PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Rohr RP, Naisbit RE, Mazza C, Bersier L-F. 2016 *Matching–centrality* decomposition and the forecasting of new links in networks. *Proc. R. Soc. B* **283**: 20152702. http://dx.doi.org/10.1098/rspb.2015.2702

Received: 9 November 2015
Accepted: 8 January 2016

Subject Areas:
computational biology, ecology, systems biology

Keywords:
complex networks, ecological networks, metabolic networks, missing links, predicting networks, social networks

Author for correspondence:
Rudolf P. Rohr
e-mail: rudolf.rohr@unifr.ch

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2015.2702 or via http://rspb.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Matching–centrality decomposition and the forecasting of new links in networks

Rudolf P. Rohr[1,2], Russell E. Naisbit[1], Christian Mazza[3] and Louis-Félix Bersier[1]

[1]Department of Biology–Ecology and Evolution, University of Fribourg, Chemin du Musée 10, Fribourg 1700, Switzerland
[2]Integrative Ecology Group, Estación Biológica de Doñana, EBD-CSIC, Calle Américo Vespucio s/n, Sevilla 41092, Spain
[3]Department of Mathematics, University of Fribourg, Chemin du Musée 23, Fribourg 1700, Switzerland

Networks play a prominent role in the study of complex systems of interacting entities in biology, sociology, and economics. Despite this diversity, we demonstrate here that a statistical model decomposing networks into *matching* and *centrality* components provides a comprehensive and unifying quantification of their architecture. The *matching* term quantifies the assortative structure in which node makes links with which other node, whereas the *centrality* term quantifies the number of links that nodes make. We show, for a diverse set of networks, that this decomposition can provide a tight fit to observed networks. Then we provide three applications. First, we show that the model allows very accurate prediction of missing links in partially known networks. Second, when node characteristics are known, we show how the *matching–centrality* decomposition can be related to this external information. Consequently, it offers us a simple and versatile tool to explore how node characteristics explain network architecture. Finally, we demonstrate the efficiency and flexibility of the model to forecast the links that a novel node would create if it were to join an existing network.

1. Introduction

The modern world is an increasingly connected place, through transport, social, and economic networks, and via our knowledge of interactions at the ecological or molecular level [1–3]. It is increasingly recognized that such systems should be studied globally, and networks of interacting entities provide us a powerful representation of their structure and function. Research on network theory parallels this growth [3]. A first body of research concentrates on the fact that observed networks are often considered to be only partially known. This would be the case in a food web, for instance, in which some real interactions may have yet to be observed, or a protein interaction network where not all pairwise combinations had been tested in the laboratory. Thus, observed links are typically considered as certain, whereas an absence of a link between a pair of nodes may reflect an absence of information rather than a real absence of interaction. Models have been devised to predict these 'missing links' and thus correct a network dataset for this incomplete sampling or direct future research towards these candidate interactions [4,5].

A second domain aims to determine if the structure of these networks exhibits basic generalities, and to uncover the processes that may generate these patterns. This aspect has been tackled with a variety of mostly comparative approaches, such as those treating the classification of networks [6,7], motifs [8], or stochastic models [9]. Progress in this undertaking could be achieved if there were general methods to relate network structure to characteristics of the nodes. For example, body size has been related to patterns in food webs [10], or country politics and trade to the organization of military conflict networks [11].

A third potential application of research is network forecasting, which aims to predict the links made by new nodes joining a network. Many current issues facing human society would benefit from the ability to forecast networks, such as for the ecological interactions of invasive species [12], the molecular

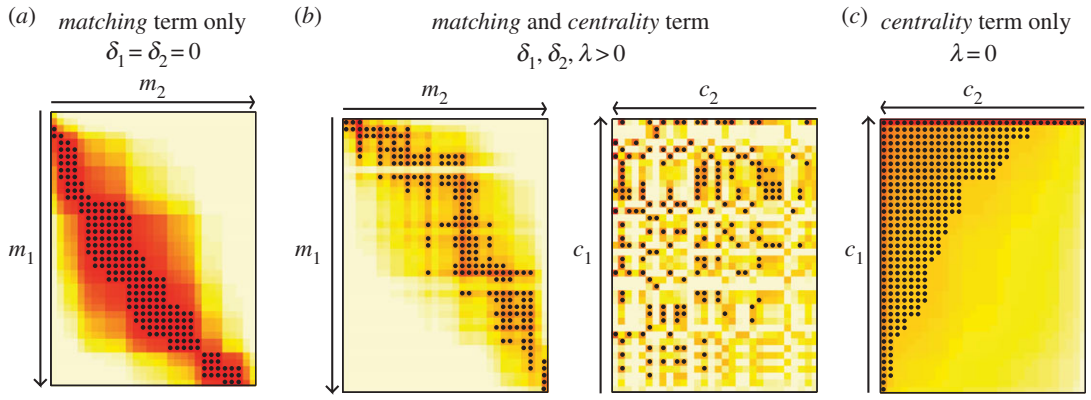

Figure 1. The *matching–centrality* model. The probability of a link between two nodes is decomposed into a *matching* term, quantifying assortative structure in who makes links with whom (*a*), and a *centrality* term, capturing the fact that nodes can vary considerably in their degree (*c*). The aim is to simultaneously quantify latent traits of the nodes that are responsible for the *matching* (m_1, m_2) and *centrality* (c_1, c_2) in the network (*b*). Panels (*a–c*) show adjacency matrices of three simulated networks, where a black dot represents a link and colours, from yellow to red, represent increasing linking probability computed with the model. The nodes of the matrices are ordered according to their *matching* or *centrality* traits. If a network exhibits a modular structure [3,25,26], this will be captured by the *matching* term, whereas variation in node degree [3,9,27,28] is captured by the *centrality* traits (see also figure 3).

interactions of a newly discovered protein [13], or the links of a subversive social group [14]. However, there exists no general framework for this forecasting. Here we provide a general model that can be applied to all three domains of network research.

