Microfoundations of Link Formation in Supply Networks

Kashish Arora
CUHK Business School, The Chinese University of Hong Kong, Hong Kong, kashisharora@cuhk.edu.hk

Nikolay Osadchiy
Goizueta Business School, Emory University, Atlanta, GA, USA. nikolay.osadchiy@emory.edu

Supply networks are endogenous constructs that are shaped by firms' strategic decisions under economic incentives and frictions. Credible empirical research on non-experimental networks, therefore, requires carefully modeling how supply networks form. In this paper, we examine how links form in supply networks and what implications these have for the evolution of the network. We begin by noting the challenges in estimating the drivers of network formation, including endogeneity, dependence across dyads, and the computational burden of vast graph spaces. Traditional approaches to modeling supply network formation, such as edge-independent models (e.g., Erdős–Rényi, random geometric graphs), overlook interdependencies like capacity constraints and signaling. Similarly, while preferential-attachment models introduce dependence, they lack firm agency and exogenous heterogeneity. To overcome these gaps, we develop a dynamic, random-utility framework that (a) microfounds link formation in exogenous (productivity, geography, size) and endogenous (local structure, homophily) factors, (b) allows correlated decisions across dyads, and (c) is estimated on temporal network data from FactSet Revere. We employ a dyad-level logit with correlated random effects estimated on 170 million dyads, network instruments, and a structural dynamic model that distinguishes between formation/dissolution to address selection, endogeneity, and non-independence, respectively. Our results show that both supplier and buyer productivity raise formation probabilities, with buyer productivity being more influential, consistent with the link initiation being primarily buyer-driven and more likely after buyer-side shocks; familiarity via shared partners increases link formation likelihood with diminishing returns that plateau around five partners; geographic proximity aids formation; and dissimilar firm sizes are more likely to match. Using the model, we conduct counterfactual analyses (e.g., boosting supplier productivity in Arizona through state-targeted programs, such as the semiconductor push, and seeding intra-state links in California motivated by EV supply-chain initiatives), demonstrating how targeted interventions reshape network structure and enhance productivity. To our knowledge, this is the first operations-management study to estimate supply-network formation at scale, offering a tractable framework that addresses selection and endogeneity and yields actionable guidance for supply chain partner search and brokerage.

Key words: Empirical Operations Management, Supply Chain Management, Supply Networks, Link Formation, Structural Estimation, Productivity

1. Introduction

Supply networks form the backbone of modern economies. They facilitate the flow of goods, services, and information that sustain daily life, from manufacturing to distribution and beyond. With the increasing availability of detailed granular data on firm-to-firm relationships, researchers can now analyze economic and operational phenomena on supply networks at scale. For example, prior empirical work has quantified how disruptions propagate across buyer–supplier links (e.g., Carvalho et al. 2021), how diversification and redundancy shape resilience (e.g., Sheffi 2005), and how supply networks affect firm performance and risk (e.g., Wang et al. 2021). However, much of the existing empirical literature treats these networks as fixed entities, without exploring their origins. In reality, a supply network is shaped by firms' strategic decisions in response to economic incentives,

Authors acknowledge financial support for this research from their respective institutions. Authors have no financial relationships or other conflicts of interest that apply to this research.

geographic constraints, and the local network structure itself, which makes the entire network an *endogenous construct*. In observational settings lacking exogenous assignment or experimental design, this endogeneity can bias causal inferences. Hence, we must understand the mechanisms driving network formation to conduct credible empirical research. To do so, we ask a first-order question in this paper: *how do supply networks form and evolve over time?*

We begin by highlighting that estimating the drivers of supply network formation is challenging for many reasons. First, due to the high costs of collecting high-quality network data, obtaining repeated observations of a network over time is often difficult. Second, as we explain below, the linking decisions between firms can be highly correlated throughout the network, rather than being independent. Third, it is crucial that the model accurately captures meaningful patterns in the data that correspond to underlying operational and economic phenomena. Fourth, issues such as endogeneity and unobserved heterogeneity can complicate identification, as the structure of the network both influences and is influenced by unobserved factors or external shocks. Finally, even when only a single large network is observed, estimation can be computationally demanding due to the enormous number of possible graph configurations, which amounts to $2^{\binom{n}{2}}$.

A starting point for modeling network formation is the use of conditional edge-independent models, which treat dyads as independent given certain covariates. Classic examples include the Erdős-Rényi model, where links form independently with a fixed probability, yielding thin-tailed degree distributions and negligible clustering. Random geometric graphs incorporate geography into these models by placing nodes in a metric space and connecting them only if their distance is below a specified threshold. However, conditional on locations, they still assume independent dyads. As a result, these models overlook key features of supply networks, including interdependent payoffs from capacity and common partners, as well as tiered and clustered structures. To illustrate this dependence, suppose Toyota establishes a long-term contract with Denso for fuel valves. Once Denso commits its production capacity to Toyota, it may have fewer resources to serve other automakers, such as Honda or Nissan, which may steer their sourcing decisions toward alternative suppliers. At the same time, Toyota's endorsement signals that Denso meets stringent quality and reliability standards, which can make smaller firms more likely to initiate links with Denso despite tighter terms. Thus, the single link between Toyota and Denso both constrains Denso's downstream choices and reshapes the incentives of other firms to partner with Denso, illustrating the interdependency in link formation decisions in supply networks—a feature that conditional edge-independent models miss.

The preferential attachment model (Barabási and Albert 1999) relaxes conditional edge independence by allowing new nodes to attach to existing ones with a probability proportional to their current degree, such that each new edge immediately alters future link probabilities. While being

parsimonious and elegant, preferential attachment treats link formation in a recursive manner, where the network at a previous period predicts which links are formed next. As such, it lacks input from exogenous factors that could provide insight into the mechanisms and actions through which nodes shape their network neighborhoods. Moreover, it overlooks agency: nodes have no control over the number of links they have, and the evolution of their connectivity is largely predetermined. This recursive probability adjustment, while introducing dependence, fails to capture the multifaceted interdependencies in supply networks, where edges evolve among existing nodes influenced by firm-specific attributes, such as productivity or geography, rather than solely by degree-based accumulation.

To overcome these gaps, we develop a dynamic model of network formation that (a) is microfounded in exogenous and endogenous factors, (b) accounts for correlation between linking decisions across dyads, and (c) is estimated on multiple temporal observations of a large supply network constructed from the FactSet Revere dataset. The key microfoundations of the model are: (i) productivity, (ii) local network structure, and (iii) homophily, which we detail in §1.1 below. Firms are allowed to trade these off in a random-utility framework. We estimate the effect of these factors on the likelihood of link formation, explicitly addressing empirical issues of sample selection, endogeneity, and non-independence. Using the estimated parameters, we conduct counterfactual analyses to evaluate how raising supplier productivity or seeding intra-state ties in targeted locations reshape network clusters, state-to-state flows, and buyers' access to supplier TFP. Before discussing our models and results in detail, we give an overview of the mechanisms we consider.

1.1. Link Formation Mechanisms and the Literature

Productivity: The first mechanism we consider is productivity. Supplier productivity can significantly impact link formation because more productive suppliers offer superior economic value to buyers through cost efficiency, higher quality, innovation, and reliability. Empirical evidence shows that highly productive suppliers are chosen by a disproportionate number of buyers (Bernard et al. 2018, Lim 2018). Buyer productivity is just as important: as buyers scale, more suppliers may be required to support their increased level of output. This results in a broader and more geographically diverse supplier base (Bernard et al. 2018), which in turn fosters greater organizational capabilities and stronger management expertise, enabling buyers to effectively coordinate a larger number of suppliers.[1] An interesting finding in the literature is that every buyer-supplier relationship requires at least one large or productive partner (Blum et al. 2010, Bernard and Moxnes 2018). Overall, productivity drives link formation by enhancing economic value and facilitating network expansion within supply chains.

[1] These buyers often manage extensive supplier networks and frequently source internationally, capitalizing on profit margins in global sourcing, thus contributing to a virtuous cycle of productivity (Antràs et al. 2017).

Degrees: Next, we consider node degrees, which play a role in our framework as proxies for visibility, scale, and other unobserved factors driving network formation. Building on the seminal preferential attachment model, Bianconi and Barabási (2001) introduced a generalized version that incorporates the concept of "fitness", transitioning from the idea that "the big get bigger" to "the fit get bigger". The notion of fit can be made concrete in a specific context. For example, a recent random graph model of supply networks by **?** proposes specialization as a determinant of link formation. In our context, fitness is a combination of observable attributes, such as productivity, industry segment, location, and unobservable ones, captured by degree. For instance, suppliers with high out-degrees (many downstream buyers) can attract additional buyers through their perceived value, reliability, and economies of scale. Likewise, buyers with large in-degrees (many upstream suppliers) expand their supply bases to sustain operations, benefiting from enhanced network visibility and established sourcing practices.

Homophily: Homophily, or the tendency for similar nodes to connect, is another key determinant of link formation. It can remedy limitations in the preferential attachment model (Barabási and Albert 1999) by incorporating exogenous, similarity-based preferences into linking probabilities, thereby introducing agency and real-world factors. Firms often prefer partners from the same country or industry due to factors such as cultural alignment, reduced communication barriers, compatible business practices, and shared operational scales (Currarini et al. 2009, Serpa and Krishnan 2018). Geographical proximity also fosters homophily, as increased distance can raise frictions and search costs.[2] Empirical studies show that firms are more likely to form supply relationships with nearby partners (Bernard et al. 2018), or those close to existing connections, promoting regional clusters and localized supply chains (Baldwin and Venables 2013). Prior research also suggests assortative matching on scale, as similarly sized firms tend to align in terms of bargaining power, quality expectations, and capacity. At the same time, the presence of complementarities can lead to size heterophily, where larger buyers seek to pool resources with smaller suppliers or smaller buyers look to larger firms for scale. Therefore, whether supply networks exhibit size homophily or heterophily is an empirical question that requires careful investigation.

Transitivity: Transitivity occurs when two firms share common partners, such as suppliers or buyers, or intermediaries (firms that connect them by buying from one and selling to the other), forming an undirected triadic closure upon linking. When firms have shared connections, they often develop direct ties because these common partners help build trust, facilitate better information exchange, and reduce transaction costs. As a result, this accelerates the formation of links between firms that

[2] These costs may include higher transportation costs, logistics complexity, longer lead times, and reduced communication efficiency.

are indirectly connected (Atalay et al. 2011, Chaney 2014). We incorporate this key network feature into our model to capture clustering tendencies. We highlight that transitivity also introduces dyadic dependence, where the probability of one link can influence others through shared partners, thereby addressing interdependence in the formation of networks.

1.2. Models and Results

Building on this literature, we investigate how productivity, local network structure, homophily, and geography influence link formation, while accounting for degree effects and partner commonality. While prior literature has examined these factors individually or in pairs, such as the role of productivity in attracting partners, our work is novel in combining them to reveal interpretable trade-offs (which we describe later) and resolve empirical tensions, including those regarding size homophily versus heterophily. Before moving on to estimation, we first construct a measure of firm-level total factor productivity (TFP) using the Olley and Pakes (1996) method. We restrict the supply network to firms with observed TFP, merging these firms with the FactSet Revere relationships data to create a *dyad-complete panel*. Our analysis proceeds in three steps. In the first step, we estimate a dyad-level logit model, which is a conditional edge-independent model, on the set of all possible dyads for each period. This exhaustive set of dyads allows us to circumvent the sample selection problem. Overall, we analyze approximately 170 million dyads. Given the significant heterogeneity among dyads, we include dyad-level correlated random effects (CRE) as well as time effects in the model. The CRE control for unobserved time-invariant dyad-level heterogeneity while maintaining computational feasibility on this large dataset. Additionally, CRE allow for the random effects to correlate with covariates, providing a more flexible modeling approach for dyad-specific factors.

While the dyad-level logit accounts for sample selection and heterogeneity, it does not address potential endogeneity in productivity due to unobserved time-varying factors simultaneously influencing productivity and link formation. Hence, to address remaining endogeneity, we use network instruments inspired by Bramoullé et al. (2009). Specifically, we instrument TFP with lagged characteristics from second-degree neighbors, such as the average productivity of a supplier's tier-2 suppliers and a buyer's tier-2 customers. These instruments are relevant for firm productivity through network spillovers and are plausibly exogenous to dyad-specific, time-varying shocks.

Finally, we develop a structural model of network formation that goes beyond traditional dyad-wise conditional edge-independent models, treating the entire directed buyer-supplier network as the subject of analysis. Our model differentiates between the processes of link formation for absent ties and link dissolution for existing ties. It takes into account dyad attributes, such as frictions and similarities, as well as firm fundamentals like productivity, size, and local network features, including

transitivity through shared partners. We treat network transitions as conditionally independent formation and dissolution processes given prior network structure and covariates, allowing for interdependence where one link's formation can influence others' utilities. The model is estimated at the network level rather than the dyad level, highlighting why reduced-form approaches, such as standard logit, are insufficient as they overlook network dependence. Since the likelihood of transition is defined jointly over the entire network, rather than as a product of independent dyads, we estimate parameters using simulation-based maximum likelihood.

We establish the following facts. Results that are new in the paper (to the best of our knowledge) are marked with an asterisk.

1. *Both supplier and buyer productivity increase the probability of link formation, but buyer productivity has a higher impact on the probability of forming a link with a given supplier.
2. *Link formation more likely follows buyer-side productivity shocks, consistent with buyers initiating search, vetting, and onboarding.
3. Familiarity established through the presence of common partners increases the likelihood of link formation, *with diminishing marginal returns that level off once there are approximately five shared partners.
4. Geographical proximity positively affects the likelihood of link formation, *while dissimilar firm sizes promote connections.
5. Firms derive utility from homophily or distance proximity, which they trade off for productivity. *The reduction of geographical distance by 10% derives the same utility as an increase in suppliers' productivity by 28.8%.
6. Degree-based popularity (preferential attachment) terms remain predictive of formation even after accounting for productivity, homophily, distance, and local structure, indicating additional mechanisms beyond those we model explicitly.

Our structural model enables us to perform counterfactuals to evaluate the impact of targeted policies on network evolution. To do so, we hold the estimated model parameters fixed and simulate the network's evolution under targeted policy changes to covariates or link eligibility. We study two policies. First, a productivity boost for all suppliers in Arizona, mimicking state-targeted programs like the semiconductor push. Second, a state-sponsored seeding program that inserts a small set of in-state links among high-connectivity firms in California, motivated by California's EV supply-chain push. We find that for Arizona, the localized shock rewires the entire national supply network as enhanced supplier attractiveness prompts buyers to reoptimize sourcing toward complementary, tech-heavy states like Virginia, crowding out less overlapping states like Ohio. Interestingly, we also find that gains in the number of buyers are highly concentrated among well-embedded Arizona

suppliers due to triadic closure and homophily amplifying uneven propagation. For California, we find that a small amount of careful local seeding can foster dense intra-state clusters with new edges and triangles touching the state. The policy can also lead to increased productivity access for buyers as more local ties connect to high-productivity suppliers, without the adverse effects of out-of-state spillovers.

We contribute to the literature on network formation and productivity. To the best of our knowledge, our paper is the first to empirically study supply network formation in the Operations Management (OM) literature. Methodologically, we develop an estimable framework that addresses selection into observed links and the endogeneity of productivity by combining correlated random effects, network-based instruments, and a dynamic network model with separate formation and dissolution processes. Structurally, we analyze link creation and dissolution at the network level, allowing a single relationship to shift payoffs beyond the focal dyad through local dependence, such as shared partners, which the model's temporal dynamics capture. Substantively and managerially, our findings reveal that buyer productivity has a more significant impact on link formation than supplier productivity. This supports the idea that buyers are typically the ones initiating searches and onboarding new partners. Additionally, we find that geographic distance and a lack of familiarity create substantial frictions, causing buyers to overlook high-capability matches. Interestingly, a small number of shared partners provides most of the benefits related to familiarity. We also observe that links tend to form disproportionately between firms of different sizes, indicating size heterophily, which aligns with the principles of complementarities in capacity and risk sharing. Our results suggest that firms can unlock capability and productivity gains by screening for complementary scale, by selling to or sourcing from nodes embedded in clusters of potential partners, and by reducing search and distance-related frictions.

