EPJ Data Science
a SpringerOpen Journal

RESEARCH

Open Access

Efficiency and resilience: key drivers of distribution network growth

Ambra Amico[1*] iD, Giacomo Vaccario[1] and Frank Schweitzer[1]

*Correspondence:
ambra.amico@smart.mit.edu
[1] Chair of Systems Design, ETH
Zürich, Weinbergstrasse 56/58,
Zürich, 8092, Switzerland

Abstract

Networks to distribute goods, from raw materials to food and medicines, are the backbone of a functioning economy. They are shaped by several supply relations connecting manufacturers, distributors, and final buyers worldwide. We present a network-based model to describe the mechanisms underlying the emergence and growth of distribution networks. In our model, firms consider two practices when establishing new supply relations: centralization, the tendency to choose highly connected partners, and multi-sourcing, the preference for multiple suppliers. Centralization enhances network efficiency by leveraging short distribution paths; multi-sourcing fosters resilience by providing multiple distribution paths connecting final buyers to the manufacturer. We validate the proposed model using data on drug shipments in the US. Drawing on these data, we reconstruct 22 nationwide pharmaceutical distribution networks. We demonstrate that the proposed model successfully replicates several structural features of the empirical networks, including their out-degree and path length distributions as well as their resilience and efficiency properties. These findings suggest that the proposed firm-level practices effectively capture the network growth process that leads to the observed structures.

Keywords: Network growth model; Supply chains; Distribution networks; Resilience; Efficiency; Multi-sourcing; Centralization; Data-driven model

1 Introduction

Our daily lives depend on numerous essential goods, such as food, clothes, and medicines. Before reaching their final buyers, goods follow long journeys: they are first produced by manufacturers, they travel vast geographical distances while passing through multiple distributors. The interactions of manufacturers and distributors give rise to intricate distribution networks that grow more complex every year [1, 2]. This has recently revealed some significant downsides. The Covid-19 pandemic and the conflict in Ukraine have highlighted that even local shortages can be amplified through the supply linkages and ultimately affect millions of people [3–6]. These events have called for a deeper understanding of distribution networks' structures and how these structures affect their resilience [7].

Traditionally, scholars of supply chain management and operations logistics have conceptualized distribution systems as linear chains. Using this perspective implies that supply chains in principle can be fully designed by a single manufacturer [8–10]. However,

Springer

nowadays, this conventional approach falls short. While firms could choose their partners, they have limited control over the business relations of those partners [11]. In other words, the connections within the distribution system extend beyond the control of a single entity, and the resulting structure strongly deviates from a simple chain. Thus, today's distribution systems should be better viewed as self-organized systems emerging from the interactions of several firms [12, 13].
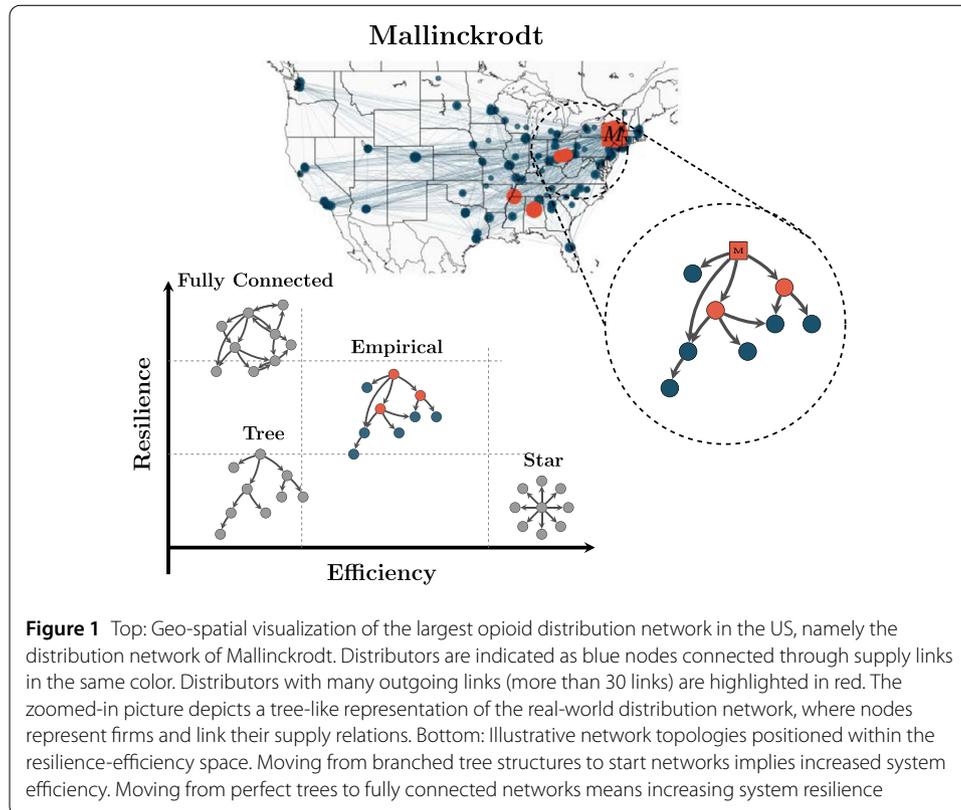
As recently highlighted [11, 14–16], these self-organized systems can be suitably represented as complex networks. Network science has provided tools to move beyond the oversimplified chain perspective. Yet, research in this direction has been limited by a lack of comprehensive data. Given this limitation, previous research has been constrained to small-scale case studies [17, 18] or simulations without empirical validation [9, 19–21]. Network models validated on large-scale distribution systems are still missing [11].

We advance this research by proposing and validating a network growth model for large-scale distribution networks. The model is parsimonious and considers only two fundamental necessities of these networks: efficiency and resilience. Efficiency is the ability to deliver goods to final buyers in a timely and cost-effective manner [22, 23]. Resilience, instead, is the ability to withstand, adapt and recover from disruptions [24, 25]. These are systemic properties and depend on the entire network structure. Since no single entity has control over the whole structure, these properties are not imposed top-down. Instead, they emerge from the aggregation of supply relations established between pairs of firms.

To what extent firms' interactions, at the micro-level, translate into an increase in efficiency and enhancement of resilience is an open question that we investigate. We model firms' interactions by considering two practices strictly related to efficiency and resilience: centralization and multi-sourcing. Centralization is the tendency of firms to link to other firms with the most connections [22]. Multi-sourcing, instead, is the tendency of firms to source products from multiple suppliers to decrease their exposure to single failures [26]. We formalize these practices as interaction rules for link formation and use them to explain the growth of large-scale distribution networks. Moreover, we explore their impact on the efficiency and resilience of empirical distribution networks.

Finally, we validate the model using data on over twenty pharmaceutical distribution networks spanning the whole US. These networks are reconstructed from the unique ARCOS dataset [27], that comprises all legal shipments of opioid drugs recorded between 2006 and 2014. Precisely, it lists 499,534,836 shipping records involving about 2,000 distributors and manufacturers firms and serving over 200,000 final buyers, such as pharmacies, hospitals, and practitioners. Drawing on these large-scale data, we obtain stylized facts of the empirical structures and check whether the proposed model reproduces them.

The remainder of the paper is organized as follows. In Sect. 2, we frame our study and its contribution in the supply chain management, and operational logistics literature. Further, we elucidate the concepts of efficiency and resilience as systemic properties. In Sect. 3, we introduce the model and formalize the centralization and multi-sourcing practices as interaction rules for link formation. Then, in Sect. 4, we first introduce the data used for calibration. Subsequently, we provide a comprehensive analysis and interpretation of the model parameters, followed by the validation of the model in reproducing key features of the empirical networks. Finally, Sect. 5 presents the concluding remarks.

Figure 1 Top: Geo-spatial visualization of the largest opioid distribution network in the US, namely the distribution network of Mallinckrodt. Distributors are indicated as blue nodes connected through supply links in the same color. Distributors with many outgoing links (more than 30 links) are highlighted in red. The zoomed-in picture depicts a tree-like representation of the real-world distribution network, where nodes represent firms and link their supply relations. Bottom: Illustrative network topologies positioned within the resilience-efficiency space. Moving from branched tree structures to start networks implies increased system efficiency. Moving from perfect trees to fully connected networks means increasing system resilience

2 Efficiency and resilience as systemic properties

The aim of this paper is to reconstruct and to explain the structure of large-scale distribution networks. In Fig. 1, we visualize the largest empirical network of opioid distributions in the US, namely the distribution network of Mallinckrodt. This network encompasses 1132 supply relationships involving 417 distributors across all 50 US states. On the right-hand side of the figure, we also provide a schematic representation of this network to highlight its tree-like structure. In this structure, the manufacturer serves as the root node, connecting to distributors, which in turn connect to other distributors or final buyers further down the tree.