The key feature of our methodology is the development of a model for the probability of interaction between nodes based on the decomposition of network architecture into *matching* and *centrality* terms. The *matching* term aims to quantify assortative structure in who makes links with whom [15–17], whereas the *centrality* term captures variation in the number of links that nodes make. Typically, research on network structure has focused on patterns in either assortativeness or *centrality*. However, the architecture of empirical networks is usually a product of both features simultaneously [18]. Here, we take into consideration both patterns. Specifically, the decomposition is implemented at the node level, with each node characterized by latent traits of *matching* and *centrality*. Latent traits are variables whose values are unknown *a priori*, but can be estimated *a posteriori* from the network adjacency matrix itself [19,20]. The model, called the *matching–centrality* model, is implemented in such a way that the closer the *matching* traits of two nodes, the greater the probability that they are linked, and the higher the *centrality* trait of a node, the greater the probability that this node makes links. This model belongs to the general class of 'hidden' variables models [5,19–24], for which some variables are unknown *a priori*, but can be estimated from the data *a posteriori*. In our model, these 'hidden' variables are the latent traits of *matching* and *centrality*.

Based on a dataset of 86 networks from disparate fields, we show that this decomposition into *matching* and *centrality* terms can provide a precise fit to observed networks. Then, we provide three applications of the *matching–centrality* model. First, we show that the model can be used to accurately predict missing links. Second, we show that the latent traits of *matching* and *centrality* are not just abstract traits, but can be linked to external information about the nodes and thus provide a means to study network organization. For example, in a food web, the latent traits are related to the body size and phylogeny of the interacting species. Finally, by placing latent traits as intermediates between the network structure and characteristics of the nodes, the model offers the possibility to forecast the interactions made by novel nodes when joining the network. For example, in a spatial network of mammal communities on mountains, we show that we can accurately predict the mammal fauna of unsampled mountains, or in a food web the trophic links for a new incoming species.

2. Methods

We formulate our *matching–centrality* model for undirected bipartite networks, but it can be applied to any kind of undirected or directed network, as explained below. Bipartite networks are made of two sets of nodes (S_1 and S_2) with connections only between them and not within; plant–pollinator networks provide a classical example. Let A be the adjacency matrix of the network, i.e. $a_{ij} = 1$ if there is a link between nodes i and j, and zero otherwise. The model characterizes each node i in set S_1 by a latent trait of *centrality* denoted $c_{1,i}$, and by $d \geq 1$ latent traits of *matching* denoted $m^1_{1,i}, \ldots, m^d_{1,i}$ [15–17], and similarly, each node j in set S_2 by a *centrality* trait $c_{2,j}$ and *matching* traits $m^1_{2,j}, \ldots, m^d_{2,j}$. The value of d gives the number of *matching* space dimensions and can be tuned to improve the goodness of fit of the model. We take a statistical approach, in which the probability of existence of a link between a pair of nodes i and j (hereafter the linking probability $P(a_{ij} = 1)$) is modelled through its logit [24]. Our model is given by

$$\log\left(\frac{P(a_{ij}=1)}{1-P(a_{ij}=1)}\right) = \underbrace{-\sum_{k=1}^{d} \lambda_k (m^k_{1,i} - m^k_{2,j})^2}_{\textit{matching}\ \text{term}} + \underbrace{\delta_1 c_{1,i} + \delta_2 c_{2,j}}_{\textit{centrality}\ \text{term}} + m, \tag{2.1}$$

where δ_1, δ_2, $\lambda_1, \ldots, \lambda_d$ are positive constants that scale the relative importance of the *matching* and *centrality* terms and m the common intercept. Figure 1 depicts patterns in interaction networks that the model is able to capture. When only the *matching* term is present, the model is perfectly tailored to fit a modular structure (figure 1*a*) [3,25,26]. In turn, the *centrality* term can perfectly fit highly nested networks [29] (figure 1*c*), because it captures the variation in node degree [3,9,27,28]. The latent traits of *centrality* are usually highly correlated to node degree. Therefore, when both the *matching* and the *centrality* terms are present, the model can fit simultaneously the modular and nested components of network structure (figure 1*b*).

For a given dimension d, the model parameters and latent trait values for each node can be estimated using a simulated

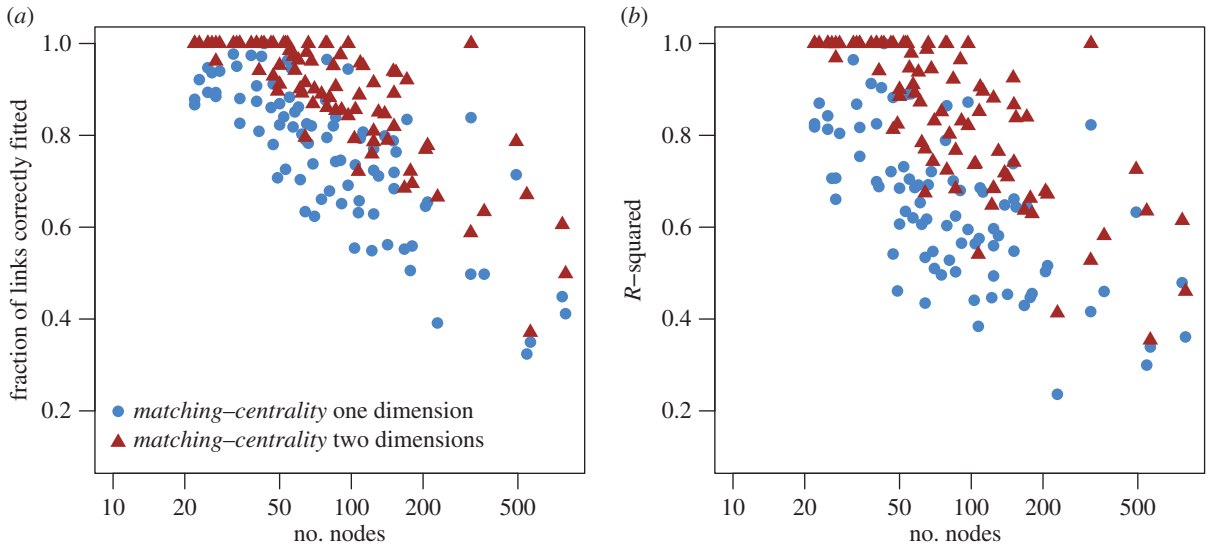

Figure 2. Performance of the *matching–centrality* model in inferring a dataset of 86 networks. Panels (*a*) and (*b*) show the fraction of correctly fitted links and the *R*-squared as a function of the number of nodes in the network, respectively. The figure shows that the higher the number of *matching* dimensions, the better the quality of fit of the model. As it can be expected, both the fraction of correctly fitted links and *R*-squared decrease with the number of nodes in the network. (Online version in colour.)

