

A Methodological supplementary material

A.1 Applied node- and network-level indicators

A.1.1 Applied node-level indicators

Degree centrality is defined as the number of links incident upon a node (i.e., the number of ties that a node has). In the case of a directed network (where ties have direction), there are usually two separate measures of degree centrality, namely, in-degree (DI) and out-degree (DO) centrality. The in-degree centrality is a count of the number of ties directed to the node, and the out-degree centrality is the number of ties that the node directs to others. The degree centrality of a country i is defined as:

$$DA_i = \deg(i), \quad (3)$$

$$DI_i = \text{in-deg}(i), \quad (4)$$

$$DO_i = \text{out-deg}(i). \quad (5)$$

Betweenness centrality (BC) quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Vertices that have a high probability of occurring on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness.

$$BC_j = \sum_{i \neq j \neq k} \frac{n_{ik}^{\text{SP}}(j)}{n_{ik}^{\text{SP}}} \quad (6)$$

For a given graph $G := (V, E)$ with V vertices, let $\mathbf{A} = (a_{v,t})$ be the adjacency matrix, i.e., $a_{v,t} = 1$ if vertex v is linked to vertex t , and $a_{v,t} = 0$ otherwise. The relative centrality score, EC_v , of vertex v can be defined as:

$$EC_v = \frac{1}{\lambda} \sum_{t \in M(v)} EC_t = \frac{1}{\lambda} \sum_{t \in V} a_{v,t} EC_t \quad (7)$$

where $M(v)$ is the set of neighbors of v and where λ is a constant. For the given weighted graph $G := (V, E, W)$, W is the weight matrix (Segarra and Ribeiro, 2015).

$$EC_v = \frac{1}{\lambda} \sum_{t \in V} w_{v,t} EC_t \quad (8)$$

where n_{ik}^{SP} is the total number of shortest paths (SPs) from node i to node k and where $n_{ik}^{\text{SP}}(j)$ is the number of paths that pass through node j .

A.1.2 Applied network-level indicators

The (degree) *assortativity* refers to the tendency of nodes in a network to be connected to other nodes that are similar to them in some way.

In general, the level of assortativity in a collaboration network can provide insights into the underlying structures and patterns of collaboration between organizations. High levels of assortativity may suggest that organizations collaborate primarily with other organizations that are similar to them.

Mathematically, the assortativity coefficient (Newman, 2002) can be written as:

$$AC = \frac{1}{\sigma_q^2} \left[\left(\sum_{jk} jk e_{j,k} \right) - \mu_q^2 \right] \quad (9)$$

Intuitively, assortativity measures the extent to which nodes with similar degrees tend to be connected to each other (when assortativity is positive) or tend to be connected to nodes with different degrees (when assortativity is negative). A value of assortativity close to 0 indicates no assortativity in the network, whereas values close to +1 or -1 indicate strong assortativity in the network. In addition to degree correlation, other values that can be used to create an assortativity value include the following:

1. **Attribute Assortativity:** This measures the tendency of nodes to connect on the basis of categorical attributes, such as the same department or research area.
2. **Numeric assortativity:** This measures the correlation between the numerical attributes of nodes, such as the number of publications or years of experience.
3. **Strength Assortativity:** In weighted networks, this measures the correlation between the strengths (sum of weights of edges) of connected nodes.

Since structural concentration was the focus of the study, degree assortativity is applied as an assortativity value.

The γ value is a parameter that measures the degree of scale-free property in a collaboration network. A γ value close to 1 indicates that the distribution of node degrees is approximately uniform, whereas a gamma value much greater than 1 indicates a highly skewed distribution with a few hubs dominating the network.

We attempt to fit the power law distribution to each collaboration network. If the p value is greater than 0.05, the γ values are stored in each network. where e_{jk} is the joint probability distribution of the excess degrees of the two nodes at either end of a randomly chosen link. Here, μ_q and σ_q are the expected value or mean and standard deviation of the excess degree distribution q_k , respectively.

Centralization (Z) is a measure of the degree to which the flow of resources or information within a network is concentrated among a small number of nodes or edges rather than being distributed evenly across the network. This concept can be applied to various types of networks, including collaboration networks, trade networks, social networks, transportation networks, and information networks.

Mathematically, centralizations can be measured via various centrality metrics. In this way, we can obtain degree centralization (DZ), betweenness centralization (BZ), or eigenvector centralization (EZ). These metrics quantify the relative importance or influence of each node or edge in the network on the basis of factors such as its degree of connectivity, its position in the shortest paths between nodes, or its relationship to other high-scoring nodes.