Building on previous theoretical insights [14, 23, 28], we consider efficiency and resilience as the key drivers for the formation of these networks, as explained in the following. A supply chain is considered *efficient* if goods traverse a low number of firms on their way from the production to the consumption side [14, 23, 29]. Short distribution paths connecting manufacturers to final buyers favour a rapid transfer of information, thus facilitating more efficient material and financial flows [23]. Hence, with a few distributors operating along such paths, lead times are shortened, and inventory costs are kept down, thus enhancing the system's efficiency [22]. Then, maximum efficiency is reached when a single distributor, e.g., a central warehouse, manages all the shipments to final buyers. In this case, a fully centralized structure, i.e., a star network, is achieved. As the number of intermediary steps between manufacturers and final buyers increases, the network branches out, becoming a tree, and its efficiency decreases (see diagram in Fig. 1). We note that this concept of efficiency is different from the microeconomic definition where an efficient network maximizes the total utility of all firms [30, 31].
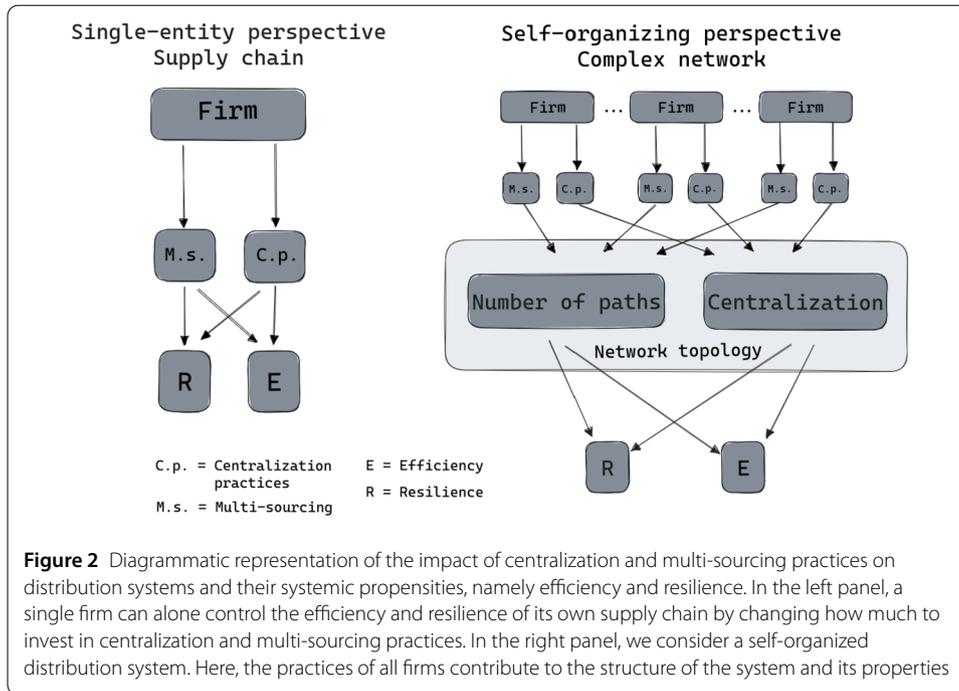
Firms tend to implement centralization practices, i.e., they try to connect to *central* firms with the most connections. These practices are appealing because they enable firms to reduce transportation, inventory and handling costs and also enhance communication [22, 32, 33]. Network centralization as a systemic property depends on the propensity of individual firms to implement centralization practices and their position in the network. The observation that different levels of network centralization emerges leads us to the question: *How do centralization practices of the single firm favour the efficient and centralized structure of empirical distribution networks?*

Efficiency is not the only concern of a distribution system; resilience also has a crucial role. In general, resilience implies the ability to withstand, adapt, and recover from disruptions [24, 25]. Resilient distribution networks must maintain functionality and ensure a minimum supply level during disruptions. One key characteristic of resilient distribution networks is the presence of multiple distribution paths connecting manufacturers to final buyers [34]. They showed that in case of disruptions, alternative paths not used under usual operations represent a crucial resource to mitigate supply shortages and reduce the supply deficit of final buyers. Within this path-based view, the least resilient network topology has single paths connecting the manufacturer to final buyers. This corresponds to a tree or a star structure with the manufacturers positioned at the top of the tree or the center of the star, respectively. Then, structures with higher levels of resilience are attained as nodes acquire incoming connections, and more distribution paths are available, e.g., in a fully connected network as schematically visualized in the diagram in Fig. 1.

From a firm's perspective, multi-sourcing is a valuable strategy to withstand disruptions. That means, firms source products from multiple suppliers to decrease their exposure to single failures [26]. This increases the number of distribution paths connecting manufacturers to final buyers and, hence, also improves the system's resilience. This increase, however, is not linear. Because firms are embedded in complex distribution networks, the effectiveness of their actions strongly depends on the actions of other firms and the overall network topology. This observation brings us to the question: *How do multi-sourcing practices of single firms favour the resilient and path-redundant structure of the whole distribution network?*

In Fig. 2, we use an illustrative diagram to better clarify the approach taken in this paper and the two questions we address. In the left panel, we consider a single firm that can decide about its own supply chain. Here, the firm can balance the trade-off between efficiency and resilience by changing how much to invest in centralization and multi-sourcing. In other words, the properties of the supply chain are controlled by a single firm. In contrast, in the right panel, we consider a self-organized distribution system that emerges from decision of all firms. Their centralization and multi-sourcing practices collectively impact the structure and the efficiency and resilience of the distribution network. Hence, no single firm controls the entire network and its systemic properties. By addressing our question, we aim to understand the impact of firms' decisions on the entire structure.

The above two questions lead to the central question of this work: *To what extent are the centralization and multi-sourcing practices of firms sufficient to reproduce the emerging structure of empirical distribution networks?* To answer this question, we propose a parsimonious model that captures firms' tendency towards centralization and multi-sourcing practices using only two parameters. By exploring the parameters' space, we reproduce a wide range of network structures, from fully centralized to very branched ones and from

Figure 2 Diagrammatic representation of the impact of centralization and multi-sourcing practices on distribution systems and their systemic propensities, namely efficiency and resilience. In the left panel, a single firm can alone control the efficiency and resilience of its own supply chain by changing how much to invest in centralization and multi-sourcing practices. In the right panel, we consider a self-organized distribution system. Here, the practices of all firms contribute to the structure of the system and its properties

perfect trees to almost fully connected ones. However, we do not limit ourselves to simply exploring the parameters space. We also estimate the model parameters using data from over twenty US pharmaceutical distribution networks, reconstructed from the ar-cos dataset. By doing so, we study how centralization and multi-sourcing practices affect the efficiency and resilience of real-world distribution networks. Moreover, we quantify the role of these practices in the growth of real-world distribution networks.

3 Model

Set-up   Let us consider a distribution network comprising N + 1 nodes, representing one manufacturer and N distribution firms; and E links, representing supply relations. Links are directed according to the direction of the shipments, from senders to receivers. Given a direct link $i \rightarrow j$, from sender i to receiver j, we define i as the *source* node and j as the *target* node. Then, d_i^{out} is i's out-degree, indicating the number of its target partners; and d_i^{in} is its in-degree, meaning the number of source partners i relies on.

We model the growth dynamic of this system, where new nodes join the network, and new links are established over time. At the initial time, $t = 1$, a simple chain of three nodes exists: $M \rightarrow i \rightarrow j$, where M is the root node representing the manufacturer. Then, at every time, a new supply link is formed between a source and a target node. Thus, the evolution of E is given by $E(t) = t + 1$.

The source node is selected among the existing nodes in the network, while the target node is either a newcomer or an existing node. Specifically, at a $1 - \alpha$ rate, the target node is a newcomer and forms a link with an existing node. At a rate α, a new link is established between two existing nodes in the network. The selection rules for the source and target node are clarified in the paragraphs below.

Note that, in this study, we exclude the root node M from forming new links. This is based on the observation from our data that manufacturers are mainly connected to a

single distributor (e.g., their warehouses), which then links to numerous other distribution firms. This setup does not constrain the model's generalizability.