annealing algorithm [30]. The likelihood of the model is computed by

$$L = \prod_{i,j} P(a_{ij} = 1)^{a_{ij}} (1 - P(a_{ij} = 1))^{(1-a_{ij})}, \tag{2.2}$$

i.e. we assume the presence/absence of links follows a multi-Bernoulli distribution and that the probabilities are conditionally independent given the parameters and the latent traits. Moreover, to make the parameters and the latent traits of the model uniquely defined, we have to impose some constraints

(1) All vectors of latent traits $\boldsymbol{m}_1^k$, $\boldsymbol{m}_2^k$, $\boldsymbol{c}_1$, and $\boldsymbol{c}_2$ are orthogonal to the unit vector,
(2) All vectors of *matching* traits $\boldsymbol{m}_1^k (k = 1, \cdots, d)$ are pairwise orthogonal, and similarly for the vectors $\boldsymbol{m}_2^k$, and
(3) The length of the vectors $\boldsymbol{m}_1^k$, $\boldsymbol{c}_1$ is set to $\sqrt{n_{S_1}}$, where n_{S_1} is the number of nodes in set S_1. Similarly for set S_2, the length of the vectors $\boldsymbol{m}_2^k$, $\boldsymbol{c}_2$ is set to $\sqrt{n_{S_2}}$.

Application to other types of network requires simple modifications: for directed unipartite networks, e.g. food web or military conflict networks, the two sets of nodes reflect the function of the nodes (consumer and resource species, initiator and target nations, respectively); thus, a node can appear in both sets. For undirected unipartite networks such as social networks, the adjacency matrix is symmetric; we have to impose $m_{1,i}^k = m_{2,i}^k$, $c_{1,i} = c_{2,i}$, $\delta_1 = \delta_2$, and $P(a_{ii} = 1) = 0$ (the probability of a self-link is equal to zero). For more complex networks like directed or undirected multipartite networks, linking probabilities must be set to 0 for pairs of nodes that, by definition, cannot be linked. Note that our model is designed here for qualitative networks; its application to weighted networks would require modifying the likelihood function (equation (2.2)) to include the probability density for the weights of the links, in a similar way as in zero-inflated models [31].

3. Results

(a) Performance of the model

We illustrate the ability of the *matching–centrality* model to capture network architecture using a set of 86 examples from disparate fields: social interactions in Zachary's karate club network [32], co-appearance of characters in the novel *Les Misérables* [33], United States college football (USCF) games of the USCF teams [34], social networks of long-lasting association between 62 dolphins [35], associations between the terrorists involved in the September 11 attacks [14], military conflicts between countries [11], a subset of the network of physical interactions between the nuclear proteins in *Saccharomyces cerevisiae* [2,36], the neural network of *Caenorhabditis elegans* [37], 18 food webs, 59 mutualistic ecological networks [38], and the presence/absence of data of mammal species on peaks within the southern Rocky Mountains [39] (see electronic supplementary material for details; data can be downloaded from Dryad [40]). Once fitted, the model provides a new visual representation of the network in the latent trait space (see electronic supplementary material, figures S1–S11).

We fit the model for one and two dimensions of *matching* traits, and calculate the fraction of correctly fitted links and the McFadden's pseudo-*R*-squared [41] as metrics for its performance. The fraction of correctly fitted links is defined as the true positive rate after having classified the presence/absence of the links in the following way. We choose a cut-off point in linking probability and then classify a link to be present between a given pair of nodes if its linking probability is higher than the cut-off point; otherwise, the link is classified as absent. The level of the cut-off point is chosen so that the false-positive rate is equal to the false-negative rate. The McFadden's pseudo-*R*-squared is given by $R^2 = 1 - \ln(L)/\ln(L_n)$, where L is the likelihood of the model (equation (2.2)) and L_n is the likelihood of the model without the latent traits, i.e. only with the intercept m.

Figure 2 shows the performance of the model as a function of the number of nodes for one and two *matching* dimensions. We observed that the performance of the model decreases with the number of links, whereas it increases with the number of matching dimensions. This behaviour is expected, as increasing the number of nodes leads usually to more complex networks, and more matching dimensions are needed to reach a high level of fit. This is somewhat akin to a principal component analysis where increasing the number of dimensions leads to a higher fraction of explained variance. Theoretically, by increasing the number of *matching*

