communications physics

ARTICLE

https://doi.org/10.1038/s42005-022-00856-9 OPEN

Meta-validation of bipartite network projections

Giulio Cimini [1,2]✉, Alessandro Carra[3], Luca Didomenicantonio[3] & Andrea Zaccaria [2,4]

Monopartite projections of bipartite networks are useful tools for modeling indirect interactions in complex systems. The standard approach to identify significant links is statistical validation using a suitable null network model, such as the popular configuration model (CM) that constrains node degrees and randomizes everything else. However different CM formulations exist, depending on how the constraints are imposed and for which sets of nodes. Here we systematically investigate the application of these formulations in validating the same network, showing that they lead to different results even when the same significance threshold is used. Instead a much better agreement is obtained for the same density of validated links. We thus propose a meta-validation approach that allows to identify model-specific significance thresholds for which the signal is strongest, and at the same time to obtain results independent of the way in which the null hypothesis is formulated. We illustrate this procedure using data on scientific production of world countries.

[1] Physics Department and INFN, University of Rome Tor Vergata, 00133 Rome, Italy. [2] Enrico Fermi Research Center, 00184 Rome, Italy. [3] Physics Department, Sapienza University of Rome, 00185 Rome, Italy. [4] Institute for Complex Systems (CNR) UoS Sapienza, 00185 Rome, Italy.
✉email: giulio.cimini@roma2.infn.it

Networks are simplified yet effective models for a large class of natural, socio-economic and technological systems described by complex interaction patterns. Independently of the nature of the underlying interactions, the network representation allows capturing the emergent features of these systems as well as their dynamical patterns[1–5]. As such, network science has gained increasing popularity in the last twenty years[6–8].

A network is labeled as *bipartite* when its elements (the nodes) can be split in two disjoint sets, such that links can only exist between nodes of different sets[9]. Bipartite networks are the natural representation for several systems, such as: social affiliation and collaboration networks, where individuals connect to the groups they are member of[10,11]; financial and commercial ownership networks, where entities are linked to the goods they own or consume[12,13]; trade networks, where economies connect to the products they export[14,15]; ecological networks, where species connect to the habitat they live in[16,17]; biological and medical networks connecting, *e.g.*, patients and diseases[18,19]. Mathematically speaking, a bipartite network is defined as a graph with two sets L and Γ of nodes, and a $|\mathrm{L}| \times |\Gamma|$ matrix of connections $\mathbf{M}$ called *bi-adjacency matrix*. The generic element of this matrix is

$$M_{i\alpha} = \begin{cases} 1 \text{ if nodes } i \in \mathrm{L} \text{ and } \alpha \in \Gamma \text{ are connected,} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

The number of connections or *degree* of a node $i \in \mathrm{L}$ is then defined as $k_i = \sum_{\alpha\in\Gamma} M_{i\alpha}$, while the degree of a node $\alpha \in \Gamma$ is $\kappa_\alpha = \sum_{i\in\mathrm{L}} M_{i\alpha}$. The total number of links in the network is $E = \sum_{i\in\mathrm{L}} k_i = \sum_{\alpha\in\Gamma} \kappa_\alpha$.

The indirect relation between two nodes belonging to the same set of a bipartite network can be measured through their *co-occurrences* (or common neighbors), namely how many nodes of the other set they are both connected to. For instance the co-occurrences of nodes i and j of set L are given by

$$C_{ij} = \sum_{\alpha\in\Gamma} M_{i\alpha} M_{j\alpha}. \quad (2)$$

The $\mathrm{L} \times \mathrm{L}$ square matrix $\mathbf{C}$ represents a monopartite network obtained as the projection of the original bipartite network onto the set L (L-projection)[12]. Analogously, one can project the bipartite network onto the set Γ to obtain the co-occurrences between nodes of that set (Γ-projection).

The main problem in studying bipartite network projections is that they are often very dense and thus difficult to handle with the tools of network theory. This happens because any two nodes are connected in the projected network as soon as they have a single co-occurrence. Moreover, co-occurrences can be influenced by single node variables, thus understanding whether they indicate an effective interdependence between nodes may be difficult. For example, nodes that have high degree in the bipartite network naturally tend to have more co-occurrences than low-degree nodes (more generally, the degree sequence of the network projection is highly dependent on the degree sequence of the two sets from the original bipartite structure[20]). It is thus useful to extract representative links of the projected network; this can be achieved using several filtering techniques, from unconditional thresholding to Minimal Spanning Trees[21] and Planar Maximally Filtered Graphs[22]. Yet in order to identify the most informative co-occurrences, the statistically-grounded approach consists in performing link validation using a null network model.

Statistical validation of network patterns is a common approach in the literature (the classical applications being motifs expression analysis[23,24], network backbone extraction[25] and community detection[26–28]). The goal is to identify the empirical patterns that deviate from a benchmark null model, in order to ensure that those patterns are indeed a salient feature of the network and not a mere consequence of some of its other properties (given the potentially strong interdependence between structural network quantities[29–33]). Following the prescription of information theory[34], the null model is thus obtained by constraining some network properties and randomizing everything else. In this way, the formulated null hypothesis is that these constraints are the only explanatory variables for the network at hand; when the null hypothesis is rejected, we can state that the observed network patterns are not a mere consequence of the imposed constraints.

Going back to our context of bipartite network projections, the statistical significance of each observed co-occurrence value $C_{ij} > 0$ can be quantified through its p-value:

$$p[C_{ij}] = 1 - \sum_{x=0}^{C_{ij}-1} \pi(x|i,j), \quad (3)$$

where $\pi(\cdot\,|i,j)$ is the probability distribution of the expected co-occurrences between i and j under the null model. The right-hand side of Eq. (3) is the probability that i and j have no less than C_{ij} co-occurrences in the null model. This quantity can be used to build a *validated* (or filtered) projection of the original bipartite network, containing only the most significant links according to the null model. For each C_{ij}, if the p-value of Eq. (3) is smaller than a significance threshold (or confidence level) p^*, the link i,j is placed on the monopartite validated network; otherwise, it is discarded. In other words, the comparison is deemed statistically significant if the observed co-occurrences are an unlikely realization of the null hypothesis according to the significance level p^* (in particular, we are interested in detecting the co-occurrences that are significantly larger than their null model expectation; significantly smaller values can be obtained in a similar fashion — see Supplementary Note 1). In this way the original amount of links is drastically reduced, and the result is a much sparser validated network with a clearer meaning.

Naturally, statistical validation has some intrinsic degrees of freedom: the choice of the null model, its specific formulation, and the value of the significance level p^*. In particular, the choice of the model is a step that should be handled with care, as a bad choice may lead to wrong conclusions about the structural and functional features of the network[35,36]. For instance, using a (bipartite) Erdös-Rényi model[37], *i.e.*, a random network preserving only the density of the original bipartite graph, leads to an identical distribution $\pi(\cdot\,|i,j)$ for each node pair i,j and thus to an unconditional global threshold to select the most significant C_{ij} values[38,39]. However this choice does not solve the bias problem for high degree nodes, which is very important in networks given that degree distributions are typically very broad[40]. A natural way to take this aspect into account is given by the popular *configuration model* (CM)[41–43], which generates random networks with a given degree sequence.

In the context of bipartite networks, the first model formulation of this family was obtained by constraining the degrees of nodes in one set (say L)[44]. In this case, the co-occurrences probability can be computed exactly as a hypergeometric distribution[45,46]. Yet this model solves the degree bias only partially, since it assumes nodes of the other set (say Γ) to be equivalent and interchangeable. The alternative approach is to model random bipartite networks preserving the degrees of both node sets L and Γ, and then use these networks to obtain the null model for the projected network. *Degree sequence models* follow this approach, however they either require multiple observations of the empirical network[47] or are based on computational link swap methods[48] that are typically impractical and biased[49]. An exception is represented by the recently proposed *Curveball* algorithm[50–52], a link swap method that is extremely efficient in generating network configurations and is ergodic (*i.e.*, it can

sample uniformly over the set of all possible network configurations). The alternative route to Monte Carlo sampling is represented by maximum-entropy models. The *Bipartite Configuration Model* (BiCM)[53] allows generating an ensemble of bipartite networks where node degrees of both sets L and Γ are preserved as ensemble expectations. The null model for the network projections is then obtained by projecting BiCM-generated networks[54,55]. This latter approach allows computing the co-occurrences distributions both numerically and analytically, and simplifies as a *Bipartite Partial Configuration Model* (BiPCM) when degree constraints are imposed only on one set of nodes. At last we note that, in principle, the projection of a bipartite network can be statistically validated also using a null model for monopartite weighted networks[25,33,56]. That is, instead of defining the null model on the bipartite network and then deriving its formulation for the network projection, the null model can be directly defined on the monopartite projection. However this approach discards the information contained in the original bipartite network, and as such typically leads to completely different and not significant outcomes (see Supplementary Note 2).

To sum up, the four main CM-based null models for bipartite network projections proposed in the literature (Hypergeometric, Curveball, BiPCM and BICM) can differ under two aspects. The first aspect concerns which constraints are imposed, whether the degrees of one set or both sets of the bipartite network. We can thus speak of "partial" models (Hypergeometric and BiPCM) and "full" models (Curveball and BiCM). The second aspect concerns how these constraints are imposed, either exactly (hard constraints) or as ensemble expectations (soft constraints). Using the analogy with statistical physics[34], we can refer to these approaches respectively as "microcanonical" models (Hypergeometric and Curveball) and "canonical" models (BiPCM and BiCM). Table 1 summarizes this classification (see the Methods section for the formal definition of the four null models). A fundamental point that has not been addressed so far is whether these formulations lead to different validated networks, and thus how to interpret and compare results of the various studies in the literature.