Source node: centralization   The source node i is selected among the existing nodes with a probability $p_i(t)$, given by:

$$p_i(t) = q_s \frac{d_i^{\text{out}}(t)}{t} + (1 - q_s)\frac{1}{N(t)} \tag{1}$$

where the parameter q_s is used to interpolate between two mechanisms: a preferential attachment and a random one [35, 36]. Specifically, the first term on the right-hand side describes a preferential attachment mechanism, where the probability of being selected as a source node is proportional to the number of target partners the node has. Using the equivalences $E(t) = t + 1$, and $\sum_{i,i\neq M} d_i^{\text{out}}(t) = E(t) - 1$, we simply write t as normalization factor. The second term, instead, describes a random attachment mechanism where all existing nodes have an equal probability of being selected as source nodes.

Thus, by tuning q_s from zero to one, we can adjust the weight of the two mechanisms. In the extreme case with $q_s = 1$ (i.e., preferential attachment is fully dominant), a star network is attained. A single node in the centre connects to all other nodes, thus leading to the most centralized structure. By decreasing the value of q_s, we obtain less centralized networks and more branched out. In other words, q_s is the centralization rate of the network. It quantifies the probability that, in a given time, the source is a central node, namely a node with high out-degree.

Target node: multi-sourcing   The target node j is selected as a newcomer node at a rate $1 - \alpha$ and among the existing nodes at a rate α. In other words, at a rate α, an existing node j in the network forms a link with a new source partner, thus implementing multi-sourcing.

The probability for an existing node j to implement multi-sourcing is given by:

$$p_j(t) = q_t \frac{d_j^{\text{out}}(t)}{t - n_i(t)} + (1 - q_t)\frac{1}{N(t) - n_i(t)} \tag{2}$$

where $n_i(t) = d_i^{\text{out}} + 1$ and q_t is a model parameter. The factor n_i accounts for non-eligible targets, such as nodes already connected to the source and the source node itself. By this, we do not allow for self-loops and multi-edges. Thus, the normalization of Eq. (2) is equal to that of Eq. (1), apart from the correcting factors that exclude $n_i(t)$ non-eligible targets from the count.

The model parameter q_t is used to interpolate between two mechanisms: a preferential (first term) and a random mechanism (second term). According to the preferential mechanism, nodes with more target partners are more likely to implement multi-sourcing. On the other hand, the random mechanism assumes that every node has an equal probability of adopting multi-sourcing. Hence, q_t controls the *relative* propensity of nodes towards multi-sourcing. It interpolates between two scenarios: one where all nodes share the same propensity (the second term only) and the other where nodes exhibit maximum diversity in their propensity towards multi-sourcing (first term only).

In addition, tuning q_t has also a systemic effect. Higher q_t values result in an increase in the number of paths from root to leaves. If we consider a topology where all nodes have

only a single source partner, i.e., a perfect tree, a single path links the root to each leaf. Thus, the total number of paths matches the number of leaves. However, as we deviate from the perfect tree structure and allow nodes to have multiple parents, the number of paths starts to grow, and this growth can be controlled by the parameter q_t.

Finally, note that the preferential attachment mechanism in Eq. (2) is based on the out-degree of the target rather than on its in-degree or total degree, as proposed in previous studies [37–39]. Thus, the multi-sourcing strategy is more likely to be implemented by nodes with higher out-degrees than in-degrees. This choice reflects the idea that firms with more customers (i.e., target nodes) have, on average, higher demand, and they may need to source products from more suppliers to meet their demand.

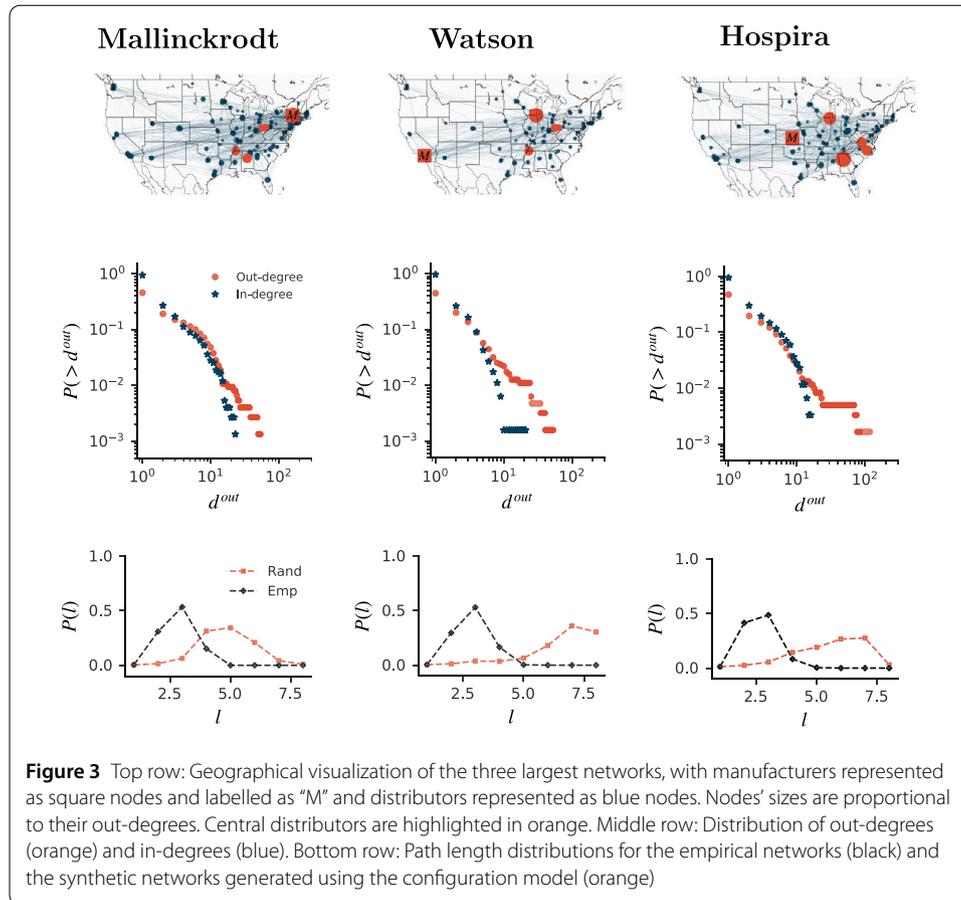## 4 Results

4.1 The US opioid distribution networks

Dataset   To test our model, we use the ARCOS, a dataset maintained by the Drug Enforcement Administration and recently made public by the US court [27]. This dataset represents the largest collection of shipping records available to date. It comprises *all legal shipments* of opioid drugs recorded in the United States between 2006 and 2014. There are 499,534,836 records involving approximately 1,928 distributing and manufacturing firms and serving over 200,000 final buyers across the United States, such as pharmacies, hospitals, and practitioners.

These records represent shipments of opioid drugs that are uniquely identified by their national drug code (NDC). Each NDC comprises 11 digits, with the first five serving as a unique manufacturer identifier. Grouping the records based on these initial five digits, we obtain all shipments of drugs produced by the same manufacturer. By this, we reconstruct and analyze the distribution networks of individual manufacturers. In the present study, we focus on the largest 22 opioid manufacturers, whose products appear in more than 70% of the shipping records.

Network representation   We study the annual snapshots of the 22 distribution networks. We represent manufacturers, distributors, and final buyers as nodes; and supply relations as links. Specifically, we consider a link between two nodes if at least one shipment has been observed between them in the given year. Although the dataset contains information on the quantity of shipped drugs, we do not consider it in the network representation because our focus is on the network structure.

Note that the number of final buyers in the dataset is two orders of magnitude larger than that of distributors. To ensure that the more numerous distributor-final buyer interactions don't mask distributors' interactions, we represent multiple final buyers connected to the same distributor as a single node. For the remainder of this paper, we do not distinguish between manufacturers, distributors, and final buyers and refer to them as nodes.

Stylized facts   In Fig. 3, we visualize the three largest networks on the US map and present their key macroscopic features, focusing on the out-degree and in-degree distributions (CCDF), as well as the distribution of path lengths. In particular, we analyze all paths linking the root to the leaves. These paths represent possible distribution routes to deliver products from the manufacturer to the final buyers.