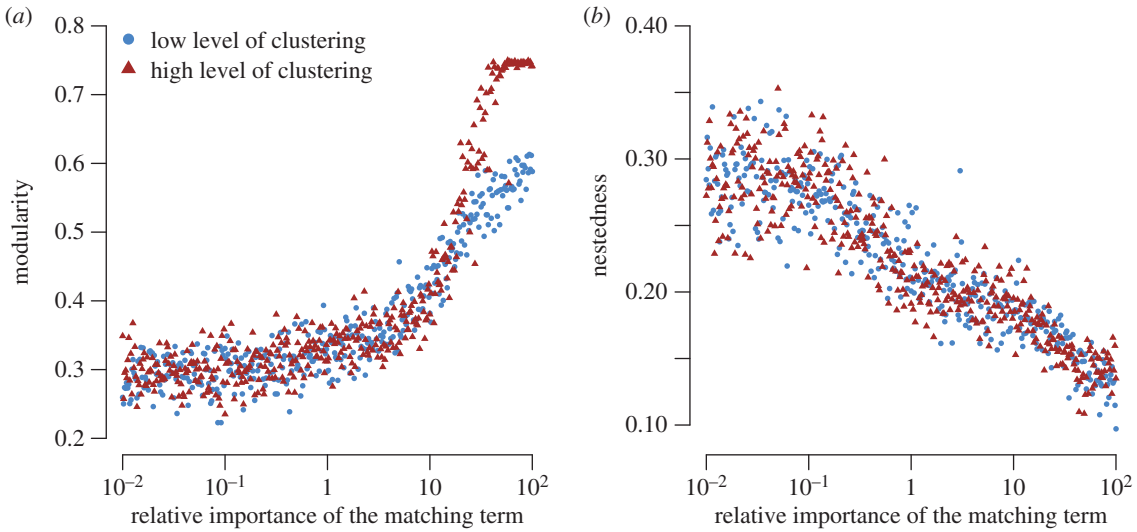

Figure 3. Relationship between network architecture and the *matching* and *centrality* terms. Panels (*a*,*b*) show the effect of the relative importance of the *matching* term on the modularity and nestedness metric, respectively, for 800 simulated networks. Increasing the importance of the *matching* term in the simulated networks results in higher modularity and lower nestedness. Moreover, increased clustering of the *matching* traits results in increased modularity for a high relative importance of *matching*.

dimensions d, it would be possible to obtain a fraction of correctly fitted links and R-squared equal to 1. However, an increase in the number of *matching* dimensions has drawbacks. First, it will lengthen computation time. More fundamentally, it can lead to an overfitting of the network. In such a situation, the prediction of missing links becomes meaningless, in a way analogous to increasing the degree of the polynomial for interpolation in a regression context. Ultimately, the choice of the number of *matching* dimensions d must be made in accordance with the application of the model.

(b) Relation between the *matching–centrality* model and network architecture

We explore how the relative importance of the *matching* term (for $d = 1$) over the *centrality* term (i.e. λ_1/δ; $\delta = \delta_1 = \delta_2$) affects the architecture of the network. We simulated bipartite networks of $n_{S_1} = 20$ and $n_{S_2} = 30$ nodes, with a connectance of 0.15, along a gradient of relative importance of the *matching* term. For each sampled network, we computed its level of nestedness [29] and modularity [3,26]. Figure 3 shows that, with greater importance of the *matching* term, modularity, indeed, increases, while nestedness decreases. Moreover, increasing the clustering in the matching traits (from a uniform distribution to four clusters) also increases the level of modularity (figure 3*a*). This result is qualitatively robust to change in network connectance and size (results not shown).

4. Applications of the *matching–centrality* model

(a) Prediction of missing links

The *matching–centrality* model can be used for the prediction of 'missing' links in partially known networks, where the absence of an interaction may in fact reflect an absence of information [4,5,42]. Here, we demonstrate its performance by simulating missing links in a subset of eight networks (figure 4). We simulate missing links by removing a given percentage of links and attempting to recover them.

Specifically, we simulated networks with missing links by setting a given fraction of 1 s to 0 s in the adjacency matrix. We removed at random 2, 5, 15, 30, 50, 75, and 90% of the 1 s, replicated 100 times for each fraction. Then, the *matching–centrality* model was fitted to the incompletely observed networks, and latent traits were estimated for each node. We fitted the model with only one dimension of *matching*. These *matching* and *centrality* traits were then used to estimate linking probabilities for each pair of nodes (equation (2.1)). We judge the performance by comparing the matrix of estimated linking probabilities and the true network using the area under the receiver operating characteristic curve (AUC) criterion. Here, the AUC can be interpreted as the probability that missing or observed links are given higher linking probabilities than real absences. For comparison, we give the performance of four other methods: the stochastic block model [5] with its extension to directed and bipartite networks [44–46], the Jaccard index, the common neighbours, and the degree product [42]. Note that some methods were not designed for directed or bipartite networks, and were not applied in these specific cases. In practical applications, the absent links ($a_{ij} = 0$) with the highest predicted linking probabilities are considered to be missing links. These are the candidate interactions that should come under the scrutiny of researchers, thus serving as a guide for cost-effective analysis of complex systems.

Figure 4 shows that the *matching–centrality* model performs very well in recovering missing links. There is one notable exception for the football game network, for which the stochastic block model performs better (figure 4*c*). This is not surprising given that this specific network exhibits a strong block structure (see electronic supplementary material, figure S3). Note that we also tested the method with two *matching* dimensions. We encountered overfitting with increasing percentages of removed links, and therefore, a decrease in the ability to recover missing links (results not shown). For larger networks, however, more than one *matching* dimension may be needed.