Here we provide, for the first time to our knowledge, a systematic comparison of validation results obtained with the various CM formulations for bipartite network projections. We find that albeit based on very similar null hypothesis, the different formulations lead to very different filtered networks even for the same value of validation threshold p^*. However we show that a reconciliation of results is possible within a region of model-specific thresholds p^* such that the densities of links validated by the null models overlap. In particular we show that a common community structure may emerge in this region. This criterion provides a quantitative approach to build a meta-validated network projection that is independent on the specific implementation of the null model.

Table 1 Classification of configuration models (CM) for bipartite network projections by number and type of constraints.

	Partial (1 set)	Full (2 sets)
Hard (microcanonical)	Hypergeometric[44–46]	Curveball[50,51]
Soft (canonical)	BiPCM[55]	BiCM[54,55]

Partial models only constraints the degrees of nodes belonging to the projection set, while Full models constraints the degrees of nodes in both sets. Hard or microcanonical models impose exact constraints, while Soft or canonical models impose them as ensemble averages. BiPCM stands for Bipartite Partial CM and BICM for Bipartite CM.

Results and discussion

We perform the comparison of validation outcomes in the context of co-occurrences for country production networks, following the recent stream of works on economic fitness and complexity[44,57–60]. In particular, our empirical bipartite system is defined by two set of nodes, scientific fields (set L) and world countries (set Γ), and by links that connect countries with the scientific fields they have a comparative advantage on. The L-projection of this bipartite network is a monopartite network of scientific fields, whose generic link C_{ij} is the co-occurrence of fields i and j worldwide. Figure 1 summarizes how the validation procedure is applied to this network. For a description of raw data and pre-processing, see the Methods section.

We start by recalling the key assumption underlying economic complexity studies on co-occurrences: if two scientific fields feature significant co-occurrences (in terms of an appropriate null model) then we can assume that there is an overlap between the capabilities required to achieve proficient level (*i.e.*, competitive advantage) in both fields[60]. The need for statistical validation arises in this context since both countries and scientific fields are heterogeneous (if nothing, by their size): two scientific fields may happen to co-occur in many countries just because they are popular worldwide. Therefore, a reasonable baseline choice of null model is the (bipartite) CM, for which degrees (*i.e.*, the ubiquity of fields and possibly the diversification of countries) sum up all the information. The corresponding null hypothesis is thus that fields are independent and there is no capability structure behind the network: co-occurrences between scientific fields happen at random, some more likely than others just because of their ubiquities or countries' diversification. Therefore, any specific observed link i, j for which we can reject such null hypothesis is interpreted as the signal of some real interdependence between the specific capabilities required to make proficient scientific research in fields i and j.

Different models, different validated networks. In order to better understand how the validation procedure works, we begin by comparing in Fig. 2 the empirical value of the co-occurrences C_{ij} with the respective null model distribution $\pi(\cdot|i,j)$ for some representative pairs i,j of scientific fields. We recall that C_{ij} is validated if it satisfies the condition $p[C_{ij}] \leq p^*$, that is, if the area under the distribution starting from the empirical value is smaller than the threshold p^*. The three example we report are the co-occurrences of: (a) *Mathematical Physics - Geometry and Topology*, which are likely validated, in accordance with our expectations that the two fields are related by requiring common skills and capabilities; (b) *Mathematical Physics - Aquatic Science*, which are likely not validated, again as we can expect that the two fields are unrelated; and (c) *Finance - Applied Psychology*, whose relatedness is plausible but the outcome of the validation procedure is uncertain, as it strongly depends not only on the choice of the threshold p^* but also on the choice of the null model. This is a first evidence that different models may lead to different validated networks.

These plots also highlight some important symmetries between the various null model distributions (see also Fig. 3). On one hand, the peaks coincide for the two partial models (Hypergeometric and BiPCM), since they have the same average value $\langle C_{ij} \rangle = k_i k_j / |\Gamma|$ (see Eqs. (6) and (14)), but the same also happens for the two full models (Curveball and BiCM). This means that the average value of the co-occurrences in the null model depends on the set(s) on which degree constraints are imposed. Besides, such average is higher for full models as they capture the heterogeneity of both sets. This can be readily seen by taking the sparse limit of the BiCM, for which Eq. (9) becomes

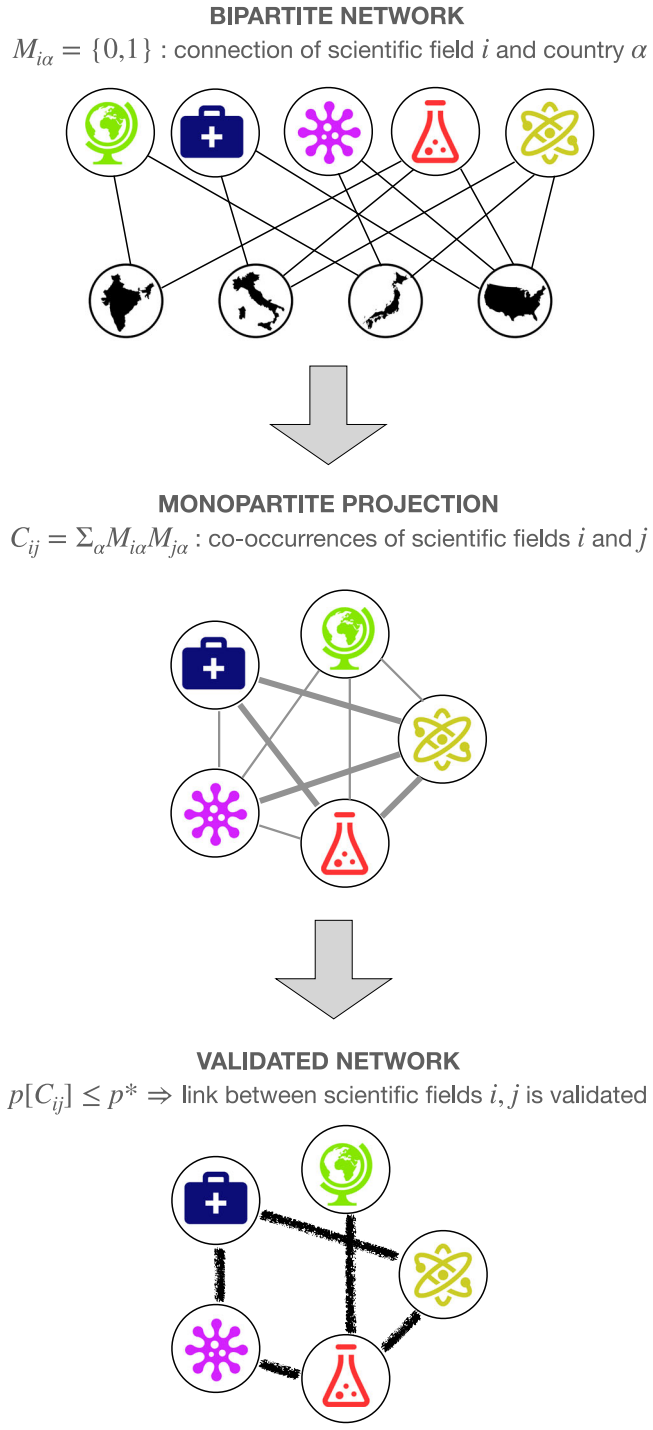

Fig. 1 Schematic illustration of the validation procedure for the bipartite network of scientific fields and world countries. We start from the bipartite network **M** of scientific fields (in this example: *Earth Sciences*, *Medicine*, *Biology*, *Chemistry* and *Physics*[111]) and word countries (here: India, Italy, Japan, United States[112]). In this network, links connect countries with the scientific fields they have a comparative advantage on. From this bipartite structure we create a monopartite projected network **C** of scientific fields, whose weighted links represent the co-occurrences of field pairs in the various countries. Finally we assess the statistical significance of each observed co-occurrence against its null model expectation: we place a link on the validated network only when the *p*-value is smaller than the significance threshold p^*. Note that this procedure is general and applies to any bipartite network.