2. Data and Measurement: Productivity Estimation and Network Construction

2.1. TFP Estimation

To estimate firm-level Total Factor Productivity (TFP), we adopt the Olley and Pakes (1996) framework, which addresses simultaneity and selection biases in production function estimation. We consider a panel of firms indexed by $i \in \{1, 2, \ldots, N\}$, observed over years $t \in \{1, 2, \ldots, T\}$. We assume that firms convert inputs to outputs using the following Cobb-Douglas production function:

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_k k_{it} + \omega_{it} + \epsilon_{it} \tag{1}$$

where y_{it} is the log of output (e.g., sales or value-added) for firm i at time t, l_{it} is the log of labor input (e.g., employment), k_{it} is the log of capital, ω_{it} is productivity, observed by the firm, but not researchers, and ϵ_{it} is the idiosyncratic error term.

A note on endogeneity: The estimation of (1) via OLS is problematic because it suffers from (a) the simultaneity problem, and (b) the selection bias. The simultaneity problem exists because firms observe ω_{it} when choosing l_{it}, leading to a correlation between the labor input and ω_{it}. Selection bias exists because low-productivity firms may exit the market, resulting in a non-random sample of surviving firms. These issues make the OLS estimator of β_l and β_k biased and inconsistent. We follow the approach of Olley and Pakes (OP, 1996) to correct for these biases. In what follows, we first list down the standard assumptions made in the production function literature and then describe our estimation procedure.

Assumption 1 (OP, Productivity Dynamics). *Firm-level productivity follows a first-order Markov process, i.e.,* $\omega_{it} = g(\omega_{i,t-1}) + \xi_{it}$*, where* $g(\cdot)$ *is an unknown function capturing the persistence of productivity, and* ξ_{it} *is an innovation term uncorrelated with past inputs.*

Assumption 2 (OP, Investment Proxy). *Investment is determined by a function of capital and productivity, such that,* $i_{it} = f_t(k_{it}, \omega_{it})$*, where* k_{it} *is the capital stock. The function* f_t *is strictly increasing in* ω_{it}*, allowing inversion to express productivity as* $\omega_{it} = h_t(i_{it}, k_{it})$*.*

Assumption 3 (OP, Input Timing). *Labor and capital inputs have distinct adjustment dynamics. Labor,* l_{it}*, is a flexible input, chosen by firm* i *at time* t *after observing productivity* ω_{it}*. Capital,* k_{it}*, is a fixed input, predetermined at time* t*. The capital stock evolves according to:*

$$k_{i,t+1} = (1 - \delta)k_{it} + i_{it}$$

where δ *is the depreciation rate, and* i_{it} *is investment.*

Assumption 4 (OP, Exit Decision). *Firms make an exit decision based on their productivity relative to a threshold. Firm* i *survives at time* t *(i.e.,* $\chi_{it} = 1$*) if its productivity satisfies:*

$$\omega_{it} \geq \bar{\omega}_t(k_{it})$$

Otherwise, the firm exits ($\chi_{it} = 0$*), where* $\bar{\omega}_t(k_{it})$ *is a productivity threshold that depends on the firm's capital stock* k_{it}*.*

Estimation Procedure: Our estimation proceeds in the following two steps:

Step 1: Controlling for Simultaneity: In step 1, we use the investment proxy to express productivity as $\omega_{it} = h_t(i_{it}, k_{it})$. Substituting into (1), we get:

$$y_{it} = \beta_l l_{it} + \phi_t(i_{it}, k_{it}) + \epsilon_{it} \tag{2}$$

where $\phi_t(i_{it}, k_{it}) = \beta_0 + \beta_k k_{it} + h_t(i_{it}, k_{it})$. Then, we estimate (2) non-parametrically, approximating $\phi_t(i_{it}, k_{it})$ with a flexible polynomial in i_{it} and k_{it}. This yields $\hat{\beta}_l$ and $\hat{\phi}_t$. This step corrects for simultaneity by proxying ω_{it} with investment, addressing the endogeneity of labor.

Step 2: Estimating the Capital Coefficient and Productivity: From step 1, we obtain $\hat{\phi}_t = \hat{\beta}_0 + \beta_k k_{it} + \hat{\omega}_{it}$. Thus, productivity can be expressed as $\hat{\omega}_{it} = \hat{\phi}_t(i_{it}, k_{it}) - \hat{\beta}_0 - \beta_k k_{it}$. To estimate β_k, we use the Markov assumption for productivity and model expected productivity as:

$$E[\omega_{it}|\omega_{i,t-1}, \chi_{it} = 1] = g(\omega_{i,t-1}, P_{it})$$

where P_{it} is the probability of survival, estimated via a probit model:$P_{it} = \Pr(\chi_{it} = 1|i_{i,t-1}, k_{i,t-1})$. Next, we use the following moment condition:

$$E[\xi_{it}|k_{it}, \chi_{it} = 1] = 0,$$

where $\xi_{it} = \omega_{it} - E[\omega_{it} \mid \omega_{i,t-1}, \chi_{it} = 1]$. Finally, we regress $y_{it} - \hat{\beta}_l l_{it} = \beta_k k_{it} + g(\hat{\omega}_{i,t-1}, \hat{P}_{it}) + \xi_{it} + \epsilon_{it}$ to estimate β_k. This step corrects for the selection bias by modeling survival, thus ensuring that β_k is unbiased. Once the coefficients $\hat{\beta}_l$ and $\hat{\beta}_k$ are obtained, TFP is computed as the residual:

$$w_{it} = y_{it} - \hat{\beta}_l l_{it} - \hat{\beta}_k k_{it} \tag{3}$$

which represents the firm-specific productivity after accounting for the contributions of labor and capital.

2.1.1. Summary statistics and estimated TFP: To estimate TFP, we compile firm fundamentals from Compustat annual files. Our sample contains 20,033 unique firms and 122,649 firm–year observations spanning 1971 to 2023, with a coverage of 66 two-digit and 396 four-digit SIC industries. The working dataset keeps firm IDs (`gvkey`, `cusip`), industry codes (`sic`, `naics`), calendar year (`yr`), and the variables required for the Olley–Pakes procedure: sales (`sale`), employment (`emp`), property, plant, and equipment (`ppent`, `ppegt`), acquisitions and capital expenditures (`aqc`), a constructed capital stock (`capstock`), net investment (`netinv`), and price deflators for output and investment (`gdpdefl`, `nrinvdefl`). We deflate sales with `gdpdefl` and investment with `nrinvdefl`; construct `capstock` via a perpetual inventory update from `ppent` and `ppegt`; and use the log of the inputs `lsale`, `lemp`, `lcapstock`, and `lnetinv`, dropping observations with nonpositive output, labor, capital, or investment. TFP is the OP residual from the Cobb–Douglas specification described in §2.1, with labor as the freely varying input, capital as the state variable, and investment as the proxy. Table 1 reports summary statistics of the sample, the logged inputs, and the estimated TFP, which has a mean of around 4.51 and a standard deviation of 0.96.

2.2. Construction of Supply Networks

We build yearly directed supply networks from the FactSet Revere supply chain relationships database, which records supply chain relationships with start and end dates using multiple public

Table 1 Descriptive statistics for TFP and OP inputs

Variable	Mean	SD	p10	p50	p90
log Capital stock	3.265	2.792	-0.240	3.103	7.083
log Employment	-0.400	2.393	-3.507	-0.411	2.797
log Net investment	1.423	2.677	-2.054	1.406	4.970
log Sales	4.768	2.705	1.287	4.874	8.189
TFP (OP residual)	4.508	0.964	3.434	4.578	5.564

Table 2 Coverage of Full Sample

Metric	Value
Unique firms	20,033
Firm-years	122,649
Years covered	1971 to 2023
SIC 2-digit groups	66
SIC 4-digit industries	396

sources, such as public filings, investor presentations, press releases, and other public sources. Its coverage is broad across firms, countries, and sectors and is widely used to study supply networks in operations (Osadchiy et al. 2021, Agrawal and Osadchiy 2024, **?**).

We restrict our network to the set of firms with estimated TFP and a valid CUSIP.[3] The TFP availability requirement is reasonable as it implies that firms have sufficient information about their potential supply chain partners. We construct a dyad-year-level panel from 2003 through 2023. We mark a supplier→customer link as active in year t if any day between its start and end dates falls in year t. For each year t, we generate the full set of dyads (i, j) among the firms in the productivity sample for that year and set:

$$\texttt{link}_{ij,t} = \begin{cases} 1 & \text{if FactSet reports a supplier} \to \text{customer relationship from } i \text{ to } j \text{ in } t, \\ 0 & \text{otherwise.} \end{cases}$$

This yields a *dyad-complete panel* that includes both realized and non-realized links. In other words, if there are N_t firms with productivity data in year t, we have a total of $\sum_{t \in T} N_t (N_t - 1)$ observations in our sample over T years.

We use both node and dyad-level covariates in our analyses. At the node level we include supplier and customer productivity. At the dyad level we include: (i) homophily indicators $\mathbf{1}\{x_{i,t}^{(m)} = x_{j,t}^{(m)}\}$ for categorical attributes, for example same industry or same country; (ii) dissimilarity in continuous attributes such as size or capital;[4] (iii) degree measures computed for year t: the supplier's out-degree and the customer's in-degree,

$$\text{outdeg}_{i,t} = \log\Big(1 + \sum_{k \neq i} \text{link}_{ik,t}\Big), \qquad \text{indeg}_{j,t} = \log\Big(1 + \sum_{k \neq j} \text{link}_{kj,t}\Big);$$

(iv) log of the geographical separation between the firms' headquarters using the distance formula by Vincenty (1975), represented as d_{ij}; and (v) transitivity via the presence of undirected two–paths from i to j,

$$\text{Common Partners}_{ij,t} = \mathbf{1}\{\text{link}_{(ik)|(ki),t} = 1 \text{ and } \text{link}_{(kj)|(jk),t} = 1\}, \text{ for some } k.$$

[3] By using CUSIPs (6 character), we follow the prior literature in emphasizing relationships disclosed by U.S. firms to keep the coverage consistent through time.

[4] We use absolute differences after a monotone transform $z(\cdot)$, $|\, z(x_{i,t}^{(m)}) - z(x_{j,t}^{(m)})\,|$

Table 3 Descriptive statistics for supply-network variables.

	N	Mean	SD	p10	p50	p90
Continuous & count variables						
Supplier TFP	170,370,907	4.714	1.136	3.539	4.836	5.832
Customer TFP	170,370,907	4.714	1.136	3.539	4.836	5.832
Supplier out-degree (log)	170,370,907	0.467	0.733	0.000	0.000	1.609
Customer in-degree (log)	170,370,907	0.420	0.715	0.000	0.000	1.386
Common partners	170,370,907	0.076	0.416	0.000	0.000	0.000
Abs. diff. log Capital	170,370,907	3.180	2.384	0.503	2.710	6.514
Abs. diff. log Assets	170,370,907	2.642	2.019	0.412	2.222	5.453
Log Distance	115,826,730	7.303	1.232	5.928	7.540	8.326
Binary variables						
Link	170,370,907	0.001	0.031			
Same industry	170,370,907	0.038	0.192			
Same country	170,370,907	0.646	0.478			

Table 3 reports the descriptive statistics for all variables used in the analysis over the full dyad–complete panel. The total size of the dataset is around 170 million observations. We note that the network is sparse: the link indicator averages 0.001, about 0.1% of dyads. The mean supplier and customer TFP are both 4.714, and the average log distance is 7.302, which corresponds to roughly 1,485 km.

Remark 1. Because we focus on economically and managerially relevant covariates, such as TFP, the availability of data dictates that our observations come mostly from large public firms, about 3000 firms per year. Even with this restriction, our dyad-complete panel includes more than 170 million observations and is one of the largest firm-to-firm dyadic datasets analyzed in the empirical operations and trade literature. Expanding it to the whole FactSet universe with 40,000 firms per year on average would result in a data set with more than 30 billion observations, which would be borderline infeasible to analyze.

Remark 2. Tariffs are relatively stable during our sample period. As supply chains rewire in response to recent trade policy uncertainty and that data becomes available, bilateral tariffs can be incorporated in our model.

3. Reduced Form Model with Correlated Random Effects

We model the formation of directed supplier→customer links over time. Let i index suppliers, j index customers, and $t \in \{1, 2, \ldots, T\}$ denote years. Let $g_{ij}^t \in \{0, 1\}$ be a binary variable (an element of the year t adjacency matrix g^t) equal to one if a directed link from supplier i to customer j is observed in year t, and zero otherwise. We assume that the link formation is governed by an unobserved latent variable y_{ijt}^*, which captures the following net utility of the supplier–customer match in a given period:

$$y_{ijt}^* = \mathbf{x}_{ijt}'\boldsymbol{\beta} + \alpha_{ij} + \delta_t + \varepsilon_{ijt}, \tag{4}$$

where, $\mathbf{x}_{ijt} \in \mathbb{R}^K$ is a vector of observed time-varying dyad-level covariates. These covariates include supplier and customer TFP, network centrality measures such as supplier out-degree and customer

in-degree, a transitivity indicator whether the link would complete a triad or not (i.e., whether there are common suppliers or customers to the nodes), homophily indicators for the same industry and the same country, and the logarithm of geographic distance between the firm's headquarters. The term δ_t denotes year fixed effects, and α_{ij} represents unobserved dyad-specific effects that capture time-invariant heterogeneity in the underlying match quality. The error term ε_{ijt} is assumed to follow a standard logistic distribution, yielding a logit specification. The resulting observed link outcome is:

$$g_{ij}^t = \begin{cases} 1 & \text{if } y_{ijt}^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Accordingly, the probability of observing a link between supplier i and customer j at time t is:

$$\Pr\big(g_{ij}^t = 1 \mid \mathbf{x}_{ijt}, \alpha_{ij}, \delta_t\big) = \Lambda\big(\mathbf{x}_{ijt}'\beta + \alpha_{ij} + \delta_t\big)\,,$$

where $\Lambda(u) = \dfrac{e^u}{1+e^u}$ is the logistic CDF.

We highlight that accounting for dyad-specific heterogeneity α_{ij} is important in our context. Specifically, the term α_{ij} captures persistent, time-invariant *unobserved* attributes of the supplier-customer pair that are not captured in $\mathbf{x}_{ijt}$. These attributes may include strategic fit between the partners, accumulated relationship capital, legacy contracts, or organizational frictions. These unobservables plausibly affect link formation and may be correlated with observed covariates in $\mathbf{x}_{ijt}$. For instance, dyads that share several common partners or operate in the same industry may have unobserved compatibility in standards, contracting routines, or information flows, inducing correlation between α_{ij} and the common-partners and homophily covariates in $\mathbf{x}_{ijt}$.

The fact that α_{ij} is correlated with $\mathbf{x}_{ijt}$ has direct consequences for identification of $\boldsymbol{\beta}$. A common baseline is a random-effects logit that models α_{ij} as i.i.d. $\mathcal{N}(0, \sigma_\alpha^2)$ and imposes $\mathbb{E}[\alpha_{ij} \mid \mathbf{x}_{ij1}, \ldots, \mathbf{x}_{ijT}] = 0$. This specification is consistent and efficient only if the orthogonality condition holds. However, since this condition is not met in our application, random-effect estimates of $\boldsymbol{\beta}$ would be inconsistent. Fixed effects (FE) estimation solves this issue in linear models by differencing out α_{ij} via within-transformation. However, in nonlinear models like logit, introducing a parameter α_{ij} for each dyad creates an incidental-parameters problem: as the number of dyads N_{IJ} grows large with T fixed, the maximum-likelihood estimator of $\boldsymbol{\beta}$ is biased and inconsistent (Neyman and Scott 1948, Hahn and Newey 2004, Fernández-Val and Weidner 2016). This concern is first-order here, since N_{IJ} in our application is of the order of 10^7, while T is modest (around 5). In addition to the incidental parameters problem, fixed effect models eliminate all between-dyad variation, which is often substantively meaningful in link formation settings, e.g., systematic differences in match quality across dyads due to long-term characteristics.