Figure 3 Top row: Geographical visualization of the three largest networks, with manufacturers represented as square nodes and labelled as "M" and distributors represented as blue nodes. Nodes' sizes are proportional to their out-degrees. Central distributors are highlighted in orange. Middle row: Distribution of out-degrees (orange) and in-degrees (blue). Bottom row: Path length distributions for the empirical networks (black) and the synthetic networks generated using the configuration model (orange)

We see that all three networks examined share several topological features. First, all networks are characterized by a few hub firms, highlighted as orange nodes in the network visualization (top row of Fig. 3). These firms connect the manufacturer to numerous smaller firms, potentially retailer distributors. The heterogeneity of these firms becomes evident when evaluating the out-degree distributions (orange line in the middle row of Fig. 3). In all three cases, we observe a relatively small average (about 1.6), a larger standard deviation, and pronounced heavy tails, thus indicating the presence of a few hubs and many small firms.

Second, different from the out-degree, the in-degree distributions are usually narrower, with standard deviations ranging from 0.5 to 4.8 (blue line in the middle row of Fig. 3). Nonetheless, around 30% of the nodes have an in-degree value different from 1. This means that a non-negligible number of firms engage in multi-sourcing by establishing connections with multiple source partners.

Third, all the examined networks exhibit short path lengths. The average path length is 2.6, and the majority of them have a maximum length of 4 (bottom row of Fig. 3). This observation suggests that while the networks may not perfectly resemble star-like structures, the distance between manufacturers and consumers remains notably short. To provide a more quantitative assessment, we compare the empirical distributions with the ones derived from the configuration model [40]. This comparison reveals that the maximum path length generated by the random model (orange dashed line in the Figure) is nearly twice the value observed in the empirical networks.

Table 1 Key network features across 22 networks analyzed: Number of nodes $\tilde{N}$, number of links $\tilde{E}$, average in-degree and out-degree $\langle d^{\text{out/in}} \rangle$, average path length $\langle l \rangle$, and standard deviations of out and in-degrees $\sigma_{d}\text{out}$, $\sigma_{d}\text{in}$

Manufacturer	$\tilde{N}$	$\tilde{E}$	$\langle d^{\text{out/in}} \rangle$	$\langle l \rangle$	$\sigma_{d}\text{out}$	$\sigma_{d}\text{in}$	Manufacturer	$\tilde{N}$	$\tilde{E}$	$\langle d^{\text{out/in}} \rangle$	$\langle l \rangle$	$\sigma_{d}\text{out}$	$\sigma_{d}\text{in}$
Mallinckrodt	750	1467	1.96	2.83	8.36	2.67	Ucb	425	540	1.27	3.07	6.56	0.65
Watson	637	1017	1.60	2.86	7.76	1.54	Ortho-mcneil	425	790	1.86	2.81	9.77	1.58
Hospira	606	1243	2.05	2.64	10.63	2.46	Amneal	421	786	1.87	2.78	9.36	1.92
Mikart	547	1707	3.12	2.77	11.18	4.83	Watson NY	385	539	1.40	2.71	6.80	1.04
Generics	536	940	1.75	2.68	8.31	3.17	Pf Lab	382	714	1.87	2.70	7.97	1.75
Vintage	498	848	1.70	2.69	7.96	2.88	Janssen	382	644	1.69	2.81	8.51	1.47
Actavis	489	566	1.16	2.79	6.95	0.70	Mylan	378	411	1.09	2.87	5.68	0.49
Teva	481	573	1.19	2.98	5.49	0.88	Associates	376	621	1.65	2.64	7.63	2.95
Hospira NC	456	688	1.51	2.65	8.24	1.34	Novartis	363	644	1.77	2.74	7.05	1.73
Boehringer	453	931	2.06	2.76	7.96	2.61	Actavis FL	354	395	1.12	2.90	5.26	0.76
Baxter	449	662	1.47	2.78	6.88	1.55	Actavis NJ	312	361	1.16	2.72	5.98	0.87

So far, we presented our findings for the three largest networks. However, they remain valid for all the 22 networks examined. Table 1 summarizes the key network features for these networks.

4.2 Optimal parameters: estimation and interpretation

To validate the proposed model, we first identify the parameters that best fit our data, denoting them as the *optimal* parameters. We then interpret these parameters and show that the networks generated by feeding the model with the optimal parameters exhibit the distinctive features of the empirical ones.

Estimation   The model has three free parameters: α, q_s, and q_t. The parameter α can be determined analytically. Recalling that $E(t) = t + 1$, and that, on average, $N(t) = (1 - \alpha) \times t + 1$, we have:

$$\alpha = 1 - \frac{N(t) - 1}{E(t) - 1} \tag{3}$$

Thus, the optimal α is obtained by setting $N(t)$ and $E(t)$ to their corresponding empirical values, $\tilde{N}$ and $\tilde{E}$, respectively.

After estimating α, we are left with the two parameters q_s and q_t. To obtain their optimal values, we perform a grid search in the bi-dimensional parameters' space. We consider values of q_s and q_t ranging from 0 to 1, with an interval of 0.02. For each pair (q_s, q_t), we run the model 100 times and assign a *fitting score*. Thus, we perform a total of 250,000 computer simulations (for each year of observation) and stop every simulation when the generated network reaches the same number of links as the empirical one. Following the approach proposed by Tomasello et al. [41], we design a fitting score normalized between zero and one. Specifically, we compute the relative error, δ_Ω, for each generated network and for different network quantities, Ω:

$$\delta_\Omega(q_s, q_t) = \frac{\Omega_e - \Omega_s(q_s, q_t)}{\Omega_e} \tag{4}$$
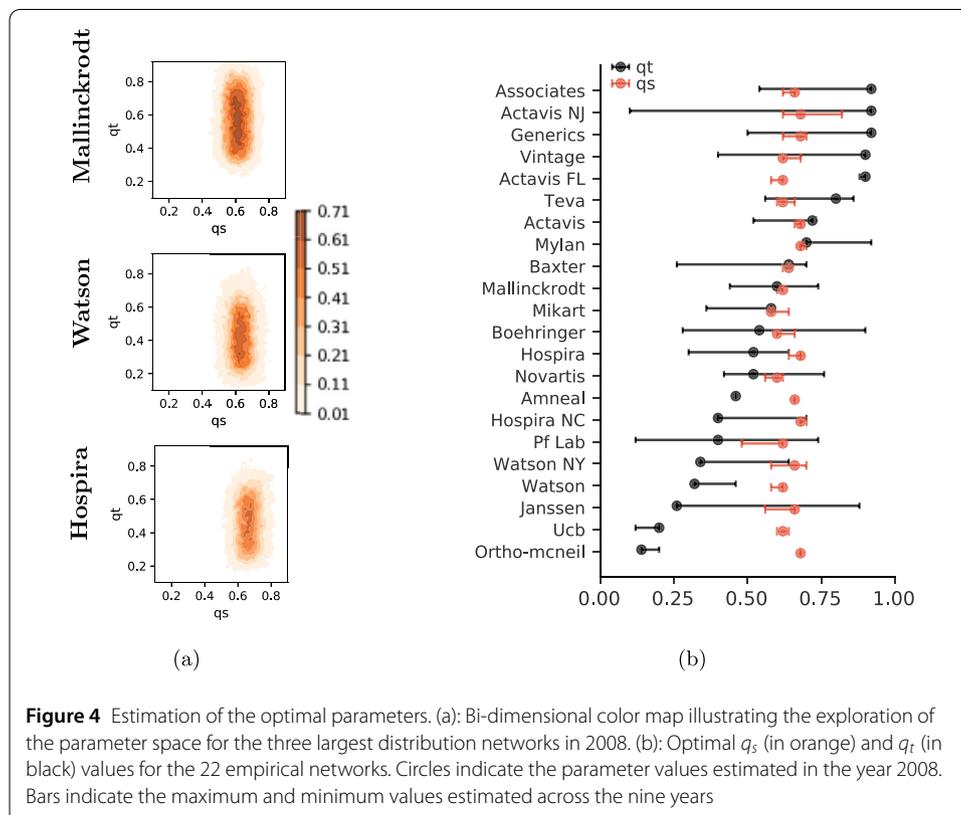
where the subscript e stands for empirical, the subscript s stands for simulated. We select five network quantities: the first and second moments of the distribution of out-degrees

and in-degrees and the average path length. We choose these quantities in a way that they include the minimum amount of information that, when used as model input, would allow us to replicate the real-world features. Besides the most straightforward choice of first moments, we include the second moments because the first moments alone are not very informative about highly-skewed distributions. In the validation section below, we test whether the proposed quantities to fit, combined with the model principles, are indeed sufficient to reproduce the key features of the empirical networks.