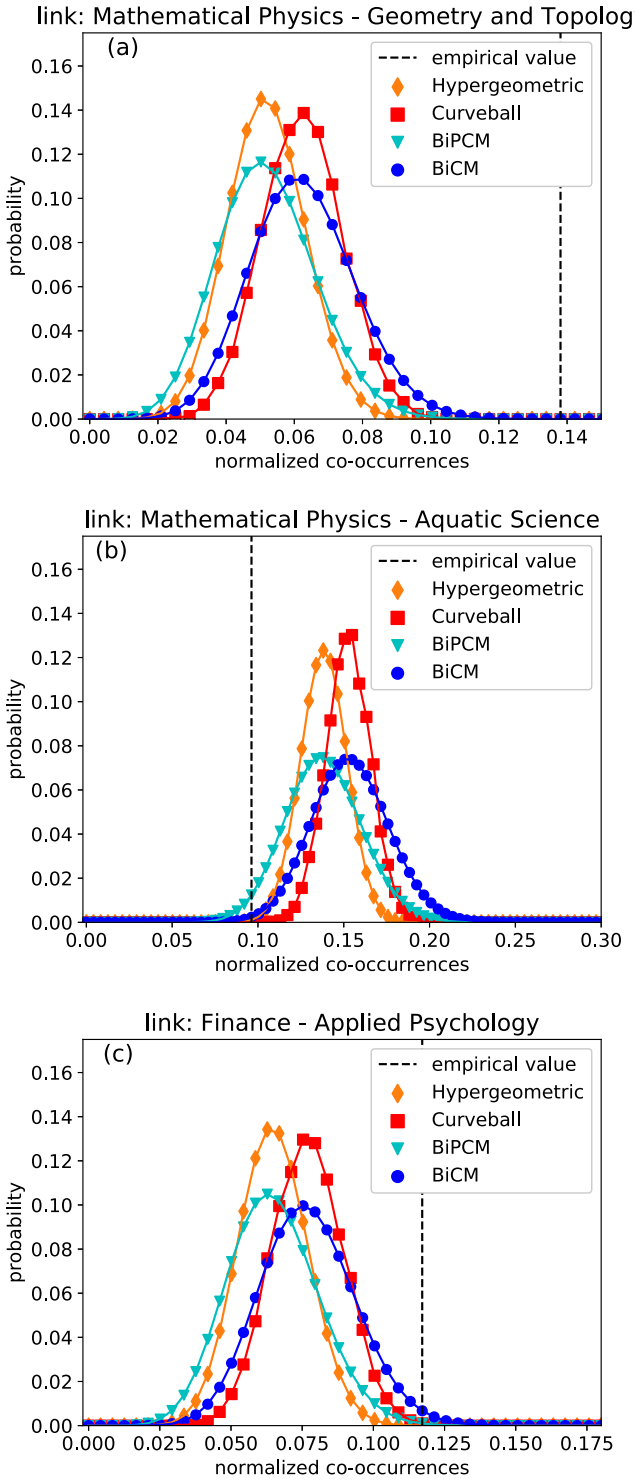

Fig. 2 Comparison of empirical co-occurrences and their null model distributions for representative scientific field pairs. a *Mathematical Physics - Geometry and Topology*, **b** *Mathematical Physics - Aquatic Science*, **c** *Finance - Applied Psychology*. For each pair (i, j) of scientific fields we report the empirical value of the (normalized) co-occurrences $C_{ij}/|\Gamma|$ and the respective null model distributions $\pi(\cdot\,|i, j)$. The *p*-value is given by the area under the distribution starting from the empirical value, hence the outcome of the validation procedure strongly depends both on the significance threshold p^* and the choice of the null model.

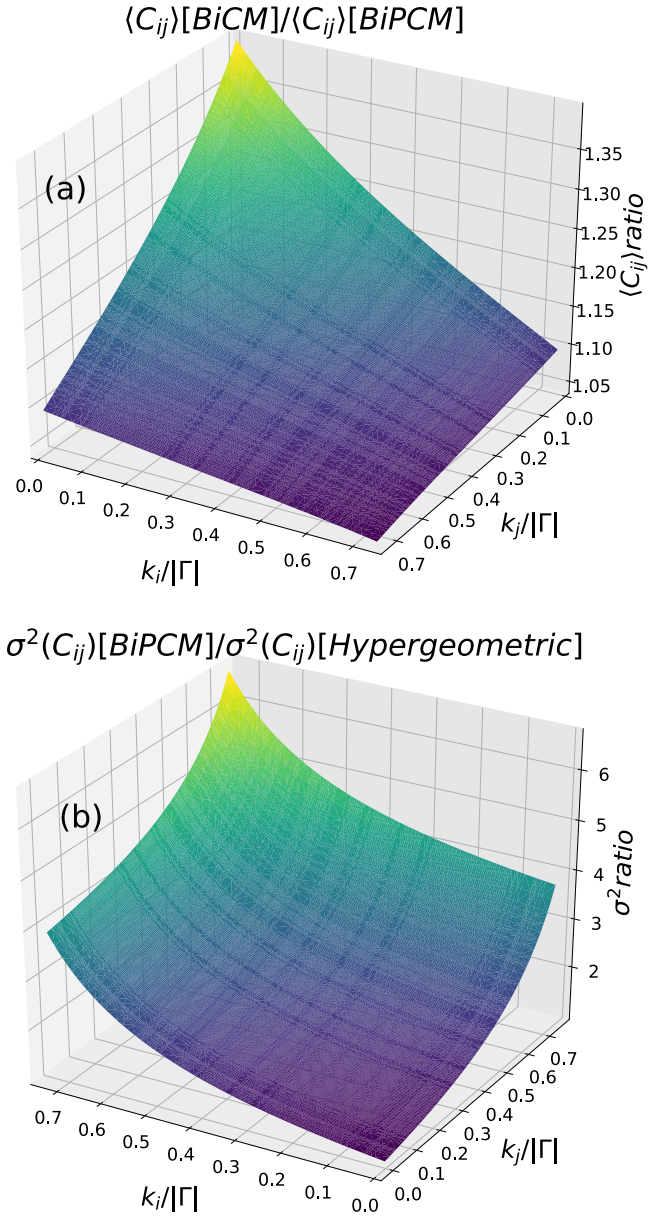

Fig. 3 Comparison of null model features. a Ratio of mean co-occurrences $\langle C_{ij}\rangle$ for BiCM and BiPCM, and **b** ratio of variances $\sigma^2(C_{ij})$ for BiPCM and Hypergeometric, as a function of the normalized degrees $k_i/|\Gamma|$ and $k_j/|\Gamma|$ of the corresponding nodes. Both quantities are strongly dependent on the specific model formulations, especially for high values of the degrees.

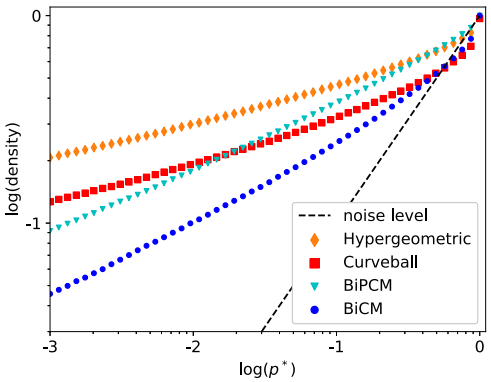

Fig. 4 Density ρ of links validated by the various null models as a function of the significance threshold $p^\star$. The dashed bisector denotes the noise level, namely the probability to statistically validate a link generated by the null model. The four null models leads to different results even for the same p^* value.

$p_{i\alpha} \simeq e^{-\theta_i + t_\alpha} = k_i\kappa_\alpha/E$. Inserting this expression into Eq. (11) we get $\langle C_{ij}\rangle \simeq k_i k_j \sum_\alpha \kappa_\alpha^2/E^2$, which is greater than $\langle C_{ij}\rangle = k_i k_j/|\Gamma|$ of the BiPCM (and equal only when set Γ has no heterogeneity, that is, $\kappa_\alpha = E/|\Gamma|\ \forall\,\alpha$). On the other hand, the width of the distribution looks similar for microcanonical models (Hypergeometric and Curveball) and for canonical models (BiPCM and BiCM), implying that the standard deviation of the co-occurrences depends on the types of constraints. As expected, the choice of hard constraints leads to a narrower distribution while the choice of soft constraints leads to a broader distribution. This is easily seen by taking the ratio of variances for the BiPCM (Binomial) and Hypergeometric model, which after some simple algebra can be written as $(|\Gamma|^2 - k_i k_j)/[(|\Gamma| - k_i)(|\Gamma| - k_j)] > 1$. Further insights on model comparison are provided in Supplementary Note 3. Overall, these differences between the null model distributions are likely to produce strong disagreement between the validated networks.

After focusing on individual co-occurrences, we ask in general how many co-occurrences are validated by each null model. We thus measure the link density $\rho(p^*)$ of the validated network, defined as the fraction of i,j pairs such that $p[C_{ij}] \le p^*$, as a function of the significance threshold p^*. Results for the various null models, reported in Fig. 4, show patterns that are consistent with the above observations. The width of the null model distributions sets the slope of the curves, so that BiCM is the model that validates the least by having longer tails and higher mean, whereas Hypergeometric validates the most by having shorter tails and lower mean. Note also that the Hypergeometric ρ does not tend to 0 for $p^* \to 0$: this is due to the distribution not vanishing within its finite support, whereas the effective support of the canonical models distribution is much larger due to the softness of the imposed constraints. Upstream of these considerations, we observe a very large difference in density between the various null models, even of one order of magnitude for the same p^* values. We can thus conclude that the different null models unavoidably lead to different filtered network structures, even when correcting for multiple hypothesis testing (see Supplementary Note 4).