To address these limitations of random- and fixed-effects models, we adopt a Correlated Random Effects (CRE) specification in the spirit of Chamberlain (1984). Specifically, rather than imposing that α_{ij} be uncorrelated with $\mathbf{x}_{ijt}$, we explicitly model the dyad effects as a linear function of the dyad-level means of the covariates:

$$\alpha_{ij} = \overline{\mathbf{x}}_{ij}'\boldsymbol{\gamma} + a_{ij}, \tag{5}$$

where $\overline{\mathbf{x}}_{ij} = T_{ij}^{-1}\sum_{t=1}^{T_{ij}} \mathbf{x}_{ijt}$ is the dyad-specific time average computed over the observed periods, $\boldsymbol{\gamma} \in \mathbb{R}^K$ captures how long-run dyad characteristics affect match quality, and a_{ij} is a residual dyad effect with zero mean conditional on $\mathbf{x}_{ijt}$ and $\overline{\mathbf{x}}_{ij}$, i.e. $\mathbb{E}[a_{ij} \mid \mathbf{x}_{ijt}, \overline{\mathbf{x}}_{ij}] = 0$. Substituting (5) into (4) yields

$$y_{ijt}^{*} = \mathbf{x}_{ijt}'\boldsymbol{\beta} + \overline{\mathbf{x}}_{ij}'\boldsymbol{\gamma} + \delta_t + a_{ij} + \varepsilon_{ijt}, \tag{6}$$

where ε_{ijt} follows a standard logistic distribution. The conditional link probability is therefore

$$\Pr\big(g_{ij}^{t} = 1 \mid \mathbf{x}_{ijt}, \overline{\mathbf{x}}_{ij}, \delta_t, a_{ij}\big) = \Lambda\big(\mathbf{x}_{ijt}'\boldsymbol{\beta} + \overline{\mathbf{x}}_{ij}'\boldsymbol{\gamma} + \delta_t + a_{ij}\big), \tag{7}$$

and the likelihood integrates over the distribution of a_{ij} (for example $a_{ij} \sim \mathcal{N}(0, \sigma_a^2)$). We provide more details of the estimation procedure in E-Companion EC.1.

The CRE formulation has several advantages. First, it preserves within-dyad variation through $\mathbf{x}_{ijt}$ and between-dyad variation through $\overline{\mathbf{x}}_{ij}$, allowing us to estimate the effects of short-run changes (via $\boldsymbol{\beta}$) and persistent cross-dyad differences (via $\boldsymbol{\gamma}$). This is in contrast to FE models, which discard all between-dyad information. Second, it avoids the incidental parameters problem because it does not estimate a separate α_{ij} for each dyad. Finally, it relaxes the strong orthogonality requirement of a standard random-effects logit.[5]

3.1. Reduced-form Results

We estimate model (7) via maximum likelihood on the dyad-complete panel, constructed from the period-t adjacency matrices. We highlight that the estimation is computationally intensive due to the large network size, requiring careful data management, over 200 GB of RAM, and approximately 12 hours on a high-performance cluster with 16 CPU cores utilizing parallelization. Additionally, a direct computation of robust standard errors is challenging due to prohibitive memory requirements. Therefore, we compute the standard errors from the observed information matrix (OIM) and estimate robust standard errors by comparing the robust and OIM standard errors on smaller samples, where such computations are feasible. For further details, please refer to EC.2. The estimates are reported in Table 4. A joint test on the coefficients of the dyad-level

[5] Moreover, the explicit inclusion of dyad-level means enables formal testing of the RE assumption: if the coefficients on $\overline{\mathbf{x}}_{ij}$ are jointly insignificant, the RE model may be appropriate. Thus, CRE not only yields consistent estimates in the presence of correlated heterogeneity but also nests the RE model as a testable special case.

Table 4 Logit model with dyadic CRE and year FE, OIM SE in parentheses; estimated robust SE clustered by dyad in brackets. Statistical significance is evaluated based on the robust SE.

	$\mathbb{1}(link)$	$\mathbb{1}(link)$	$\mathbb{1}(link)$
Supplier TFP	0.1986***	0.1312***	0.1049***
	(0.0094) [0.0113]	(0.0131) [0.0171]	(0.0136) [0.0185]
Customer TFP	0.1457***	0.0763***	0.0392**
	(0.0105) [0.0119]	(0.0145) [0.0177]	(0.0149) [0.019]
Supplier Outdegree		1.5985***	1.5407***
		(0.0072) [0.0101]	(0.0074) [0.0106]
Customer Indegree		1.0076***	0.9665***
		(0.0071) [0.0081]	(0.0073) [0.0093]
Common Partners		0.8183***	0.6979***
		(0.0134) [0.0185]	(0.0135) [0.0194]
Same Industry			3.1198***
			(0.0281) [0.0412]
Same Country			-0.6723***
			(0.0324) [0.0522]
ln(Distance)			-0.3269***
			(0.0057) [0.0082]
Abs.Diff ln(Capital)			-0.0034*
			(0.0094) [0.014]
Abs.Diff ln(AT)			0.0769***
			(0.0107) [0.0162]
$\ln \sigma_a^2$	2.8911***	3.8157***	3.8788***
	(0.0011) [0.0065]	(0.0021) [0.005]	(0.0024) [0.0064]
Constant	-15.9119***	-27.4578***	-25.1531***
	(0.0681) [0.0871]	(0.0947) [0.1367]	(0.1152) [0.185]
Observations	170,370,907	170,370,907	115,826,730
Number of dyad	36,885,324	36,885,324	29,460,550

*** p<0.01, ** p<0.05, * p<0.1

means $\overline{\mathbf{x}}_{ij}$ rejects the random-effects restriction, supporting the CRE specification. We build the model sequentially, starting with productivity measures, then adding network structure, and finally incorporating homophily covariates. The results show that both supplier and customer TFP are positively associated with link formation probability. The supplier out-degree, customer in-degree, and measures of shared partners are all strongly and positively associated with the link formation probability. Firms in the same industry are more likely to form a relationship. The coefficient on the absolute difference in total assets is positive, indicating disassortative matching by size: pairs with more disparate asset levels are more likely to link, consistent with Blum et al. (2010).

3.2. Endogeneity

Our reduced-form model incorporates dyad-specific unobserved heterogeneity using a Correlated Random Effects (CRE) specification. CRE addresses endogeneity that stems from time-invariant dyad effects α_{ij} being correlated with the covariates by parameterizing α_{ij} as a function of dyad-level covariate means. However, additional sources of endogeneity can remain if there are *time-varying unobservables* or simultaneity. In particular, if the period-t shock satisfies

$$\mathbb{E}[\varepsilon_{ijt} \mid \mathbf{x}_{ijt}, \overline{\mathbf{x}}_{ij}, \delta_t] \neq 0 \quad \text{(which implies, } \mathrm{Cov}(\varepsilon_{ijt}, \mathbf{x}_{ijt}) \neq 0),$$

then the CRE estimator of $\boldsymbol{\beta}$ is biased. Such correlation can arise from: (i) reverse causality, where link formation (or its anticipation) shifts firm-level variables like productivity; (ii) time-varying

omitted variables, for example contract renegotiations or local shocks that move both $\mathbf{x}_{ijt}$ and link propensity; and (iii) network spillovers, where links co-determine one another within period t.

This concern is closely related to the endogeneity of network structure discussed by Goldsmith-Pinkham and Imbens (2013). In their framework, the network is treated as a regressor in peer-effects models, while we model the network as the outcome. Nevertheless, the shared challenge is the simultaneous determination of links and covariates due to unobserved shocks, which necessitates restrictions beyond time-invariant dyad heterogeneity. In our case, while CRE accounts for $\text{Cov}(\alpha_{ij}, \mathbf{x}_{ijt}) \neq 0$ via $\overline{\mathbf{x}}_{ij}$, it does not eliminate $\text{Cov}(\varepsilon_{ijt}, \mathbf{x}_{ijt}) \neq 0$, so additional timing assumptions, valid instruments, or a dynamic or structural network specification are needed.

3.3. Instruments

To address potential endogeneity in our reduced-form model analysis, we implement an instrumental variables (IV) strategy introduced by Bramoullé et al. (2009). Specifically, we use the lagged network structure and characteristics of neighbors-of-neighbors (firms that are two links away in the directed graph) for endogenous covariates, such as supplier and customer TFP.

Specifically, let $g^t \in \{0,1\}^{N\times N}$ denote the time-t directed adjacency matrix of the supply network, where $g^t_{ij} = 1$ if firm i supplies to firm j, and 0 otherwise. For a given firm i, the second-order neighborhood at time $t-1$ is defined by the nonzero entries of $h^{t-1} \equiv (g^{t-1})^2$, with $h^{t-1}_{ki} > 0$ indicating that firm k is a supplier to a supplier of firm i at $t-1$. Thus, the matrix h^{t-1} captures indirect influence and serves as the basis for our instruments. For supplier i, the instrument for covariate x_{it} (e.g., TFP) is the average of $x_{k,t-1}$ across all firms k that are two steps away from i at time $t-1$:

$$z_{it} = \frac{1}{\left|\mathcal{N}^{(2)}_{i,t-1}\right|} \sum_{k\in\mathcal{N}^{(2)}_{i,t-1}} x_{k,t-1}, \quad \text{where } \mathcal{N}^{(2)}_{i,t-1} \equiv \{\, k : h^{t-1}_{ki} > 0 \,\}.$$

Similarly, for customer j,

$$z_{jt} = \frac{1}{\left|\mathcal{N}^{(2)}_{j,t-1}\right|} \sum_{k\in\mathcal{N}^{(2)}_{j,t-1}} x_{k,t-1}, \quad \text{where } \mathcal{N}^{(2)}_{j,t-1} \equiv \{\, k : h^{t-1}_{jk} > 0 \,\}.$$

Using the time lag for instrument construction mitigates within-period simultaneity between the instruments and the link decision at time t. The average operator summarizes conditions in the second-order neighborhood and attenuates idiosyncratic shocks to any single firm; the resulting instrument is plausibly exogenous to the dyad-specific error ε_{ijt} because these firms are not directly connected to i or j. Figure 1 shows a graphical depiction of our instruments. The validity of our instruments relies on satisfying three key conditions: relevance, exclusion restriction, and the rank condition.

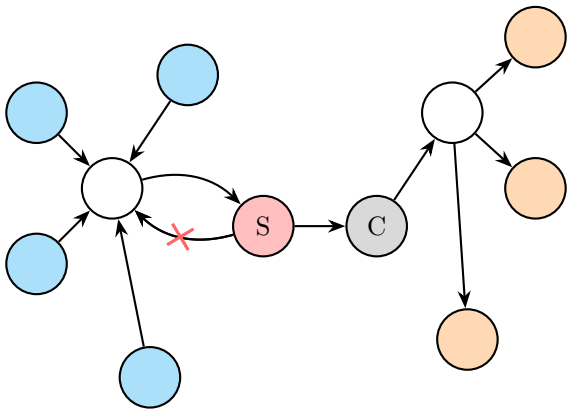


Figure 1 The figure illustrates the network surrounding the dyad S → C. The blue nodes represent the second-degree suppliers of S, which are utilized to construct supplier-side instrumental variables (IVs) for S. The orange nodes denote the second-degree customers of C, which are used to create customer-side IVs for C. Self-loops are excluded.

3.3.1. Relevance. The instruments must be correlated with the endogenous covariates, conditional on controls:

$$\mathbb{C}\mathbf{ov}\big(z_{it},\, x_{it} \,|\, \overline{\mathbf{x}}_{ij}, \delta_t\big) \neq 0, \qquad \mathbb{C}\mathbf{ov}\big(z_{jt},\, x_{jt} \,|\, \overline{\mathbf{x}}_{ij}, \delta_t\big) \neq 0.$$

This condition holds because the characteristics of second-order neighbors affect firm outcomes through network spillovers, such as diffusion, competition, or market interactions, as in Bramoullé et al. (2009).

3.3.2. Exclusion restriction condition. The instruments should be orthogonal to the dyad-specific shock in the link equation:

$$\mathbb{C}\mathbf{ov}\big(z_{it},\, \varepsilon_{ijt} \,|\, \mathbf{x}_{ijt}, \overline{\mathbf{x}}_{ij}, \delta_t, a_{ij}\big) = 0, \qquad \mathbb{C}\mathbf{ov}\big(z_{jt},\, \varepsilon_{ijt} \,|\, \mathbf{x}_{ijt}, \overline{\mathbf{x}}_{ij}, \delta_t, a_{ij}\big) = 0.$$

This is plausible because the instruments use second-order neighbors and are lagged to $t-1$, so they are not directly tied to shocks specific to the i–j link at time t. Also, note that we carefully exclude self-loops by setting $g^t_{ii} = 0$ and zeroing the diagonal of $h^{t-1} = (g^{t-1})^2$, so $i \notin \mathcal{N}^{(2)}_{i,t-1}$ and length-2 return paths $i \to \cdot \to i$ never enter the instrument.

Put simply, our instruments use lagged second-order neighbors to shift firms' productivity, which ensures relevance. Exclusion requires that, conditional on the dyad correlated random effects, year fixed effects, and rich local-structure controls (degrees, directed triadic closure/common partners, homophily, distance), these lagged tier-2 characteristics do not add any extra, pair-specific shock to the (i, j) link at time t.[6]

[6] Similar kinds of network instruments have also been used in prior OM research, please see (He et al. 2021, Arora et al. 2023).

Table 5 Logit CRE model with Bramoullé Instruments dyadic CRE and year FE, MLE estimator, First-stage results, SE clustered by dyad.

	Supplier TFP	Customer TFP	Supplier TFP	Customer TFP	Supplier TFP	Customer TFP
z^1 (Suppl-2 avg TFP)	0.0062***	0.0000	0.0061***	-0.0000	0.0054***	-0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
z^2 (Cust-2 avg TFP)	0.0003***	-0.0025***	0.0003***	-0.0026***	0.0004***	-0.0022***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
Supplier Outdegree			0.0065***	-0.0019***	0.0067***	-0.0017***
			(0.0001)	(0.0002)	(0.0001)	(0.0002)
Customer Indegree			-0.0004***	0.0028***	-0.0002***	0.0031***
			(0.0001)	(0.0000)	(0.0001)	(0.0001)
Common Partners			0.0020***	0.0020***	0.0021***	0.0020***
			(0.0002)	(0.0002)	(0.0002)	(0.0002)
Same Industry					0.0008***	0.0005***
					(0.0000)	(0.0000)
Same Country					-0.0015***	-0.0016***
					(0.0001)	(0.0001)
Abs.Diff ln(Capital)					0.0523***	0.0582***
					(0.0004)	(0.0004)
Abs.Diff ln(AT)					-0.0593***	-0.0681***
					(0.0005)	(0.0005)
ln(Distance)					0.0001***	0.0001***
					(0.0000)	(0.0000)
Constant	4.9230***	4.9636***	4.9182***	4.9633***	4.9134***	4.9580***
	(0.0012)	(0.0011)	(0.0012)	(0.0011)	(0.0017)	(0.0017)
Observations	29,740,202	29,740,202	29,740,202	29,740,202	29,546,268	29,546,268
R-squared	0.8690	0.8583	0.8690	0.8583	0.8694	0.8587

*** p<0.01, ** p<0.05, * p<0.1

3.3.3. Rank condition. Identification with second-order neighbors requires that the linear operators I, g^{t-1}, and h^{t-1} be linearly independent, where $h^{t-1} \equiv (g^{t-1})^2$. This condition ensures that the instrument adds variation beyond what is captured by individual characteristics and first-order neighbor averages. As shown in Bramoullé et al. (2009), this holds when the network is sufficiently intransitive, so that second-order neighbors are not systematically first-order neighbors. We show in E-Companion EC.3 that our empirical network meets this condition.

3.4. Reduced-form Results with Bramoullé Instruments

To account for endogeneity, we augment the CRE probit model (7) with instrumental variables as follows. Recall, each observation represents a possible dyad ij at time t. In the dyad, the two endogenous regressors are the supplier's productivity, denoted as w^1_{ijt}, and the customer's productivity w^2_{ijt}. Each of them is estimated using (3). In a double index notation (firm-time), used in (3), $w^1_{ijt} = w_{it}$ and $w^2_{ijt} = w_{jt}$. Their respective instruments are z^1_{ijt} and z^2_{ijt}, the mean of the tier-2 suppliers' productivity (relative to supplier i) and the mean of tier-2 customers' productivity (relative to customer j). Note that z^1_{ijt} varies across i (supplier's index) and t, whereas z^2_{ijt} varies across j (customer's index) and t. The first stage regressions are as follows:

$$w^1_{ijt} = x'_{ijt}\pi^1_x + z'_{ijt}\pi^1_z + v^1_{ijt},$$
$$w^2_{ijt} = x'_{ijt}\pi^2_x + z'_{ijt}\pi^2_z + v^2_{ijt},$$

where $z_{ijt} = (z^1_{ijt}, z^2_{ijt})$, is a column vector of instruments and x_{ijt} is a vector of exogenous variables. We estimate residuals $v^{1,2}$ using OLS and use them to form a control function that we add to the second-stage regression (Wooldridge 2015). Accordingly, the second stage regression is:

$$\begin{aligned}\Pr(\, g^t_{ij} = 1 \mid w^1_{ijt}, w^2_{ijt}, \mathbf{x}_{ijt}, \overline{w}^1_{ij}, \overline{w}^2_{ij}, \overline{\mathbf{x}}_{ij}, \delta_t, a_{ij}) \;=\; \\ = \Lambda\big(\hat{w}^1_{ijt}\beta_1 + \hat{w}^2_{ijt}\beta_2 + \mathbf{x}'_{ijt}\boldsymbol{\beta} + \overline{w}^1_{ij}\gamma_1 + \overline{w}^2_{ij}\gamma_2 + \overline{\mathbf{x}}'_{ij}\boldsymbol{\gamma} + \hat{v}^1_{ijt}\varrho_1 + \hat{v}^2_{ijt}\varrho_2 + \delta_t + a_{ij}\big)\,,\end{aligned}$$

where $(\beta_1, \beta_2, \boldsymbol{\beta})$ is the vector of coefficients of interest, $(\gamma_1, \gamma_2, \boldsymbol{\gamma})$ are the coefficients on the Mundlak means, $\hat{v}^1_{ijt}\varrho_1 + \hat{v}^2_{ijt}\varrho_2$ is the control function, δ_t are the time fixed effects, and a_{ij} are the dyad random effects. Table 5 reports the first stage results, and Table 6 the second stage MLE estimates.

