nature communications

Article

https://doi.org/10.1038/s41467-026-74723-4

Extracting the transitivity backbone of bipartite networks

Received: 3 November 2025

Accepted: 13 June 2026

Published online: 29 June 2026

Check for updates

Lucía S. Ramirez[1,2], Roya Aliakbarisani[1,2], M. Ángeles Serrano[1,2,3] & Marián Boguñá[1,2]

Real bipartite networks combine degree-constrained random mixing with structured connectivity balancing short and long range connections, effectively accounted for by geometric network models. We introduce a statistical filter that benchmarks node-level bipartite clustering against degree-preserving randomizations to classify nodes as geometric (signal) or degree constrained noise. In synthetic mixtures with known ground truth, the filter achieves high classification accuracy and sharpens inference of latent geometric parameters. Applied to four empirical systems –metabolism, online group membership, plant-pollinator interactions, and languages– the filter isolates recurrent neighborhoods while removing ubiquitous or weakly co-occurring entities. Filtering exposes a compact geometric backbone that disproportionately sustains connectivity under percolation and preserves downstream classifier accuracy in node-feature tasks, offering a simple, scalable way to disentangle structure from noise in bipartite networks.

Mathematics provides the language of nature, and network science provides the language of complex systems[1,2]. In complex networks, the architecture of nodes and edges–the topology–offers a window into how the system forms and operates[1]. Yet, in most real-world cases, different connectivity mechanisms overlap within the same network. A portion of connections might emerge through a random process introducing noisy connections (while still respecting certain local properties, such as preserving each node's degree)[3–5], whereas others follow structural principles, such as similarity, hierarchy, or spatial proximity[6–9]. These varied mechanisms, often coexisting, reflect the multiple forces at play in shaping a network's overall structure[10,11].

A food web highlights this coexistence of noisy interactions with structured connectivity patterns. Some species, known as generalists, can feed on a wide variety of prey, suggesting non-specific connections which can be rewired to different prey without greatly affecting the overall structure of the system[12–14]. Other species, the specialists, focus on narrow sets of closely related prey, following a specific ecological logic[12,15,16]. Similarly, in technological systems, such as user-item recommendation networks, certain items might be recommended broadly regardless of the user profile (akin to a generic non-personalized suggestion possibly subject to some general constraint), while others are recommended only to users matching specific criteria, reflecting a more structured approach[17–19].

Disentangling these connectivity types is crucial to understanding how the system's structure arises and how it might evolve. "Noise" is not merely a blurring of signals but can have beneficial or detrimental effects: in ecological scenarios, non-specific connections can bolster stability or buffer change depending on architecture[13,14,16], while in machine learning, noisy edges might obscure meaningful relationships[18,19]. For instance, in a recommendation system, non-personalized recommendations could help surface novel items but might also reduce relevance if not managed properly[17,19]. Recognizing when non-specific connectivity confers adaptability and when it constitutes unhelpful noise is essential to refining both our conceptual understanding and practical applications. One natural strategy to address this problem is to reduce the network to its most informative structural elements while filtering out connections compatible with random or non-specific interactions. Backbone extraction methods pursue precisely this goal by identifying the subset of nodes or edges that carry the most meaningful structural information.

[1]Departament de Física de la Matèria Condensada, Universitat de Barcelona, Barcelona, Spain. [2]Universitat de Barcelona Institute of Complex Systems (UBICS), Universitat de Barcelona, Barcelona, Spain. [3]Institució Catalana de Recerca i Estudis Avançats ICREA, Barcelona, Spain.
e-mail: marian.boguna@ub.edu

Backbone extraction has traditionally focused on reducing weighted or temporal networks by identifying statistically or structurally significant edges. Statistical approaches, such as the Disparity Filter[20], Marginal Likelihood Filter[21], Noise-Corrected Filter[22], and the Enhanced Configuration Model[23] assess the relevance of links by comparing observed weights with appropriate null models that preserve local constraints like degree or strength. Structural methods instead exploit topological features—such as shortest paths, planarity, or salience—to extract reduced subgraphs that retain key structural properties[24–27]. More recently, temporal extensions like the Significant Tie filter introduced activity-based null models to identify irreducible backbones of significant interactions over time[28]. Despite their differences, these approaches are largely edge-centric and are primarily designed for weighted or time-aggregated networks.

In contrast, filtering methods specifically designed for bipartite networks[29] have received comparatively less attention. Existing approaches rely on projecting the bipartite structure onto a one-mode network and then filtering the resulting edges. However, in many bipartite systems, the most relevant question concerns not the significance of individual links but the informativeness of the nodes themselves. This situation arises when one layer represents attributes, features, or items whose contribution to the network's structure is of interest. In such cases, the goal is not necessarily to remove individual links but rather to identify which nodes behave as structurally informative elements and which correspond to random or noisy components. Although this node-level perspective has been much less explored than traditional edge-centric backbone extraction, it is particularly important for understanding the organization of bipartite systems and for practical applications, such as feature selection and noise filtering in machine learning.

In this paper, we introduce a methodology for separating noise-driven connections from those that might reflect specific underlying structural patterns in bipartite networks. Bipartite structures emerge in many fields[1,9]: plant-pollinator networks, where different pollinators visit different plants[30]; metabolic interactions, linking reactions to metabolites[31,32]; scientific collaborations, pairing researchers with their publications[33]; user-item recommendation systems, matching people with products or services[34,35]; and node-feature graphs in machine learning, relating data instances to their attributes[36]. By isolating and quantifying the non-specific component in these systems, our method offers a clearer view of the underlying forces that shape network organization. This, in turn, facilitates more accurate modeling, enabling predictions of how the network might respond to changes, such as the introduction or removal of nodes– and aiding in the design of interventions or optimizations that take both processes into account[29,37–39].

As discussed above, in many real-world bipartite networks, edges emerge through multiple mechanisms that coexist and overlap[3,40,41]. Ultimately, understanding the balance between specific and non-specific connectivity can deepen our appreciation of how complex systems emerge, adapt, and function. By identifying and disentangling these dual mechanisms, we become better equipped to analyze, predict, and influence the behavior of a wide range of bipartite networks, from ecological webs to machine-learning pipelines[10,11].

Results

Here, we consider a bipartite network, where the nodes can be divided into two distinct sets, A and B, and links only connect nodes from different sets. For some nodes of type B, their connections to nodes of type A appear to follow a non-specific process, one that preserves each node's individual degree but otherwise distributes edges as freely as possible –the so-called configuration model (CM)[5,37,38,42,43]. This random mechanism explains a portion of the network in which no clear pattern dictates connections, beyond the degrees of nodes. This type of connectivity is referred to here as noisy, and does not fully capture the network's structural complexity.

The rest of the type B nodes exhibit a specific linkage pattern. That is, their connections to nodes in type A follow principles or constraints beyond the CM, producing topological features and patterns that deviate from a purely random arrangement[29,44,45]. Although we do not specify here what exact processes drive this structured behavior, we do know it must be distinguished from the degree-constrained random baseline if we are to understand the system comprehensively. The challenge, therefore, is to disentangle these two connectivity modes, specific and non-specific, within the same bipartite network. By doing so, we can determine how much of the overall structure is attributable to a non-specific mechanism and identify the specific structural signatures of the connections, illuminating the multiple forces that shape network organization.

To achieve this goal, we first select a topological property capable of distinguishing specific from non-specific or noisy connectivity in heterogeneous networks. The degree of individual nodes alone cannot serve this purpose, as degree is a local property that can be satisfied even in highly random cases. Instead, we use the bipartite clustering coefficient, which reflects the density of "squares" formed by two type-A nodes and two type-B nodes (see Appendix IV B)[9,46,47]. Clustering is crucial here because it often indicates underlying similarity (or dissimilarity), suggesting the presence of a latent similarity space shaping the network's topology[7,8,45]. Importantly, the clustering coefficient can also be measured at the level of individual nodes, allowing the identification of nodes with particularly high or low clustering values.