Null models reconciliation. We now discuss a general methodology to reconcile the four validation schemes. The idea is to find a coherence area in the space of parameters where the filtered networks show a relatively good agreement. We start by assessing the structural similarity of the validated networks using three popular metrics of graph distance[61]. The first one is the simple Jaccard coefficient, which measures the number of links in common between two graphs. Jaccard is a known node-correspondence method, *i.e.*, it requires that the two graphs have the same node set and the pairwise correspondence between nodes is known (our validated networks satisfy this requirement). We further consider: *DeltaCon*[62], another known node-correspondence method based on the comparison of l-length paths connecting each node pair (we use the approximated version of the algorithm, which restricts the computation to randomly chosen pairs); and *Portrait Divergence*[63], an unknown node-correspondence method that compares the distribution of the shortest-path lengths between graphs. Any of these methods takes as input the adjacency matrices $\mathbf{V}$ and $\mathbf{V}'$ of two networks, each validated by a different null model, and returns a measure of their similarity $s_{\mathbf{V}\,\mathbf{V}'} \in [0,1]$, where $s_{\mathbf{V}\,\mathbf{V}'} = 0$ means the two

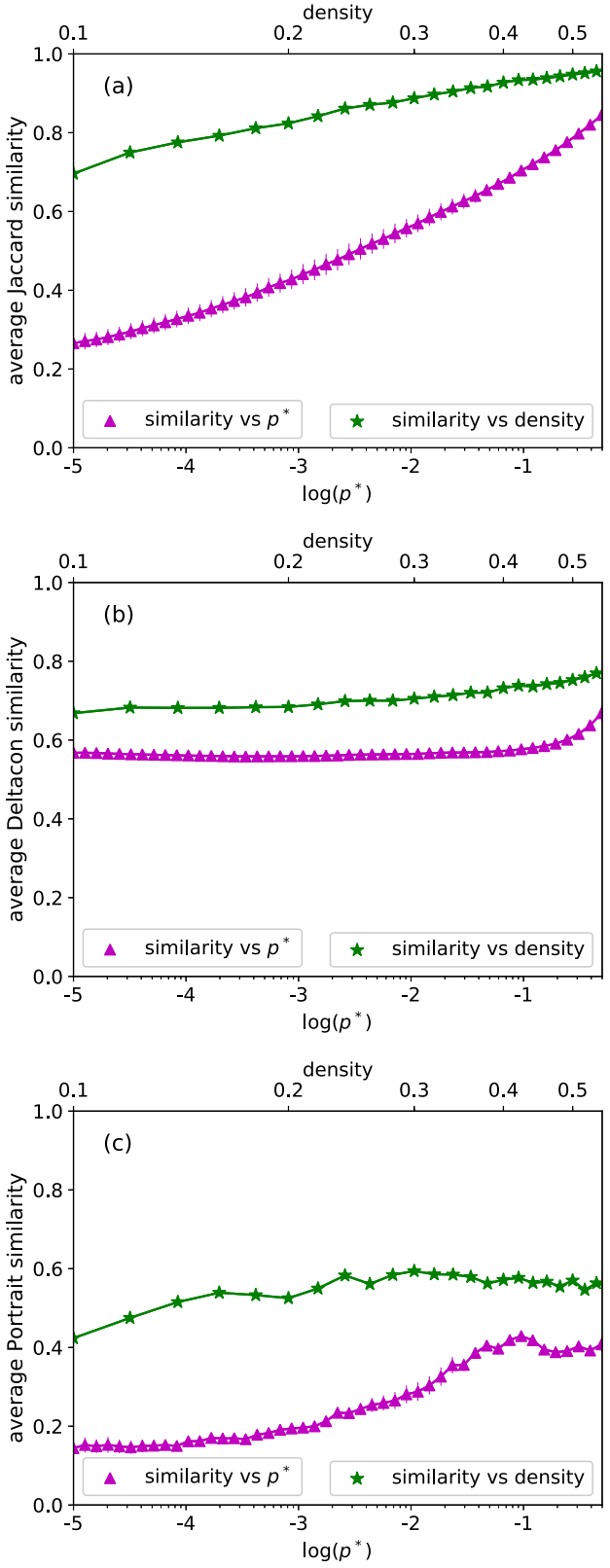

Fig. 5 Average structural similarity of the networks validated by the various null models. Error bars (not visible) represent standard deviations over choices of null model pairs. Similarity values measured through **a** Jaccard; **b** DeltaCon; **c** Portrait are plotted for filtered networks obtained with the same significance threshold p^* (magenta triangles) or of equal density (green stars). The latter option reveals a higher concordance among the null models. Note how similarity has a baseline dependence on the network density: a change of a link has more impact in lower density graphs.

networks are maximally different while $s_{\mathbf{V}\mathbf{V}'} = 1$ that they are identical. We can then obtain a mean similarity score by averaging over the six possible choices of null model pairs. A key issue in this comparison is how to choose the validated networks to match. The simplest choice is to compare networks obtained with the same significance threshold p^*, and study the average similarity as a function of p^* (Fig. 5, magenta triangles). We see that the average similarity is rather low, especially for Portrait Divergence. In order to recover some compatibility between the results of the different models, we can repeat the operations described above by taking validated networks with the same value of the link density ρ (that is, we adjust p^* for each network in order to obtain the match of ρ values). The resulting curves (green stars in Fig. 5) show that the average similarity of networks at equal ρ is always much higher than for networks at equal p^*.

We further study whether the validated networks are similar in terms of mesoscale or community structure. We choose this benchmark because statistical validation on networks is precisely meant to highlight the emergence of multiple-nodes patterns like motifs and communities. Broadly speaking, a community structure is defined by (typically non-overlapping) sets of nodes, characterized by having many more internal links—connecting nodes belonging to the same community—than external links—which connect nodes of different communities[26]. In order to find the best partition of the network nodes, a number of community detection algorithms have been proposed in the literature (we remand to[26] for a recent review of the field). Here we use the popular Louvain method[64], which is based on maximizing the quality function known as *Modularity*[65], defined as the observed fraction of links internal to communities with respect to a random benchmark. As there is no community detection method that performs best in all situations[66,67], in the Supplementary Note 5 we repeat the same analysis using community inference with Bayesian stochastic blockmodeling[68] (finding qualitatively similar results).

Figure 6 shows the results of the Louvain algorithm applied on the networks validated by the various null models. Given the previous analysis on structural similarity, we use ρ rather than p^* as independent variable with the aim of achieving a better compatibility between the results of the different models. In each plot, full circles represent the modularity of the best network partition, which increases as the network becomes more sparse, whereas the solid line marks the number of communities, which also increases for decreasing density due to the appearance of more disconnected components. A nontrivial feature that is common to all plots is the presence of a plateau at 4 communities for $0.1 \lesssim \rho \lesssim 0.3$. Additionally, modularity tends to stabilize at the lower extreme of this plateau. In order to understand whether a community structure common to all null models emerges in this region, we show in Fig. 7 the modularity as a function of the number of detected communities. Using this visualization we get rid of the trivial dependence of modularity and number of communities on ρ, and we observe a clear collapse of the curves corresponding to the four models. Additionally we see that modularity, after a fast increase, practically stops to grow after 4 communities are reached (the growth resumes only for a much larger number of communities). This observation points in the direction of a shared community structure. However we still do not know if the partitions identified in each null model setup are actually similar to each other.

To quantify the similarity between different partitions we use the *Adjusted Mutual Information* (AMI)[69]. We choose this

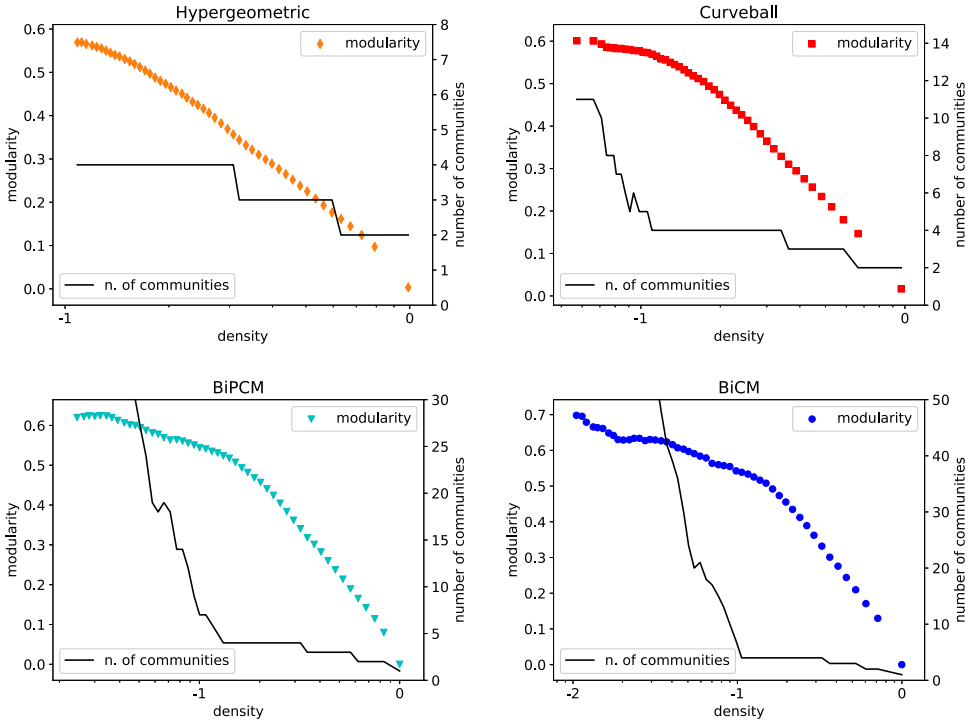

Fig. 6 Modularity and number of communities for the best partition obtained by the Louvain method on the validated network, as a function of the density ρ of validated links. Each point corresponds to the partition of highest modularity obtained in 100 runs of the algorithm with random initialization (hence it has no associated error). While modularity monotonically decreases with the density, we observe that the number of communities has a plateau at 4 for $0.1 \lesssim \rho \lesssim 0.3$, suggesting the presence of a region where an agreement between the four approaches is recovered.