In the first stage, the coefficients on the Bramoullé instruments (z^1, z^2) are highly significant for their respective productivities, in line with the relevance of the instruments. The sign on the lagged tier-2 supplier productivity (z^1) is positive, suggesting that having productive suppliers in the past is positively associated with the current productivity. The sign on the lagged tier-2 customer productivity (z^2) is negative for customers, suggesting that having productive customers in the past is negatively associated with the current productivity. Note that the coefficient on z^2 is smaller in magnitude and carries less significance for the suppliers. Similarly, the coefficient on z^1 is less relevant for customers. In other words, supply-side instruments are more relevant for suppliers, and customer-side instruments are more relevant for buyers.

In the second stage, we observe a positive and highly significant effect of Supplier TFP, as well as a more muted effect of Customer TFP, on the likelihood of link formation. The control function coefficients on $\hat{v}_{1,2}$ are jointly statistically significant, supporting the presence of endogeneity. Interestingly, compared to the uninstrumented reduced form (7), the magnitude of coefficients on TFP is greater, suggesting that endogeneity attenuates the effect of productivity on the likelihood of link formation.

We continue to observe strong positive effects of degrees and the presence of common supply chain partners, as well as industry homophily, similar capital availability, and assortative size. Geographical distance is negatively associated with the likelihood of link formation. These estimates, while instructive and isolating the effects of productivity, do not account for the endogenous effects of network structure that influence link formation and future network structure. In §4, we develop an intertemporal network formation model that addresses this shortcoming.

3.5. Motivation for the Structural Approach

The reduced-form, dyad-level analysis presented in this section serves as a natural starting point for quantifying how observable covariates relate to the probability that a supplier–customer pair is linked in a given period. It yields interpretable coefficients that benchmark the roles of geography,

Table 6 Logit model with Bramoullé instruments, dyadic correlated random effects, and year FE, 2-stage estimator with control function, second stage robust SE clustered by dyad.

	$\mathbb{1}(link)$	$\mathbb{1}(link)$	$\mathbb{1}(link)$
Supplier TFP (fitted)	21.0449***	5.4557	13.7742***
	(3.1222)	(3.8617)	(4.3670)
Customer TFP (fitted)	-0.0905	3.3367	13.3375*
	(5.2144)	(6.3589)	(7.7841)
Supplier Outdegree		1.1066***	1.0915***
		(0.0295)	(0.0326)
Customer Indegree		1.3687***	1.3487***
		(0.0195)	(0.0251)
Common Partners		1.4170***	1.2946***
		(0.0231)	(0.0258)
Same Industry			2.4375***
			(0.0282)
Same Country			-0.4376***
			(0.0417)
ln(Distance)			-0.2355***
			(0.0059)
Abs.Diff ln(Capital)			-1.4841***
			(0.5329)
Abs.Diff ln(AT)			1.8059***
			(0.6195)
Constant	-117.8099***	-62.1149	-151.0367***
	(30.8411)	(37.9072)	(46.4702)
$\hat{v}_1$	0.1205***	0.0956***	0.0960***
	(0.0220)	(0.0268)	(0.0257)
$\hat{v}_2$	0.0880***	0.0291	0.0295
	(0.0232)	(0.0264)	(0.0258)
$\ln \sigma_a^2$	3.1155***	3.1425***	3.1104***
	(0.0062)	(0.0089)	(0.0093)
Observations	29,740,202	29,740,202	29,546,268
Number of dyad	10,141,867	10,141,867	10,078,715

*** p<0.01, ** p<0.05, * p<0.1

industry proximity, productivity, and local network features, while accommodating correlated dyad heterogeneity. These results provide insight into marginal associations at the dyad level. However, even with controls for shared partners, a supply network is not simply a collection of dyads. A single new relationship in the supply network can shift visibility and attractiveness elsewhere, influencing distant connections. Capturing such system-wide interactions requires a structural model that treats the *entire network* as the object of interest rather than a set of dyads.

Our goal is to understand how the entire network reconfigures when fundamental or frictional changes occur, such as when trade costs, compliance constraints, or technological shifts occur. Addressing these questions requires a structural representation of how links *form and persist together* over time. For example, let g^t denote the directed network at date t. A structural approach helps specify transitions that map the prior configuration g^{t-1} and period-t covariates into the current network g^t. This mapping separates creation from persistence; respects the feasibility implied by the last period's links, and ties economically meaningful statistics to the payoffs from adding or retaining a link. With such a model in hand, we can simulate counterfactual paths under policy or environment changes, trace how local adjustments propagate through neighborhoods, and quantify the roles of homophily, productivity, and embeddedness in shaping reconfiguration. That model is cast at the network level respects dyad-level granularity, works with network statistics, and explicitly models the network dynamics. Thus, the network-wide perspective complements the dyad-level

evidence by producing predictions consistent with the joint evolution of links and the constraints that govern which links can form or dissolve in each period.

4. A Structural Model of Network Dynamics

In this section, we present a dynamic model of link formation and dissolution in a directed supplier–customer network. The model is set in the temporal exponential random graph (ERGM) framework, cf. Robins et al. (2007), Hanneke et al. (2010). We start by laying out the primitives. We then introduce the network statistics used in the model, present the utility specification, and construct the likelihood function. We conclude with a comparison to a standard logit.

We study a directed network on firms $\mathcal{N} = \{1, \ldots, N\}$ observed at times $t = 0, 1, \ldots, T$. Let $\mathcal{G}$ be the set of all admissible networks, $G^t \in \mathcal{G}$ be the random network at time t, and $g^t \in \mathcal{G}$ its realization with entries $g^t_{ij} \in \{0, 1\}$ indicating a link from supplier i to customer j. Our goal is to model the evolution of the network. Specifically, we are interested in the conditional distribution $p_\theta(g \mid g^{t-1}, X^t)$ of G^t given $(G^{t-1} = g^{t-1}, X^t)$.[7] Between time periods $t-1$ and t, the network g^t can evolve in two ways: new links may form for dyads with $g^{t-1}_{ij} = 0$ and existing links with $g^{t-1}_{ij} = 1$ may dissolve. We model these changes with a separable structure: a formation process on $\{(i,j) : g^{t-1}_{ij} = 0\}$ and a dissolution process on $\{(i,j) : g^{t-1}_{ij} = 1\}$.

To describe the transition from $t-1$ to t, we introduce two auxiliary networks:

$$g^+_{ij} \equiv \mathbf{1}\{\, g^{t-1}_{ij} = 1 \text{ or } g^t_{ij} = 1 \,\} \qquad \text{(present at } t-1 \text{ or } t\text{)},$$

$$g^-_{ij} \equiv \mathbf{1}\{\, g^{t-1}_{ij} = 1 \text{ and } g^t_{ij} = 1 \,\} \qquad \text{(ties that persist)}.$$

Intuitively, g^+ is the union $g^{t-1} \cup g^t$, so it flags any link that appears by the end of t, whether retained or newly formed. g^- is the intersection $g^{t-1} \cap g^t$, so it flags only survivors. On dyads with $g^{t-1}_{ij} = 0$ we have $g^+_{ij} = g^t_{ij}$ and $g^-_{ij} = 0$; on dyads with $g^{t-1}_{ij} = 1$ we have $g^-_{ij} = g^t_{ij}$ and $g^+_{ij} = 1$. Hence, g^+ carries the formation decisions. Specifically, on the formation set $\left\{(i,j) : g^{t-1}_{ij} = 0\right\}, g^+_{ij}$ is the variable that turns on if the tie forms. Likewise, g^- carries the retention decisions—on the retention set $\left\{(i,j) : g^{t-1}_{ij} = 1\right\}, g^-_{ij}$is the variable that stays 1 if the tie persists.

We characterize each process with a pairwise utility specification. At each time t, firms evaluate the utility of forming or retaining a directed link from supplier i to customer j. For dyad (i,j), the net surplus at time t is:

$$U^t_{ij} \;=\; \phi_{ij}\big(g^{t-1}, X^t; \theta\big) \;+\; \varepsilon^t_{ij},$$

[7] Timing convention: we assume that covariates X^t are observed at the start of period t.

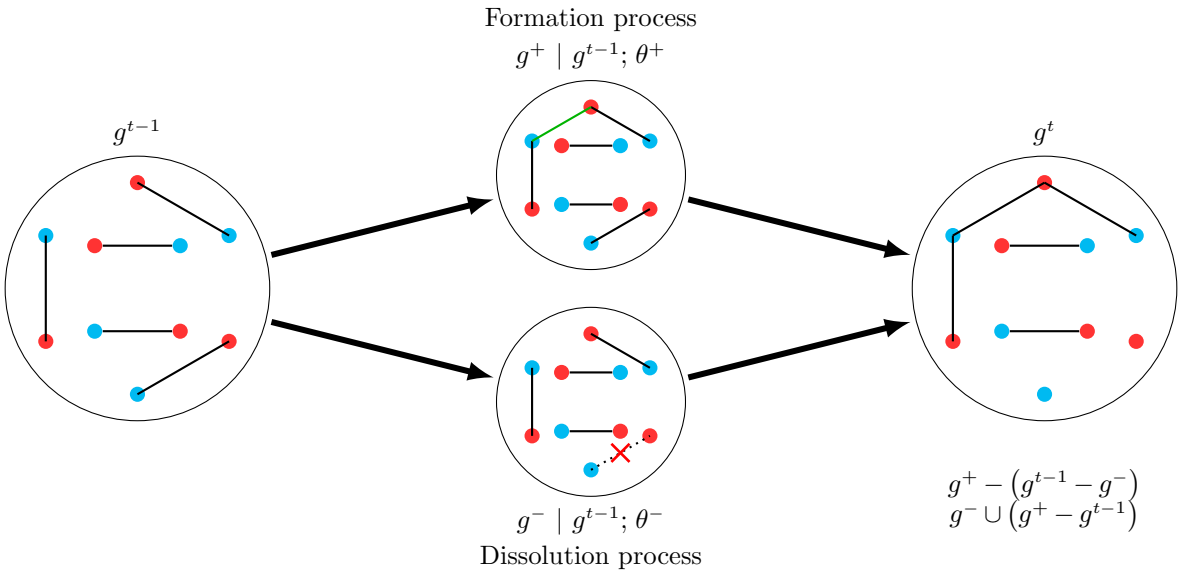


Figure 2 Illustration of the formation and dissolution processes.

where $\varepsilon_{ij}^t \sim$ i.i.d. Logistic$(0,1)$.[8] To distinguish between formation and dissolution, we use process-specific linear predictors:

$$\phi_{ij,t}^{\text{form}} = (\theta^+)^\top s_{ij}^{\text{form}}\big(g^+; g^{t-1}, X^t\big) \qquad \phi_{ij,t}^{\text{diss}} = (\theta^-)^\top s_{ij}^{\text{diss}}\big(g^-; g^{t-1}, X^t\big).$$

where $\theta^+ \in \mathbb{R}^{K^+}$ and $\theta^- \in \mathbb{R}^{K^-}$ are parameters to be estimated. The vectors s_{ij}^{form} and s_{ij}^{diss} include predetermined covariates that depend on g^{t-1} and X^t, as well as "network terms" evaluated on the period-t auxiliary networks g^+ or g^-. For each dyad, the network term is the change statistic, defined as the difference in the relevant network statistic when (i,j) is toggled on versus off in the corresponding auxiliary network. Importantly, because this change statistic depends on the surrounding links in period t, flipping any other link can alter it and therefore change the probability for (i,j). In short, dyads are interdependent rather than independent. Figure 2 depicts the formation and the dissolution processes in our model.

Functional form of the utility function: The vectors in our model include three classes of predictors: (i) dyad attributes and frictions such as same-industry and same-country indicators and log distance; (ii) firm fundamentals for i and j such as productivity and size; and (iii) local network terms that summarize how adding or retaining (i,j) embeds the relation in existing neighborhoods. A key local term is the geometrically weighted edgewise shared partners (GWESP), which captures transitivity in the network. Specifically, it reflects the tendency for two firms to form a direct link if they share common partners, thereby promoting local clustering and embeddedness. It is formally defined as:

$$\text{GWESP}(g;\alpha) = \sum_{k=1}^{n-2} e^{\alpha}\left[1-(1-e^{-\alpha})^k\right] \cdot SP_k(g),$$

[8] We use the (;) notation to separate primary variables or arguments of a function from parameters or fixed quantities, and (|) for conditioning in stochastic objects.

where $SP_k(g)$ is the number of edges with *exactly k shared partners*. Note that GWESP is a function of the entire graph g. The parameter α is a decay parameter that controls the diminishing marginal contribution of additional shared partners.[9] Formally, the formation utility then becomes:

$$\phi^{\text{form}}_{ij,t} = \underbrace{\theta^{+}_{0}}_{\text{Fixed cost}} + \underbrace{\theta^{\text{prod},+}_{s}\,\text{TFP}_{i,t} + \theta^{\text{prod},+}_{c}\,\text{TFP}_{j,t}}_{\text{Productivity}} + \underbrace{\sum_{k=1}^{K}\theta^{\text{hom},+}_{k}\,f_k(x_{i,t},x_{j,t})}_{\text{Homophily}} + \\ + \underbrace{\theta^{\text{net},+}_{1}\,\text{OutDeg}_{i,t-1} + \theta^{\text{net},+}_{2}\,\text{InDeg}_{j,t-1} + \theta^{\text{net},+}_{3}\,\Delta\text{GWESP}^{+}_{ij,t}(g;\alpha)}_{\text{Network features}}. \tag{8}$$

Here $f_k(x_{i,t},x_{j,t})$ collects homophily and friction terms, including $\log \text{Dist}_{ij}$, indicators for same industry and same location, $\left|\log \text{Size}_{i,t} - \log \text{Size}_{j,t}\right|$, and $|\text{Cap}_{i,t} - \text{Cap}_{j,t}|$. $\text{OutDeg}_{i,t-1}$ and $\text{InDeg}_{j,t-1}$ are degrees in g^{t-1}. An analogous specification defines $\phi^{\text{diss}}_{ij,t}$ with parameters θ^{-}.

Modeling network dependence. Note that the GWESP term that enters the utility is the *change statistic* for the (i,j) dyad on the formation auxiliary network:

$$\Delta\text{GWESP}^{+}_{ij,t}(g;\alpha) \equiv \text{GWESP}\big(g^{+,\text{on}}_{ij,t};\alpha\big) - \text{GWESP}\big(g^{+,\text{off}}_{ij,t};\alpha\big),$$

where $g^{+,\text{on}}_{ij,t}$ and $g^{+,\text{off}}_{ij,t}$ are the period-t formation networks with the (i,j) tie toggled on and off, respectively. Put simply, this is the change in the *network-level* GWESP statistic induced by toggling (i,j). Because $\Delta\text{GWESP}^{+}_{ij,t}(g;\alpha)$ depends on the surrounding links in period t, flipping other edges can alter it and therefore change the probability for (i,j). Hence dyads are interdependent rather than independent in a given period.