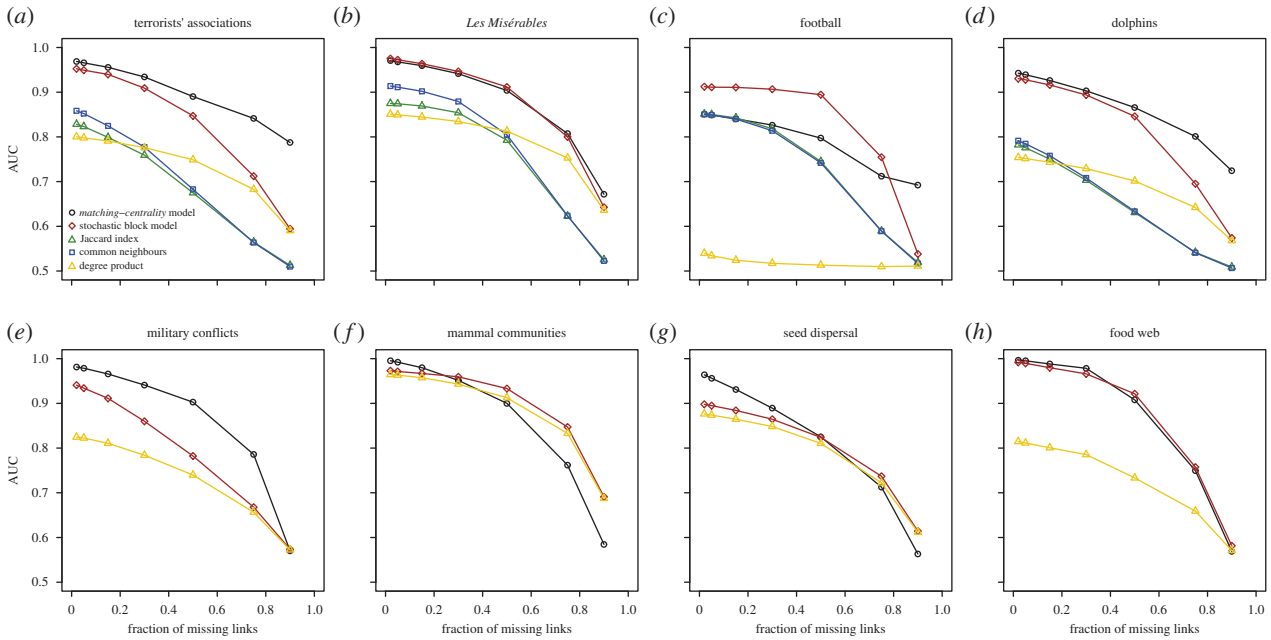

Figure 4. Prediction of missing links in eight partially known networks. We present the performance of the *matching–centrality* model (with one dimension) in predicting missing links for: (*a*) terrorists' associations [14], (*b*) character co-appearance in the novel *Les Misérables* [33], (*c*) United States college football games [34], (*d*) association between 62 dolphins [35], (*e*) military conflicts [37], (*f*) presence/absence data of mammal species [39], (*g*) seed dispersal [38], (*h*) trophic interaction in Tuesday Lake [43]. We compared the *matching–centrality* model to several alternative methods [5,42]. The average AUC statistic (the probability that an existing or missing link is given a higher linking probability than a true negative) is represented as a function of the fraction of simulated missing links created by removing links in the observed network.

(b) Linking latent traits to node characteristics

As shown above, the model can be fitted and used for prediction simply based on the network itself. However, it also offers an intuitive tool to gain insights into possible processes underlying network structure, as the *matching* and *centrality* traits of nodes can be related to independent information about the nodes, using standard analyses such as linear models or Mantel tests.

For example, in the food web of Tuesday Lake [43,47], both the *matching* and *centrality* traits of predators and prey can be related to their body size and phylogeny. Specifically, we used phylogenetic regression [48] to relate the *matching* and *centrality* traits to species' body size and phylogeny. Therefore, we assume that the latent traits follow a multivariate normal distribution (MVN), where the linear term is given by the logarithm of the body size and where the correlation structure is induced by the phylogeny, i.e.

$$\boldsymbol{m}_1^1, \boldsymbol{m}_1^2, \boldsymbol{m}_2^1, \boldsymbol{m}_2^2, \boldsymbol{c}_1, \boldsymbol{c}_2 \sim \mathrm{MVN}(\alpha + \beta \log(\boldsymbol{bs}), \Sigma(\lambda)), \tag{4.1}$$

where $\boldsymbol{m}_1^1, \boldsymbol{m}_1^2, \boldsymbol{m}_2^1, \boldsymbol{m}_2^2, \boldsymbol{c}_1, \boldsymbol{c}_2$ denote the vectors of the *matching* and *centrality* traits of resources and consumers, respectively, $\boldsymbol{bs}$ the body sizes, $\Sigma(\lambda)$ is the variance–covariance matrix induced by the phylogeny, and α, β, and λ are the parameters of the phylogenetic regression. We use Pagel's-λ [49] structure for the variance–covariance matrix, i.e.

$$\Sigma(\lambda)_{ij} = \begin{cases} \sigma^2 \cdot t_{ij} \cdot \lambda & \text{if} \quad i \neq j \\ \sigma^2 & \text{if} \quad i = j \end{cases}, \tag{4.2}$$

where σ^2 is the common variance, t_{ij} is the proportion of time that species i and j spent in common before speciation on the phylogenetic tree, and λ is the control parameter for the strength of the phylogenetic correlation ($\lambda = 0$ is equivalent to no correlation). The p-values of the parameters α and β are computed with the usual z-test, whereas the p-value associated with the correlation structure is computed using a log-likelihood ratio test between models with and without correlation. The analyses were done in R [50] with the libraries ape [51] and nlme [52].

For this specific network, the result shows that both the *matching* and *centrality* traits are related to the species' body size and phylogeny (electronic supplementary material, table S2). Although in this specific example we used only phylogeny and body size, any other relevant ecological and behavioural traits, or environmental conditions, could be used [53–56]. The latent traits are thus not just an abstract characterization of the nodes, but provide a versatile method to unravel factors underlying the different aspects of network structure.

(c) Forecasting the links of new nodes

Finally, a significant feature of the *matching–centrality* model is the possibility to forecast the links that new nodes would create when joining an existing network. This might be applied, for example, to forecast the interactions of an invasive species entering a food web or pollination network, the contacts of a non-surveyed individual in a terrorist network, or the biota of an unsampled mountain. The procedure is as follows: from the adjacency matrix, we first estimate the latent traits of *matching* and *centrality* for each node in the existing network, and verify that our model provides an accurate fit. Then, using appropriate statistical models, we relate the latent traits of the nodes to external information about them (as done in the previous section), and ensure that the model provides a good fit to the data. If both conditions are met, we can forecast the *matching* and *centrality* traits of the new node(s) using the external information, and finally their linking probability with each of the existing nodes.

