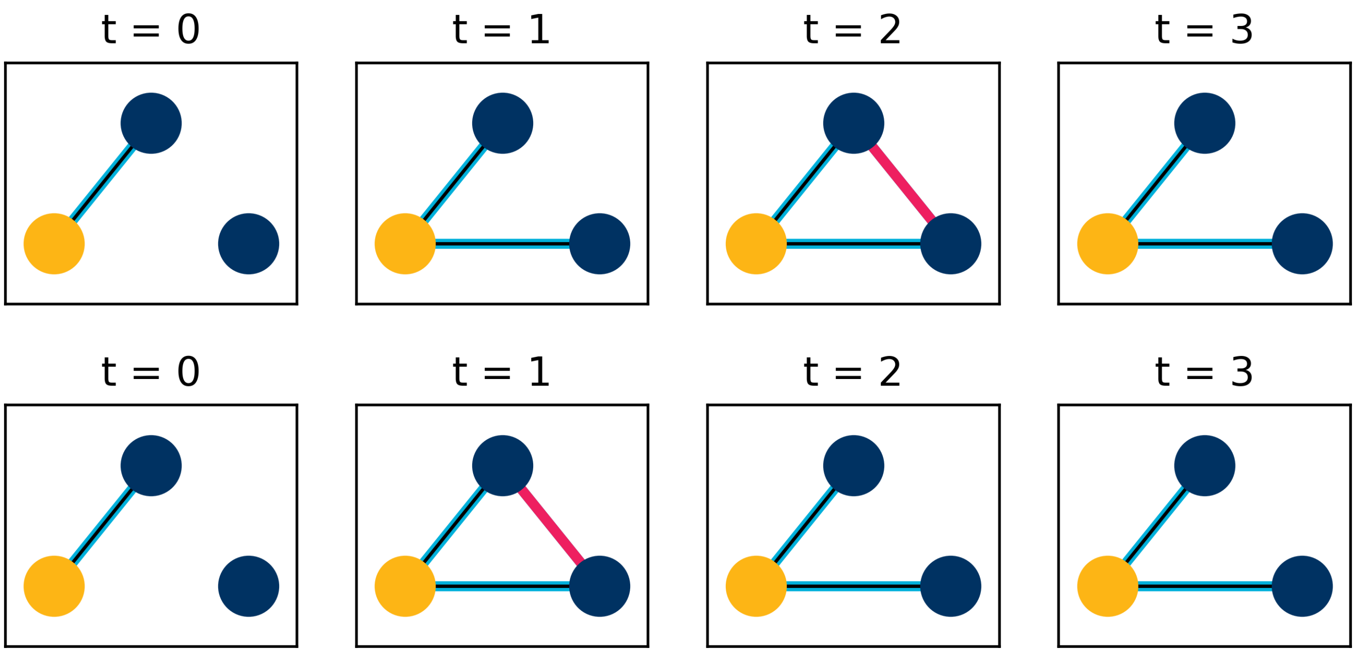
$$L_N(\theta) = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} l_{ij}(\theta)$$