Network transition probability: We model the transition from $t-1$ to t as conditionally independent *formation* and *dissolution* processes. Each process admits an ERGM-type probability on a constrained sample space,[10] which gives the following formation and dissolution probabilities:

$$P_{\theta^{+}}\big(g^{+} \mid g^{t-1}, X^{t}\big) = \frac{\exp\big((\theta^{+})^{\top} S^{\text{form}}(g^{+};\, g^{t-1}, X^{t})\big)}{Z^{+}\big(\theta^{+} \mid g^{t-1}, X^{t}\big)}, \qquad g^{+} \in \mathcal{G}^{+}(g^{t-1}),$$

$$P_{\theta^{-}}\big(g^{-} \mid g^{t-1}, X^{t}\big) = \frac{\exp\big((\theta^{-})^{\top} S^{\text{diss}}(g^{-};\, g^{t-1}, X^{t})\big)}{Z^{-}\big(\theta^{-} \mid g^{t-1}, X^{t}\big)}, \qquad g^{-} \in \mathcal{G}^{-}(g^{t-1}),$$

where Z^{+} or Z^{-} are normalizing constants, and S is a summation of dyadic covariates s over sets of eligible dyads, see Remark 4. Assuming the two processes are independent conditional on (g^{t-1}, X^{t}),

$$P\big(g^{t} \mid g^{t-1}, X^{t}; \theta^{+}, \theta^{-}\big) = P_{\theta^{+}}\big(g^{+} \mid g^{t-1}, X^{t}\big) \cdot P_{\theta^{-}}\big(g^{-} \mid g^{t-1}, X^{t}\big),$$

[9] This decay is necessary to model realistic network structures where the incremental benefit of more shared partners saturates, preventing the overemphasis on extremely dense clusters that could lead to unrealistic network densities or instability in estimation.

[10] $\mathcal{G}^{+}(g^{t-1}) = \{h \in \mathcal{G} :\ g^{t-1}_{ij} = 1 \Rightarrow h_{ij} = 1\}$ holds all existing links fixed and lets only missing dyads vary, so formation can only turn non-edges into edges. $\mathcal{G}^{-}(g^{t-1}) = \{h \in \mathcal{G} :\ g^{t-1}_{ij} = 0 \Rightarrow h_{ij} = 0\}$ holds all missing links fixed and lets only existing dyads vary, so dissolution can only turn edges into non-edges. Think "+" for turning on missing links, and "−" for turning off existing links.

we can recover g^t by $g^t_{ij} = g^+_{ij}$ if $g^{t-1}_{ij} = 0$ and $g^t_{ij} = g^-_{ij}$ if $g^{t-1}_{ij} = 1$. Within a period, eligible dyads are coupled because statistics in S^{form} or S^{diss} depend on g^+ or g^-.

Likelihood, normalizing constants, and estimation: Given the time slices of network $\{g^t\}_{t=0}^T$, the log-likelihood sums period-by-period transitions:

$$\ell(\theta^+, \theta^-) = \sum_{t=1}^{T} \Big[(\theta^+)^\top S^{\text{form}}\big(g^+;\, g^{t-1}, X^t\big) - \log Z^+\big(\theta^+ \mid g^{t-1}, X^t\big) + (\theta^-)^\top S^{\text{diss}}\big(g^-;\, g^{t-1}, X^t\big) - \log Z^-\big(\theta^- \mid g^{t-1}, X^t\big) \Big].$$

We highlight here that each normalizer Z^+ or Z^- sums over the constrained sample space that grows exponentially in the number of eligible dyads. Specifically, when S^{form} or S^{diss} include dyad-dependent network terms, the exponent cannot be separated by individual dyad. Therefore the likelihood does not have a closed form representation and we have to estimate the model by simulation-based maximum likelihood. Specifically, we use MCMC–MLE that replaces the intractable expectations in the score with MCMC averages under $p^+(\cdot \mid g^{t-1}, X^t; \theta^+)$ and $p^-(\cdot \mid g^{t-1}, X^t; \theta^-)$. A toy example in E-Companion EC.5 illustrates the failure of factorization once dependence terms enter and shows how the MCMC targets the required expectations.

Remark 3 (Interdependence across time). Note that even though we assume that the formation and dissolution processes are independent of each other within the same time step, this does not imply independence across time. In fact, for any dyad-dependent model, interdependence persists across time steps. While formation and dissolution remain independent within a given time step, their outcomes influence subsequent time steps, introducing temporal dependence.

Remark 4 (Special case: dyad-independent statistics). If there are no statistics that couple different dyads through g^+ or g^- (for example GWESP), then the per-period likelihood decomposes into a product over eligible dyads (equivalently, the log-likelihood is a sum of independent dyad terms). This holds, for instance, when the utility in (8) does not depend on network features. In this case,

$$S^{\text{form}} = \sum_{(i,j):\, g^{t-1}_{ij}=0} s^{\text{form}}_{ij}(g^{t-1}, X^t)\, g^t_{ij}, \quad S^{\text{diss}} = \sum_{(i,j):\, g^{t-1}_{ij}=1} s^{\text{diss}}_{ij}(g^{t-1}, X^t)\, g^t_{ij}.$$

The normalizers are then separable across eligible dyads. For example,

$$Z^+\big(\theta^+ \mid g^{t-1}, X^t\big) = \prod_{(i,j):\, g^{t-1}_{ij}=0} \Big\{ 1 + \exp\big((\theta^+)^\top s^{\text{form}}_{ij}(g^{t-1}, X^t)\big) \Big\},$$

and similarly for Z^-. Each component becomes a product of logistic probabilities, so the likelihood has closed-form contributions and can be estimated by standard MLE without MCMC. We now show the relationship of this dyad-independent specification to the standard logit model.

4.1. Comparison with Standard Logit

By now it is evident that although the dyadic transition probabilities in our model resemble a logistic regression, the model cannot in general be estimated by a standard dyad-wise logit. A typical conditional probability is

$$\Pr\left(g_{ij} = 1 \mid g_{-ij}\right) = \frac{\exp\left(\theta^\top s_{ij}(g)\right)}{1 + \exp\left(\theta^\top s_{ij}(g)\right)},$$

which looks like a logit with covariates $s_{ij}(g)$. However, three issues prevent a dyad-wise logit approach:

1. *Network dependence.* Many regressors, such as degrees and transitivity terms, are functions of the entire network g. Toggling one edge (i, j) changes $s_{kl}(g)$ for other dyads (k, l), so dyads are not conditionally independent and the independence assumption required by dyad-wise logit fails.
2. *Normalizing constants and lack of factorization.* ERGMs define probabilities over networks, not independent dyads, so the likelihood is a not a product of dyad-level terms:

$$\mathcal{L}(\theta \mid g) \neq \prod_{(i,j)} \Pr(g_{ij} \mid s_{ij}).$$

   In our model, the per-period contribution takes the form

$$\mathcal{L}_t(\theta^+, \theta^-) = \frac{\exp\left((\theta^+)^\top S^{\text{form}}(g^+; g^{t-1}, X^t)\right)}{Z^+(\theta^+ \mid g^{t-1}, X^t)} \cdot \frac{\exp\left((\theta^-)^\top S^{\text{diss}}(g^-; g^{t-1}, X^t)\right)}{Z^-(\theta^- \mid g^{t-1}, X^t)},$$

   where Z^+ and Z^- sum over constrained sample spaces defined by g^{t-1}. When statistics couple dyads through g^+ or g^-, these normalizers are intractable. This rules out closed-form evaluation and motivates simulation-based methods such as MCMC–MLE.[11]
3. *Dynamic feedback.* Statistics are evaluated on g^{t-1}, so choices at t determine g^t and thereby the regressors that enter period $t+1$. Pooled or CRE binary-choice models can absorb time-invariant match quality, but they do not propagate network feedbacks through the transition law.

Absent network dependence, TERGM estimates coincide with the standard logit on a dyad-complete panel but without random effects, i.e., without controls for unobserved match quality. CRE models of §3 address that issue. In comparison, an advantage of TERGM is in addressing the network dependence and explicitly modeling network dynamics.

[11] As shown in Remark 1, If the specification contains only dyad-independent terms that are functions of (g^{t-1}, X^t), the transition likelihood factorizes over eligible dyads and separate logistic regressions for formation and dissolution deliver the MLE.

Table 7 Parameter Estimates for Link Formation and Dissolution Models

	Formation Model				Dissolution Model			
Term	**Estimate**	**Sd. Err**	$\Pr(>\|z\|)$	**Sig.**	**Estimate**	**Sd. Err**	$\Pr(>\|z\|)$	**Sig.**
Baseline								
Edges	-8.717	0.254	< 0.0001	***	-1.437	0.260	< 0.0001	***
Productivity								
Customer TFP	0.083	0.022	0.0002	***	-0.000	0.028	0.9930	
Supplier TFP	0.037	0.022	0.0935	.	-0.025	0.028	0.3680	
Homophily and Firm Similarities								
Distance	-0.119	0.026	< 0.0001	***	0.015	0.023	0.5200	
$\mathbb{1}_{\{\text{Industry}_i=\text{Industry}_j\}}$	1.201	0.138	< 0.0001	***				
$\mathbb{1}_{\{\text{Location}_i=\text{Location}_j\}}$	0.399	0.117	0.0007	***				
$\lvert\text{Capital}_i - \text{Capital}_j\rvert$	-0.048	0.019	0.0132	*	-0.022	0.031	0.4650	
$\lvert\log(\text{Firm Size})_i - \log(\text{Firm Size})_j\rvert$	0.136	0.022	< 0.0001	***	-0.012	0.035	0.7260	
Structural Dependencies								
GWESP (OTP)	0.756	0.043	< 0.0001	***				
GWESP (OTP Decay)	0.919	0.138	< 0.0001	***				
GWESP (OTP Fixed 0.1)					-0.117	0.045	0.0099	**
Customer In-Degree	0.008	0.002	0.0009	***	0.001	0.003	0.8650	
Supplier Out-Degree	0.008	0.002	0.0001	***	-0.000	0.002	0.9330	

Notes: Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$
Model fit: Null deviance = 6,101,847 on 4,401,552 df; Residual deviance = 23,200 on 4,401,531 df
AIC = 23,242; BIC = 23,521; MCMC Std. Error = 2.109

5. Results

In this section, we present the estimates from our structural model, as described in §4. We distinguish between drivers of new link creation (the formation model) and those governing the persistence of existing ties (the dissolution model). Table 7 reports the parameter estimates for both these processes, along with standard errors, p-values, and significance levels. The table highlights both strong structural and covariate-based factors driving new link creation. Structural drivers include endogenous network features, such as GWESP, which reflect how the existing network topology affects link formation. In contrast, covariate-based drivers are effects stemming from firm attributes or dyadic characteristics, including productivity (e.g., TFP) and homophily (e.g., industry matching).

From Table 7, we note that the coefficient for the *Edges* variable is large, negative, and highly significant. This suggests a baseline level of network *sparsity*, indicating that link formation occurs infrequently without additional reinforcing factors. However, when we take into account firm-level characteristics, homophily, and the endogenous structure of the network, we can identify systematic patterns in the linking decisions made by firms. We explain the role of these factors next.

Firm-Level Productivity: We note that both customer and supplier TFP coefficients are positive, indicating that productive firms are more likely to form new ties. The customer TFP coefficient (0.083) is nearly twice as large as the supplier TFP coefficient (0.037) and is statistically significant at the 0.001 level, while the supplier TFP is marginally significant at the 0.1 level. This suggests that link creation is more strongly influenced by customer-side productivity than supplier-side

productivity. This finding aligns with the idea that customers bear the costs of initiating ties, and high-productivity customers are more likely to possess the organizational capacity, scale, and sophistication required to initiate and manage supplier relationships. Conversely, supplier productivity appears to play a more limited role in link formation compared to customer productivity. These findings suggest that, in the absence of proactive customer interest, even productive suppliers may not form new links, as link formation appears to be more demand-driven than supply-driven. To summarize, although directed links flow from suppliers to customers in supply networks, the initiation often comes from the customer's side, and customer TFP plays a more significant role in link formation.

Homophily Effects: We also find that observable firm similarities significantly facilitate link formation. The same-industry coefficient (1.201) is large and positive, indicating a strong tendency for intra-industry connections. Geographic homophily, based on shared location, also has a positive and significant effect (0.399), though smaller in magnitude. Additionally, the negative and highly significant coefficient on distance (-0.119) shows that greater geographic separation reduces the likelihood of link formation, possibly due to longer lead times and higher transportation costs, highlighting the role of proximity in overcoming barriers to connection. These patterns reflect reduced search and coordination frictions in geographically proximate locations. Notably, these homophily effects persist even after controlling for productivity, suggesting match-specific complementarities beyond efficiency.

In contrast to the homophily observed in industry and location, the positive and significant coefficient on the absolute difference in log firm size (0.136) indicates *heterophily*, where firms of dissimilar sizes are more likely to form ties. This pattern reflects complementarities in buyer-supplier networks, where large customers link with smaller, specialized suppliers to gain flexibility, cost advantages, or niche inputs. In contrast, smaller suppliers benefit from access to broader markets and stability (Bernard et al. 2022). Such disassortative matching contributes to core-periphery network structures, where large firms connect to many smaller ones, enhancing overall efficiency and diversification. This is consistent with empirical evidence from production networks, which shows that heterogeneous links based on size differences can drive innovation through knowledge spillovers (Bernard et al. 2018) and help explain fat-tailed firm size distributions in interconnected economies (di Giovanni et al. 2014).

Structural Dependencies: The positive GWESP coefficient (0.756) provides evidence of transitivity, indicating that links form more readily when firms share common partners, which supports theories and evidence of embeddedness and triadic closure (Mosleh et al. 2025). The positive decay term (0.919) indicates that the marginal value of additional shared partners diminishes quickly, implying

that few shared partners activate informational or trust spillovers, but redundancy follows after a few partners. We demonstrate this impact of the decay parameter in detail in §5.1.

In the dissolution model, the negative *Edges* coefficient suggests a baseline tendency for ties to persist. The GWESP term (fixed at 0.1 decay) has a negative and significant coefficient (-0.117). This indicates that ties in dense neighborhoods are more robust to dissolution, reinforcing the role of embeddedness in network stability. In addition to these factors, the dissolution model reveals few significant predictors. Most coefficients, including those for productivity, homophily, and firm attributes, are insignificant and have smaller magnitudes than their counterparts in the formation model.

Model Fit: Finally, our model demonstrates a strong overall fit, as evident by the substantial reduction in deviance from the null model (6,101,847 on 4,401,552 degrees of freedom) to the residual deviance (23,200 on 4,401,531 degrees of freedom). This corresponds to approximately 99.62 percent of the deviance explained, indicating that the included predictors account for the vast majority of the variation in observed link patterns compared to a baseline model with no covariates.

5.1. Marginal Contribution of the k-th Shared Partner:

Recall, the GWESP term is formally defined as $\text{GWESP}(g;\alpha) = \sum_{k=1}^{n-2} e^{\alpha}[1-(1-e^{-\alpha})^k] \cdot SP_k(g)$, where $SP_k(g)$ is the number of edges with *exactly k shared partners*. Mathematically, the marginal contribution of the k-th shared partner is derived as the difference in the cumulative weighting function:

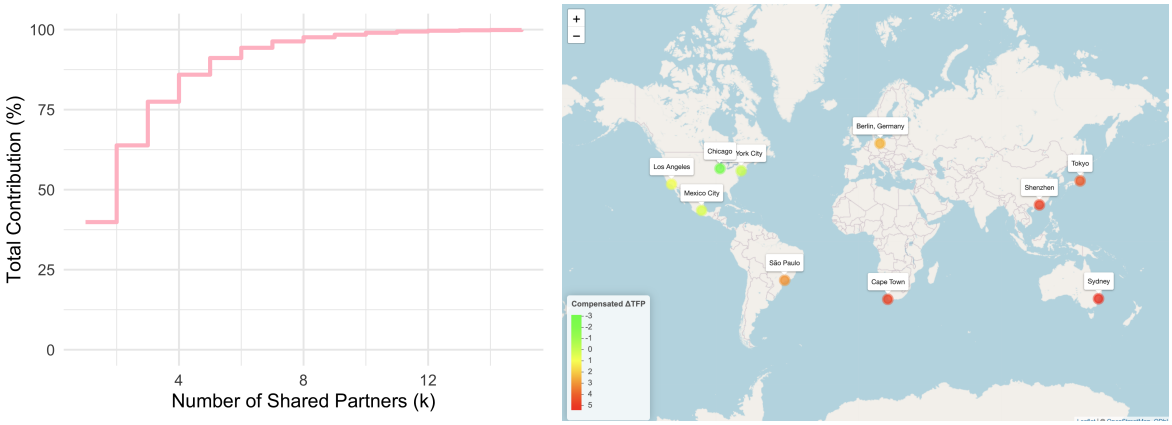
$$e^{\alpha}\left[1-(1-e^{-\alpha})^k\right] - e^{\alpha}\left[1-(1-e^{-\alpha})^{k-1}\right] = e^{\alpha}\left(1-e^{-\alpha}\right)^{k-1} \cdot e^{-\alpha}$$

From Table 7, the estimated decay parameter for GWESP in the formation process is $\alpha = 0.919$. Using this value, we calculate the cumulative contribution as a function of the number of shared partners (please see Figure 4). We note from Figure 4 that only four shared partners suffice to achieve 80% of the maximum GWESP contribution, beyond which additional partners add little marginal value. This saturation point underscores the model's ability to capture clustering in supplier-customer networks without overemphasizing higher-order shared connections.[12].