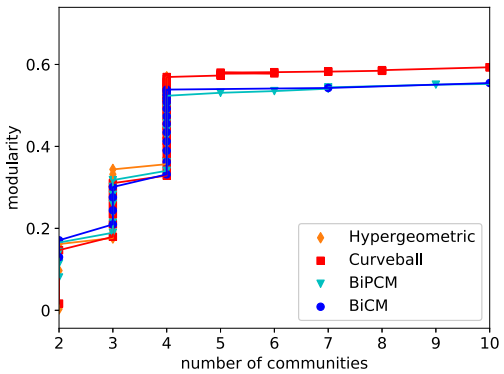

Fig. 7 Modularity vs number of communities for the best partition obtained on the validated networks. Note how the curves of the various models collapse onto each other. Besides, the growth of modularity suddenly stops at 4 communities, suggesting the presence of a robust partition shared between all null models.

metrics as it discounts the agreement between partitions solely due to chance, and it is also relatively stable with respect to the presence of disconnected components[70]. Given two network partitions U_1 and U_2, their AMI is defined as

$$\mathrm{AMI}(U_1, U_2) = \frac{\mathrm{MI}(U_1, U_2) - \mathcal{E}\{\mathrm{MI}(U_1, U_2)\}}{\max\{\mathrm{H}(U_1), \mathrm{H}(U_2)\} - \mathcal{E}\{\mathrm{MI}(U_1, U_2)\}} \quad (4)$$

where $\mathrm{MI}(U_1, U_2)$ is the mutual information between U_1 and U_2[71] while $\mathrm{H}(U_1)$ and $\mathrm{H}(U_2)$ is the Shannon entropy associated with U_1 and U_2, respectively. The adjustment consists in discounting the expected value $\mathcal{E}\{\mathrm{MI}(U_1, U_2)\}$ of the mutual information between two random partitions with the same number of nodes per community as U_1 and U_2. This correction is needed since the baseline value of mutual information between two random partitions is not constant but grows with the number of communities[69]. AMI varies between 0 (if the observed partitions are consistent with a random labeling) and 1 (if the two partitions coincide).

We can thus take a pair of networks each validated using a given null model, extract the respective best partitions (of highest modularity) and compute their AMI. In analogy to what we did for structural similarity metrics, we compare networks that are either validated with the same significance threshold p^*, or have the same density (that is, we adjust p^* for each network in order to obtain the match of ρ values). We perform this operation for the six possible choices of null model pairs to obtain an average AMI value. Results as a function of p^* or ρ (Fig. 8, magenta triangles or green stars, respectively) confirm that the density ρ, and not p^*, is the right knob to turn for finding an agreement among the models. Indeed the average AMI computed for networks at equal ρ is almost always higher than AMI for networks at equal p^*—the only exception being the region around $\rho \simeq 0.3$ where the number of detected communities switches from 4 to 3 but not simultaneously for all models. We can also identify a maximum AMI for $\rho \sim 0.2$, corresponding to the shared community structure illustrated in Fig. 9. Hence the four filtering techniques, which in general produce very different statistically validated networks, can be reconciled by choosing model-specific p^* such that the resulting densities of the validated networks are approximately equal and the AMI is maximum. This strategy is rather general and, in principle, can be applied to any bipartite to monopartite projection in order to produce a "meta-validated" filtered network that maximizes the agreement between the different filtering techniques. Even more importantly, it can resolve the arbitrariness in the choice of the significance threshold p^*.

To further support the general applicability of our framework, we show in the Supplementary Note 6 the same analysis performed to several other bipartite networks belonging to totally

different contexts. In all cases we find that both the structural similarity and the AMI of the network partition are consistently higher when computed at equal ρ than when obtained at equal p^*. Additionally, when modularity is high enough, a robust community structure shared among the null models emerges.

Conclusions

The increasing availability of complex data we are experiencing nowadays calls for techniques to extract meaningful information from large-scale networks of interactions. The statistical validation of networks is based on comparing empirically observed patterns with their distributional expectation under a null network model. This allows performing a statistical test of whether empirical data is explained by the model or represent additional information. A statistically validated network is built by retaining only the structurally relevant interactions for which the null hypothesis is rejected. This can be of crucial importance to obtain simpler and clearer descriptions of complex systems[34].

For instance, statistical validation of bipartite network projections has been used to detect important patterns in financial markets, such as preferential or avoided relationships[72,73], clusters of investors characterized by the same investment profile[74–76], and overlapping portfolios bearing the highest riskiness for fire sales liquidation[54]. In the context of economic and innovation systems, validated network projections have been used to detect modules of countries with similar industrial profile and the hierarchical structure of products and services[55,77], traces of specializations emerging from the baseline diversification strategy of countries[78], and predictive innovation patterns involving the interplay of scientific, technological and economic activities[60]. In the context of mobile communications, validated networks were shown to be more resilient than ordinary networks to errors[79].

Naturally, any null model hinges on a definition of what type of information represents a signal as opposed to noise. As a result, the validated networks obtained through different filtering techniques carry different meanings and highlight different properties. Even different constructions of the same null model may yield different outcomes. This latter issue has been recently shown in the context of nestedness in ecological systems[17]: whether the degree sequence is responsible for the nestedness of a bipartite network[80,81] depends on the choice of the CM-based null model ensemble (microcanonical or canonical)[82]. The non-equivalence between the microcanonical and canonical ensembles is due to the extensive number of constraints (one for each network node) and holds also in the thermodynamic limit[83–85].

In this work we have reviewed (using a unified notation) the CM-based null model formulations for bipartite network projections, and performed a systematic comparison in terms of null model characteristics and validation outcomes within the same contexts. We showed that the different model formulations lead to different validation results, both at the level of individual links and of macroscale network properties (see also this recent preprint[86]). However we do provide a recipe to reconcile the

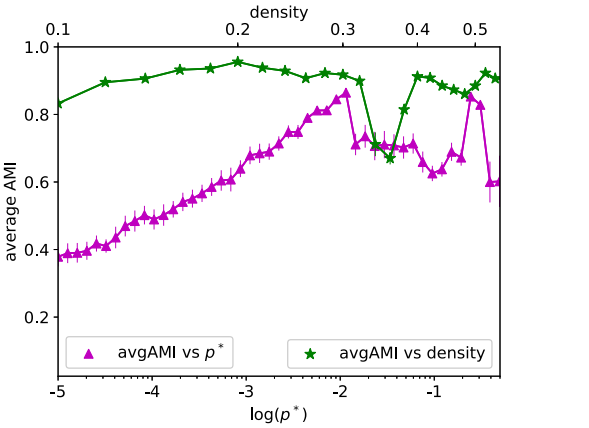

Fig. 8 Average Adjusted Mutual Information (AMI) between the best partitions of the network validated by the various null models. Error bars represent standard deviations over choices of null model pairs. Values are plotted for filtered networks obtained with the same significance threshold p^* (magenta triangles) or of equal density (green stars). The latter option reveals a higher concordance among the null models.

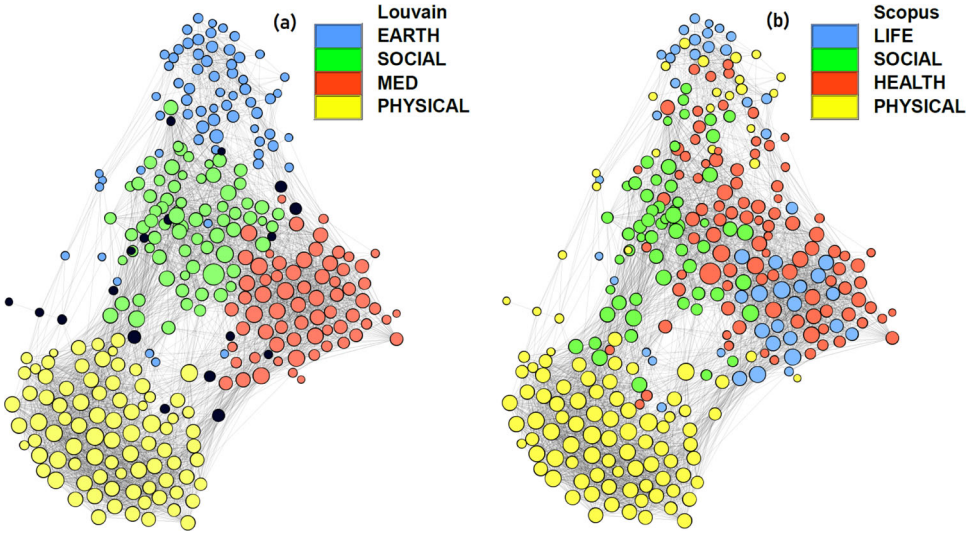

Fig. 9 Community structure of the network of co-occurrence between scientific fields, validated by the Hypergeometric null model. The link density is $\rho = 0.2$, corresponding to the maximum adjusted mutual information (AMI) value of Fig. 8. **a** Each color identifies a community (hand-labeled by us as in the legend) shared among the four validated networks, whereas the few black nodes denote the mismatches. The network is represented using a force atlas layout. **b** As a comparison, we report the same network with scientific fields labeled according to their ASJC (All Science Journal Classification) subject area assigned by Scopus' in-house experts. Visually, the community structure defined by such classification is not much coherent with the network partition induced by our meta-validation approach based on significant co-occurrences, which we recall links two scientific fields when they require common capabilities.

validation outcomes, by comparing networks obtained with model-specific significance thresholds such that the density of validated links becomes comparable. Additionally this comparison may allow to identify the region of density values where the agreement between models is maximum. On one hand this solves the arbitrariness in the choice of the significance threshold. On the other hand, it offers a meta-validation approach to identify the filtered configurations with the highest signal-to-noise ratio.

We have included in our comparative study CM-based null models defined on one or both sets of the bipartite network, as well as those defined directly on the monopartite projection, considering in all cases both hard and soft constraints. Note that, in principle, "softer" constraints could be imposed by fixing the functional form of the degree distribution rather than the degree sequence, as in the *hypersoft* CM[87–89]. This approach may be more adequate in the case of dynamic networks, in which degree sequences are never fixed but their distributions are often stable. Developing a hypersoft CM for bipartite networks and projections, and adding it to our meta-validation framework represents a promising research direction. Additional challenges for future research are represented by the development of suitable models for weighted bipartite networks and their projections[90,91], as well as the extension of validation methods beyond pairwise interactions[92,93].