However, a single node's clustering coefficient is strongly influenced by its degree, making such a selection difficult when the network has heterogeneous degree distributions. To address this, we compare clustering within degree classes[48]. Specifically, we generate randomized versions of the bipartite network by rewiring edges while preserving the overall degree sequence[5,42,43]. Then, for each degree class, we compute the distribution of clustering coefficients across a large number of randomized networks and assess each node's observed clustering relative to this random baseline[29]. If a type B node's clustering coefficient significantly exceeds the random expectation for its degree class, we classify its connections as specific; otherwise, the node is classified as compatible with the null model and non-specific.

Node filtration

Figure 1 sketches the pipeline of our proposal to filter out noisy nodes of a real bipartite network. It is divided into four main components that we describe below.

- Step 0: identify type A and type B nodes. The identification process relies on the specific characteristics of the system. For example, in the machine learning applications discussed in Sec. 1.6, we analyze bipartite networks composed of nodes and their associated features. In this framework, features are treated as type B nodes, as they can represent any attribute assigned to the nodes without carrying information that could be used to infer similarities among them.
- Step 1: compute node-level bipartite clustering in the original network for nodes of type B. Multiple clustering definitions exist for bipartite graphs; here we adopt the measure in the Appendix IV B (3), which is well suited to distinguishing geometric (signal) from random-like (noise) structure. We compute c_i for all nodes in set B and plot clustering versus degree (Step 1 in Fig. 1). In the illustrative synthetic example –a mixture of the CM and the geometric $\mathbb{S}^1$ model[6,49] which generates significant non-vanishing clustering– points labeled as noise (red) lie systematically below geometric points (blue).
- Step 2: rewiring the network. We compare the observed clustering with degree-preserving nulls built by random edge swaps. At each

https://doi.org/10.1038/s41467-026-74723-4

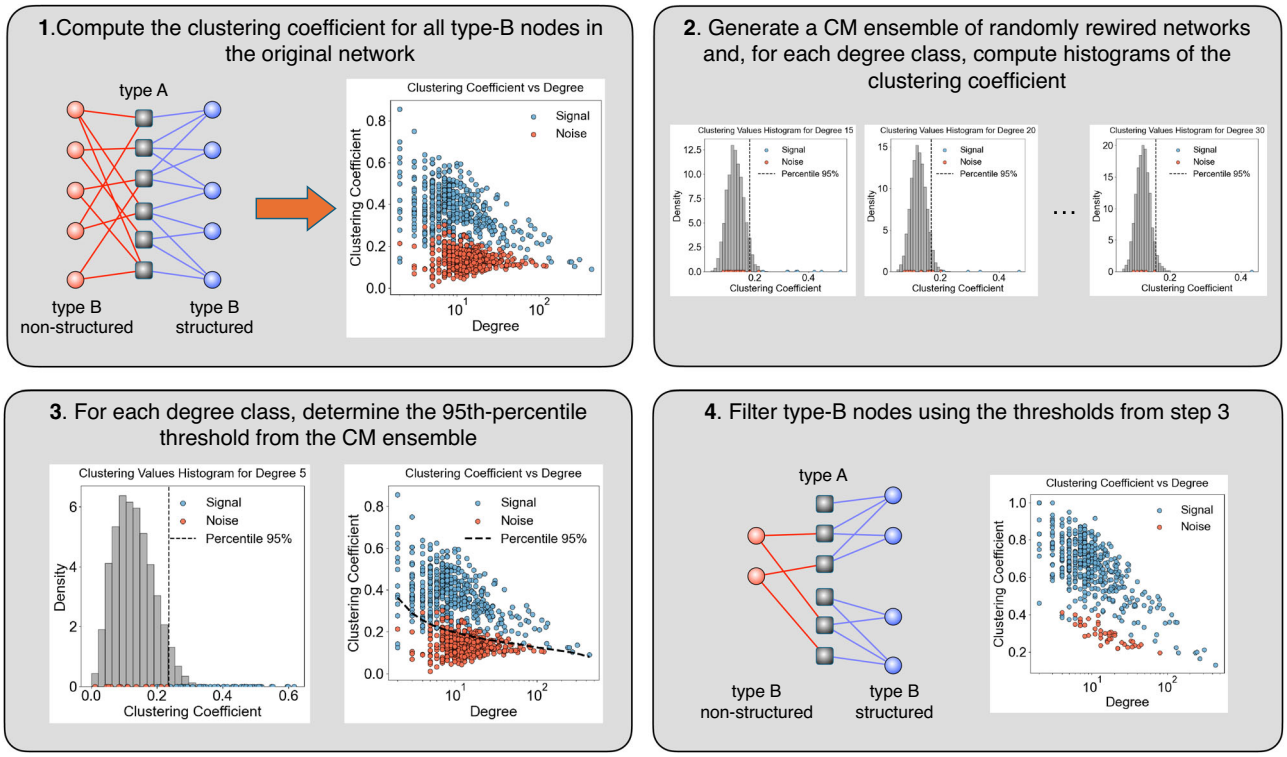

Fig. 1 | Pipeline for filtering noisy nodes in bipartite networks. (1) Compute the clustering coefficient for every type-B node in the original network. (2) Create an ensemble of degree-preserving random networks using the configuration model (CM) and, for each type-B degree class, build the histogram of clustering coefficients under the CM. (3) For each degree class, take the 95th-percentile value of this distribution as the threshold (black dashed line). (4) Filter nodes accordingly: those above the threshold are retained as signal; those below are classified as noise.

step, choose two A–B edges (i, j) and (k, ℓ) and, if no multiedge would form, rewire to (i, ℓ) and (k, j); this preserves both degree sequences and realizes the CM. We perform enough swaps to rewire every edge at least once and generate 1000 CM surrogates. For each surrogate, we measure the clustering of type-B nodes and, across surrogates, compile the degree-conditioned empirical null distributions (histograms) used in Step 3 of Fig. 1.

- Step 3: comparing the original network with the CM. Using the degree-conditioned histograms from the CM surrogates as the null, we set an upper-tail clustering threshold for each degree class at a chosen significance level p (e.g., $p = 0.05$, corresponding to the 95th percentile; Fig. 1). Nodes whose observed clustering exceeds this threshold are labeled signal (geometric in the $\mathbb{S}^1$ model); those below are compatible with the CM and labeled noise. The CM serves as the random-mixing baseline—we make no assumptions about the signal beyond exhibiting clustering in excess of the null. The resulting per-degree thresholds trace a separating curve in the clustering–degree plane (Step 3 in Fig. 1).

Experiments in synthetic bipartite networks

We first test our method in a fully controlled environment where we have access to the network's ground truth. We then define an ensemble of synthetic bipartite graphs by combining two well-known random graph ensembles: the soft version of the bipartite configuration model (SCM)[37] and the bipartite-$\mathbb{S}^1$ model[45,50]. The SCM is the "canonical ensemble" version of the CM, where the degree sequence is fixed on average—each node is assigned an expected degree κ rather than its actual degree. Hence, the SCM is fully characterized by the hidden degree distributions for both node types, $\rho_A(\kappa_A)$ and $\rho_B(\kappa_B)$.

The bipartite-$\mathbb{S}^1$ model[45,50] is the bipartite counterpart of the $\mathbb{S}^1$ model[6,49]. In this model, nodes are assigned expected degrees according to given distributions and lie on a metric space (a D-sphere with $D = 1$ in this case). The probability of a connection between a type A node and a type B node is determined by their distance in this space, rescaled by the product of their expected degrees. An additional parameter, β, acts as the inverse temperature of the ensemble: large β values yield networks that are closely aligned with the underlying metric space, whereas smaller values increase randomness and recover the SCM in the limit $\beta \to 0$. Consequently, β controls the level of clustering. In particular, the model exhibits a topological phase transition at $\beta = 1$ in the one-dimensional case considered here, or more generally at $\beta = D$ in a D-dimensional similarity space. Below this threshold, clustering vanishes in the thermodynamic limit; above it, it converges to a constant, reaching its maximum as $\beta \to \infty$. Both models are described in detail in the Appendix IV A.