5.2. Marginal Rates of Substitution

The model also allows us to quantify the trade-offs between different covariates by computing marginal rates of substitution (MRS). Specifically, the MRS measures how much one covariate must change to offset a change in another while keeping the log-odds of link formation constant. For

[12] To find where the remaining contribution from additional shared partners falls below 20% of the maximum, we solve $(1-e^{-\alpha})^k \leq 0.2$ with $\alpha = 0.919$, where $e^{-\alpha} \approx 0.3989$ and $1-e^{-\alpha} \approx 0.6011$. This yields $k \geq \frac{\ln(0.2)}{\ln(0.6011)} \approx 3.168$, so $k = 4$ (ceiling to nearest integer), where the cumulative contribution reaches approximately 86.95%, indicating the sum of marginal contributions beyond four shared partners is less than 20%



(a) Cumulative Contribution (b) Distance vs TFP Tradeoff

Figure 3 Panel (a) shows the cumulative contribution of a shared partner to log odds. Panel (b) shows the required Increase in Supplier TFP to Offset Geographic Distance from New York City. This map shows selected global cities, with markers colored by the compensating ΔTFP (green for lower, red for higher), derived from the structural formation model coefficients.

two continuous covariates v and w with coefficients β_v and β_w, the condition for no net change in log-odds is

$$\beta_v \, \Delta v + \beta_w \, \Delta w = 0 \quad \Rightarrow \quad \Delta w = -\frac{\beta_v}{\beta_w} \, \Delta v.$$

Specializing this to the distance covariate (which has a negative coefficient $\beta_{\text{distance}} < 0$), the compensating change in covariate v that offsets a change in log-distance is

$$\Delta v \;=\; \frac{|\beta_{\text{distance}}|}{\beta_v} \, \Delta \log(\text{distance}).$$

For instance, for supplier TFP, $\beta_{\text{supplier TFP}} \approx 0.037$ and $|\beta_{\text{distance}}| \approx 0.119$, and the MRS is $|\beta_{\text{distance}}|/\beta_{\text{supplier TFP}} \approx 3.2162$. Thus, a one-unit increase in log-distance must be offset by a 3.2162-unit rise in supplier TFP to maintain the formation probability constant. For binary homophily terms with coefficient β_{match}, the interpretation reverses: the number of log-distance units offset by switching from mismatch to match is $\beta_{\text{match}}/|\beta_{\text{distance}}|$.[13] Table 8 reports MRS values for the key covariates using these definitions, providing a compact way to compare economic trade-offs implied by the model. We describe some of these below.

Distance and TFP trade-off. Using the supplier TFP MRS from above (3.2162), Figure 3 panel (b) maps the compensating increase in supplier TFP needed to neutralize distance from New York City for selected global cities. Each city's compensating change is computed relative to the sample median log-distance of 8 (unlogged median $\approx$ 2,981 km):

$$\Delta\text{TFP}_{\text{city}} \;=\; \big(\log(\text{distance})_{\text{city}} - 8\big) \times 3.2162.$$

[13] All references to distance units are in *log-distance* because the regressor is log(distance).

A positive value indicates the additional supplier TFP required for farther locations. For example, Mexico City's log-distance is about 8.1195, so $\Delta \log(\text{distance}) \approx 0.1195$ and $\Delta\text{TFP} \approx 0.1195 \times 3.2162 \approx 0.3844$. This quantifies how distance raises the productivity threshold for link formation. In plain terms, a greater distance raises the productivity bar a supplier must clear for a link to form, and closer distances lower the bar, thereby forgoing productivity for geographical proximity.

GWESP and diminishing returns. The GWESP term captures transitivity with diminishing marginal effects. In our formation estimates (Table 7), the GWESP coefficient is $\theta \approx 0.756$ and the estimated decay is $\alpha \approx 0.919$. For exposition, define the geometric ratio $\gamma \equiv 1 - e^{-\alpha} \approx 0.601$. Then the marginal increase in log-odds from the kth shared partner is $\theta\,\gamma^{k-1}$. Converting that gain into equivalent log-distance units by the same MRS logic yields

$$\text{log-distance units offset by partner } k = \frac{\theta\,\gamma^{k-1}}{|\beta_{\text{distance}}|}.$$

Numerically, the first shared partner offsets about $0.756/0.119 \approx 6.35$ log-distance units (see Table 9), the second offsets about $0.4544/0.119 \approx 3.82$, and so on. Returning to our example of Mexico City, the distance penalty to neutralize is $|\beta_{\text{distance}}| \times 0.1195 \approx 0.0142$ in log-odds. Hence, starting from zero shared partners, a single shared partner more than offsets Mexico City's additional distance.

Table 8 MRS b/w distance and other covariates

Variable	Compensating Change (Δv)
Supplier TFP	3.2162
Customer TFP	1.4337
Supplier out-degree	14.8750
Customer in-degree	14.8750
Abs. diff. Capital	-2.4792
Abs. diff. Log Firm Size	0.8750

Table 9 MRS b/w distance and shared partners

# Partners	Δ Log-Odds	Distance Units Offset
1	0.7560	6.35
2	0.4547	3.82
3	0.2735	2.30
4	0.1645	1.38
5	0.0989	0.83
6	0.0595	0.50

A note on sample selection and robustness: Our model requires a common set of nodes (balanced panel) that appear in every year. This persistence requirement creates a trade-off: lengthening the time window raises the number of periods T that we observe the firm for, but lowers the number of persistent firms N_{persist}, while shortening the window does the opposite. For a directed network, the number of dyad–period observations is $\text{DF} = N_{\text{persist}}(N_{\text{persist}} - 1)T$. We adopt the 2016–2022 period as the baseline window because it preserves a large N_{persist} while providing sufficient time variation to identify dynamics.

To assess the robustness of our results, we vary the start year between 2013 and 2018 (holding the end year at 2022). In Appendix EC.6, we report the resulting N_{persist} and DF, and the estimates from the model estimated on each window. The main coefficients and their interpretation remain stable across these alternatives, demonstrating the robustness of our estimates.

6. Counterfactuals

A key strength of our structural model is its ability to conduct policy counterfactuals. Unlike reduced-form regressions, whose parameters capture local, conditional associations under a given policy environment, our structural model encodes a stochastic transition mechanism that governs link formation and dissolution given covariates and the current network state. This lets us perturb network primitives such as node and edge characteristics or link eligibility rules, simulate how the entire network evolves under these changes, and compute network-wide outcomes under alternative policy environments.

Specifically, we formalize a counterfactual as replacing the observed environment (base policy P_0) with a target policy P_1. Let $g^t \in \{0,1\}^{n\times n}$ denote the directed network at time t, and let X_t collect node/edge covariates (e.g., productivity, size, industry, location, pairwise distance). Holding the estimated parameters $\theta = (\theta^+, \theta^-)$ fixed, we alter the primitives implied by P_1; simulate the network forward from the observed baseline; and summarize network-wide outcomes under P_1. Our policies enter in two transparent ways: (i) covariate transformations $X_t \mapsto X_t(P)$ (e.g., a productivity boost for a subset of firms or locations), and/or (ii) feasibility constraints via an eligibility set $\mathcal{A}_P \subseteq \{1,\dots,n\}^2$ (eligible seed dyads receive a large persistent utility boost).

Simulation protocol: We initialize the system with the last observed network g^{T_0} and covariates X_{T_0}. For each policy $P \in \{P_0, P_1\}$, we simulate M independent Monte Carlo sample paths, each of length H, using the estimated parameters θ and policy-specific primitives:

$$\{\, g^{T_0+h,(m,P)} \,\}_{h=1}^{H} \quad \text{via} \quad g^{T_0+h,(m,P)} \sim \mathcal{T}\big(g^{T_0+h-1,(m,P)};\, \theta,\, X_{T_0+h}(P),\, \mathcal{A}_P\big), \quad m = 1,\dots,M.$$

Note that this involves full-network simulations: at each step, transition probabilities depend on the current network state g^{t-1}, allowing shocks to propagate through degree effects, triadic closure, and other model mechanisms.

Measurement: Let $f(\cdot)$ be a statistic of interest computed on the terminal network g. We estimate policy effects as Monte Carlo differences in expectations:

$$\widehat{\Delta f} = \frac{1}{M}\sum_{m=1}^{M} f\big(g^{T_0+H,(m,P_1)}\big) \;-\; \frac{1}{M}\sum_{m=1}^{M} f\big(g^{T_0+H,(m,P_0)}\big).$$

We track four outcomes in the counterfactuals: (i) node-level metrics, e.g., supplier out-degree $\deg_i^{\text{out}}(g) = \sum_j g_{ij}$ and customer in-degree $\deg_j^{\text{in}}(g) = \sum_i g_{ij}$; (ii) average shipping distance for supplier i, $D_i(g) = \deg_i^{\text{out}}(g)^{-1} \sum_j g_{ij} d_{ij}$, where d_{ij} is exogenous distance; (iii) a state-to-state flow matrix $F_{rs}(g) = \sum_{i:\,\text{state}(i)=r} \sum_{j:\,\text{state}(j)=s} g_{ij}$ (so F_{rr} is within-state and F_{rs} for $r \neq s$ is cross-state; with binary g_{ij} these are link counts; with weights they are totals); and (iv) network-structure measures such as triangle counts and average shortest-path lengths. Next, we describe our two policy interventions and evaluate their impacts.

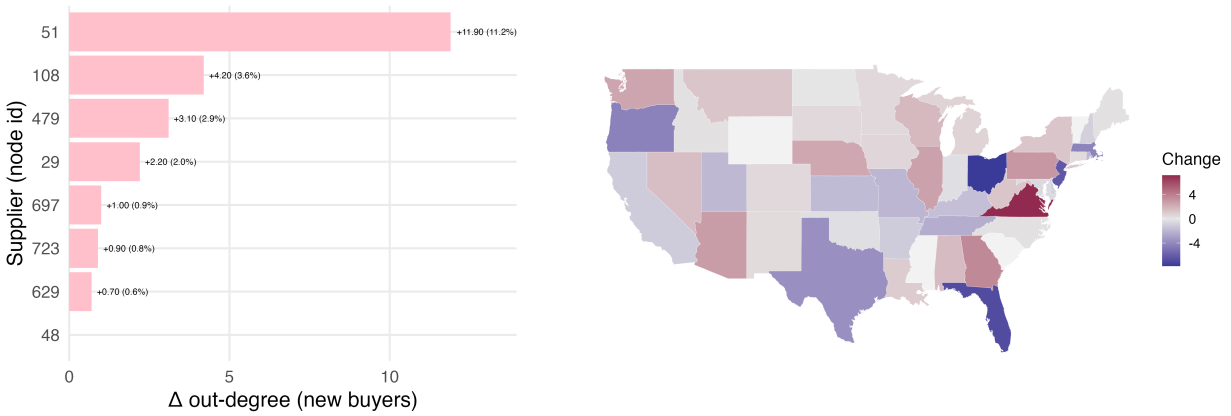
6.1. Targeted Productivity Shocks: The Case of Semiconductor Push in Arizona

States can induce productivity shocks to firms through targeted industrial policies, including subsidies, tax incentives, infrastructure investments, and workforce development programs that attract foreign direct investment and anchor firms. For example, the Arizona semiconductor push, primarily driven by the CHIPS and Science Act of 2022, involves federal funding to bolster domestic semiconductor manufacturing. The exact intervention includes direct subsidies and grants, such as up to $6.6 billion awarded to TSMC Arizona for constructing advanced fabrication facilities in Phoenix, $8.9 billion invested in Intel for expanding its Arizona operations, and $325 million to Hemlock Semiconductor for a new facility.[14] These policies boost local supplier productivity through subsidizing equipment and facility investments, facilitating technology transfer from anchor firms for improved yield and quality control, and fostering learning spillovers as local suppliers adopt advanced practices from global leaders such as TSMC and Intel. Prior literature demonstrates that such mechanisms can significantly boost domestic firm productivity. For instance, Chou et al. (2014) found that in Shanghai, Suzhou, and Wuxi, FDI clustering, supported by government policies, boosted semiconductor productivity through technology transfer and industrial chain formation. Additionally, Moretti (2021) shows that larger high-tech clusters increase inventor productivity by approximately 12% when inventors relocate to larger clusters, due to knowledge spillovers. Motivated by these real-world interventions, we simulate a policy that raises the productivity of Arizona-headquartered suppliers and examine its impact on supply network restructuring.

Procedure: We evaluate this policy using the estimates from our structural model. Specifically, we raise supplier productivity for AZ-headquartered suppliers by a fixed amount (our baseline uses a 0.25 log-point "ΔTFP") and then simulate the entire network forward for five periods under two regimes: the base policy (status quo) and the target policy (with ΔTFP). We then compare the following metrics (averaged over sample paths) across two policies: (i) buyers gained per AZ supplier (change in out-degree), and (ii) the AZ→destination-state output flows (counts of AZ→state links).

We first analyze the impact of the policy on the local network (i.e., the state network). Interestingly, we find that the productivity boost to Arizona suppliers leads to unequal and concentrated gains in buyer connections. The leading supplier adds nearly 12 buyers on average (see Figure 4, panel a), the next few add about four or fewer, and many others lose connections. This concentration follows from the network dependence in our estimated formation model. Triadic closure, captured by GWESP, and homophily steer new links toward suppliers that are already embedded in dense, compatible neighborhoods. At the same time, additional productivity reallocates buyers' attention

[14] In addition, auxiliary initiatives encompass the establishment of the National Semiconductor Technology Center with $5 billion for research and development, supplier-development programs, and state-level tax credits and workforce training grants to build local talent pools.



(a) AZ Suppliers: largest gain in buyers

(b) AZ→destination-state flow change

Figure 4 Panel (a) identifies Arizona suppliers with the largest gains in buyer connections under the simulated productivity boost. Panel (b) maps the changes in supply flows from Arizona to destination states resulting from the policy intervention.

away from less well-positioned peers. The result is an uneven propagation of a uniform shock. This result is reminiscent of the inequality of gains from trade in economics (see, e.g., Helpman et al. 2016), but in our context, the pattern arises from the interaction of closure and similarity with the common productivity shift.

At the national level, the policy triggers significant restructuring within the network. Figure 4 Panel (b) shows the interstate flows induced by the policy. We find that Arizona suppliers form notably more links to buyers in Virginia (an increase of over 7) and Georgia, while diverting connections away from states like Ohio and Florida (decreases of around 8 and 7, respectively). Note that the productivity boost in this counterfactual uniformly enhances the attractiveness of all Arizona suppliers, but it does not lead to uniform gains across all potential buyers. Instead, buyers reoptimize their sourcing, prioritizing high-value matches with complementary and productive partners in similar industries, while letting weaker ties lapse. For instance, Virginia's tech-heavy economy, with information services and data centers reliant on semiconductors, draws more ties via high-productivity buyers and amplified signaling or shared partner effects. By contrast, Ohio's manufacturing base in automotive, aerospace, and heavy industry shows less overlap, leading to competitive crowding through interdependence and triadic closure, as suppliers favor tech fits like Virginia and divert their business away from Ohio. This pattern suggests that boosted Arizona suppliers tend to preferentially connect with distant, productive buyers in tech-heavy or logistically favorable regions, while reducing their links to less complementary areas. Overall, the counterfactual reveals how a localized productivity shock propagates nationally as links are reallocated toward embedded highly productive neighborhoods where the gains outweigh distance penalties.

6.2. Seeding Interventions to Strengthen Local Supply Chains in California

Policymakers often aim to strengthen local industry clusters rather than create long-distance links. Examples of such interventions include state procurement fairs that pre-match certified suppliers with anchor firms; industry association networking within metro areas, and subsidies that reduce onboarding costs when firms share vetted partners. In a similar vein, California's Next-Gen Battery Manufacturing Activation Plan of January 2025 aims to create integrated in-state clusters to complete the EV battery supply chain.[15] These programs use grants, matchmaking, and partner-based incentives to foster in-state ties, reduce dependence on out-of-state or international sources, boost resilience, and spur growth from local proximity and shared resources. Motivated by these real-world interventions, we simulate a policy that adds seed edges within California firms to examine its impact on network structure, productivity access, and spillovers.

Procedure: We evaluate this policy using the estimates from our structural model. First, we identify potential seed edges within California by selecting pairs of high-degree California firms, defined by their total degree (in-degree + out-degree), that have many common neighbors (meaning the two firms share a large number of distinct partners in the current network). We then only retain pairs with no existing link between them at the start, and that are strictly intra-state. From this candidate pool, we add 50 directed seed edges at time 1 and simulate the network's evolution for five periods in both baseline and policy settings. For evaluation, we take the terminal snapshot from each sample path and average across sample paths to construct outcome measures.