In the following, we illustrate the method using the network describing the presence/absence of mammal species

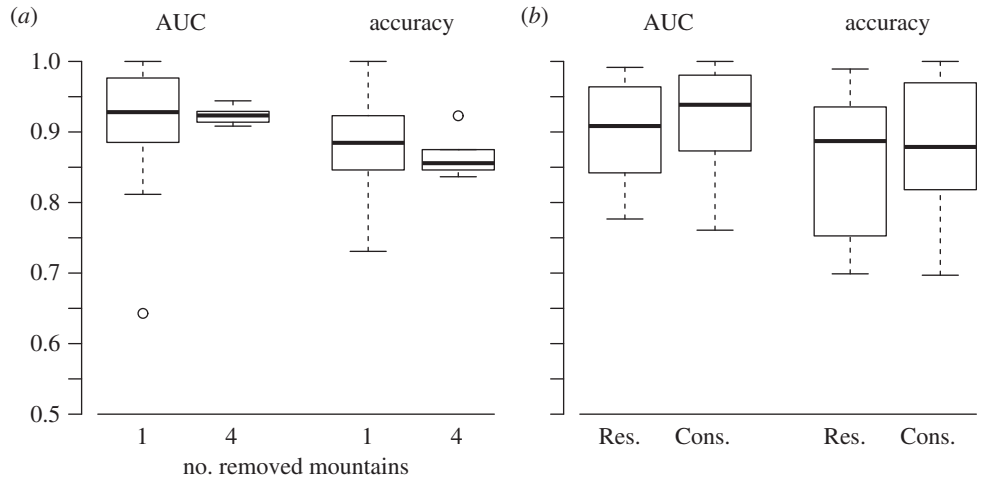

Figure 5. Performance of the *matching–centrality* model in forecasting (*a*) the mammal communities on unsampled mountains [39] and (*b*) trophic interactions of new species in the Tuesday Lake food web [43]. We illustrate the performance of the model using an out-of-sample test, removing mountains singly and in groups of four, and species in groups of three, respectively. After fitting the model to the known network, we use a statistical analysis to predict the latent traits of the unsampled mountains and removed species, based on their geographical characteristics and on their body size and phylogeny, respectively. Then, we use these predicted traits in the *matching–centrality* model to forecast mammal communities and trophic links, respectively. Graphs show box-plots for the AUC and for the accuracy of the forecasts (i.e. the percentage of correctly forecasted 0 s and 1 s). Res., resource; Cons., consumer species.

on mountains within the Rocky Mountains [39], and the food web of Tuesday Lake [43]. The model fits these two networks perfectly, with a fraction of correctly fitted links equal to 1 for two dimensions of *matching*. We first present the technique of forecasting applied to the presence/absence network of the Rocky Mountains. Then we give the modifications that have to be made to apply this technique to the food web of Tuesday Lake. The latent traits of the mountains ($\boldsymbol{m}_1^1$, $\boldsymbol{m}_1^2$, and $\boldsymbol{c}_1$) can be related to their area, elevation, and geographical position using a generalized least-squares linear model with a spatial correlation structure [31]. Specifically, we assume that the *matching* and *centrality* traits follow an MVN, where the linear part is given by the longitude, latitude, area, and elevation of the mountains, and that the spatial correlation structure follows an exponential law, i.e.

$$\boldsymbol{m}_1^1, \boldsymbol{m}_1^2, \boldsymbol{c}_1 \sim \mathrm{MVN}(\alpha + \beta_1 \cdot \mathrm{area} + \beta_2 \cdot \mathrm{elevation} + \beta_3 \cdot \mathrm{latitude} + \beta_4 \cdot \mathrm{longitude}, \Sigma(r)), \quad (4.3)$$

with the elements of the variance–covariance matrix given by

$$\Sigma(r)_{ij} = \sigma^2 \mathrm{e}^{-d_{ij}/r}. \quad (4.4)$$

The parameters α, β_1, β_2, β_3, β_4 are the intercept and slope; r is a parameter tuning the exponential decay of the spatial correlation; σ^2 is the common variance; and d_{ij} is the distance between mountains i and j. For the *matching* traits, we found that only the spatial correlation structure was significant, whereas for the *centrality* traits, area, elevation, and latitude were significant (electronic supplementary material, table S3). Note that the *matching* and *centrality* traits for the mammals could not be accurately related to species traits (electronic supplementary material, table S3).

Because the *matching* and *centrality* traits of the mountains are significantly related to several covariates and to the correlation structure given by the between-mountain distances, it should be possible to forecast the mammal communities in unsampled mountains for which the covariates are known, given the information provided by the covariates of the sampled mountains and the observed presence/absence network. We test the forecasting performance of the *matching–centrality* model by removing, one-by-one and by sets of four, each mountain from the dataset and attempting to recover its mammal community. This yields the following out-of-sample test

(1) We remove one or four mountains, denoted by k, from the network and then estimate the *matching* and *centrality* traits of the mammals and of the remaining mountains using the *matching–centrality* model,
(2) We fit the statistical model (4.3) on the estimated *matching* and *centrality* traits of step 1,
(3) Using the fitted parameters from step 2, we compute the conditional expectation for the *matching* and *centrality* traits of mountain(s) k; for the *centrality* trait, this value is given by

$$\hat{c}_{1,k} = \hat{\alpha} + \hat{\beta}_1 \cdot \mathrm{area} + \hat{\beta}_2 \cdot \mathrm{elevation} + \hat{\beta}_3 \cdot \mathrm{latitude}, \quad (4.5)$$

where $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{r}$ are the fitted parameters from step 2. The conditional expectation for the *matching* trait is given by

$$\hat{m}_{1,k} = \hat{\alpha} + \Sigma(\hat{r})_{-kk} \Sigma(\hat{r})_{kk}^{-1} (\boldsymbol{m}_{1,-k} - \boldsymbol{m}_{1,-k}), \quad (4.6)$$

where $\Sigma(\hat{r})_{-kk}$ is the kth column(s) without the kth row(s) (indicated by subscript $-k$) of the variance–covariance matrix estimated using equation (4.4); $\Sigma(\hat{r})_{kk}$ is the (k, k) element(s) of the estimated variance–covariance matrix; $(\boldsymbol{m}_{1,-k} - \boldsymbol{m}_{1,-k})$ is the row vector of residuals obtained from step 2. The last term of equation (4.6) represents the deviation from the linear prediction that is introduced by knowledge of the spatial correlation structure,
(4) With the predicted *matching* and *centrality* traits, we can predict the linking probabilities (equation (2.1)) between the removed mountain(s) and the mammals. The presence/absence of the mammals is determined from these linking probabilities and the cut-off point obtained with the fit of step 1. The cut-off point is defined as the linking probability chosen such that the number of false-positives is equal to the number of false-negatives, and
(5) We repeat steps 1–4 for all mountains in turn.