For each pair, (q_s, q_t), the fitting score is then given by the fraction of simulated networks for which the relative error δ_Ω is smaller than a given threshold ϵ_Ω for all network measures. We consider a 5% threshold on the first moments and a 25% threshold on the second moments.

We expect that the (q_s, q_t) pairs with higher fitting scores are those with high values for q_s and q_t. Low values of q_t would imply uniform in-degree distributions, whereas low q_s values would produce networks with long distribution paths. As discussed in Sect. 4.1, these features do not characterize the empirical networks that instead exhibit right-skewed in-degree distributions and short path lengths.

The exploration of the bi-dimensional space is visualized in Fig. 4a via the 2D color map for the three reference networks. As expected, low values of q_s and q_t return a very low fitting score while the optimal parameter pairs are $(0.62, 0.60)$, $(0.62, 0.32)$, $(0.68, 0.52)$ for the three networks. Moreover, within this parameter space, we observe a distinct optimal region represented by the dark orange color, where the fitting score reaches its maximum value of 0.71. This means that 71% of the networks generated by the model are close to the empirical one, within the threshold ϵ. The optimal region is narrow along the q_s di-

Figure 4 Estimation of the optimal parameters. (a): Bi-dimensional color map illustrating the exploration of the parameter space for the three largest distribution networks in 2008. (b): Optimal q_s (in orange) and q_t (in black) values for the 22 empirical networks. Circles indicate the parameter values estimated in the year 2008. Bars indicate the maximum and minimum values estimated across the nine years

mension and slightly broader along the q_t dimension. This suggests that, compared to q_s, this region has relatively more suboptimal values for q_t. Similar patterns are observed for all the networks analyzed (not shown).

Interpretation   The optimal q_s and q_t values for all the networks are reported in Fig. 4b. The circles mark the values obtained for the reference year (2008), while the bars denote the maximum and minimum values recorded over the nine years (2006-2014). We see that q_t values (in black) have broader variations ranging from 0.1 to 0.92. On the other hand, the values for the centralization rates q_s (in orange) are clustered within a relatively narrow range between 0.55 and 0.75.

Still, small variations of q_s can lead to high variations in terms of network centralization. To illustrate this, we measure the level of network centralization [42], $\mathcal{X}$, by taking into account the out-degree of each node, as:

$$\mathcal{X} = \sum_{i=1}^{N} \frac{d_{i*}^{\text{out}} - d_i^{\text{out}}}{N - 1} \tag{5}$$

where $i*$ is the node with the highest out-degree. Then, to obtain a value of $\mathcal{X}$ ranging from zero to one, we normalize the expression in Eq. (5) against the highest possible centralization value that is attained in a star configuration.

From Fig. 5a, we see that the level of network centralization increases with q_s. Yet, this increase is not linear and mainly occurs in the range $[0.5, 0.9]$. Interestingly, this range includes the optimal q_s for the analyzed data, highlighting diversity in centralization levels within empirical networks. These networks exhibit both medium and high degrees of centralization.

To further clarify the role of the centralization rate q_s, we present two network snapshots generated with $q_s = 0.9$ and $q_s = 0.1$. The network has a star-like configuration in the first case, while a more branched structure is observed in the second case. This topological
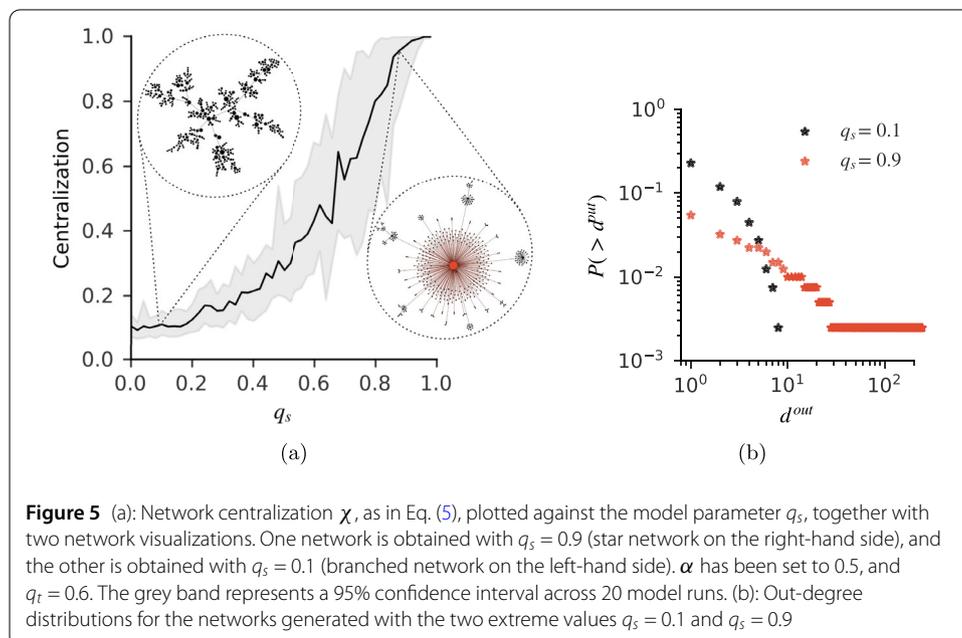
Figure 5 (a): Network centralization χ, as in Eq. (5), plotted against the model parameter q_s, together with two network visualizations. One network is obtained with $q_s = 0.9$ (star network on the right-hand side), and the other is obtained with $q_s = 0.1$ (branched network on the left-hand side). α has been set to 0.5, and $q_t = 0.6$. The grey band represents a 95% confidence interval across 20 model runs. (b): Out-degree distributions for the networks generated with the two extreme values $q_s = 0.1$ and $q_s = 0.9$

difference is also evident in the out-degree distributions, where larger values of q_s result in heavier tails, as depicted in Fig. 5b.

While q_s controls the network centralization, the parameter q_t controls the relative propensity of nodes to adopt multi-sourcing. Specifically, q_t enhances the diversity among nodes in their propensities toward multi-sourcing. With low values of q_t, nodes exhibit comparable propensities, leading to uniform and lower in-degrees across all nodes. Conversely, as q_t increases, certain nodes exhibit considerably higher in-degrees, while most nodes still maintain smaller values.

We show the effect of q_t at the node level by visualizing networks generated with $q_t = 0.1$ and 0.9 on the left and right sides of Fig. 6a, respectively. In the left network, only two nodes have in-degrees exceeding a given threshold, i.e., $d^{in} > 7$. These are depicted in orange. In the right network, instead, many more nodes surpass this threshold. This distinction is also highlighted in the distribution of in-degrees plotted in Fig. 6b: the higher q_t value results in a broader distribution.
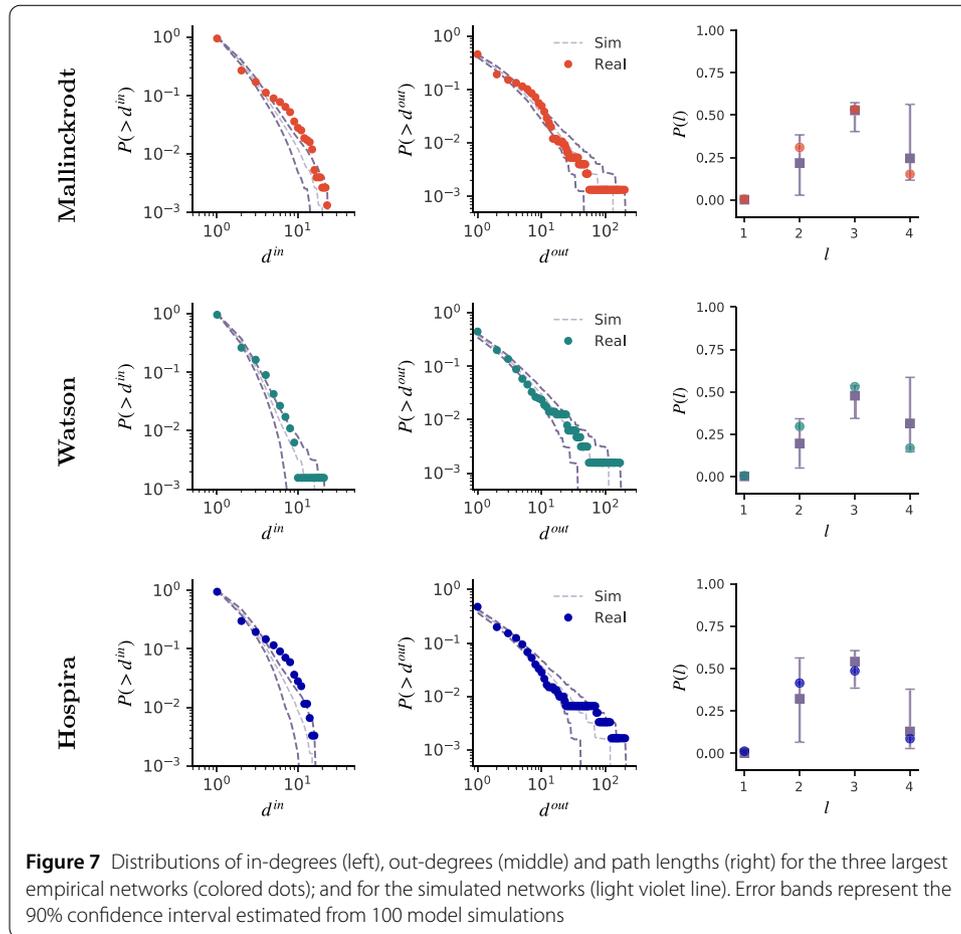
Finally, q_t does not only have node-level effects but also systemic ones. As discussed in Sect. 3, higher values of q_t lead to an increase in the total number of paths. To measure this increase, we define the *path increment* as the ratio by which the number of paths in a given network increases compared to those in a perfect tree of equivalent size, meaning with the same number of links.