As long-term goal we plan to investigate whether the proposed meta-validation approach allows not only to capture the most relevant structural properties of the network projection, but also to help in predicting its evolution – namely, which links will appear in the future. This could be important in various contexts, from link prediction for recommender systems based on collaborative filtering[12,94] to assessing prices trend and systemic risk in financial networks of portfolios and assets[54,95] and forecasting development patterns in economic and innovation systems[60,96,97].

Methods

Null models of bipartite network projections. Here we provide the mathematical definitions of the null models used in our analysis.

Microcanonical partial model: Hypergeometric. For the projection of a bipartite network on set L, an analytic null hypothesis can be formulated by assuming random connections between nodes of the two sets L and Γ that preserve the degree heterogeneity of set L[44,46]. Under this hypothesis, the probability that nodes i and j have x co-occurrences is given by the hypergeometric distribution

$$\pi(x|i,j) = \binom{k_i}{x}\binom{|\Gamma| - k_i}{k_j - x}\Big/\binom{|\Gamma|}{k_j} \tag{5}$$

and the mean value of the co-occurrences is

$$\langle C_{ij}\rangle = k_i k_j/|\Gamma|. \tag{6}$$

This probability is exact only when nodes of set Γ have the same degree. A tentative extension to deal with the degree heterogeneity of set Γ consists in splitting the original bipartite network into subnetworks each consisting of set Γ nodes with the same degree and of all set L nodes linked to them, so that the null hypothesis can be properly cast for each subnetwork[46]. However, when set Γ is highly heterogeneous these subnetworks are many in number and very sparse, causing severe resolution issues (see the discussion in[54]).

Microcanonical full model: Curveball. Building a null bipartite network model where only configurations with a given degree sequence for both sets of nodes are allowed has not been tackled analytically up to now (exact results exist only in the thermodynamic limit concerning the count of bipartite graphs with given degree sequences[98]). This model is hard to deal with because, differently from the canonical case (see below), link probabilities are not pairwise independent. Therefore the model must be defined through an ensemble of bipartite network configurations that are generated numerically by swapping links iteratively so to preserve degrees exactly. The *Curveball* algorithm[50–52] works as follows: Starting from the empirical bipartite network $\mathbf{M}$, these steps are repeated n times:

1. Select at random a pair of nodes i, j in set L;
2. Check that the neighborhoods of the nodes are not perfectly overlapping (otherwise start again);
3. Take the set of uncommon neighbors $\delta(i,j) = \{\alpha \in \Gamma \,|\, M_{i\alpha} \oplus M_{j\alpha} = 1\}$ and remove them from the neighborhood of both;
4. Assign $k_i - \sum_\alpha M_{i\alpha}M_{j\alpha}$ new neighbors to node i, chosen at random from $\delta(i,j)$, and the rest of the nodes in $\delta(i,j)$ to node j.

The result is a randomized bipartite network configuration $\tilde{\mathbf{M}}$ (here and in what follows we use the tilde symbol to denote matrix configurations of the null model). This procedure is repeated iteratively to generate an ensemble $\{\tilde{\mathbf{M}}_q\}_{q=1}^{Q}$ of Q independent randomizations of the bipartite network. The null model ensemble $\{\tilde{\mathbf{C}}_r\}_{r=1}^{R}$ of projected networks is then obtained by projecting pairs of different instances of bipartite randomizations (that is, a generic configuration is obtained as $\tilde{\mathbf{C}}_r = \tilde{\mathbf{M}}_q\tilde{\mathbf{M}}_{q'}^T$ with $q \neq q'$). The null model distributions $\pi(\,\cdot\,|i,j)\ \forall\, i,j$ are then computed numerically by sampling from such an ensemble. Here we use $n = 5\min\{|\,\mathrm{L}\,|, |\Gamma|\}$ and $Q = R = 10000$. Note that for a numerically-generated ensemble of R network configurations, the minimum p-value that can be used for statistical testing is $1/R$.

Canonical full model: BiCM. Generally speaking, canonical models of networks[43,99–101] (also known as *exponential random graphs*[102–104]) define an ensemble Ω of networks using a constrained entropy maximization procedure, which leads to assuming the utmost ignorance about the unconstrained degrees of freedom of the system[34,105]. The Bipartite Configuration Model (BiCM)[53] applies to bipartite networks by constraining the ensemble average of the degree sequence for both node sets. The ensemble probability distribution that maximizes the Shannon entropy under these constraints is

$$P(\tilde{\mathbf{M}}|\{\theta_i\},\{t_\alpha\}) = e^{-H(\tilde{\mathbf{M}},\{\theta_i\},\{t_\alpha\})}/Z(\{\theta_i\},\{t_\alpha\}) \tag{7}$$

where $\{\theta_i\}$ and $\{t_\alpha\}$ are the sets of Lagrange multipliers associated to the constraints $\{k_i\}$ and $\{\kappa_\alpha\}$ respectively, $Z(\{\theta_i\},\{t_\alpha\}) = \sum_{\mathbf{M}\in\Omega} e^{-H(\tilde{\mathbf{M}},\{\theta_i\},\{t_\alpha\})}$ is the partition function and the Hamiltonian $H(\tilde{\mathbf{M}},\{\theta_i\},\{t_\alpha\}) = \sum_{i\in\mathrm{L}}\theta_i k_i(\tilde{\mathbf{M}}) + \sum_{\alpha\in\Gamma} t_\alpha \kappa_\alpha(\tilde{\mathbf{M}})$ sums up the imposed constraints. Note that $P(\tilde{\mathbf{M}}|\{\theta_i\},\{t_\alpha\})$ depends on $\tilde{\mathbf{M}}$ only through $k_i(\tilde{\mathbf{M}})$ and $\kappa_\alpha(\tilde{\mathbf{M}})$: network configurations with the same value of the constraints are equiprobable, which implies that the canonical ensemble is maximally non-committal (or the least biased) with respect to the properties that are not enforced on the system. Since degrees are linear constraints the partition function can be computed analytically, so the ensemble probability factorizes as

$$P(\tilde{\mathbf{M}}|\{\theta_i\},\{t_\alpha\}) = \prod_{i,\alpha} p_{i\alpha}^{\tilde{M}_{i\alpha}}(1-p_{i\alpha})^{1-\tilde{M}_{i\alpha}} \tag{8}$$

where $p_{i\alpha}$ is the existence probability of the link connecting nodes i and α:

$$p_{i\alpha} = (e^{\theta_i + t_\alpha} + 1)^{-1}. \tag{9}$$

The numerical values of the link probabilities (*i.e.*, of the Lagrange multipliers) are determined by maximizing the likelihood of the empirical bipartite network $\mathbf{M}$ in the ensemble, which implies solving the constraints equations

$$\begin{cases} k_i = \sum_{\alpha\in\Gamma} p_{i\alpha} & i \in \mathrm{L} \\ \kappa_\alpha = \sum_{i\in\mathrm{L}} p_{i\alpha} & \alpha \in \Gamma \end{cases} \tag{10}$$

Once link probabilities have been found, the expected co-occurrences between any two nodes $i \neq j$ are

$$\langle C_{ij}\rangle = \sum_{\alpha\in\Gamma} p_{i\alpha}p_{j\alpha}, \tag{11}$$

and the probability distribution $\pi(\,\cdot\,|i,j)$ of this quantity is the distribution of the sum of Γ independent Bernoulli trials, each with success probability $p_{i\alpha}p_{j\alpha}$. This is a Poisson-Binomial distribution, which can be computed numerically[54] or analytically[55] as

$$\pi(x|i,j) = \sum_{\gamma_x}\left[\prod_{\alpha\in\gamma_x} p_{i\alpha}p_{j\alpha}\prod_{\beta\notin\gamma_x}(1-p_{i\beta}p_{j\beta})\right] \tag{12}$$

where γ_x denotes all possible x-tuples of nodes in set Γ.

Canonical partial model: BiPCM. The "partial" version of the BiCM, named BiPCM in ref. [55], is defined as the canonical model that constrains only the degree sequence of set L. As such, it is a special case of the BICM described above where all Lagrange multipliers $\{t_\alpha\}$ associated with degrees of set Γ are "switched off" (*i.e.*, set equal to zero). The Hamiltonian is thus $H(\tilde{\mathbf{M}},\{\theta_i\}) = \sum_{i\in\mathrm{L}}\theta_i k_i(\tilde{\mathbf{M}})$ and the link probability of generic link (i,α) becomes $p_{i\alpha} = (e^{\theta_i}+1)^{-1}$. Using the constraint equations $k_i = \sum_{\alpha\in\Gamma} p_{i\alpha}\ \forall\, i \in \mathrm{L}$ we get the explicit expression

$$p_{i\alpha} = k_i/|\Gamma| \qquad \forall i \in \mathrm{L}\,. \tag{13}$$

Therefore the expected value of the co-occurrence between any two nodes i and j is

$$\langle C_{ij}\rangle = k_i k_j/|\Gamma| \tag{14}$$

and its distribution has a simple Binomial form

$$\pi(x|i,j) = \binom{|\Gamma|}{x}\left(\frac{k_i k_j}{|\Gamma|^2}\right)^x\left(1-\frac{k_i k_j}{|\Gamma|^2}\right)^{|\Gamma|-x} \quad (15)$$

Data, RCA filter and projection. To build the bipartite network of countries and scientific fields, we use data on scientific productivity and impact of countries collected from the SCIMAGO platform (based on Scopus). The database contains the corpus of scientific publications in journals, book series, conference proceedings, and books in the various scientific fields, covering the time interval from 1996 to 2018. Data are then aggregated at the level of countries and scientific fields (in total there are $|\mathrm{L}| = 307$ scientific fields and $|\Gamma| = 239$ countries), so that $W_{i\alpha}$ is the total number of scientific documents produced by country α in scientific field i during the time span of the data.