We applied the same procedure for the food web of Tuesday Lake, but we used the phylogenetic regression given by equation (4.1) to relate the *matching* and *centrality* traits to the covariables. The model performs remarkably well: on average, 87% of the data were correctly forecasted for the Rocky Mountain network (figure 5*a*). For the food web of Tuesday Lake, recall that the *matching* and *centrality* traits are closely related to the body size and phylogeny of the species (electronic supplementary material, table S2). Based on these predictors, on average, 86% of the trophic links were correctly forecasted (figure 5*b*).

In a real-case situation for the forecasting of unsampled nodes, the first condition for application is that the *matching–centrality* model provides a good fit to the sampled network. Second, there must exist a strong relationship between the *matching* and *centrality* traits and the independent information on the nodes; in our first example, forecasting the mountains occupied by a new mammal would not be possible. Once these conditions are met, one can proceed to the forecasting of the links connected to the unsampled nodes. We note three final technical points. First, as validation, we recommend performing a complete out-of-sample test as applied above. Second, the new nodes have to belong to the same statistical population as the original ones (it would obviously make no sense to forecast the mammal community present in a completely different region). Third, in our examples, we related the *matching* and *centrality* traits to the characteristics of the nodes using linear models; however, models of any form can be applied at this stage.

To the best of our knowledge, only one other method has been devised to forecast the links made by a new node in a network, for host–parasitoid networks based on the phylogenies of the species [57]. Our approach has two decisive advantages that make it extremely versatile: first, it is not necessary to include information about both sets of nodes of the bipartite network to make the forecast; second, by placing latent traits as intermediates between the network adjacency matrix and node characteristics, we provide an entirely flexible way to incorporate external information about nodes, for any conceivable statistical model could be used to relate the latent traits to external variables. Both features are perfectly illustrated in the above examples.

5. Conclusion

By translating an adjacency matrix into a set of quantitative traits for the nodes, the *matching–centrality* model represents a powerful and general tool for network analysis. It allows the reconstruction of missing information and forecasting of the links of entirely novel nodes, and opens the door to comparative analyses to shed light on the factors underlying network structure across disciplines.

Data accessibility. Network data: Dryad http://dx.doi.org/10.5061/dryad.5fn84.

Authors' contributions. All authors designed the research; R.P.R. performed the simulations; R.P.R., R.E.N., and L.F.B. analysed the results; all authors wrote the paper.

Competing interests. The authors declare no competing financial interests.

Funding. The work was supported by the National Centre of Competence in Research 'Plant Survival', the Swiss National Science Foundation grant 31003A-138489 (both to L.-F. Bersier), by SystemsX.ch, the Swiss Initiative in Systems Biology (to L.-F. Bersier and C. Mazza), and by the FP7-REGPOT-2010-1 programme (project 264125 EcoGenes), and an ERC Advanced Grant (both to J. Bascompte).

Acknowledgements. We thank J. Bascompte, A. Davison, and S. Saavedra for critical discussion.

References

1. Cohen JE. 1978 *Food webs and niche space*. Princeton, NJ: Princeton University Press.
2. Maslov S, Sneppen K. 2002 Specificity and stability in topology of protein networks. *Science* **296**, 910–913. (doi:10.1126/science.1065103)
3. Newman MEJ. 2010 *Networks: an introduction*. Oxford, UK: Oxford University Press.
4. Clauset A, Moore C, Newman MEJ. 2008 Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101. (doi:10.1038/nature06830)
5. Guimera R, Sales-Pardo M. 2009 Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl Acad. Sci. USA* **106**, 22 073–22 078. (doi:10.1073/pnas.0908366106)
6. Guimera R, Sales-Pardo M, Amaral LAN. 2007 Classes of complex networks defined by role-to-role connectivity profiles. *Nat. Phys.* **3**, 63–69. (doi:10.1038/nphys489)
7. Onnela J-P, Fenn DJ, Reid S, Porter MA, Mucha PJ, Fricker MD, Jones NS. 2012 Taxonomies of networks from community structure. *Phys. Rev. E* **86**, 036104. (doi:10.1103/PhysRevE.86.036104)
8. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002 Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827. (doi:10.1126/science.298.5594.824)
9. Barabasi AL, Albert R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)
10. Petchey OL, Beckerman AP, Riede JO, Warren PH. 2008 Size, foraging, and food web structure. *Proc. Natl Acad. Sci. USA* **105**, 4191–4196. (doi:10.1073/pnas.0710672105)
11. Ward MD, Siverson RM, Cao X. 2007 Disputes, democracies, and dependencies: a reexamination of the kantian peace. *Am. J. Polit. Sci.* **51**, 583–601. (doi:10.1111/j.1540-5907.2007.00269.x)
12. Memmott J. 2009 Food webs: a ladder for picking strawberries or a practical tool for practical problems? *Phil. Trans. R. Soc. B* **364**, 1693–1699. (doi:10.1098/rstb.2008.0255)
13. Garcia-Garcia J, Bonet J, Guney E, Fornes O, Planas J, Oliva B. 2012 Networks of protein–protein interactions: from uncertainty to molecular details. *Mol. Inf.* **31**, 342–362. (doi:10.1002/minf.201200005)
14. Krebs VE. 2002 Mapping networks of terrorist cells. *Connections* **24**, 43–52.
15. Rossberg AG, Matsuda H, Amemiya T, Itoh K. 2006 Food webs: experts consuming families of experts. *J. Theor. Biol.* **241**, 552–563. (doi:10.1016/j.jtbi.2005.12.021)
16. Rossberg AG, Brannstrom A, Dieckmann U. 2010 How trophic interaction strength depends on traits. *Theor. Ecol.* **3**, 13–24. (doi:10.1007/s12080-009-0049-1)
17. Rossberg AG. 2013 *Food webs and biodiversity*. Sussex, UK: John Wiley and Sons.
18. Fortuna MA, Popa-Lisseanu G, Ibanez C, Bascompte J. 2009 The roosting spatial network of a bird-predator bat. *Ecology* **90**, 934–944. (doi:10.1890/08-0174.1)
19. Hoff PD, Raftery AE, Handcock MS. 2002 Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* **97**, 1090–1098. (doi:10.1198/016214502388618906)
20. Rohr R, Scherer H, Kehrli P, Mazza C, Bersier L. 2010 Modeling food webs: exploring unexplained structure using latent traits. *Am. Nat.* **173**, 170–177. (doi:10.1086/653667)