The mixture ensemble is defined as follows: Type B nodes are split into two groups, B^{CM} and $B^{\mathbb{S}^1}$, with sizes N_B^{CM} and $N_B^{\mathbb{S}^1}$, respectively. Group B^{CM} is assigned expected degrees drawn from $\rho_B^{CM}(\kappa_B)$, and its nodes connect to Type A nodes using the SCM connection probability given by (2). Similarly, group $B^{\mathbb{S}^1}$ is assigned expected degrees drawn from $\rho_B^{\mathbb{S}^1}(\kappa_B)$, and its nodes connect to Type A nodes according to the bipartite-$\mathbb{S}^1$ model given by (1). Type A nodes are given an expected degree sequence drawn from $\rho_A(\kappa_A)$. In our simulations, we used power-law distributions for the expected degrees of the form $\rho(\kappa) = (\gamma - 1)\kappa_0^{\gamma-1}\kappa^{-\gamma}$ for $\kappa \geq \kappa_0$, with mean degree $\langle k \rangle = (\gamma - 1)\kappa_0/(\gamma - 2)$.

Figure 2 summarizes noise-detection performance on synthetic bipartite networks across parameter settings and significance levels p. In all experiments we fix $N_A = 1000$, $\langle k_A \rangle = 16$, and $\gamma_A = 5$, yielding a comparatively homogeneous degree distribution for Type A nodes, as often observed in empirical bipartite systems. For Type B we set $N_B^{CM} = 500$ and $N_B^{\mathbb{S}^1} = 500$. The values of γ_B^{CM}, $\gamma_B^{\mathbb{S}^1}$, and the geometric parameter β used for Type B are indicated in the figure.

https://doi.org/10.1038/s41467-026-74723-4

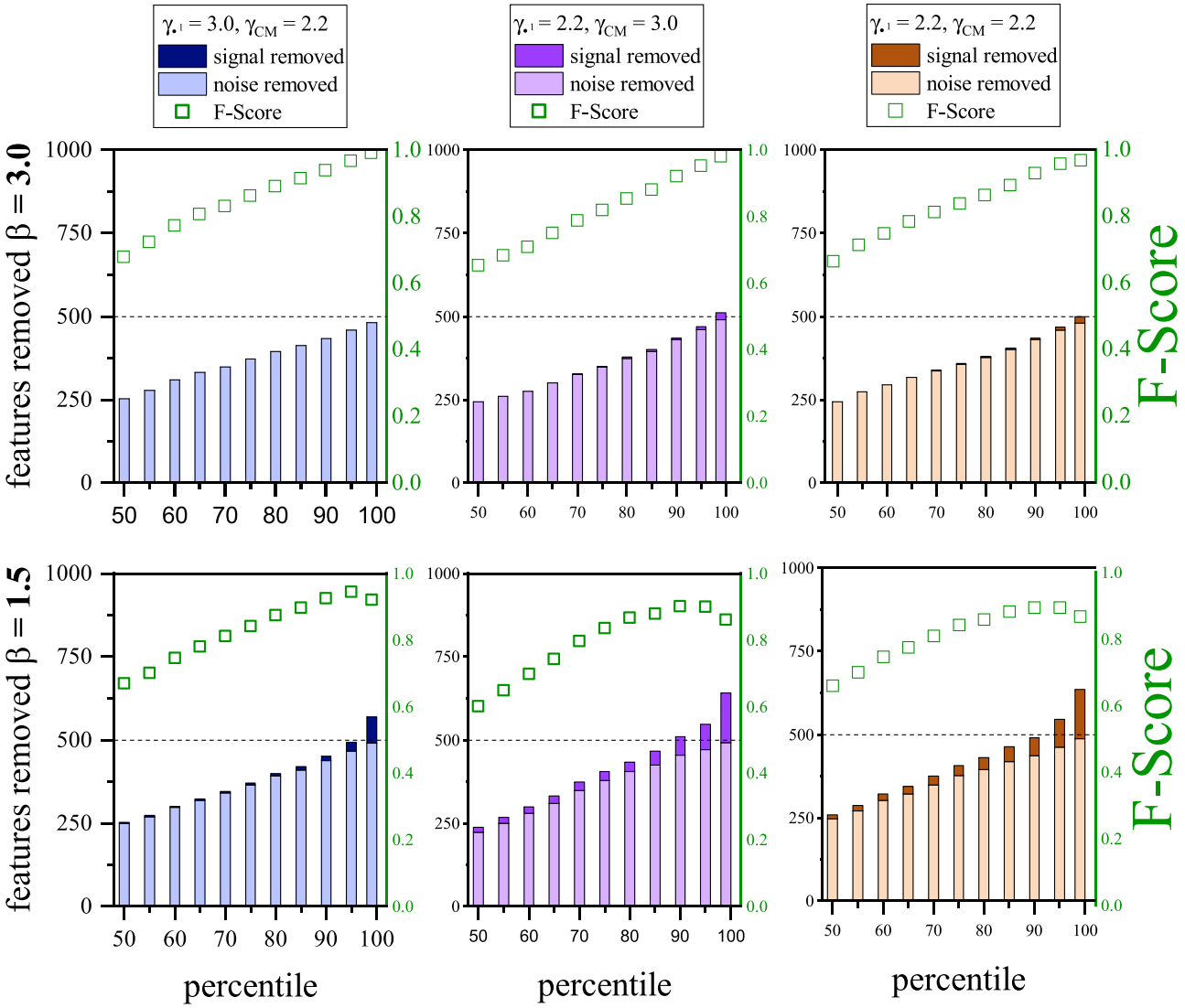

Fig. 2 | Performance of noise detection in synthetic bipartite networks. The panels report the number of nodes classified as noise across different percentiles --corresponding to different significance levels p. Light bars denote true positives --noisy nodes correctly identified--whereas dark bars denote false positives --signal nodes misclassified as noise. Each row corresponds to a different value of β (indicated in the figure), and each column represents combinations of $\gamma_{\mathbb{S}^1}$ and γ_{CM} for the degree distributions of sets $B^{\mathbb{S}^1}$ and B^{CM}. Green squares show the F-Score at each percentile, capturing the precision-recall trade-off. For percentiles above 0.85%, F-Score values are typically $\gtrsim 0.9$, demonstrating the method's robustness.

The bars in Fig. 2 report the number of Type B nodes labeled as noise. Light bars correspond to true positives (noisy nodes from B^{CM} correctly identified), whereas dark bars correspond to false positives (signal nodes from $B^{\mathbb{S}^1}$ misclassified as noise), shown for different significance levels. Increasing p generally increases sensitivity of the filter, leading to higher true-positive counts; this effect is especially pronounced for the 85% percentile, where the method correctly retrieves more than 80% of noisy nodes even under challenging conditions (low β and highly heterogeneous Type B degree distributions). Pushing percentiles closer to 100% yields detection of nearly all noisy nodes but at the cost of a higher false-positive rate.

To quantify this trade-off, we report the F-Score, a standard metric for binary classification that balances precision and recall[51,52]. It is defined in the range (0, 1), with values close to 1 indicating better balance between identifying noise while avoiding false positives (see Appendix IV C). The green squares in Fig. 2 show the F-Score for each configuration. The filter performs robustly, with $F \gtrsim 0.9$ for percentiles above 85% at $\beta = 1.5$. For $\beta = 3$, the F-Score approaches 1 as the percentile approaches 100%. Additional sweeps over ensemble parameters and over different mixtures of B^{CM} and $B^{\mathbb{S}^1}$ yield consistent trends; see the Supplementary Information for details.