Figure 6a depicts the path increment as a function of q_t. Notably, this increment ranges from 1 to 10^5 for a network comprising 400 links. This implies that, with high values of q_t, the number of paths expands by five orders of magnitude compared to a tree configuration.

Such a substantial increase in paths have also practical implications. With more distribution paths, firms may rely on multiple routes to supply products to the final buyer. In scenarios of disruption, even if some paths become unavailable, products can still reach their final buyers. Overall, increasing q_t may lead to the emergence of more resilient distribution networks. More details in the discussion.
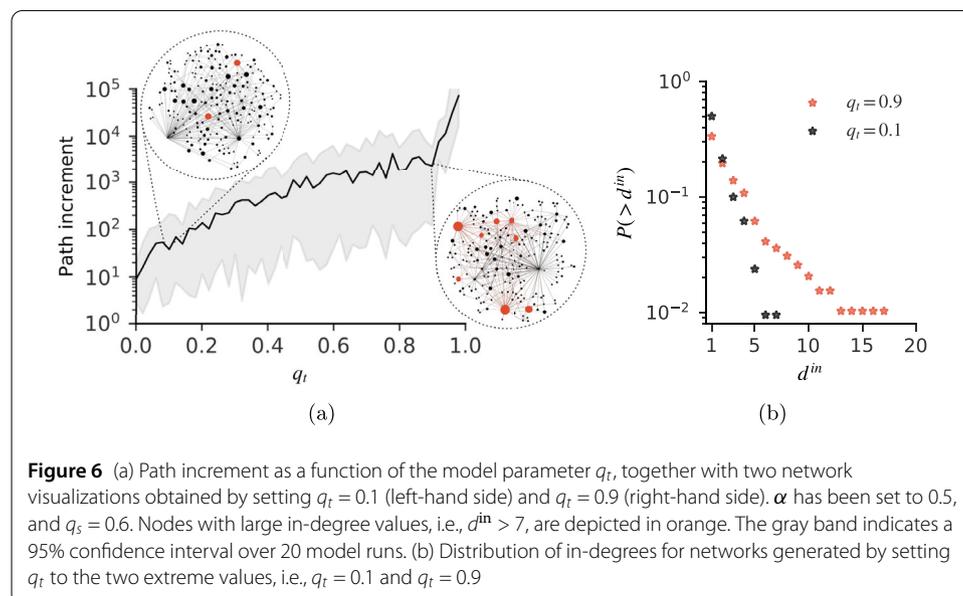
Figure 6 (a) Path increment as a function of the model parameter q_t, together with two network visualizations obtained by setting $q_t = 0.1$ (left-hand side) and $q_t = 0.9$ (right-hand side). α has been set to 0.5, and $q_s = 0.6$. Nodes with large in-degree values, i.e., $d^{in} > 7$, are depicted in orange. The gray band indicates a 95% confidence interval over 20 model runs. (b) Distribution of in-degrees for networks generated by setting q_t to the two extreme values, i.e., $q_t = 0.1$ and $q_t = 0.9$

Figure 7 Distributions of in-degrees (left), out-degrees (middle) and path lengths (right) for the three largest empirical networks (colored dots); and for the simulated networks (light violet line). Error bands represent the 90% confidence interval estimated from 100 model simulations

4.3 Optimal parameters: validation

We validate our model by assessing its ability to replicate key characteristics of the empirical networks. These include the stylized facts described in Sect. 4.1, as well as network features related to efficiency and resilience.

Stylized facts   For the stylized facts, we look at the distributions of in-degrees, out-degrees, and paths. Note that this information was not utilized to estimate the optimal model parameters. Throughout the parameter estimation, we solely considered the first and second moments of these distributions. In Fig. 7, we compare the empirical distributions (colored dots) and the ones obtained from the model simulations (light violet lines). Error bands represent the 90% confidence interval estimated from 100 simulations.

In the left column, we show the in-degree distributions and see that most of the empirical data fall within the error band generated by the simulations. This indicates that our model effectively captures the characteristic right-skewed nature of the distributions, including the presence of outliers in the tails. In the middle column, we examine the out-degree distributions. The model replicates the typical heavy-tail pattern observed in real-world networks. In the right column, we assess the path length distributions. The model accurately reproduces the peaked shape observed in the empirical data. Moreover, it captures the maximum distance of four steps between the manufacturer and the final buyers. We perform the Kolmogorov-Smirnov test on the out-degrees and in-degrees to compare

the distributions quantitatively. In 92% of the simulated out-degree distributions, we do not find significant differences with their empirical counterparts ($p < 0.01$). This is not true for the in-degree distributions, where we find statistical differences for most simulations. This result may be due to the mismatch observed in the correspondence of medium and low in-degree values. The model tends to underestimate the number of nodes with these in-degree values. All empirical values fall within the confidence interval for the path-length distribution, indicating a good match between the empirical and simulated data. This highlights the model's ability to reproduce not only the stylized facts of the networks but also the details of two of the distributions analyzed.

Efficiency   Here we consider two measures. The first one is the centralization index, $\mathcal{X}$, as expressed by Eq. (5). The second measure is the global network efficiency, introduced by Latora and Marchiori [43], and defined as the mean value of the inverse of the distances between all pair of nodes in a network. However, different from the original definition, we do not consider the distances between all pairs of nodes; instead, we only focus on the paths connecting the root (manufacturer) to the leaf nodes (final buyers). Thus, we obtain the following expression for global efficiency:

$$\mathcal{E} = \frac{1}{|J|} \sum_{j \in J} \frac{1}{d_{ij}} \tag{6}$$

where i is the root node, J is the set of leaf nodes, and d_{ij} is the topological distance between i and j. Note that centralization and global efficiency relate to each other: As the network becomes more centralized, the distances between the root and the leaf nodes decrease, thus increasing global efficiency.

In the left panel of Fig. 8, we compare the centralization (top-left) and efficiency (bottom-left) of the 22 empirical networks with the simulated ones. Orange dots represent the empirical networks, while gray dots represent the expected values from 100 model simulations and the error bars their estimated 90% confidence interval. We see that the centralization and efficiency of the empirical network follows within the error bars, indicating a good match between the empirical data and the model.

Resilience   To assess resilience, we maintain our path-based view [34] and consider two measures. First, we measure the average number of paths available to every leaf node to connect to the root. We call this number *paths available*. The higher the number of paths available, the higher the network's resilience. We complement this first measure with a random attack simulation [44, 45]. For every simulated and empirical network, we remove 10% of the nodes at random and compute the fraction of paths that remain available from the root to the leaf nodes, thus still allowing the network to function. We define this second measure as the *undisrupted path fraction*. Our path-based resilience measures have similarities with the *availability* and *accessibility* measures introduced by [44] but with some differences. The difference between *accessibility* and *paths available* is that the former focuses on the length of the paths, while the latter focuses on the number of paths. The shift in focus is because path lengths have already been discussed when looking at the networks' efficiency. The difference between *availability* and *undisrupted path fraction* lies in the fact that Zhao et al. [44] focus on the number of leaf *nodes* that stay connected to the root. In contrast, we focus on the number of *paths* linking leaves to the root [34].