with

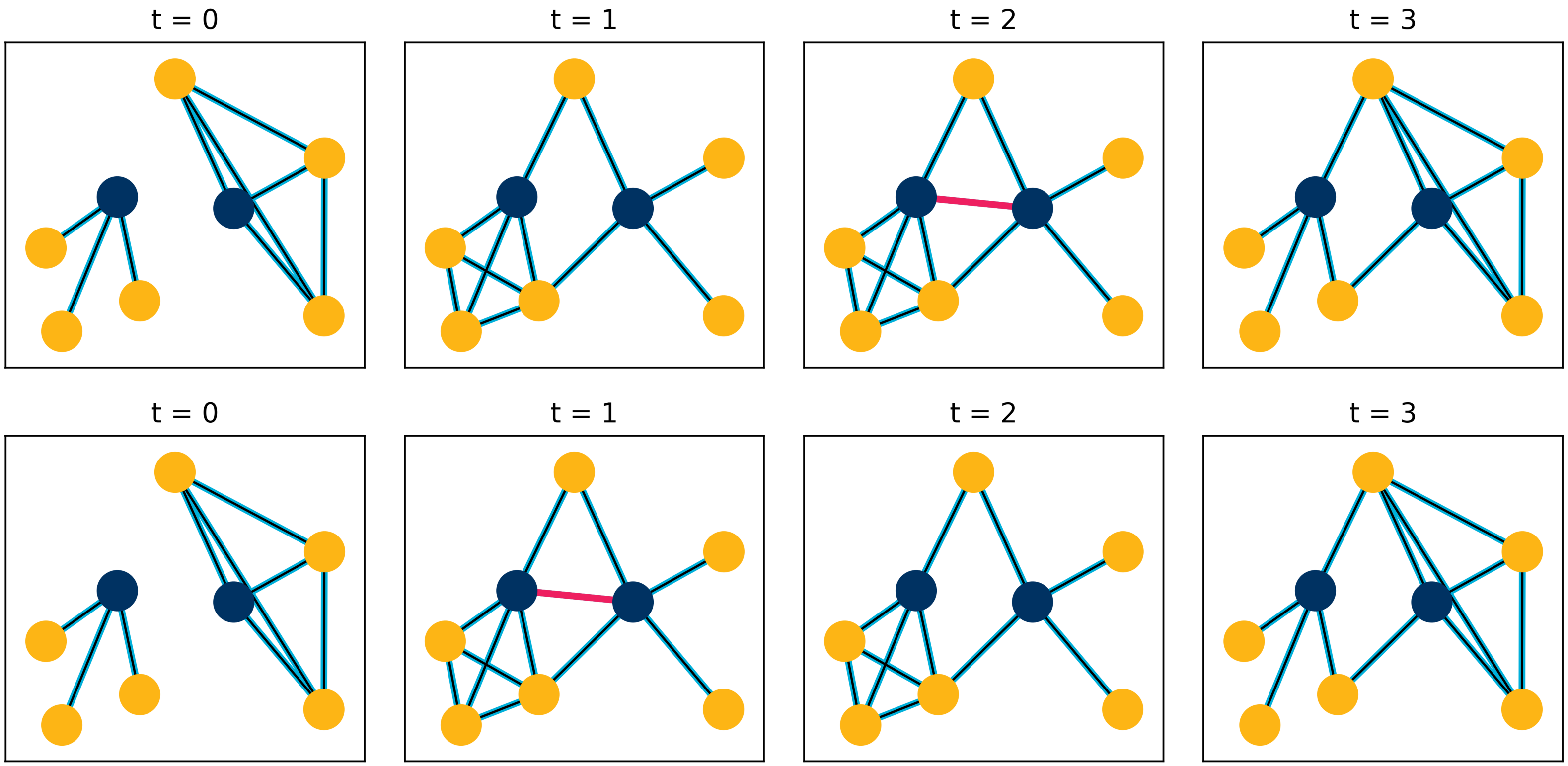
$$l_{ij}(\theta) = Z_{ij}\left\{S_{ij}X_{ij}'\theta - \ln\left[1 + \exp\left(S_{ij}X_{ij}'\theta\right)\right]\right\}.$$

Summation over a random set of dyads...