It is worth noting that the filter performs particularly well in identifying and correctly classifying high-degree nodes, which provide richer local connectivity patterns and therefore clearer clustering signals. In contrast, nodes with very small degree can be more difficult to classify reliably, as their limited number of connections provides fewer opportunities to form local cycles. As a result, low-degree signal nodes may occasionally fall below the clustering threshold and be misclassified as noise. This effect reflects a natural resolution limit associated with degree-dependent information in the network; see the Supplementary Information for details.

We also analyze the scaling of the noise-detection performance with system size and increasing mean degree. As expected, increasing the system size improves performance, since the clustering coefficient of the CM goes to zero in the thermodynamic limit. Similarly, in the sparse-network regime of interest, performance increases as the mean degree grows. In both cases, the F-score increases. Further details are provided in the Supplementary Information.

Table 1 | Summary of the real networks analyzed, including the number of nodes in Set A and Set B, and the sizes of the filtered subsets for different thresholds (p85, p95, and p99)

Network	Set A	$\langle k \rangle_A$	Set B	$\langle k \rangle_B$	Av.Clustering Set B	Filtered p85	Filtered p95	Filtered p99
Human Metabolic	2212	4.8	1214	8.6	0.3822	508 (41.5%)	648 (52.77%)	849 (69.36%)
YouTube group membership	37981	3.9	12009	12.3	0.1723	5366 (44.68%)	7126 (59.34%)	8900 (74.11%)
Plant Pollinator	456	33.5	1044	14.6	0.1399	288 (33.26%)	364 (42.03%)	467 (53.43%)
World Languages	246	3.8	166	5.6	0.3987	79 (47.59%)	123 (74.1%)	159 (95.78%)
CORA	2708	18.18	1431	34.41	0.0987	587 (38.25%)	754 (52.73%)	948 (66.29%)
Citeseer	3312	31.75	3702	28.4	0.1184	608 (16.42%)	868 (23.44%)	1121 (30.27%)
BlogCatalog	5196	71.1	8181	45.11	0.2606	2877(35.16%)	3969(48.51%)	5004(61.16%)

As a final analysis of these synthetic experiments, we compare the parameter β –as estimated by the B-Mercator embedding tool[53]– for the original network, treating it as a realization of the bipartite-$\mathbb{S}^1$ model, with the corresponding estimate for the filtered network. When the full network (including both geometric and noisy nodes) is considered, the inferred parameter β_i is systematically lower than the ground-truth value β used to generate the connections of nodes in the $B^{\mathbb{S}^1}$ set. After filtering, the inferred β_i shifts closer to the true value, despite the presence of residual noise. This behavior is illustrated in Supplementary Fig. S10, where we report the ratio β_i/β across different network configurations.

Filtering real networks

We analyze four real bipartite systems: (i) the human metabolic network from the BiGG database[54], linking reactions and metabolites in human cells; (ii) a YouTube group-membership network[55], linking users and groups; (iii) a plant-pollinator network[56], linking plants to visiting insect species; and (iv) a world-language network, linking regions to their spoken languages[57]. Table 1 summarizes basic properties after removing degree-one nodes (clustering coefficients are undefined for such nodes); the reported sizes refer to the post-processing networks. Further details about the datasets are provided in the Supplementary Information.

Figure 3 plots, for each dataset, the bipartite clustering coefficient of all type-B nodes versus degree, together with the 85th, 95th, and 99th percentiles from degree-preserving randomizations used as filtering thresholds (Section 1.1), where type-B nodes are, respectively, metabolites, groups, pollinators, and languages. Across systems –metabolism, online group membership, plant-pollinator interactions, and world languages– the same pattern emerges: many low-degree type-B nodes fall below the thresholds and are removed, consistent with near-non-specific attachment, whereas nodes that participate in recurrent neighborhoods rise above the thresholds and are retained. Table 1 reports the fraction of surviving type-B nodes at each percentile and shows the expected monotonic decrease with stricter thresholds.

The qualitative behavior is consistent across domains yet reveals domain-specific structure. In biochemical data, some ubiquitous components with very high degree (e.g., H_2O, coenzyme A) are filtered out because their broad participation lacks recurrent co-occurrence, while other high-degree metabolites (e.g., ATP/ADP, NAD/NADP) remain because they concentrate within tightly reused reaction neighborhoods. In social affiliation, many small groups behave in a non-specific way and are removed, whereas large, structured communities persist due to strong co-membership clustering. In ecological interactions, filtering pollinators reveals a clear pattern: taxa with more structured visitation patterns, such as bees, are preferentially retained in the backbone (180 out of 299), whereas more opportunistic visitors, such as flies, are less frequently part of the backbone (140 out of 398). This difference is consistent with the tendency of bees to exhibit more consistent co-visitation patterns, as they are typically more specialized pollinators, whereas many flies interact more diffusely across plant communities[58–60]. As a result, bees contribute more strongly to the structured backbone, whereas flies more often generate weaker or less recurrent links. In linguistic data, widely dispersed languages are often filtered because cross-territorial ubiquity does not imply local co-use, but languages embedded in multilingual territories survive the filter due to elevated local clustering.

Together, these results show that the filter does more than suppress noise: it isolates geometrical signal, recurrent, locality-like structure, while flagging non-geometrical components whose connections resemble random mixing. This integrated view holds across disparate bipartite systems, supporting the method's utility as a general tool for separating structure from noise.

Impact of the choice of filtered node set

The filtering procedure requires selecting the node set that contains both structured and non-structured nodes, which is then taken as the set to be filtered (set B). While in many applications this choice arises naturally from the interpretation of the data, it is instructive to analyze the outcome of applying the filter after interchanging the roles of sets A and B.

To this end, we apply the filtering procedure to both layers of the plant–pollinator network (Fig. 4). The figure shows the clustering coefficient as a function of degree, together with the upper and lower bounds obtained from the degree-preserving randomization. The upper and lower bounds delimit the region where the clustering of randomly connected nodes is expected to lie, containing 95% of the values under the random baseline.

When filtering pollinators, the lower bound closely follows the data across the degree range, effectively delimiting the region where nodes compatible with the random baseline are expected. After swapping the node sets and filtering plants, the lower bound lies systematically below the bulk of the data, and clustering values appear to be shifted upwards with respect to this region. This shift suggests that nodes in that set inherit a mixture of noisy and geometrically structured connections from the opposite layer. The geometric links increase the clustering that a purely random node could have.

This difference indicates that pollinators constitute the appropriate choice for the filtered set, while plants behave as the complementary layer. A more complete description, together with experiments on swapping sets A and B in the synthetic benchmark, is provided in the Supplementary Information (Sec. V).

Implications for network vulnerability

The two connectivity modes have clear implications for global structure. To assess structural robustness, we performed a percolation analysis that tracks the size of the giant connected component as we remove type-B nodes classified as geometric or as noise. After applying the filtering procedure, we partitioned type-B nodes into two sets: the *geometric* set (nodes retained by the filter) and the *noise* set (nodes