Figure 8 Centralization (top-left), Efficiency (bottom-left), Paths available (top-right), and Undisrupted paths (bottom-right) measured for the 22 empirical networks (orange dots) and the simulated networks (gray dots and error bars). The gray dots represent the mean value over 100 simulations, and the error bars represent the 90% confidence interval
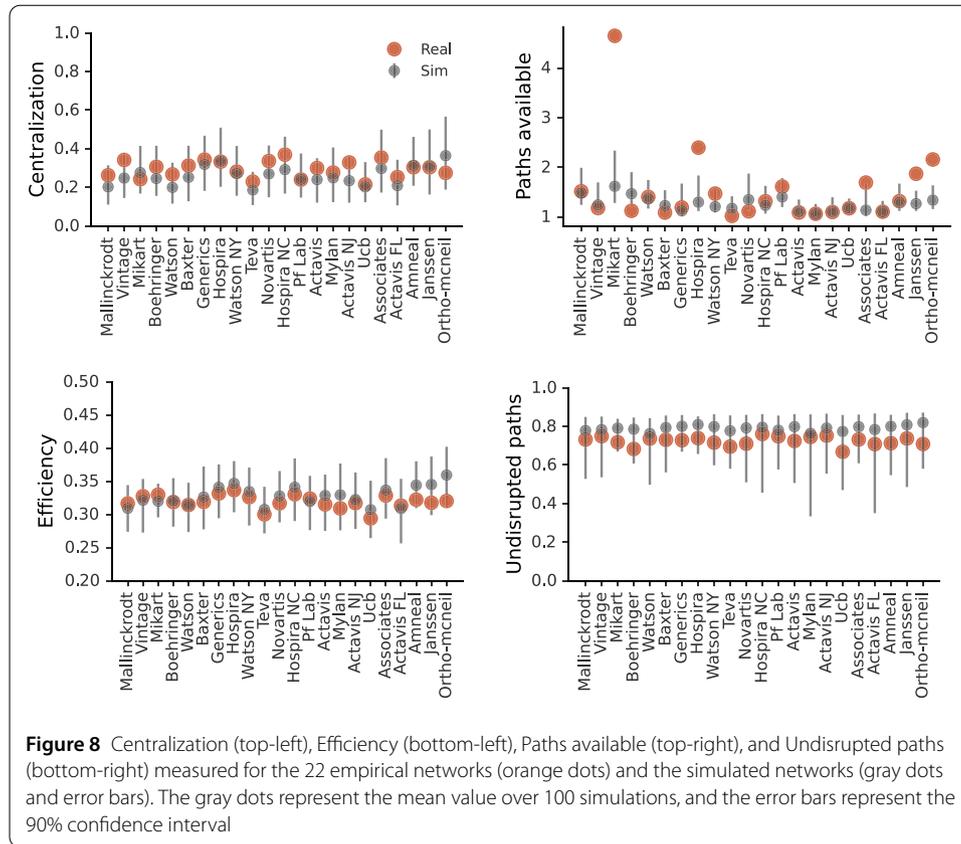
In the right panel of Fig. 8, we compare the paths available (top-right) and undisrupted paths fraction (bottom-right) of the 22 empirical networks with the simulated ones. Again, for the number of paths available, we see that the model is able to replicate the empirical number for the majority of the networks. Exceptions are the networks of Hospira, Janssen, and Ortho-mcneil where the model significantly underestimates this number. This may be due to the presence of more intermediary distributors with medium-low in-degree in the empirical networks compared to the simulated ones, as already discussed in relation to the in-degree distribution in Fig. 7. The presence of such distributors may provide additional paths and increase resilience, but do not affect the distance of the leafs nodes to the root (see bottom-right Fig. 7 and left panels in Fig. 8). For the undisrupted path fraction, the values of empirical networks always follows within the simulation confidence intervals, indicating a good match between the data and the model.

Overall, the comparisons between the simulations and real-world data, as illustrated in Fig. 7 and Fig. 8, demonstrate the strong ability of our model to replicate the topology of the distribution networks under study, as well as their resilience and efficiency.

5 Discussion and conclusions

Our economy crucially relies on distribution networks. These networks grow in size as new firms join and new supply relations are formed. Here, we argue that the growth of distribution networks is primarily driven by two necessities: efficiency and resilience.

In our view, efficiency and resilience are systemic properties not controlled by a single entity but emerging from the interactions between manufacturers, distributors, and final

buyers. Achieving an efficient and resilient distribution network depends on the collective decisions of all firms rather than on an individual choice.

The goal of this paper is to clarify how these firm-level decisions influence the growth of distribution networks and their systemic properties. To achieve this, we introduce a network growth model where firms select their partners by implementing multi-sourcing and centralization practices. We fine-tune and validate the model using data from 22 nationwide pharmaceutical distribution networks in the US.

We find that these real-world networks exhibit a high centralization rate. As they grow, approximately 60% of supply relations are formed with central firms. Although this percentage varies among different networks, it consistently remains above 50%. We conclude that the majority of supply relationships are formed through centralization practices. However, despite firms' dedicated efforts to implement centralization, the resulting networks do not closely resemble fully centralized and efficient structures. Instead, we find medium centralized networks, where up to 60% of the outgoing links are *not* established by the most central firm. Also, the global efficiency [43] measured in the empirical networks exhibits low values, consistently below 0.5, indicating the difference between the firm level and the systemic level.

Next, our research demonstrates that multi-sourcing increases the number of available distribution paths, which in turn enhances network resilience, as shown in [34]. However, the effectiveness of multi-sourcing practices depends considerably on specific firms. If firms are selected *randomly* to implement multi-sourcing, our simulation results show that the number of paths increases by one order of magnitude compared to the case where multi-sourcing is not implemented. Instead, if firms are selected depending on their *size*, the number of paths goes up to *five orders of magnitude.* Thus, our simulations indicate that firm heterogeneity is crucial when implementing this practice.

We confirm that this heterogeneity is indeed present in the data. In Fig. 4 we show that in most examined networks firms have very different propensities in implementing multi-sourcing. Specifically, firms with high out-degree tend to perform multi-sourcing, small firms perform single-sourcing. Only a smaller number of networks exhibit a low heterogeneity in multi-sourcing. Also, we show that the average number of paths available to each final buyer to connect to the manufacturer is below two for most networks, thus suggesting low resilience and the potential to enhance it.

Finally, the validation step confirms a good match between simulated and empirical data. The model can reproduce stylized facts of the empirical data, such as the broad in-degree and out-degree distributions and the peaked path length distributions. Beyond these stylized facts, the model successfully reproduces efficiency and resilience features of the empirical networks, namely the centralization level, the global efficiency, the number of available paths, and the number of undisrupted paths under random nodes' removal. Recovering these properties at the macro-level indicates that the proposed micro-rules are valuable explanatory mechanisms for the observed network structures. This allows us to bridge firm-level practices with the systemic properties of real-world distribution networks.

To what extent the proposed model can reproduce stylized facts of other distribution networks remains an open question worth investigating in future studies. The primary challenge to address in this direction is data availability. Currently, most of the available data regard production networks [46–48], with very few datasets collecting information

on the supply relations between firms in distribution networks [34, 49]. The available data are currently protected by stringent policy agreements, limiting their open usage within the scientific community. Thus, the first important step is establishing secure infrastructure for storing and processing firms' sensitive information [50]. This would enable the safe utilization of such data by researchers that can support firms and institutions in their decision-making processes.

As there is currently more data on production than distribution networks, an immediate question is whether the proposed model can be extended to production networks. Recent studies [47] on the nationwide production networks in Hungary and Ecuador have identified similar network features to those discussed for the opioid distribution networks, e.g., a peaked distribution of path lengths and heavy-tailed distributions of out-degree and in-degree. Moreover, the resilience and efficiency principles underlying our model are broad enough to apply to firm-level interactions within production networks. Thus, it is reasonable to speculate that firms may adopt similar strategies in seeking suppliers, and the proposed model can be generalized to production networks. Yet, a significant distinction between the two networks can already be pointed out: distributor firms typically do not rely on access to raw materials to commence operations. They predominantly handle finished products. In contrast, in production networks, the availability of raw materials can influence firms' choice in selecting their suppliers. This may lead to network features for distribution networks that differ from those discussed in this paper and represent an interesting venue for future research.