The mathematical form of centralization (Z) is as follows:

$$Z = \sum_i \frac{\max_j C_j - C_i}{(n-1) \cdot (n-2)} \quad (10)$$

where Z is the centralization value of the network, C_i is the centrality value of node i , $\max_j C_j$ is the maximal centrality value, and n is the number of nodes.

Modularity-based community detection algorithms minimize Eq. (11).

$$M = \frac{1}{2L} \sum_{i,j} (w_{ij} - \gamma \hat{w}_{ij}) \delta(C_i, C_j), \quad (11)$$

where M is the *modularity value*, w_{ij} is the edge weight between node i and node j , $\hat{w}_{ij}$ is the expected weight between node i and node j (the so-called null model), L is the total number of weights in the network, γ is constant (the default value is 1), and the Kronecker δ is 1 if the community of variable i and the community of variable j are the same and 0 otherwise.

We used Leiden's (Traag et al, 2019) algorithm, which is an improvement over the Louvain algorithm. Louvain's algorithm is a greedy optimization method that maximizes the modularity of the network. The Leiden algorithm improves upon the Louvain algorithm by refining the community (Traag et al, 2019), by using local moves to optimize the modularity, and by using a smart initialization that prevents empty communities.

A.1.3 Countrywise aggregation of network-level indicators

For each institution, the country of the institution's location is known, and as a result, its connections can be aggregated from the institutional collaboration network to determine the connections between countries (see Figure 9).

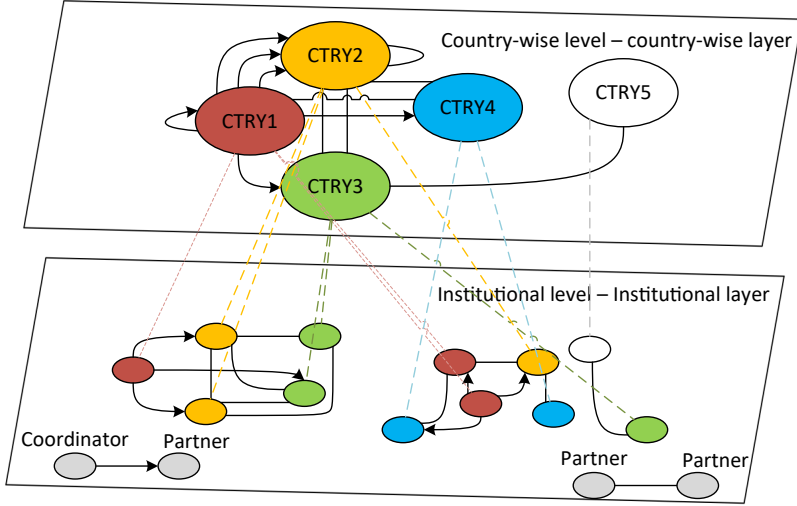


Fig. 9 Countrywise aggregation of collaboration structures

Formally, $G = (V, E)$ is the organizational level of the collaboration network, where $V = \{v_1, v_2, \dots, v_n\}$ is the set of vertices (nodes) (i.e., the set of organizations) and where n is the number of nodes. $E \subset V \times V$ is the set of edges (i.e., projects implemented in a collaboration). The collaboration network is directed because, in most cases, the coordinator (tail) and the partner (head of the edge) are known. Where it is not, we mark it as undirected edges, but in fact, we indent the edges in both directions. The vertices can be aggregated in many ways, but we only aggregate them by country. The aggregated network is denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{\nu_1, \nu_2, \dots, \nu_k\}$ and where k is the number of countries. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges in a countrywise collaboration network. $(\nu_\kappa, \nu_\iota \in \mathcal{E} \Leftrightarrow (v_i, v_j) \in E \wedge v_i \in \nu_\kappa, v_j \in \nu_\iota$.

A.2 Calculation of institutional concentration

The earned contribution (EC) of organization i where its location in country j is denoted as EC_{ij} . The total contribution of country i is defined as $TC_i = \sum_j EC_{ij}$. The countrywise *institutional concentration* $IC_i^{x\%}$ is the percentage of organizations in country i that contribute to $x\%$ of the total contribution.

For the calculation of concentration, let $EC_{i(1)}, EC_{i(2)}, \dots, EC_{i(n_i)}$ be the sorted contributions for country i such that $EC_{i(1)} \geq EC_{i(2)} \geq \dots \geq EC_{i(n_i)}$, where n_i is the total number of organizations in country i . We can find the smallest integer o such that $\sum_{m=1}^o EC_{i(o)} \geq x \times TC_i$. The number of organizations required is o . The institutional concentration $IC_i = \left(\frac{o}{n_i}\right) \times x$.