In order to determine whether a given country α shows a comparative advantage in field i, both with respect to other countries as well as to other fields, the *revealed comparative advantage* (RCA)[106] filter comes at hand. While originally developed in the economic context, this metric has also found use in studies of scientific production[60,107,108]. RCA is an intensive indicator computed as the ratio between the weight of field i in the scientific basket of country α and the weight of field i in the total world science. As a comparative advantage is revealed if RCA > 1, we binarize the raw matrices to obtain new matrices

$$M_{i\alpha} = \begin{cases} 1 \text{ if } \frac{W_{i\alpha}}{\sum_j W_{j\alpha}} \Big/ \frac{\sum_\beta W_{i\beta}}{\sum_{j\beta} W_{j\beta}} \geq 1, \\ 0 \text{ otherwise.} \end{cases} \quad (16)$$

Note that the RCA filter is properly normalized by making quantities related to different countries and fields comparable[109].

Once the binary bipartite matrix is defined, we build the projected network of co-occurrences between scientific fields, whose generic connection between fields i and j is $C_{ij} = \sum_\alpha M_{i\alpha} M_{j\alpha}$. Note that for the sake of having a a clearer picture and more analytical insights on the various null models, we do not employ here more refined formulations of co-occurrences that use additional normalization by degrees[57,59,60].

Data availability

The dataset about scientific productivity of countries analyzed during the current study can be obtained from SCImago, (n.d.). SJR - SCImago Journal & Country Rank [Portal], which can be retrieved at https://www.scimagojr.com/countryrank.php. The other datasets analyzed in the Supplementary Note 6 are available at the web addresses indicated in the same document.

Code availability

The code to run the Curveball algorithm can be retrieved from[51], while the code to run BiCM is available at https://github.com/tsakim/bicm(Hypergeometric and BiPCM have analytic formulas). Codes for computing network distance metrics can be retrieved from[61] while the code for computing Modularity and AMI are available respectively at https://github.com/taynaud/python-louvain and https://scikit-learn.org/stable/modules/clustering.html#mutual-info-score[110]. Network visualizations have been generated using Gephi https://gephi.org.

Received: 14 June 2021; Accepted: 10 March 2022;
Published online: 05 April 2022

References

1. Song, C., Havlin, S. & Makse, H. A. Self-similarity of complex networks. *Nature* **433**, 392–395 (2005).
2. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).
3. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. Critical phenomena in complex networks. *Rev. Mod. Phys.* **80**, 1275–1335 (2008).
4. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925–979 (2015).
5. Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353**, 163–166 (2016).
6. Barabási, A.-L. The network takeover. *Nat. Phys.* **8**, 14–16 (2012).
7. Newman, M. Networks (Oxford university press, 2018).
8. Caldarelli, G. A perspective on complexity and networks science. *J. Phys. Complexity* **1**, 021001 (2020).
9. Holme, P., Liljeros, F., Edling, C. R. & Kim, B. J. Network bipartivity. *Phys. Rev. E* **68**, 056107 (2003).
10. Faust, K. Centrality in affiliation networks. *Soc. Netw.* **19**, 157–191 (1997).
11. Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl Acad. Sci. USA* **101**, 5200–5205 (2004).
12. Zhou, T., Ren, J., Medo, M. & Zhang, Y.-C. Bipartite network projection and personal recommendation. *Phys. Rev. E* **76**, 046115 (2007).
13. Bardoscia, M. et al. The physics of financial networks. *Nat. Rev. Phys.* **3**, 490–507 (2021).
14. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proc. Natl Acad. Sci. USA* **106**, 10570–10575 (2009).
15. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. A new metrics for countries' fitness and products' complexity. *Sci. Rep.* **2**, 723 (2012).
16. Ings, T. C. et al. Review: Ecological networks - beyond food webs. *J. Anim. Ecol.* **78**, 253–269 (2009).
17. Mariani, M. S., Ren, Z.-M., Bascompte, J. & Tessone, C. J. Nestedness in complex networks: observation, emergence, and implications. *Phys. Rep.* **813**, 1–90 (2019).
18. Goh, K.-I. et al. The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
19. Pavlopoulos, G. A. et al. Bipartite graphs in systems biology and medicine: a survey of methods and applications. GigaScience 7, https://doi.org/10.1093/gigascience/giy014 (2018).
20. Vasques Filho, D. & O'Neale, D. R. J. Degree distributions of bipartite networks and their projections. *Phys. Rev. E* **98**, 022307 (2018).
21. Kruskal, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* **7**, 48–50 (1956).
22. Tumminello, M., Aste, T., Di Matteo, T. & Mantegna, R. N. A tool for filtering information in complex systems. *Proc. Natl Acad. Sci. USA* **102**, 10421–10426 (2005).
23. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
24. Milo, R. et al. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
25. Serrano, M. A., Boguñá, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *PNAS* **106**, 6483–6488 (2009).
26. Fortunato, S. & Hric, D. Community detection in networks: a user guide. *Phys. Rep.* **659**, 1–44 (2016).
27. MacMahon, M. & Garlaschelli, D. Community detection for correlation matrices. *Phys. Rev. X* **5**, 021006 (2015).
28. Bongiorno, C., London, A., Miccichè, S. & Mantegna, R. N. Core of communities in bipartite networks. *Phys. Rev. E* **96**, 022321 (2017).
29. Vázquez, A. et al. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc. Natl Acad. Sci.* **101**, 17940–17945 (2004).
30. Foster, D. V., Foster, J. G., Grassberger, P. & Paczuski, M. Clustering drives assortativity and community structure in ensembles of networks. *Phys. Rev. E* **84**, 066117 (2011).
31. Colomer-de Simón, P., Serrano, M. Á., Beiró, M. G., Alvarez-Hamelin, J. I. & Boguñá, M. Deciphering the global organization of clustering in real complex networks. *Sci. Rep.* **3**, 2517 (2013).
32. Orsini, C. et al. Quantifying randomness in real networks. *Nat. Commun.* **6**, 8627 (2015).
33. Marcaccioli, R. & Livan, G. A pólya urn approach to information filtering in complex networks. *Nat. Commun.* **10**, 745 (2019).
34. Cimini, G. et al. The statistical physics of real-world networks. *Nat. Rev. Phys.* **1**, 58–71 (2019).
35. Colizza, V., Flammini, A., Serrano, M. A. & Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110 (2006).
36. Nunes Amaral, L. A. & Guimera, R. Lies, damned lies and statistics. *Nat. Phys.* **2**, 75–76 (2006).
37. Erdős, P. & Rényi, A. On random graphs. *Publicationes Mathematicae Debrecen* **6**, 290–297 (1959).
38. Latapy, M., Magnien, C. & Vecchio, N. D. Basic notions for the analysis of large two-mode networks. *Soc. Netw.* **30**, 31–48 (2008).
39. Neal, Z. Identifying statistically significant edges in one-mode projections. *Soc. Netw. Anal. Mining* **3**, 915–924 (2013).
40. Serafino, M. et al. True scale-free networks hidden by finite size effects. *Proc. Natl Acad. Sci. USA* **118**, https://doi.org/10.1073/pnas.2013825118 (2021).
41. Chung, F. & Lu, L. Connected components in random graphs with given expected degree sequences. *Ann. Combinatorics* **6**, 125–145 (2002).
42. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001).
43. Squartini, T. & Garlaschelli, D. Analytical maximum-likelihood method to detect patterns in real networks. *N. J. Phys.* **13**, 083001 (2011).
44. Teece, D. J., Rumelt, R., Dosi, G. & Winter, S. Understanding corporate coherence: theory and evidence. *J. Econ. Behav. Organization* **23**, 1–30 (1994).
45. Goldberg, D. S. & Roth, F. P. Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA* **100**, 4372–4376 (2003).
46. Tumminello, M., Miccichè, S., Lillo, F., Piilo, J. & Mantegna, R. N. Statistically validated networks in bipartite complex systems. *PLoS ONE* **6**, e17994 (2011).