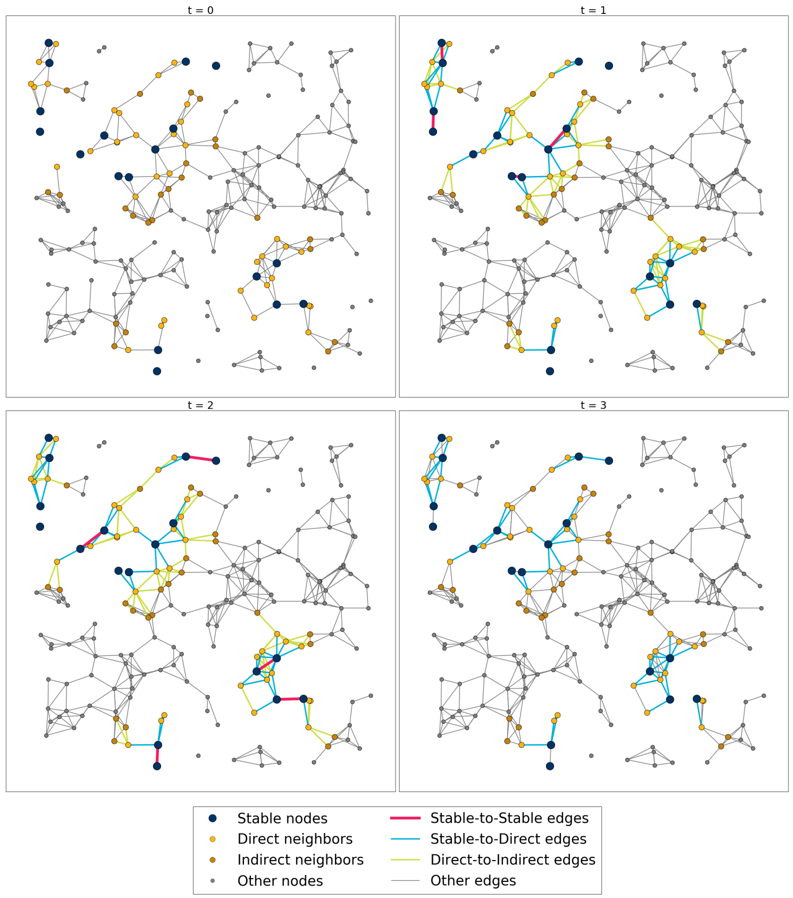
Stable neighborhood example



Stable neighborhood example



Stable neighborhoods in large network



Monte Carlo

Agents are scattered uniformly on the two-dimensional plane

$$\left[0, \sqrt{N}\right] \times \left[0, \sqrt{N}\right].$$

Initial network is generated according to

$$D_{ij0} = \mathbf{1}\left(A_{ij} - U_{ij0} \geq 0\right),$$

with U_{ij0} logistic and A_{ij} taking one of two values.

Monte Carlo (continued)

1. If the Euclidean distance between i and j is less than or equal to r, then $A_{ij} = \ln\left(\frac{0.75}{1-0.75}\right)$, otherwise $A_{ij} = -\infty$.

2. Agents less than r apart link with probability 0.75, while those greater than r apart link with probability zero.

Network in $t = 1, 2, 3$ generated using link rule with $\beta = \gamma = 1$ and U_{ijt} logistic.

Properties of simulated networks

Asymptotic Degree	4		
Period	$(N-1)\,\mathbb{E}\left[D_{it}\right]$	T	GC
$t=0$	3.94	0.44	0.58
$t=1$	4.98	0.58	0.83
$t=2$	5.12	0.59	0.84
$t=3$	5.14	0.59	0.85

Notes: The table reports period-specific network summary statistics across the $B=1,000$ Monte Carlo simulations for each design ($N=5,000$). See paper for other design details. The $(N-1)\,\mathbb{E}\left[D_{it}\right]$ column gives the average degree, T the global clustering coefficient or transitivity index and GC the fraction of agents that are part of the largest giant component.

Sampling properties of SN logit

Asymptotic Degree	4	
$N = 5,000$	β	γ
Mean	1.0438	1.0456
Median	1.0410	1.0133
Std. Dev.	0.4575	0.2976
Mean Std. Err.	0.4493	0.2917
Coverage	0.9620	0.9650
Avg. # of Stable Dyads	110.6	
# of cvg. failures	1	

Rates of convergence

Consistent estimation using a single (large sparse) network requires that $n\alpha_N \to \infty$ where $\alpha_N = \Pr\left(Z_{ij} = 1\right)$ and $n = \binom{N}{2}$.

α_N is at most $O\left(N^{-1}\right)$; since rate of convergence is $\sqrt{n\alpha_N} \Rightarrow$ it will be no faster than $\sqrt{N}$.

Empirical researcher just counts number of stable dyads pre-estimation.

cf., Andersen (1970), Chamberlain (1980)

Extension to Directed Networks

By adaption the definition of a stable dyad, it is possible to extend the main results to directed networks.

This is important for modeling buyer-supplier networks (e.g., Atalay et al., 2011), trade flows (e.g., Melitz et al., 2008) etc.

Main challenge is increase in the number of types of likelihood terms.

Extension to Directed Networks

Agents i *directs* a link towards j in periods $t = 1, \ldots, 3$ according to the rule

$$D_{ijt} = \mathbf{1}\left(\beta D_{ijt-1} + \gamma R_{ijt-1} + \delta D_{jit-1} + A_{ij} - U_{ijt} > 0\right)$$

Directed model includes a *reciprocity* parameter (δ), in addition to those for state dependence (β) and transitivity (γ).

Final Thoughts

The availability of multiple observations of a network over time is potentially very informative.

Fruitful to compare the relative frequency of certain sequences of link formation for a given pair, holding the link history of other pairs fixed.

Final Thoughts (continued)

'Fixed effect' identification analysis can also help formulate more realistic random effects models (cf., Goldsmith-Pinkham and Imbens, 2013).

Computational challenge: efficient algorithm to find all stable dyads.

Covariates, efficiency questions, empirical application...