A.3 Calculation of implementation structures

The calculation of the implementation structures and the analysis of the effects are performed through five main steps: (1) development of distance measures, (2) calculation of membership values, (3) data collection for simulation, (4) DSM generation and simulation and (5) analysis of the effects. Figure 10 shows the framework of the calculation.

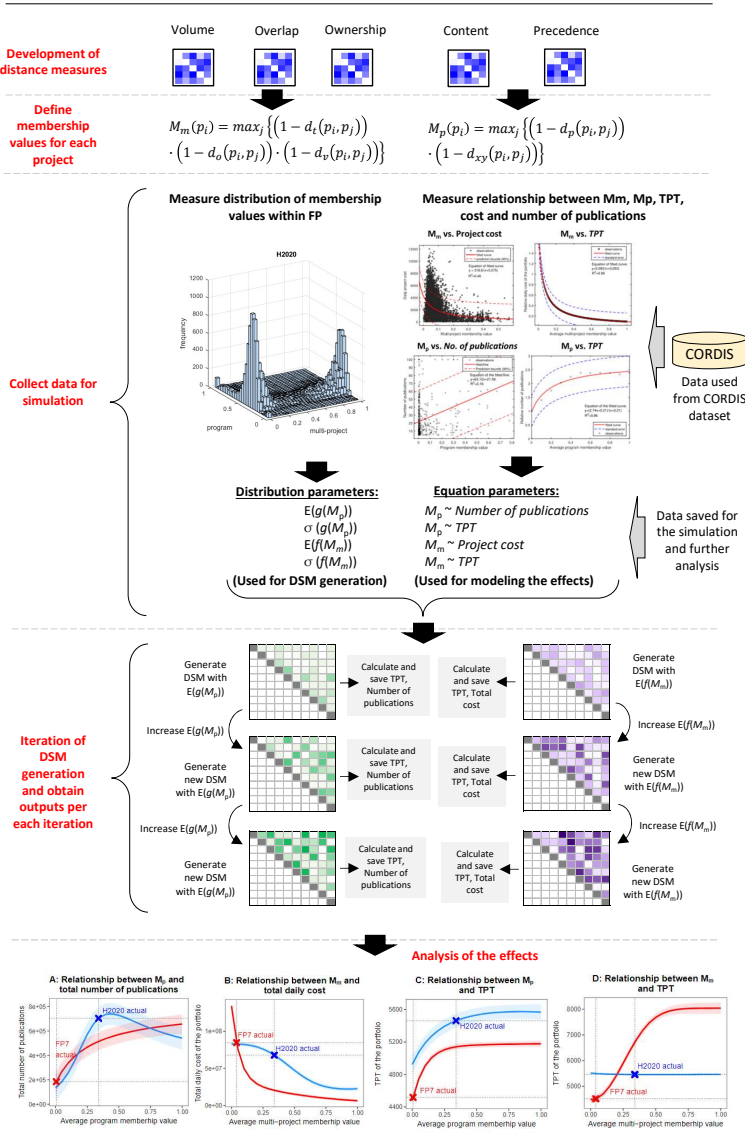


Fig. 10 Framework of the implementation structure calculation and modeling

The steps are detailed below.

(1) Development of distance measures

In the first step, the distance measures among projects must be specified. As proposed by Kosztyán et al (2022), five distance measures are used, such as overlap, volume, and ownership, to describe the multiproject characteristics, and precedence and content are used to describe the program-related characteristics. These distances are calculated pairwise. The distance of overlapping duration (d_t), distance of volume (d_v) and distance of ownership (d_o) are used to characterize the distance between projects p_i and p_j with respect to multiproject characteristics:

$$d_t(p_i, p_j) = 1 - \frac{t(p_i) \cap t(p_j)}{t(p_i) \cup t(p_j)}, d_t(p_i, p_j) \in [0, 1]. \quad (12)$$

where $t(p_i)$ is the time interval for project p_i .

$$d_v(p_i, p_j) = 1 - \min(c(p_i), c(p_j)) / \max(c(p_i), c(p_j)) \quad (13)$$

where $c(p_i)$ denotes the cost associated with project p_i .

$$d_o(p_i, p_j) = 1 - \frac{o(p_i) \cap o(p_j)}{o(p_i) \cup o(p_j)}, d_o(p_i, p_j) \in [0, 1]. \quad (14)$$

where $o(p_i)$ denotes the set of owners regarding project p_i .