47. Neal, Z. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Soc. Netw.* **39**, 84–97 (2014).
48. Zweig, K. A. & Kaufmann, M. A systematic approach to the one-mode projection of bipartite graphs. *Soc. Netw. Anal. Mining* **1**, 187–218 (2011).
49. Gionis, A., Mannila, H., Mielikäinen, T. & Tsaparas, P. Assessing data mining results via swap randomization. ACM Trans. Knowl. Discov. Data 1, https://doi.org/10.1145/1297332.1297338 (2007).
50. Verhelst, N. D. An efficient mcmc algorithm to sample binary matrices with fixed marginals. *Psychometrika* **73**, 705 (2008).
51. Strona, G., Nappo, D., Boccacci, F., Fattorini, S. & San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat. Commun.* **5**, 4114 (2014).
52. Carstens, C. J. Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm. *Phys. Rev. E* **91**, 042812 (2015).
53. Saracco, F., Di Clemente, R., Gabrielli, A. & Squartini, T. Randomizing bipartite networks: the case of the world trade web. *Sci. Rep.* **5**, 10595 (2015).
54. Gualdi, S., Cimini, G., Primicerio, K., Di Clemente, R. & Challet, D. Statistically validated network of portfolio overlaps and systemic risk. *Sci. Rep.* **6**, 39467 (2016).
55. Saracco, F. et al. Inferring monopartite projections of bipartite networks: an entropy-based approach. *N. J. Phys.* **19**, 053022 (2017).
56. Mastrandrea, R., Squartini, T., Fagiolo, G. & Garlaschelli, D. Enhanced reconstruction of weighted networks from strengths and degrees. *N. J. Phys.* **16**, 043022 (2014).
57. Hidalgo, C. A., Klinger, B., Barabási, A.-L. & Hausmann, R. The product space conditions the development of nations. *Science* **317**, 482–487 (2007).
58. Klimek, P., Hausmann, R. & Thurner, S. Empirical confirmation of creative destruction from world trade data. *PLoS ONE* **7**, e38924 (2012).
59. Zaccaria, A., Cristelli, M., Tacchella, A. & Pietronero, L. How the taxonomy of products drives the economic development of countries. *PLoS ONE* **9**, e113770 (2014).
60. Pugliese, E. et al. Unfolding the innovation system for the development of countries: co-evolution of science, technology and production. *Sci. Rep.* **9**, 16440 (2019).
61. Tantardini, M., Ieva, F., Tajoli, L. & Piccardi, C. Comparing methods for comparing networks. *Sci. Rep.* **9**, 17557 (2019).
62. Koutra, D., Shah, N., Vogelstein, J. T., Gallagher, B. & Faloutsos, C. Deltacon: principled massive-graph similarity function with attribution. ACM Trans. Knowl. Discov. Data 10, https://doi.org/10.1145/2824443 (2016).
63. Bagrow, J. P. & Bollt, E. M. An information-theoretic, all-scales approach to comparing networks. *Appl. Netw. Sci.* **4**, 45 (2019).
64. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
65. Newman, M. E. Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA* **103**, 8577–8582 (2006).
66. Peel, L., Larremore, D. B. & Clauset, A. The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, e1602548 (2017).
67. Ghasemian, A., Hosseinmardi, H. & Clauset, A. Evaluating overfit and underfit in models of network community structure. *IEEE Trans. Knowl. Data Eng.* **32**, 1722–1735 (2020).
68. Peixoto, T. P. Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Phys. Rev. E* **89**, 012804 (2014).
69. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).
70. Romano, S., Bailey, J., Nguyen, V. & Verspoor, K. Standardized mutual information for clustering comparisons: One step further in adjustment for chance. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, II-1143-II-1151, https://doi.org/10.5555/3044805.3045020 (JMLR.org, 2014).
71. Strehl, A. & Ghosh, J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2003).
72. Hatzopoulos, V., Iori, G., Mantegna, R. N., Miccichè, S. & Tumminello, M. Quantifying preferential trading in the e-mid interbank market. *Quant. Finance* **15**, 693–710 (2015).
73. Musciotto, F., Piilo, J. & Mantegna, R. N. High-frequency trading and networked markets. *Proc. Natl Acad. Sci. USA* **118**, https://doi.org/10.1073/pnas.2015573118 (2021).
74. Tumminello, M., Lillo, F., Piilo, J. & Mantegna, R. N. Identification of clusters of investors from their real trading activity in a financial market. *N. J. Phys.* **14**, 013041 (2012).
75. Musciotto, F., Marotta, L., Miccichè, S., Piilo, J. & Mantegna, R. N. Patterns of trading profiles at the nordic stock exchange. a correlation-based approach. *Chaos, Solitons & Fractals* **88**, 267–278 (2016).
76. Musciotto, F., Marotta, L., Piilo, J. & Mantegna, R. N. Long-term ecology of investors in a financial market. *Palgrave Commun.* **4**, 92 (2018).
77. Zaccaria, A., Mishra, S., Cader, M. Z. & Pietronero, L. Integrating services in the economic fitness approach. World Bank Policy Research Working Paper (2018).
78. Straka, M. J., Caldarelli, G. & Saracco, F. Grand canonical validation of the bipartite international trade network. *Phys. Rev. E* **96**, 022306 (2017).
79. Li, M.-X. et al. Statistically validated mobile communication networks: the evolution of motifs in european and chinese data. *N. J. Phys.* **16**, 083038 (2014).
80. Jonhson, S., Domínguez-García, V. & Muñoz, M. A. Factors determining nestedness in complex networks. *PLoS ONE* **8**, e74025 (2013).
81. Payrató-Borràs, C., Hernández, L. & Moreno, Y. Breaking the spell of nestedness: the entropic origin of nestedness in mutualistic systems. *Phys. Rev. X* **9**, 031024 (2019).
82. Bruno, M., Saracco, F., Garlaschelli, D., Tessone, C. J. & Caldarelli, G. The ambiguity of nestedness under soft and hard constraints. *Sci. Rep.* **10**, 19903 (2020).
83. Barré, J. & Gonçalves, B. Ensemble inequivalence in random graphs. *Phys. A* **386**, 212–218 (2007).
84. Anand, K. & Bianconi, G. Entropy measures for networks: toward an information theory of complex topologies. *Phys. Rev. E* **80**, 045102 (2009).
85. Squartini, T., de Mol, J., den Hollander, F. & Garlaschelli, D. Breaking of ensemble equivalence in networks. *Phys. Rev. Lett.* **115**, 268701 (2015).
86. Neal, Z. P., Domagalski, R. & Sagan, B. Comparing models for extracting the backbone of bipartite projections. https://arxiv.org/abs/2105.13396 (2021).
87. Anand, K., Krioukov, D. & Bianconi, G. Entropy distribution and condensation in random networks with a given degree distribution. *Phys. Rev. E* **89**, 062807 (2014).
88. van der Hoorn, P., Lippner, G. & Krioukov, D. Sparse maximum-entropy random graphs with a given power-law degree distribution. *J. Stat. Phys.* **173**, 806–844 (2018).
89. Voitalov, I., van der Hoorn, P., Kitsak, M., Papadopoulos, F. & Krioukov, D. Weighted hypersoft configuration model. *Phys. Rev. Res.* **2**, 043157 (2020).
90. Garlaschelli, D. & Loffredo, M. I. Generalized bose-fermi statistics and structural correlations in weighted networks. *Phys. Rev. Lett.* **102**, 038701 (2009).
91. Gabrielli, A., Mastrandrea, R., Caldarelli, G. & Cimini, G. Grand canonical ensemble of weighted networks. *Phys. Rev. E* **99**, 030301 (2019).
92. Battiston, F. et al. Networks beyond pairwise interactions: Structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).
93. Musciotto, F., Battiston, F. & Mantegna, R. N. Detecting informative higher-order interactions in statistically validated hypergraphs. https://arxiv.org/abs/2103.16484 (2021).
94. Kobayashi, T., Takaguchi, T. & Barrat, A. The structured backbone of temporal social ties. *Nat. Commun.* **10**, 220 (2019).
95. Vodenska, I., Dehmamy, N., Becker, A. P., Buldyrev, S. V. & Havlin, S. Systemic stress test model for shared portfolio networks. *Sci. Rep.* **11**, 3358 (2021).
96. Tacchella, A., Zaccaria, A., Miccheli, M. & Pietronero, L. Relatedness in the era of machine learning. https://arxiv.org/abs/2103.06017 (2021).
97. Straccamore, M., Pietronero, L. & Zaccaria, A. Which will be your firm's next technology? comparison between machine learning and network-based algorithms. https://arxiv.org/abs/2110.02004 (2021).
98. Liebenau, A. & Wormald, N. Asymptotic enumeration of digraphs and bipartite graphs by degree sequence. https://arxiv.org/abs/2006.15797 (2020).
99. Park, J. & Newman, M. E. J. Statistical mechanics of networks. *Phys. Rev. E* **70**, 066117 (2004).
100. Bianconi, G. The entropy of randomized network ensembles. *Europhys. Lett.* **81**, 28005 (2008).
101. Garlaschelli, D. & Loffredo, M. I. Maximum likelihood: extracting unbiased information from complex networks. *Phys. Rev. E* **78**, 015101(R) (2008).
102. Holland, P. W. & Leinhardt, S. An exponential family of probability distributions for directed graphs. *J. Am. Stat. Assoc.* **76**, 33–50 (1981).
103. Strauss, D. On a general class of models for interaction. *SIAM Rev.* **28**, 513–527 (1986).
104. Snijders, T. A. B., Pattison, P. E., Robins, G. L. & Handcock, M. S. New specifications for exponential random graph models. *Sociol. Methodol.* **36**, 99–153 (2006).
105. Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
106. Balassa, B. Trade liberalisation and "revealed" comparative advantage. *Manchester School* **33**, 99–123 (1965).
107. Bowen, H. P. On the theoretical interpretation of indices of trade intensity and revealed comparative advantage. *Weltwirtschaftliches Archiv* **119**, 464–472 (1983).
108. Guevara, M. R., Hartmann, D., Aristarán, M., Mendoza, M. & Hidalgo, C. A. The research space: Using career paths to predict the evolution of the research output of individuals, institutions, and nations. *Scientometrics* **109**, 1695–1709 (2016).

109. Radicchi, F., Fortunato, S. & Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl Acad. Sci. USA* **105**, 17268–17272 (2008).
110. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
111. Icons from Iconmonstr https://iconmonstr.com.
112. Icons from Linseed Studio, NounProject https://thenounproject.com.

Acknowledgements

We thank Benedetta Castagna and Aurelio Patelli for useful discussion. We acknowledge the CREF project "Complessità in Economia" and the ISC-CNR project "CompLang".

Author contributions

G.C. and A.Z. designed the research. A.C. and L.D. performed research. A.C. realized the figures. G.C. wrote the manuscript. G.C., A.C. and A.Z. reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-022-00856-9.

Correspondence and requests for materials should be addressed to Giulio Cimini.

Peer review information *Communications Physics* thanks Tzu-Chi Yen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.