Outcomes: The outcomes we report are: changes in mean degree (out-degree and in-degree) for CA firms; the long-haul share for CA (defined as the fraction of CA-origin links whose distance exceeds a fixed threshold); triangle counts, including the total number of cross-state triangles and triangles that touch at least one CA node;[16] spillovers, defined as changes in the number of outgoing edges from California firms to each destination state relative to the baseline; buyer access to supplier total factor productivity (TFP) in California, computed as the average (across California buyers) of the sum of their suppliers' TFP values; and the network-wide sum of distance-weighted links, proxying aggregate transport costs. This set of metrics isolates how targeted intra-state seeding in California can reshape the local supply network. Table 10 reports these outcomes.

The results demonstrate the effectiveness of local supply chain formation in California. Mean out-degree and in-degree increase by 0.305, as seeded edges create new shared-partner paths that lead to new direct intra-state links guided by transitivity.[17] Triangles touching California (those

[15] https://www.allhomeca.org/wp-content/uploads/2025/02/DRAFT-FINAL-BAJFC-Next-Gen-Battery-Manufacturing-Activation-Plan-January-2025.pdf

[16] these are undirected cycles of three firms spanning multiple states

[17] We force the seeded links to persist for 2 periods. Hence, the increase in degrees demonstrates how an initial seeding ultimately benefits the process.

Table 10 CA seeding counterfactual: salient outcomes. Means across 10 paired-path simulations at horizon H=5.

Metric	Change in the metric
Mean out-degree (CA)	+0.305
Mean in-degree (CA)	+0.305
Long-haul share (CA)	−0.00205
Cross-state triangle count	+1,775
Triangles touching CA	+2,204
Spillover CA→CA	+43.9
Spillover CA→AL,AZ,AR,CO	0
Mean TFP access (CA buyers)	+1.514
Distance-weighted links [km]	+168

Notes: Seeds are 50 intra-CA pairs. Long-haul threshold equals 7.3.

involving at least one California node) surge by 2,204, and cross-state triangles grow by 1,775. These increases indicate tighter integration: seeding closes two-step paths into triangles that support stable, lower-cost relationships among nearby firms, while the denser California core leverages existing interstate ties rather than establishing new long-distance links. Spillovers concentrate fully intra-state with +43.9 new California-to-California edges, while remaining zero to external nearby states like Alabama, Arizona, Arkansas, and Colorado, illustrating how homophily directs growth inward to thicken the local subgraph. Mean total factor productivity access across California buyers climbs by 1.514, underscoring efficiency gains as more local ties connect buyers to high-productivity suppliers within the cluster, enabling firms to trade homophily for capability in embedded environments. Thus, the policy can lead to better access to high-productivity partners by strengthening local networks and attracting productive firms to the cluster.

In summary, the policy seeds new links among California firms, which triggers triadic closure. The model amplifies this by rewarding closed two-step paths, favoring same-state links due to homophily and distance costs, and steering matches toward more productive partners. The result is a denser California cluster rather than long bridges. Each added link creates more opportunities for shared partnerships, so the transitivity reinforces itself, increasing connectivity and improving access to productive partners.

7. Conclusion

In this paper, we examine the formation of links between firms in supply networks. We posit that link formation is driven by firms' productivity, the local structure of the network the firms are embedded in, and homophily between potential buyers and suppliers. We formulate a utility-based model of link formation and estimate it on state-of-the-art granular buyer–supplier relationship data. We use a set of sophisticated models to address the empirical challenges presented by the problem, including sample selection, endogeneity in productivity, and non-independence of dyadic observations embedded in a network.

We find a causal relationship between buyers' and suppliers' productivities and the likelihood of forming a link. Moreover, buyer productivity matters more, consistent with the buyer typically

initiating the relationship. Productivity can be substituted for familiarity, measured by the number of shared network connections; however, the marginal effect of familiarity decreases and virtually vanishes after five shared connections. Links are more likely to be formed between firms in the same industry and the same country, as well as between firms of dissimilar sizes. Geographic distance represents a significant friction, effectively raising the productivity threshold required for link formation across space. These patterns persist even after controlling for observed heterogeneity and network dependence, underscoring the economic importance of both fundamentals and embeddedness in shaping supply networks.

Our structural model also supports counterfactuals. Holding estimated parameters fixed, we simulate the network's evolution under targeted policy changes. Two experiments illustrate how mechanisms translate into network reconfiguration. A localized productivity shock for Arizona suppliers reconfigures the national network, as buyers reoptimize their purchasing across states, with gains concentrating on well-established Arizona firms, as triadic closure and visibility amplify early advantages. A California seeding policy that inserts a small number of in-state links among high-connectivity firms thickens local clusters, increases in-state triangles, and improves buyers' access to high-TFP suppliers through more shared-partner paths, with effects concentrated inside the state.

These findings have practical and policy implications. For managers, investing in search, certification, and brokerage technologies that surface high-productivity suppliers can substitute for slow familiarity-building, while cultivating a small set of shared partners can unlock significant gains at a low cost. For policymakers and platform designers, reducing effective distance through logistics infrastructure, interoperable registries, and transparent reputation systems can expand feasible matches and improve aggregate productivity.

References

Agrawal, D., N. Osadchiy. 2024. Inventory productivity and stock returns in manufacturing networks. *Manufacturing & Service Operations Management* **26**(2) 573–593.

Antràs, P., T. C. Fort, F. Tintelnot. 2017. The margins of global sourcing: Theory and evidence from us firms. *American Economic Review* **107**(9) 2514–2564. doi:10.1257/aer.20141685.

Arora, K., F. Zheng, K. Girotra. 2023. Private vs. pooled transportation: Customer preference and design of green transport policy. *Manufacturing & Service Operations Management* doi:10.1287/msom.2022.0569.

Atalay, E., A. Hortaçsu, J. Roberts, C. Syverson. 2011. Network structure of production. *Proceedings of the National Academy of Sciences* **108**(13) 5199–5202. doi:10.1073/pnas.1015564108.

Baldwin, R., A. J. Venables. 2013. Spiders and snakes: offshoring and agglomeration in the global economy. *Journal of International Economics* **90**(2) 245–254.

Barabási, A.-L., R. Albert. 1999. Emergence of scaling in random networks. *Science* **286**(5439) 509–512. doi:10.1126/science.286.5439.509.

Bernard, A. B., A. Moxnes. 2018. Networks and trade. *Annual Review of Economics* **10** 65–85. doi:10.1146/annurev-economics-080217-053506.

Bernard, A. B., E. Dhyne, G. Magerman, K. Manova, A. Moxnes. 2022. The origins of firm heterogeneity: A production network approach. *Journal of Political Economy* **130**(7) 1765–1804. doi:10.1086/719759. URL `https://doi.org/10.1086/719759`.

Bernard, A. B., A. Moxnes, Y. U. Saito. 2018. Production networks, geography, and firm performance. *Journal of Political Economy* **127**(2) 639–688. doi:10.1086/700764.

Bianconi, G., A.-L. Barabási. 2001. Competition and multiscaling in evolving networks. *Europhysics letters* **54**(4) 436.

Blum, B. S., S. Claro, I. Horstmann. 2010. Facts and figures on intermediated trade. *American Economic Review* **100**(2) 419–423.

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

Carvalho, V. M., M. Nirei, Y. U. Saito, A. Tahbaz-Salehi. 2021. Supply chain disruptions: Evidence from the great east japan earthquake. *The Quarterly Journal of Economics* **136**(2) 1255–1321. doi:10.1093/qje/qjaa044.

Chamberlain, G. 1984. Panel data. Z. Griliches†, M. D. Intriligator, eds., *Handbook of Econometrics*, vol. 2, 1st ed., chap. 22. Elsevier, 1247–1318. URL `https://EconPapers.repec.org/RePEc:eee:ecochp:2-22`.

Chaney, T. 2014. The network structure of international trade. *American Economic Review* **104**(11) 3600–3634. doi:10.1257/aer.104.11.3600.

Chou, T.-L., J.-Y. Chang, T.-C. Li. 2014. Government support, fdi clustering and semiconductor sustainability in china: Case studies of shanghai, suzhou and wuxi in the yangtze delta. *Sustainability* **6**(9) 5655–5681. doi:10.3390/su6095655.

Currarini, S., M. O. Jackson, P. Pin. 2009. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica* **77**(4) 1003–1045. doi:10.3982/ECTA7528.

di Giovanni, J., A. A. Levchenko, I. Mejean. 2014. Firms, destinations, and aggregate fluctuations. *Econometrica* **82**(4) 1303–1340. doi:https://doi.org/10.3982/ECTA11041.

Fernández-Val, I., M. Weidner. 2016. Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics* **192**(1) 291–312.

Goldsmith-Pinkham, P., G. W. Imbens. 2013. Social networks and the identification of peer effects. *Journal of Business Economic Statistics* **31**(3) 253–264. URL `http://www.jstor.org/stable/43702718`.

Hahn, J., W. K. Newey. 2004. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* **72**(4) 1295–1319.

Hanneke, S., W. Fu, E. P. Xing. 2010. Discrete temporal models of social networks. *Electronic Journal of Statistics* **4** 585–605.

He, P., F. Zheng, E. Belavina, K. Girotra. 2021. Customer preference and station network in the london bike-share system. *Management Science* **67**(3) 1392–1412. doi:10.1287/mnsc.2020.3620.

Helpman, E., O. Itskhoki, M.-A. Muendler, S. J. Redding. 2016. Trade and inequality: From theory to estimation. *The Review of Economic Studies* **84**(1) 357–405.

Lim, K. 2018. Endogenous production networks and international business cycles. Tech. rep., Working Paper.

Moretti, E. 2021. The effect of high-tech clusters on the productivity of top inventors. *American Economic Review* **111**(10) 3328–3375. doi:10.1257/aer.20191277.

Mosleh, M., D. Eckles, D. G. Rand. 2025. Tendencies toward triadic closure: Field experimental evidence. *Proceedings of the National Academy of Sciences* **122**(27) e2404590122. doi:10.1073/pnas.2404590122.

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

Olley, G. S., A. Pakes. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* **64**(6) 1263–1297.

Osadchiy, N., W. Schmidt, J. Wu. 2021. The bullwhip effect in supply networks. *Management Science* **67**(10) 6153–6173.

Robins, G., P. Pattison, Y. Kalish, D. Lusher. 2007. An introduction to exponential random graph (p*) models for social networks. *Social networks* **29**(2) 173–191.

Serpa, J. C., H. Krishnan. 2018. The impact of supply chains on firm-level productivity. *Management Science* **64**(2) 511–532.

Sheffi, Y. 2005. *The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage*. MIT Press.

Vincenty, T. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review* **23**(176) 88–93.

Wang, Y., J. Li, D. Wu, R. Anupindi. 2021. When ignorance is not bliss: An empirical analysis of subtier supply network structure on firm risk. *Management Science* **67**(4) 2029–2048. doi:10.1287/mnsc.2020.3645. URL `https://pubsonline.informs.org/doi/10.1287/mnsc.2020.3645`.

Wooldridge, J. M. 2015. Control function methods in applied econometrics. *Journal of Human Resources* **50**(2) 420–445.

E-Companion to: "Microfoundations of Link Formation in Supply Networks"

EC.1. MLE for the correlated random effects logit

For each supplier–customer dyad (i,j) observed over periods $t=1,\dots,T$, let $y_{ijt}\in\{0,1\}$ denote the link indicator and $\mathbf{x}_{ijt}$ the covariates. The latent index is

$$y^*_{ijt}=\mathbf{x}'_{ijt}\boldsymbol{\beta}+\alpha_{ij}+\varepsilon_{ijt},\qquad \varepsilon_{ijt}\sim\text{Logistic}(0,1),\qquad y_{ijt}=\mathbb{1}\{y^*_{ijt}>0\}. \tag{EC.1}$$

We use a correlated random effects (CRE) specification

$$\alpha_{ij}=\mu+\bar{\mathbf{x}}'_{ij}\boldsymbol{\gamma}+a_{ij},\qquad a_{ij}\sim\mathcal{N}(0,\sigma_a^2),\qquad a_{ij}\perp\mathbf{X}_{ij}, \tag{EC.2}$$

where $\bar{\mathbf{x}}_{ij}\equiv T^{-1}\sum_{t=1}^{T}\mathbf{x}_{ijt}$ and $\mathbf{X}_{ij}=\{\mathbf{x}_{ij1},\dots,\mathbf{x}_{ijT}\}$. Define

$$\eta_{ijt}\equiv\mathbf{x}'_{ijt}\boldsymbol{\beta}+\mu+\bar{\mathbf{x}}'_{ij}\boldsymbol{\gamma}. \tag{EC.3}$$

Conditional on a_{ij}, outcomes are independent over t with

$$\Pr(y_{ijt}=1\mid\mathbf{X}_{ij},a_{ij})=\Lambda(\eta_{ijt}+a_{ij}),\qquad \text{where } \Lambda(u)=\frac{e^u}{1+e^u}. \tag{EC.4}$$

Dyad likelihood and marginalization over the random effect. Let $\mathbf{y}_{ij}=(y_{ij1},\dots,y_{ijT})'$. The conditional likelihood for dyad (i,j) given a_{ij} is

$$\mathcal{L}_{ij}(\boldsymbol{\theta}\mid a_{ij})=\prod_{t=1}^{T}\Big[\Lambda(\eta_{ijt}+a_{ij})\Big]^{y_{ijt}}\Big[1-\Lambda(\eta_{ijt}+a_{ij})\Big]^{1-y_{ijt}}, \tag{EC.5}$$

with parameter vector $\boldsymbol{\theta}=(\boldsymbol{\beta},\mu,\boldsymbol{\gamma},\sigma_a^2)$. Since a_{ij} is unobserved, the dyad contribution to the marginal likelihood integrates over its density:

$$L_{ij}(\boldsymbol{\theta})=\int_{-\infty}^{\infty}\mathcal{L}_{ij}(\boldsymbol{\theta}\mid a)\,\phi(a;0,\sigma_a^2)\,\mathrm{d}a=\int_{-\infty}^{\infty}\prod_{t=1}^{T}\Big[\Lambda(\eta_{ijt}+a)\Big]^{y_{ijt}}\Big[1-\Lambda(\eta_{ijt}+a)\Big]^{1-y_{ijt}}\phi(a;0,\sigma_a^2)\,\mathrm{d}a, \tag{EC.6}$$

where $\phi(\cdot;0,\sigma_a^2)$ is the $\mathcal{N}(0,\sigma_a^2)$ density.

Sample log-likelihood and MLE. Let $\mathcal{D}$ be the set of observed dyads. The sample log-likelihood is

$$\ell(\boldsymbol{\theta})=\sum_{(i,j)\in\mathcal{D}}\log L_{ij}(\boldsymbol{\theta}), \tag{EC.7}$$

and the maximum likelihood estimator is

$$\hat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}\ \ell(\boldsymbol{\theta}). \tag{EC.8}$$

Computation. We evaluate the one-dimensional integral by adaptive Gauss–Hermite quadrature with Q nodes. In practice we estimate the model by adding Mundlak averages for the set of time-varying regressors and running `xtlogit, re` in Stata (with 6 `intpoints`), which uses adaptive Gaussian quadrature.

EC.2. Supplementary Tables

Table EC.1 Estimated ratios between robust SE clustered by dyad and OIM SE. Estimation is done on 50 samples, each randomly drawing 2% of dyads without replacement. Mean ratios are reported.

	(1)	(2)	(3)
Supplier TFP	1.2023	1.3064	1.3590
Customer TFP	1.1322	1.2220	1.2749
Supplier Out-degree		1.4008	1.4325
Customer In-degree		1.1475	1.2772
Common Partners		1.3829	1.4339
Same Country			1.6112
Same Industry			1.4662
ln(Distance)			1.4326
Abs. Diff ln(AT)			1.5132
Abs. Diff ln(Capital)			1.4898
Constant	1.2792	1.4435	1.6060
lnsig2u	5.9524	2.3696	2.6750

EC.3. Linear Independence of I, g, and g^2 in the Production Network

Let g^{t-1} denote the *row-normalized* directed adjacency matrix built from links at time $t-1$ (rows index customers, columns index suppliers). We use tier-2 instruments constructed from $h^{t-1} \equiv (g^{t-1})^2$, for example, the maximum TFP among second-degree suppliers. See § 3.3 for instrument definitions and first-stage equations. Following Bramoullé et al. (2009), a sufficient condition for the validity of instruments based on $h^{t-1}x$ is that the matrices $\{I, g^{t-1}, h^{t-1}\}$ are linearly independent. This ensures that $h^{t-1}x$ contributes variation that is not spanned by x and $g^{t-1}x$.