Many other research directions following up the current study can be considered. For instance, it would be interestingly to model the simultaneous and coupled growth of a supply network, considering both the increase in the number of relations and the volume of goods shipped. This expansion of scope could provide a more comprehensive understanding of the functioning of these networks. Lastly, the present study focused on single distribution networks around specific manufacturers. In real-world scenarios, manufacturers often share distributors and final buyers which results in interconnected distribution networks. Exploring the mutual dependency of these growth processes is another compelling area for future investigation.

In summary, we propose a network growth model to explain the emergence and growth of distribution networks. The model is parsimonious, and its parameters are interpretable. Despite its simplicity, we can calibrate and validate it against real-world data and find a surprising ability to reproduce stylized facts. Hence, with our data-driven modeling approach, we showcase how to capture the complexity of real-world distribution systems.

Declarations

References

1. Bode C, Wagner SM (2015) Structural drivers of upstream supply chain complexity and the frequency of supply chain disruptions. J Oper Manag 36:215–228
2. Akın Ateş M, Suurmond R, Luzzini D, Krause D (2022) Order from chaos: a meta-analysis of supply chain complexity and firm performance. Supply Chain Manag 58(1):3–30
3. Trump BD, Linkov I (2020) Risk and resilience in the time of the covid-19 crisis. Environ Syst Decis 40(2):171–173
4. Ivanov D (2021) Supply chain viability and the covid-19 pandemic: a conceptual and formal generalisation of four major adaptation strategies. Int J Prod Res 59(12):3535–3552
5. Allam Z, Bibri SE, Sharpe SA (2022) The rising impacts of the covid-19 pandemic and the Russia–Ukraine war: energy transition, climate justice, global inequality, and supply chain disruption. Resources 11(11):99
6. Farrell H, Newman AL (2022) Weak links in finance and supply chains are easily weaponized. Nature 605(7909):219–222
7. Whitehouse T (2021) National Strategy for a Resilient Public Health Supply Chain. Technical report
8. Schwartz F, Voß S (2007) Distribution network design with postponement. Wirtsch Proc 2007:78
9. Wang G, Gunasekaran A, Ngai EW (2018) Distribution network design with big data: model and analysis. Ann Oper Res 270(1):539–551
10. Altiparmak F, Gen M, Lin L, Karaoglan I (2009) A steady-state genetic algorithm for multi-product supply chain network design. Comput Ind Eng 56(2):521–537
11. Brintrup A, Ledwoch A (2018) Supply network science: emergence of a new perspective on a classical field. Chaos, Interdiscip J Nonlinear Sci 28(3):033120
12. Choi TY, Dooley KJ, Rungtusanatham M (2001) Supply networks and complex adaptive systems: control versus emergence. J Oper Manag 19(3):351–366
13. Pathak SD, Day JM, Nair A, Sawaya WJ, Kristal MM (2007) Complexity and adaptivity in supply networks: building supply network theory using a complex adaptive systems perspective. Decis Sci 38(4):547–580
14. Hearnshaw EJ, Wilson MM (2013) A complex network approach to supply chain network theory. Int J Oper Prod Manag 33(4):442–469
15. Ivanov D, Dolgui A (2020) Viability of intertwined supply networks: extending the supply chain resilience angles towards survivability. A position paper motivated by covid-19 outbreak. Int J Prod Res 58(10):2904–2915
16. Inoue H, Todo Y (2019) Firm-level propagation of shocks through supply-chain networks. Nat Sustain 2(9):841–847
17. Luo J, Baldwin CY, Whitney DE, Magee CL (2012) The architecture of transaction networks: a comparative analysis of hierarchy in two sectors. Ind Corp Change 21(6):1307–1335
18. Potter A, Wilhelm M (2020) Exploring supplier–supplier innovations within the toyota supply network: a supply network perspective. J Oper Manag 66(7–8):797–819
19. Spiegler VL, Naim MM, Wikner J (2012) A control engineering approach to the assessment of supply chain resilience. Int J Prod Res 50(21):6162–6187
20. Chakraborty T, Chauhan SS, Ouhimmou M (2020) Mitigating supply disruption with a backup supplier under uncertain demand: competition vs. cooperation. Int J Prod Res 58(12):3618–3649
21. Fahimnia B, Jabbarzadeh A, Sabouhi F (2017) Sustainability analysis under disruption risks
22. Schmitt AJ, Sun SA, Snyder LV, Shen Z-JM (2015) Centralization versus decentralization: risk pooling, risk diversification, and supply chain disruptions. Omega 52:201–212
23. Kajikawa Y, Takeda Y, Sakata I, Matsushima K (2010) Multiscale analysis of interfirm networks in regional clusters. Technovation 30(3):168–180
24. Hosseini S, Ivanov D, Dolgui A (2019) Review of quantitative methods for supply chain resilience analysis. Transp Res, Part E, Logist Transp Rev 125:285–307
25. Schweitzer F, Andres G, Casiraghi G, Gote C, Roller R, Scholtes I, Vaccario G, Zingg C (2022) Modeling social resilience: questions, answers, open problems. Adv Complex Syst 25(8):2250014.
26. Inderst R (2008) Single sourcing versus multiple sourcing. Rand J Econ 39(1):199–213
27. SLCG (2019) Opioid Data. https://www.slcg.com/opioid-data. Accessed 2022-09-01
28. Sheffi Y, Rice JB Jr (2005) A supply chain view of the resilient enterprise. MIT Sloan management review
29. Kim Y, Choi TY, Yan T, Dooley K (2011) Structural investigation of supply networks: a social network analysis approach. J Oper Manag 29(3):194–211
30. Jackson MO (2005) A survey of network formation models: stability and efficiency. Cambridge University Press, Cambridge, pp 11–57.
31. König MD, Battiston S, Napoletano M, Schweitzer F (2012) The efficiency and stability of r&d networks. Games Econ Behav 75(2):694–713.
32. Jaber MY, Zolfaghari S (2008) Quantitative models for centralised supply chain coordination. Supply Chain Theory Appl, 307–338
33. Treiblmaier H (2018) Optimal levels of (de) centralization for resilient supply chains. Int J Logist Manag 29(1):435–455

34. Amico A, Verginer L, Casiraghi G, Vaccario G, Schweitzer F (2024) Adapting to disruptions: managing supply chain resilience through product rerouting. Sci Adv 10(3):1194
35. Klemm K, Eguíluz VM, San Miguel M (2005) Scaling in the structure of directory trees in a computer cluster. Phys Rev Lett 95(12):128701
36. Geipel MM, Tessone CJ, Schweitzer F (2009) A complementary view on the growth of directory trees. Eur Phys J B 71(4):641–648
37. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
38. Capocci A, Servedio VD, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G (2006) Preferential attachment in the growth of social networks: the Internet encyclopedia Wikipedia. Phys Rev E 74(3):036116
39. Krapivsky PL, Rodgers GJ, Redner S (2001) Degree distributions of growing networks. Phys Rev Lett 86(23):5401
40. Newman ME, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64(2):026118
41. Tomasello MV, Perra N, Tessone CJ, Karsai M, Schweitzer F (2014) The role of endogenous and exogenous mechanisms in the formation of r&d networks. Sci Rep 4(1):1–12
42. Butts CT (2006) Exact bounds for degree centralization. Soc Netw 28(4):283–296
43. Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Phys Rev Lett 87(19):198701
44. Zhao K, Kumar A, Harrison TP, Yen J (2011) Analyzing the resilience of complex supply network topologies against random and targeted disruptions. IEEE Syst J 5(1):28–39
45. Kim Y, Chen Y-S, Linderman K (2015) Supply network disruption and resilience: a network structural perspective. J Oper Manag 33:43–59
46. Diem C, Borsos A, Reisch T, Kertész J, Thurner S (2022) Quantifying firm-level economic systemic risk from nation-wide supply networks. Sci Rep 12(1):7719
47. Bacilieri A, Borsos A, Astudillo-Estevez P, Lafond F (2022) Firm-level production networks: what do we (really) know. Technical report, mimeo, University of Oxford
48. Bernard AB, Moxnes A, Saito YU (2019) Production networks, geography, and firm performance. J Polit Econ 127(2):639–688
49. Schueller W, Diem C, Hinterplattner M, Stangl J, Conrady B, Gerschberger M, Thurner S (2022) Propagation of disruptions in supply networks of essential goods: a population-centered perspective of systemic risk. arXiv preprint arXiv:2201.13325
50. Pichler A, Diem C, Brintrup A, Lafond F, Magerman G, Buiten G, Choi TY, Carvalho VM, Farmer JD, Thurner S (2023) Building an alliance to map global supply networks. Science 382(6668):270–272

Publisher's Note