https://doi.org/10.1038/s41467-026-74723-4

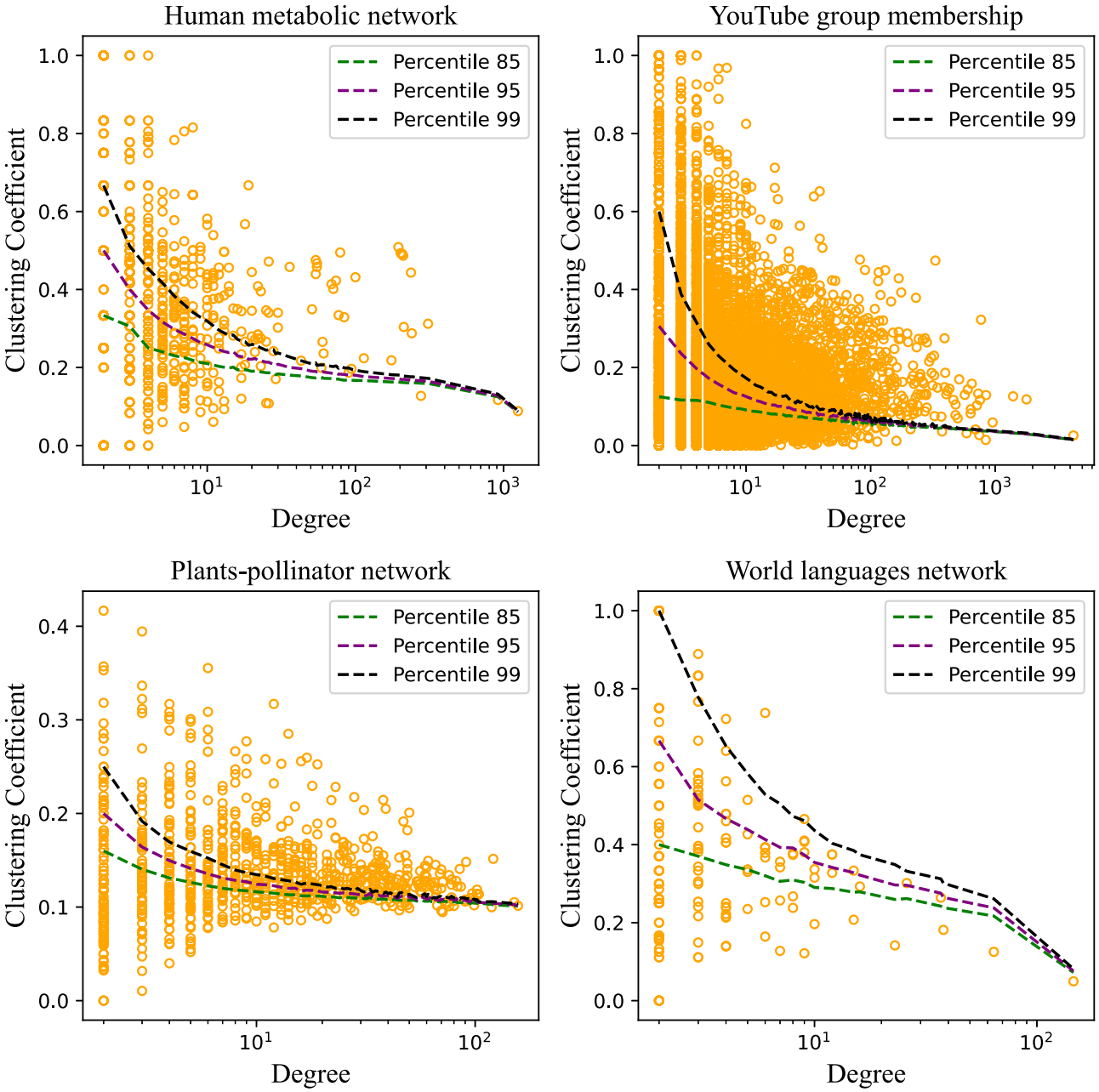

Fig. 3 | Clustering coefficient versus degree across datasets. Human metabolic network[54,66], YouTube membership network [55], plant-pollinator[56,67], and world-language networks[57]. Orange circles show clustering for type-*B* nodes; dashed lines mark the 85th (green), 95th (purple), and 99th (black) percentiles from degree-preserving randomizations. Percentiles serve as thresholds following Section 1.1: points above are classified as geometrical (signal), and points below as non-geometrical (noise).

removed by the filter). We then evaluated the resilience of the original bipartite network under targeted removals from each set and under random removals.

As expected, removing geometric nodes produces a larger decline in the size of the giant component: once these structurally organized features are deleted, the remaining network is less coherently clustered. By contrast, removing the same number of noisy nodes has a weaker effect: the surviving network, now relatively more geometric, better preserves large-scale connectivity. Figure 5 illustrates this behavior for the human metabolic and YouTube group-membership networks. Each curve reports the effect of removing nodes from the geometric, noise, or random sets. For comparability, the number of removed nodes in all cases equals the size of the geometric set at the 95% percentile.

Application to neural networks: the impact of filtration on node classification tasks

Modern neural networks are increasingly fed with graph-structured datasets in which vertices carry feature vectors and labels. Improving the signal in these features can directly affect downstream classification.

Given a graph-structured dataset with N_n nodes and N_f features associated with the same set of nodes, we construct a bipartite network between nodes and features by treating features as entities, following[61]. Each node i has a binary feature vector $\vec{f}_i \in \{0,1\}^{N_f}$, where each entry indicates the presence (1) or absence (0) of a feature. A node is therefore linked to all features for which $\vec{f}_i = 1$. We denote nodes by set A and features by set B. Our goal is to apply the proposed filter to set B to remove noisy features (A comparison of other filtering

https://doi.org/10.1038/s41467-026-74723-4

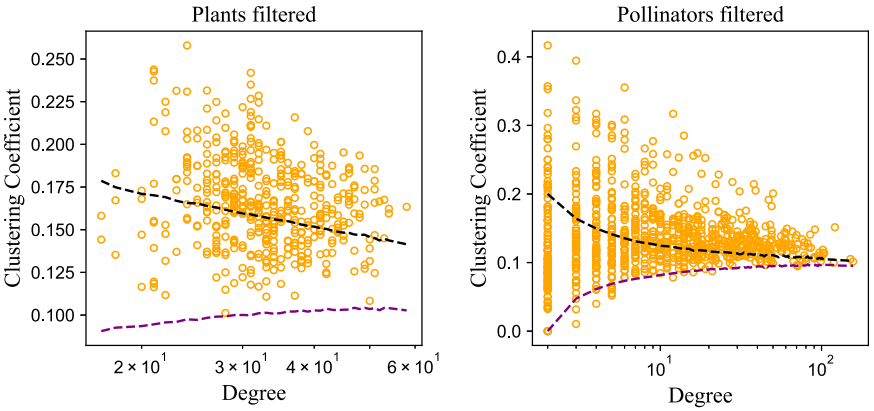

Fig. 4 | Clustering coefficient vs degree for the plant–pollinator network under the two assignments of node sets. Left: plants filtered. Right: pollinators filtered. Dashed lines show the upper and lower bounds from the degree-preserving randomization, delimiting the region where clustering values are expected under the random baseline for the 95th percentile. When filtering pollinators, the bounds closely follow the data, whereas after swapping, clustering values are shifted upwards with respect to this region.

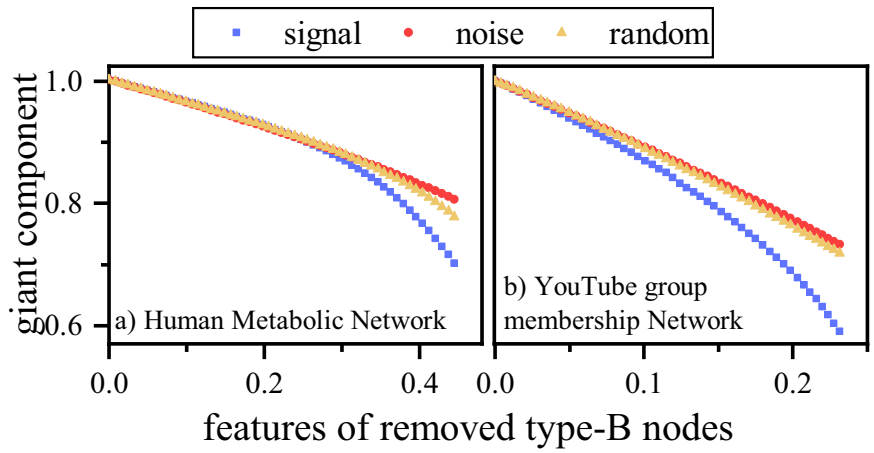

Fig. 5 | Percolation experiments. Evolution of the giant connected component as type-*B* nodes are removed in (**a**) the human metabolic network and **b** the YouTube group-membership network. Each curve corresponds to removals from the signal, noise, or random sets. For comparability, the number of removed nodes equals the size of the signal set at percentile 95%). Results are averaged over 1000 realizations.

techniques can be found in the Supplementary Information). As a baseline, we also create a randomly filtered version of the data by removing at random the same number of features as in the filtration. We then feed the original, filtered, and randomly filtered feature matrices to two feature-based neural architectures, a multilayer perceptron (MLP) and hyperbolic neural networks (HNN)[62], to perform node classification.