Lemma EC.1 (Submatrix criterion for linear independence). *Let $R, C \subset \{1, \ldots, n\}$ be subsets of row and column indices. If the three submatrices $\{I_{R,C}, (g^{t-1})_{R,C}, (h^{t-1})_{R,C}\}$ are linearly independent in $\mathbb{R}^{|R|\times|C|}$, then the full matrices $\{I, g^{t-1}, h^{t-1}\}$ are linearly independent in $\mathbb{R}^{n\times n}$.*

Define the linear map $T : \mathbb{R}^{n\times n} \to \mathbb{R}^{|R|\times|C|}$ by $T(A) = A_{R,C}$. For any scalars α, β, γ,

$$T(\alpha I + \beta g^{t-1} + \gamma h^{t-1}) = \alpha I_{R,C} + \beta (g^{t-1})_{R,C} + \gamma (h^{t-1})_{R,C}.$$

If the right-hand side equals the zero matrix only when $\alpha = \beta = \gamma = 0$, then the left-hand side does as well, which proves the claim.

1. A certifying subgraph in our data. We identify three distinct nodes $(i^\star, s_1^\star, s_2^\star)$ at $t-1$ such that there is a chain $s_2^\star \to s_1^\star \to i^\star$, with $(g^{t-1})_{i^\star s_1^\star} > 0$, $(g^{t-1})_{s_1^\star s_2^\star} > 0$, and no direct edge $s_2^\star \to i^\star$ (so $(g^{t-1})_{i^\star s_2^\star} = 0$). Take rows $R = \{i^\star, s_1^\star, s_2^\star\}$ and columns $C = \{i^\star, s_1^\star, s_2^\star\}$. Then $(I)_{R,C}$ is the 3×3

identity, $(g^{t-1})_{R,C}$ has nonzero off-diagonal entries at $(i^\star, s_1^\star)$ and $(s_1^\star, s_2^\star)$, and $(h^{t-1})_{R,C}$ has a nonzero off-diagonal entry at $(i^\star, s_2^\star)$. These patterns certify that $\{I_{R,C}, (g^{t-1})_{R,C}, (h^{t-1})_{R,C}\}$ are linearly independent, hence $\{I, g^{t-1}, h^{t-1}\}$ are linearly independent by the lemma.[18]

2. Numerical Verification. Forming $M = [\mathrm{vec}(I), \mathrm{vec}(g^{t-1}), \mathrm{vec}(h^{t-1})] \in \mathbb{R}^{n^2 \times 3}$ and checking $\mathrm{rank}(M) = 3$ is costly when n is large. Instead we compute the 3×3 Gram matrix

$$\Gamma \equiv M^\top M = \begin{bmatrix} \langle I, I\rangle_F & \langle I, g^{t-1}\rangle_F & \langle I, h^{t-1}\rangle_F \\ \langle g^{t-1}, I\rangle_F & \langle g^{t-1}, g^{t-1}\rangle_F & \langle g^{t-1}, h^{t-1}\rangle_F \\ \langle h^{t-1}, I\rangle_F & \langle h^{t-1}, g^{t-1}\rangle_F & \langle h^{t-1}, h^{t-1}\rangle_F \end{bmatrix}, \quad \text{with } \langle A, B\rangle_F \equiv \mathrm{tr}(A^\top B).$$

Since $\Gamma = M^\top M$, the columns of M are linearly independent if and only if Γ is full rank. We report $\mathrm{rank}(\Gamma)$, its eigenvalues, and the condition number $\kappa(\Gamma)$. For the g^{t-1} used in our IV construction, the eigenvalues are 740.11, 128.94, and 37.431, so $\mathrm{rank}(\Gamma) = 3$ and $\{I, g^{t-1}, h^{t-1}\}$ are linearly independent.

3. Network intransitivity diagnostic. Finally, we also summarize intransitivity by the share of two-step supplier neighbors that are not direct suppliers,

$$\Pr\{h_{ik}^{t-1} > 0 \text{ and } g_{ik}^{t-1} = 0\}.$$

In our data this share is 0.944, meaning about 94.4% of two-step connections are not direct neighbors. This indicates substantial second-order structure beyond first-order links and supports the rank condition behind using $h^{t-1}x$ as an instrument.

[18] We assume no self-loops. The "no direct edge $s_2^\star \to i^\star$" condition is satisfied in sparse, intransitive networks and can be verified empirically.

EC.4. MCMLE for the Structural Model

Algorithm 1 MCMLE for the Structural Model (formation and dissolution)

Inputs: Observed transitions $\{g^{t-1}, g^t\}_{t=1}^{T}$; initial parameters $(\theta_+^{(0)}, \theta_-^{(0)})$; Monte Carlo sample size M; burn-in B; thinning L; step sizes $\{\eta_r\}$

Outputs: $\hat{\theta}_+$, $\hat{\theta}_-$

for iteration $r = 1, 2, \ldots$ until convergence **do**

Simulate, for each $t = 1, \ldots, T$**:**

Run an MH sampler targeting $P_{\theta^{(r-1)}}(g \mid g^{t-1})$ over the constrained sample space:

$$\mathcal{G}_t \equiv \big\{ g : g_{ij}^{t-1} = 0 \text{ dyads may toggle for formation, and } g_{ij}^{t-1} = 1 \text{ dyads may toggle for dissolution} \big\}.$$

After burn-in B, collect M draws $\{G_t^{(m)}\}_{m=1}^{M}$, keeping every L-th iterate.

Monte Carlo expectations:

$$\hat{\mu}_+^{(r)} \equiv \frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{m=1}^{M} S_{\text{form}}\big(G_t^{(m)}, g^{t-1}\big), \quad \hat{\mu}_-^{(r)} \equiv \frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{m=1}^{M} S_{\text{diss}}\big(G_t^{(m)}, g^{t-1}\big).$$

Observed statistics:

$$S_{+,\text{obs}} \equiv \frac{1}{T}\sum_{t=1}^{T} S_{\text{form}}(g^t, g^{t-1}), \quad S_{-,\text{obs}} \equiv \frac{1}{T}\sum_{t=1}^{T} S_{\text{diss}}(g^t, g^{t-1}).$$

Score estimates:

$$U_+^{(r)} \equiv S_{+,\text{obs}} - \hat{\mu}_+^{(r)}, \qquad U_-^{(r)} \equiv S_{-,\text{obs}} - \hat{\mu}_-^{(r)}.$$

Optional Fisher-scoring: estimate $\hat{V}_+^{(r)}$ and $\hat{V}_-^{(r)}$ as the sample covariances of S_{form} and S_{diss} across $\{G_t^{(m)}\}$, then precondition the step.

Parameter updates:

$$\theta_+^{(r)} = \theta_+^{(r-1)} + \eta_r\, U_+^{(r)} \quad (\text{or } \eta_r\, \hat{V}_+^{(r)\,-1} U_+^{(r)}), \qquad \theta_-^{(r)} = \theta_-^{(r-1)} + \eta_r\, U_-^{(r)} \quad (\text{or } \eta_r\, \hat{V}_-^{(r)\,-1} U_-^{(r)}).$$

Stopping rule: stop if $\max\{\|U_+^{(r)}\|_\infty, \|U_-^{(r)}\|_\infty\} < \varepsilon$ and relative parameter change is below tolerance.

return $(\hat{\theta}_+, \hat{\theta}_-) = (\theta_+^{(r)}, \theta_-^{(r)})$

EC.5. Difference between Structural Model and dyad-wise logit likelihoods

This appendix contrasts a dyad-wise logit model that includes a transitivity covariate with a network model from the ERGM family. The goal is to highlight what is being assumed in a dyad-wise specification and what is captured when modeling the network jointly.

Consider an undirected network on five labeled nodes $\{A, B, C, D, E\}$ at time t. The observed edges are A–B, B–C, C–A, and D–E. There are $\binom{5}{2} = 10$ dyads in total, so we observe 4 edges and 6 non-edges. The triangle A–B–C is the only triangle in the network.

Dyad-wise Logit with a Transitivity Covariate The dyad-wise model treats dyads as conditionally independent given covariates. For dyad (i, j) we include an intercept and the number of common neighbors:

$$\Pr\left(g_{ij}^t = 1 \,\middle|\, \mathrm{CN}_{ij}^t\right) = \frac{\exp\left(\beta_0 + \beta_1 \,\mathrm{CN}_{ij}^t\right)}{1 + \exp\left(\beta_0 + \beta_1 \,\mathrm{CN}_{ij}^t\right)}.$$

Here CN_{ij}^t is the count of nodes that are adjacent to both i and j in the time-t snapshot g^t. For the observed network, the common-neighbor counts are: (i) A–B, B–C, C–A: (ii) $\mathrm{CN} = 1$ (the third vertex of the triangle); D–E: $\mathrm{CN} = 0$; and (iii) All six non-edges: $\mathrm{CN} = 0$. With $(\beta_0, \beta_1) = (-2, 1)$ define

$$p_1 \equiv \Pr(g_{ij}^t = 1 \mid \mathrm{CN} = 1) = \frac{e^{-1}}{1 + e^{-1}} \approx 0.26894, \qquad p_0 \equiv \Pr(g_{ij}^t = 1 \mid \mathrm{CN} = 0) = \frac{e^{-2}}{1 + e^{-2}} \approx 0.11920.$$

Three observed edges have $\mathrm{CN} = 1$ and one observed edge has $\mathrm{CN} = 0$, and all six non-edges have $\mathrm{CN} = 0$. The dyad-wise likelihood therefore factors as

$$P(g^t) = p_1^3 \, p_0 \, (1 - p_0)^6 \approx (0.26894)^3 (0.11920)(0.88080)^6 \approx 0.00108.$$

The independence claim is valid only when covariates are exogenous to g^t. If CN_{ij}^t is computed from the same snapshot g^t, then it depends on other dyads' outcomes, and the product likelihood above is not a valid likelihood. Two standard fixes are:

1. Compute CN_{ij} from a lagged network, for example $\mathrm{CN}_{ij}(g^{t-1})$.
2. Use covariates that do not depend on g^t, for example geography or pre-specified cluster membership.

Static ERGM An ERGM assigns a joint probability to the entire network

$$P(G^t \mid \theta) = \frac{\exp\left(\theta^\top s(g^t)\right)}{Z(\theta)}, \quad Z(\theta) = \sum_{g' \in \mathcal{G}} \exp\left(\theta^\top s(g')\right),$$

where G^t is the random network at time t, g^t its realization, and $s(\cdot)$ collects network statistics. Using two statistics, edges and triangles, $s(g^t) = \left(\mathrm{Edges}(g^t), \mathrm{Triangles}(g^t)\right)$, the observed network

has $s(g^t) = (4, 1)$. For parameters $(\theta_{\text{edges}}, \theta_{\text{tri}}) = (-2, 1)$ the unnormalized weight is $\theta^\top s(g^t) = -2 \cdot 4 + 1 \cdot 1 = -7, \qquad \exp(-7) \approx 0.00091$. Since there are $2^{10} = 1024$ labeled undirected graphs on five nodes, $Z(\theta)$ can be computed by enumeration in this toy example.[19]

Temporal Extension A temporal ERGM conditions on the lagged network and allows statistics to depend on both g^t and g^{t-1}:

$$P(G^t \mid g^{t-1}, \theta) = \frac{\exp\big(\theta^\top s(g^t, g^{t-1})\big)}{Z(g^{t-1}, \theta)}.$$

For instance, one may augment edges and triangles with a stability term, $s(g^t, g^{t-1}) = \Big(\text{Edges}(g^t), \text{Triangles}(g^t), \text{Edges}(g^t \cap g^{t-1})\Big)$, which rewards ties that persist from $t-1$ to t. The separable structural model has formation and dissolution as distinct components. Writing $s^+(g^t, g^{t-1})$ for statistics computed on ties that form between $t-1$ and t, and $s^-(g^t, g^{t-1})$ for statistics on ties that persist, the likelihood factor is

$$P(G^t \mid g^{t-1}, \theta^+, \theta^-) \propto \exp\Big((\theta^+)^\top s^+(g^t, g^{t-1}) + (\theta^-)^\top s^-(g^t, g^{t-1})\Big).$$

This separation often improves interpretability, since the same covariate can have different effects on the chance a tie forms and on the chance an existing tie survives.

The dyad-wise logit with a common-neighbors covariate approximates transitivity using local information and relies on a conditional independence assumption. ERGM-based models assign probability to the network jointly, capture dependence through global statistics such as triangles and stability, and therefore represent clustering and temporal persistence more directly. In small toy settings one can compute normalizing constants exactly, while in practice MCMC-based approximations are standard.

[19] In empirical applications, the state space is enormous, so $Z(\theta)$ is approximated with MCMC.

EC.6. Sample size and Robustness

The structural mode requires a common set of nodes across all periods, so the number of persistent firms falls as the panel window lengthens. To assess robustness, we re-estimate the model while fixing the end year at 2022 and varying the start year from 2013 to 2018. The tables below report the number of persistent firms and the resulting count of directed dyad–period observations, and compare model coefficients across windows.

Table EC.2 Persistent firms and directed dyad–period observations by start year, holding the end year at 2022.

Start year	Persistent firms (N_{persist})	DF $= N_{\text{persist}}(N_{\text{persist}}-1)T$
2013	600	3,234,600
2014	662	3,500,656
2015	740	3,828,020
2016	857	4,401,552
2017	994	4,935,210
2018	1,166	5,433,560

Across start years 2013 to 2018, the signs and magnitudes of key parameters are stable (see Table EC.3). Edges remain negative and highly significant, GWESP remains positive in the formation equation and negative in the dissolution equation, and homophily terms are consistently positive. These patterns support the robustness of our baseline results.

Table EC.3 Structural model estimates by panel start year, holding the end year at 2022. Columns report separate formation and dissolution models for each window.

	2013	2015	2016	2017	2018
Formation Model					
Edges	-9.20^{***}	-9.03^{***}	-8.67^{***}	-8.85^{***}	-9.58^{***}
	(0.35)	(0.29)	(0.25)	(0.26)	(0.32)
GWESP (OTP)	0.87^{***}	0.93^{***}	0.90^{***}	0.87^{***}	0.83^{***}
	(0.06)	(0.04)	(0.04)	(0.03)	(0.03)
$\mathbb{1}_{\{\text{Industry}_i=\text{Industry}_j\}}$	1.31^{***}	1.37^{***}	1.20^{***}	1.00^{***}	0.88^{***}
	(0.17)	(0.15)	(0.14)	(0.14)	(0.13)
$\mathbb{1}_{\{\text{Location}_i=\text{Location}_j\}}$	0.49^{***}	0.32^{*}	0.36^{**}	0.56^{***}	0.76^{***}
	(0.14)	(0.13)	(0.12)	(0.10)	(0.09)
$\lvert\text{Capital}_i - \text{Capital}_j\rvert$	-0.12^{***}	-0.08^{***}	-0.05^{**}	-0.01	-0.01
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
$\lvert\log(\text{Firm Size})_i - \log(\text{Firm Size})_j\rvert$	0.19^{***}	0.15^{***}	0.13^{***}	0.12^{***}	0.11^{***}
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Customer TFP	0.09^{**}	0.10^{***}	0.08^{***}	0.05^{*}	0.06^{**}
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Supplier TFP	0.05	0.05^{*}	0.04	0.05^{*}	0.06^{**}
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Dissolution Model					
Edges	-1.69^{***}	-1.31^{***}	-1.44^{***}	-1.27^{***}	-1.39^{***}
	(0.40)	(0.30)	(0.26)	(0.26)	(0.35)
GWESP (OTP)	-0.25^{**}	-0.25^{***}	-0.12^{*}	-0.15^{***}	-0.16^{***}
	(0.08)	(0.06)	(0.05)	(0.04)	(0.04)
$\lvert\text{Capital}_i - \text{Capital}_j\rvert$	-0.03	-0.04	-0.02	-0.02	-0.06^{*}
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
$\lvert\log(\text{Firm Size})_i - \log(\text{Firm Size})_j\rvert$	0.02	0.01	-0.01	-0.04	0.01
	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)
Customer TFP	-0.04	-0.03	-0.00	-0.01	0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Supplier TFP	-0.04	-0.02	-0.03	-0.04	-0.05^{*}
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Nodes	600	740	857	994	1166
AIC	15456	19384	23336	25670	29340
BIC	15716	19647	23602	25938	29610

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Standard errors in parentheses.