Furthermore, the distance of precedence (d_p) and distance of content (d_{xy}) are used to measure the distance between projects while considering program-related characteristics.

$$d_p(p_i, p_j) = \begin{cases} 1 & d_t(p_i, p_j) < 0.5 \\ 1 - \frac{\min\{|f(p_i) - s(p_j)|, |f(p_j) - s(p_i)|\}}{\max\{|f(p_i) - s(p_j)|, |f(p_j) - s(p_i)|\}} & d_t(p_i, p_j) \geq 0.5 \end{cases} \quad (15)$$

where $s(p_i)$ and $f(p_i)$ represent the planned start and finish dates, respectively, of project p_i .

$$d_y(p_i, p_j) = \begin{cases} 0, & \text{if } y(p_i) = y(p_j) \\ 1, & \text{if } y(p_i) \neq y(p_j) \end{cases} \quad (16)$$

$$d_x(p_i, p_j) = 1 - \cos(descr(p_i), descr(p_j)) \quad (17)$$

$$d_{xy}(p_i, p_j) = d_x(p_i, p_j) \cdot d_y(p_i, p_j) \quad (18)$$

where $descr(p_i)$ is the description of project p_i , $y(p_i)$ is the type of sub-program of project p_i and where $\cos(descr(p_i), descr(p_j))$ denotes the cosine distance between the descriptions of projects p_i and p_j .

(2) Calculation of membership values

In this step, for each project, a program membership value ($\mathcal{M}_p(p_i)$) and a multiproject membership value ($\mathcal{M}_m(p_i)$) are calculated via the distance values calculated in step 1. Since the distance measures are calculated pairwise, each project has n distance values in each measure (where n is the number of projects across the FP). To obtain one membership value for each project, the maximum value is chosen from the vector of multiplied distance measures considering the distance of overlapping duration, ownership and duration for the multiproject membership value:

$$\mathcal{M}_m(p_i) = \max_j \{(1 - d_t(p_i, p_j)) \cdot (1 - d_o(p_i, p_j)) \cdot (1 - d_v(p_i, p_j))\} \quad (19)$$

Following the same logic, the distances of precedence and content are considered for calculating the program membership value:

$$\mathcal{M}_p(p_i) = \max_j \{(1 - d_p(p_i, p_j)) \cdot (1 - d_{xy}(p_i, p_j))\} \quad (20)$$

(3) Data collection for simulation

To model the effects of changes in FPs, simulations need to be performed after the input data are collected. First, the distributions of $\mathcal{M}_p$ and $\mathcal{M}_m$ are determined, and the distribution parameters are specified for both. This is important because this information is used to increase the expected membership value (following the same distribution as detected) of the entire portfolio stepwise and generate similarity matrices. On the other hand, to account for the effect of the expected program and increase in the multiproject membership value, the relationships between the membership value and time, cost and scientific output (reflected by the number of publications as the result of each project) are measured. These relationships are measured at the project level, the best fitting curves are determined, and the equation parameters are estimated and saved. This information is used during the simulation since the output and cost are estimated at the project level and the predicted values are aggregated at the FP level to calculate the total cost and total output of the FP.

(4) DSM generation and simulation

The implementation structures are calculated through simulated design structure matrices (DSMs). In the cells of the matrices, the calculated membership values are stored, such as $\mathcal{M}_m(p_{ij})$, $\mathcal{M}_p(p_{ij})$, denoting the probability of dependencies between projects i and j . The high value of $\mathcal{M}_p(p_{ij})$ represents sequential execution between projects p_i and p_j , increasing the TPT of the portfolio, whereas the high value of $\mathcal{M}_m(m_{ij})$ denotes parallel execution between projects p_i and p_j . Using the estimated distribution parameters for $\mathcal{M}_m$, $\mathcal{M}_p$, DSMs were generated separately for programs and multiprojects.

The expected value of the membership values is increased in each iteration, and the TPT, total cost, and total number of expected publications during the execution of the entire portfolio are calculated. When the program membership value increases (left, green section of Figure 10), the trade-off between TPT and the number of publications is saved. Similarly, the effect of increasing multiprojects across the portfolio is measured by saving the TPT and total cost at each iteration (represented by the right purple section of Figure 10).

(5) Analysis of the effects

The final step is the analysis of the effects. Four different combinations are analyzed: (1) the relationship between the average program membership value across the FP and (2) the average program membership value and TPT of the FP, (3) the average multiproject membership value and total cost of the FP and (4) the average multiproject membership value and TPT of the portfolio. The input data for the analysis are obtained as the outputs of steps 3 and 4.