Figure 6 reports node-classification accuracy (fraction of correctly classified nodes) on three real datasets, the Cora[63], Citeseer[64], and BlogCatalog[65] datasets, for the two neural architectures. In all cases, as the filtration threshold p increases and features identified as noise are removed, accuracy remains nearly unchanged and close to that of the unfiltered data. In contrast, randomly filtering features produces a pronounced drop in accuracy, especially under large random removals. The Cora dataset is illustrative: node labels are highly correlated with geometric features and removing 80% of the non-specific features leaves accuracy essentially unchanged relative to the original network, whereas removing the same number of features chosen at random reduces accuracy to below 50%.

These results indicate that our filtration effectively extracts the structural backbone of the node-feature bipartite network while discarding features that do not contribute to predictive performance. Taken together, these results suggest that one can retain a small, relevant subset of features without sacrificing accuracy, thereby substantially reducing the computational cost of neural-network training.

Discussion

We introduced a statistically grounded filter that disentangles specific connectivity with geometrical structure from random mixing noise in bipartite networks by benchmarking node-level clustering against degree-preserving randomizations. Backbone extraction methods provide a natural framework for this problem, since they formalize the idea that only part of the observed connectivity reflects statistically meaningful organization. Here, we extend this perspective to unweighted bipartite networks in which the relevant distinction is not the significance of individual links, but the structural role of nodes in one layer. The resulting backbone is defined by nodes whose bipartite clustering cannot be explained by degree distribution alone. Notice that all the methodology is topology-based without assuming any specific generative model for the signal. Across synthetic and real systems, the same picture emerges: a relatively small geometric backbone concentrates recurrent, locality-like neighborhoods, while a sizable non-specific component accounts for diffuse mixing. Rather than treating this latter part as nuisance, our results show that it carries

https://doi.org/10.1038/s41467-026-74723-4

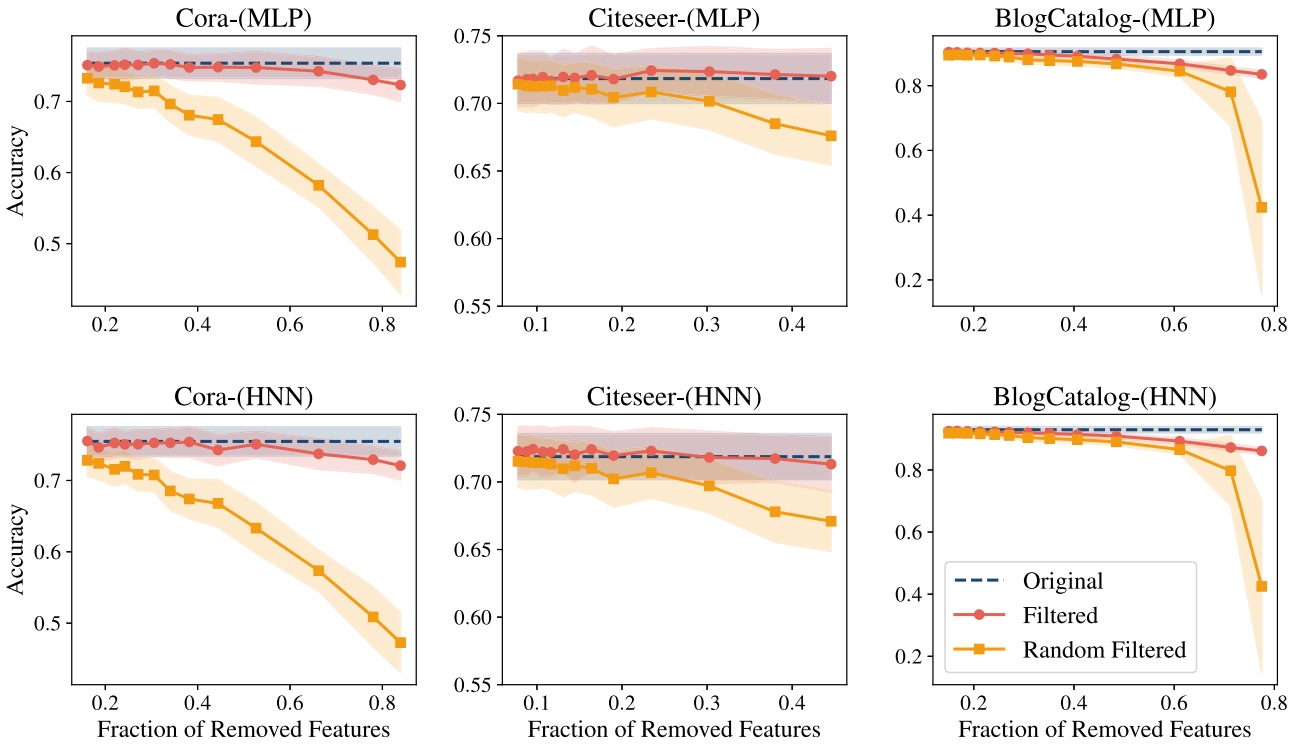

Fig. 6 | Node-classification accuracy as a function of the fraction of removed features for two feature-based neural networks (MLP and HNN). Models are trained on three versions of the node-feature bipartite data: the original features, a filtered set obtained with our node-filtration method, and a randomly filtered set in which the same number of features is removed at random. Error bars show $\pm 1\sigma$ around the mean over 100 random train/validation/test splits; for randomly filtered data they also incorporate variability across ten independent random filtrations. In all experiments, nodes are split into training (70%), validation (15%), and test (15%) sets. Models are optimized with Adam[68], use two layers with hidden dimension 16, and are trained with dropout = 0.2 and weight decay = 0.001. Class labels are assigned via a softmax over the final layer.

domain-specific meaning (e.g., broad pollinator mixing) and should be interpreted in context.

Validation on controlled mixtures (SCM + bipartite-$\mathbb{S}^1$ model) shows high detection quality for practical thresholds (F-scores typically $\gtrsim 0.9$ for percentiles above 85%) and a marked improvement in geometric parameter inference after filtering, indicating that the method sharpens latent-space structure even without assuming a specific generative model for the signal. On real networks –including metabolism, online affiliation, plant-pollinator interactions, and languages– the filter consistently retains entities embedded in recurrent neighborhoods, while broadly ubiquitous or weakly co-occurring entities fall below the randomization threshold. Together, these findings indicate that the approach extracts a compact geometric backbone that better captures the organizing principles of the data, while leaving a non-specific layer of connections whose role depends on the domain (stabilizing diversity vs. adding noise).

The plant–pollinator example also illustrates how the outcome of the filter should be interpreted at the level of network modeling. The method does not impose a symmetric decomposition of the two layers; rather, it identifies the layer in which the observed connectivity is best described as a mixture of recurrent, geometrically organized neighborhoods and degree-constrained random mixing. In this system, the pollinator layer provides the clearer decomposition: pollinators retained in the backbone tend to participate in more recurrent plant neighborhoods, whereas filtered pollinators are more compatible with diffuse or opportunistic visitation patterns. When the roles of the two layers are interchanged, plants do not show the same separation, suggesting that their clustering partly reflects the structured organization of the pollinator layer rather than an independent mixture of plant types. This asymmetry is informative for modeling: a realistic null or generative model should not necessarily treat both layers equivalently, but may instead allow one layer to contain heterogeneous interaction modes while the other acts as the substrate on which these modes are expressed. More generally, such layer asymmetries can guide which node set should be modeled as mechanistically heterogeneous, which nodes should be retained for downstream prediction or robustness analysis, and which part of the network is better interpreted as broad mixing rather than organized structure.

The separation has concrete structural and functional implications. Percolation experiments reveal that targeted removal of geometric nodes disproportionately fragments connectivity relative to removing the same number of noisy nodes, consistent with the backbone's role in holding the system together. In node-feature graphs for machine learning, discarding features flagged as noise preserves classification accuracy, whereas removing an equal number of features at random degrades performance. Thus, the filter reduces dimensionality and computational cost while keeping predictive content intact, positioning it as a complementary pre-processing tool for graph-based pipelines.

It is worth noting that the proposed filtering approach is designed for the analysis of sparse bipartite networks, which constitute the vast majority of empirical systems. In this regime, local cycles remain rare under the degree-preserving CM and clustering therefore provides a meaningful signal of non-random structural organization. The results indicate that, within the sparse regime, increasing the mean degree does not compromise the discriminative power of the method. In very dense networks, where local cycles may appear frequently even under the null model, the discriminative power of clustering-based diagnostics may naturally decrease.

Several limitations suggest directions for follow-up work. First, our discrimination relies on clustering conditioned on degree and on

the CM as a null model for noise; networks with strong constraints beyond degree may require adapted null models. Second, clustering is a sufficient but not exclusive signal of geometry; integrating additional local or mesoscopic statistics could refine the boundary between signal and noise. In particular, when clustering reflects an underlying smooth latent geometry, the transitivity backbone will correspond to the geometric backbone of the bipartite network. Finally, many systems are weighted, directed, temporal, or multilayer; extending the filter to these settings –and to both node sets simultaneously– should broaden applicability.

In sum, the method offers a simple, versatile, and scalable route to parse mixed mechanisms in bipartite graphs. By isolating a geometric backbone and characterizing the complementary random-like layer, it clarifies which components drive organization and which provide flexible mixing. This separation improves latent-space inference, strengthens robustness analyses, and streamlines downstream learning, bridging descriptive network analysis and applied modeling across ecological, biological, technological, and social systems. A natural direction for future work is to extend this framework to complex percolation and cascade dynamics, where the distinction between structured and noisy nodes in bipartite networks could prove important for understanding the onset, propagation, and containment of systemic failures.

Methods

Synthetic bipartite benchmarks

We generate bipartite graphs with node sets A and B in which only a subset of B is contaminated by noise. Set B is a mixture of geometric nodes $B^{\mathbb{S}^1}$ and random nodes B^{CM}, so that $N_B = N_B^{\mathbb{S}^1} + N_B^{\mathrm{CM}}$. Nodes in A and in $B^{\mathbb{S}^1}$ carry hidden variables (κ, θ), where κ controls expected degree and $\theta \in [0, 2\pi)$ is an angular similarity coordinate.

$\mathbb{S}^1$ (geometric) links. Edges between A and $B^{\mathbb{S}^1}$ are drawn independently with probability

$$p_{\mathbb{S}^1}(\kappa_A, \kappa_B, \Delta\theta) = \left[1 + \left(\frac{R\Delta\theta}{\mu_{\mathbb{S}^1}\kappa_A\kappa_B}\right)^{\beta}\right]^{-1}, \tag{1}$$

where $\Delta\theta$ is the angular distance, $\beta > 1$ controls geometric sharpness, and $\mu_{\mathbb{S}^1}$ sets the mean degrees. Details of both models are in the Supplementary Information.

Configuration-model (random) links. Edges between A and B^{CM} follow a soft CM,

$$p_{\mathrm{CM}}(\kappa_A, \kappa_B) = \frac{\mu_{\mathrm{CM}}\kappa_A\kappa_B}{1 + \mu_{\mathrm{CM}}\kappa_A\kappa_B}, \tag{2}$$

with μ_{CM} fixed to match the desired mean degrees in A and B^{CM}.

Degree distributions. Unless otherwise noted, hidden degrees are sampled from power laws $\rho(\kappa) \propto \kappa^{-\gamma}$ with lower cutoffs chosen so that the target mean degrees exist. Exponents for A, $B^{\mathbb{S}^1}$, and B^{CM} may differ.

Bipartite clustering coefficient

We measure local geometric organization with a bipartite clustering coefficient for node i,

$$C_3(i) = \frac{2\,\tilde{T}_i}{k_i(k_i - 1)}, \qquad \tilde{T}_i = \sum_{\alpha \neq \beta} \frac{T_{\alpha\beta}(i)}{\min(k_\alpha, k_\beta) - 1}. \tag{3}$$

Here k_i is the degree of i; $T_{\alpha\beta}(i)$ counts the number of common neighbors of the neighbor pair (α, β) in the opposite set; and the denominator normalizes by the maximum number of possible common neighbors for that pair. This definition weights pairs more heavily when they are closed through many shared neighbors.

Evaluation metric

We quantify detection performance with the F-score, the harmonic mean of precision and recall. Let TP, FP, and FN denote true positives, false positives, and false negatives, respectively:

$$F = \frac{2TP}{2TP + FP + FN}, \tag{4}$$

with $0 \leq F \leq 1$.

Data availability

Data from the real systems analyzed in this study are available from the cited references. Synthetic datasets can be generated using the code from the GitHub repository https://github.com/lucia35515/Transitivity-Filter-for-bipartite-networks.

Code availability

The code used to generate the results in this study can be downloaded at https://github.com/lucia35515/Transitivity-Filter-for-bipartite-networks.

References

1. Newman, M. E. J. Networks: An Introduction. (Oxford University Press, 2010).
2. Barabási, A.-L. Network Science. (Cambridge University Press, 2016).
3. 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).
4. Park, J. & Newman, M. E. J. Statistical mechanics of networks. *Phys. Rev. E* **70**, 066117 (2004).
5. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
6. Serrano, M. A., Krioukov, D. & Boguñá, M. Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett.* **100**, 078701 (2008).
7. Papadopoulos, F., Kitsak, M., Serrano, M. A., Boguñá, M. & Krioukov, D. Popularity versus similarity in growing networks. *Nature* **489**, 537–540 (2012).
8. Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A. & Boguñá, M. Hyperbolic geometry of complex networks. *Phys. Rev. E* **82**, 036106 (2010).
9. Latapy, M., Magnien, C. & Vecchio, N. D. Basic notions for the analysis of large two-mode networks. *Soc. Netw.* **30**, 31–48 (2008).
10. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107 (2011).
11. Larremore, D. B., Clauset, A. & Jacobs, A. Z. Efficiently inferring community structure in bipartite networks. *Phys. Rev. E* **90**, 012805 (2014).
12. Williams, R. J. & Martinez, N. D. Simple rules yield complex food webs. *Nature* **404**, 180–183 (2000).
13. May, R. M. Will a large complex system be stable? *Nature* **238**, 413–414 (1972).
14. Allesina, S. & Tang, S. Stability criteria for complex ecosystems. *Nature* **483**, 205–208 (2012).
15. Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E. & Taylor, W. W. Compartments revealed in food-web structure. *Nature* **426**, 282–285 (2003).
16. Thébault, E. & Fontaine, C. Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science* **329**, 853–856 (2010).

17. Gomez-Uribe, C. A. & Hunt, N. The Netflix recommender system: algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.* **6**, 1–19 (2015).
18. Ziegler, C., McNee, S. M., Konstan, J. A. & Lausen, G. Improving recommendation lists through topic diversification. In: *Proc. 14th International World Wide Web Conference (WWW)*, 22-32 (WWW, 2005).
19. Kaminskas, M. & Bridge, D. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Syst.* **7**, 1–42 (2016).
20. Serrano, M. A., Boguñá, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **106**, 6483–6488 (2009).
21. Dianati, N. Unwinding the hairball graph: pruning algorithms for weighted complex networks. *Phys. Rev. E* **93**, 012304 (2016).
22. Coscia, M. & Neffke, F. Network backboning with noisy data. In: *Proc. IEEE International Conference on Data Mining (ICDM)*, 425–434 (ICDM, 2017).
23. Gemmetto, V., Cardillo, A. & Garlaschelli, D. Irreducible network backbones: unbiased graph filtering via maximum entropy. *arXiv preprint arXiv:1706.00230* (2017).
24. Grady, D., Thiemann, C. & Brockmann, D. Robust classification of salient links in complex networks. *Nat. Commun.* **3**, 864 (2012).
25. Simas, T., Correia, R. B. & Rocha, L. M. The distance backbone of complex networks. *J. Complex Netw.* **9**, cnab021 (2021).
26. 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).
27. Zhang, X., Zhang, Z., Zhao, H., Wang, Q. & Zhu, J. Extracting the globally and locally adaptive backbone of complex networks. *PLOS ONE* **9**, e100428 (2014).
28. Kobayashi, T., Takaguchi, T. & Barrat, A. The structured backbone of temporal social ties. *Nat. Commun.* **10**, 220 (2019).
29. Tumminello, M., Miccichè, S., Lillo, F., Piilo, J. & Mantegna, R. N. Statistically validated networks in bipartite complex systems. *PLOS ONE* **6**, e17994 (2011).
30. Bascompte, J. & Jordano, P. Plant-animal mutualistic networks: The architecture of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **38**, 567–593 (2007).
31. Fell, D. A. & Wagner, A. The small world of metabolism. *Nat. Biotechnol.* **18**, 1121–1122 (2000).
32. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
33. Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98**, 404–409 (2001).
34. Koren, Y., Bell, R. & Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **42**, 30–37 (2009).
35. Ricci, F., Rokach, L., Shapira, B. & Kantor, P.B. (eds.) Recommender Systems Handbook. (Springer, 2011).
36. Dhillon, I.S. Co-clustering documents and words using bipartite spectral graph partitioning. In: *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269–274 (SIGKDD, 2001).
37. Saracco, F., Di Clemente, R., Gabrielli, A. & Squartini, T. Randomizing bipartite networks: the case of the world trade web. *Sci. Rep.* **5**, 10595 (2015).
38. Saracco, F., Di Clemente, R., Gabrielli, A., Caldarelli, G. & Squartini, T. Inferring monopartite projections of bipartite networks: an entropy-based approach. *N. J. Phys.* **19**, 053022 (2017).
39. Neal, Z. P. Identifying statistically significant edges in one-mode projections. *Soc. Netw. Anal. Min.* **3**, 915–924 (2013).
40. Molloy, M. & Reed, B. A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms* **6**, 161–180 (1995).
41. Molloy, M. & Reed, B. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probab. Comput.* **7**, 295–305 (1998).
42. Roberts, E. S. & Coolen, A. C. C. Unbiased degree-preserving randomization of directed binary networks. *Phys. Rev. E* **85**, 046103 (2012).
43. 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).
44. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. Network motifs: simple building blocks of complex networks. *Proc. Natl. Acad. Sci. USA* **99**, 12974–12979 (2002).
45. Kitsak, M., Papadopoulos, F. & Krioukov, D. Latent geometry of bipartite networks. *Phys. Rev. E* **95**, 032309 (2017).
46. Zhang, P., Wang, J., Li, X., Di, Z. & Fan, Y. Clustering coefficient and community structure of bipartite networks. *Phys. A:Stat. Mech. its Appl.* **387**, 6869–6875 (2008).
47. Opsahl, T. Triadic closure in two-mode networks: redefining the global and local clustering coefficients. *Soc. Netw.* **35**, 159–167 (2013).
48. Soffer, S. N. & Vázquez, A. Network clustering coefficient without degree-correlation biases. *Phys. Rev. E* **71**, 057101 (2005).
49. Serrano, M.Á. & Boguñá, M. The shortest path to network geometry: a practical guide to basic models and applications. (Cambridge University Press, 2022).
50. Serrano, M. Á, Boguñá, M. & Sagués, F. Uncovering the hidden geometry behind metabolic networks. *Mol. Biosyst.* **8**, 843–850 (2012).
51. van Rijsbergen, C.J. Information Retrieval. (Butterworth-Heinemann, 1979).
52. Manning, C.D., Raghavan, P. & Schütze, H. Introduction to Information Retrieval. (Cambridge University Press, 2008).
53. Jankowski, R., Aliakbarisani, R., Serrano, M. Á. et al. Mapping bipartite networks into multidimensional hyperbolic spaces. *Commun Phys* **9**, 35 (2026).
54. Schellenberger, J., Park, J., Conrad, T. M. & Palsson, B. Ø BiGG: a biochemical genetic and genomic knowledgebase of large-scale metabolic network reconstructions. *BMC Bioinforma.* **11**, 213 (2010).
55. Rossi, R.A. & Ahmed, N.K. The network data repository with interactive graph analytics and visualization. (AAAI, 2015).
56. Robertson, C. Flowers and insects: lists of visitors to four hundred and fifty-three flowers. (NOAA, 1929).
57. Unicode Consortium. Unicode Common Locale Data Repository (CLDR). (GitHub repository, 2024).
58. Michener, C.D. The Bees of the World (2nd ed.). (Johns Hopkins University Press, 2007).
59. Waser, N. M., Chittka, L., Price, M. V., Williams, N. M. & Ollerton, J. Generalization in pollination systems, and why it matters. *Ecology* **77**, 1043–1060 (1996).
60. Rader, R. et al. Non-bee insects are important contributors to global crop pollination. *Proc. Natl. Acad. Sci. USA* **113**, 146–151 (2016).
61. Aliakbarisani, R., Serrano, M. Á & Boguñá, M. Feature-enriched hyperbolic network geometry. *Phys. Rev. Res.* **7**, 033036 (2025).
62. Ganea, O., Becigneul, G. & Hofmann, T. Hyperbolic neural networks. In Advances in Neural Information Processing Systems. eds. Bengio, S. et al. Vol. 31, 5345–5355 (NIPS, 2018).
63. Sen, P. et al. Collective classification in network data. *AI Mag.* **29**, 93–106 (2008).
64. Giles, C.L., Bollacker, K.D. & Lawrence, S. CiteSeer: an automatic citation indexing system. In: *Proc. Third ACM Conference on Digital Libraries*, 89–98 (ACM, 1998).
65. Meng, Z., Liang, S., Bao, H. & Zhang, X. Co-embedding attributed networks. In: *Proc. Twelfth ACM International Conference on Web Search and Data Mining*, 393-401 (ACM, 2019).

66. Duarte, N. D., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R. & Palsson, B. Ø Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. USA* **104**, 1777–1782 (2007).
67. Marlin, J.C. & LaBerge, W.E. The native bee fauna of Carlinville, Illinois, revisited after 75 years: a case for persistence. *Conserv. Ecol.* **5**, 9 (2001).
68. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Acknowledgements

The authors disclose support for the research of this work from MCIN/AEI/10.13039/501100011033 and the "European Union NextGenerationEU/PRTR" [Grant TED2021-129791B-I00], MCIN/AEI/10.13039/501100011033 and by ERDF/EU [Grant PID2022-137505NB-C22]; M. B. discloses support for the research of this work from the Agency for Management of University and Research Grants of the Generalitat de Catalunya through the Acadèmia d'Excel · lència grant.

Author contributions

M.B., M.A.S., and L.S.R. designed the research. L.S.R. and R.A. performed the research. All authors discussed the results. M.B., M.A.S., and L.S.R. wrote the manuscript. All authors reviewed and edited 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/s41467-026-74723-4.

Correspondence and requests for materials should be addressed to Marián Boguñá.

Peer review information *Nature Communications* thanks Malte Schröde, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026