21. Boguñá M, Pastor-Satorras R. 2002 Epidemic spreading in correlated complex networks. *Phys. Rev. E* **66**, 047104. (doi:10.1103/PhysRevE.66.047104)
22. Boguñá M, Pastor-Satorras R, Díaz-Guilera A, Arenas A. 2004 Models of social networks based on social distance attachment. *Phys. Rev. E* **70**, 056122. (doi:10.1103/PhysRevE.70.056122)
23. Boguñá M, Krioukov D, Claffy K. 2008 Navigability of complex networks. *Nat. Phys.* **5**, 74–80. (doi:10.1038/nphys1130)
24. Kolaczyk E. 2009 *Statistical analysis of network data*. New York, NY: Springer.
25. Melian CJ, Bascompte J. 2004 Food web cohesion. *Ecology* **85**, 352–358. (doi:10.1890/02-0638)
26. Newman MEJ, Girvan M. 2004 Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 1–15. (doi:10.1103/physreve.69.026113)
27. Price DJD. 1965 Networks of scientific papers. *Science* **149**, 510–515. (doi:10.1126/science.149.3683.510)
28. Amaral LAN, Scala A, Barthelemy M, Stanley HE. 2000 Classes of small-world networks. *Proc. Natl Acad. Sci. USA* **97**, 11 149–11 152. (doi:10.1073/pnas.200327197)
29. Patterson BD, Atmar W. 1986 Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biol. J. Linn. Soc.* **28**, 65–82. (doi:10.1111/j.1095-8312.1986.tb01749.x)
30. Brémaud P. 1998 *Markov chains, Gibbs fields, Monte Carlo simulation, and queues*. New York, NY: Springer.
31. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. 2009 *Mixed effects models and extensions in ecology with R*. Berlin, Germany: Springer.
32. Zachary WW. 1977 An information flow model for conflict and fission in small groups. *J. Anthrop. Res.* **33**, 452–473.
33. Knuth DE. 1994 *The Stanford GraphBase: a platform for combinatorial computing*. New York, USA: Addison-Wesley.
34. Girvan M, Newman MEJ. 2002 Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826. (doi:10.1073/pnas.122653799)
35. Lusseau D, Schneider K, Boisseau O, Haase P, Slooten E, Dawson S. 2003 The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **54**, 396–405. (doi:10.1007/s00265-003-0651-y)
36. Melian CJ, Bascompte J. 2002 Complex networks: two ways to be robust? *Ecol. Lett.* **5**, 705–708. (doi:10.1046/j.1461-0248.2002.00386.x)
37. Watts DJ, Strogatz SH. 1998 Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442. (doi:10.1038/30918)
38. Rezende EL, Lavabre JE, Guimaraes PR, Jordano P, Bascompte J. 2007 Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature* **448**, 925–928. (doi:10.1038/nature05956)
39. Patterson BD. 1984 Mammalian extinction and biogeography in the southern rocky mountains. In *Extinctions* (ed. MH Nitecki), pp. 247–293. Chicago, IL: University of Chicago Press.
40. Rohr RP, Naisbit RE, Mazza C, Bersier L-F. 2016 Data from: matching–centrality decomposition and the forecasting of new links in networks. *Dryad Digital Repository*. (doi:10.5061/dryad.5fn84)
41. Domencich T, McFadden DL. 1975 *Urban travel demand: a behavioral analysis*. Amsterdam, The Netherlands: North-Holland Publishing Co.
42. Liben-Nowell D, Kleinberg J. 2007 The link-prediction problem for social networks. *J. Assn. Inf. Sci. Technol.* **58**, 1019–1031. (doi:10.1002/asi.20591)
43. Jonsson T, Cohen JE, Carpenter SR. 2005 Food webs, body size, and species abundance in ecological community description. *Adv. Ecol. Res.* **36**, 1–84. (doi:10.1016/S0065-2504(05)36001-6)
44. Guimerá R, Llorente A, Moro E, Sales-Pardo M. 2012 Predicting human preferences using the block structure of complex social networks. *PLoS ONE* **7**, e44620. (doi:10.1371/journal.pone.0044620)
45. Rovira-Asenjo N, Gumi T, Sales-Pardo M, Guimerá, R. 2013 Predicting future conflict between team-members with parameter-free models of social networks. *Sci. Rep.* **3**, 1999. (doi:10.1038/srep01999)
46. Larremore DB, Clauset A, Jacobs AZ. 2014 Efficiently inferring community structure in bipartite networks. *Phys. Rev. E* **90**, 012805. (doi:10.1103/PhysRevE.90.012805)
47. Naisbit RE, Rohr RP, Rossberg AG, Kehrli P, Bersier L-F. 2012 Phylogeny versus body size as determinants of food web structure. *Proc. R. Soc. B* **279**, 3291–3297. (doi:10.1098/rspb.2012.0327)
48. Grafen A. 1989 The phylogenetic regression. *Phil. Trans. R. Soc. B* **326**, 119–157. (doi:10.1098/rstb.1989.0106)
49. Freckleton RP, Harvey PH, Pagel M. 2002 Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160**, 712–726. (doi:10.1086/343873)
50. R Core Team. 2015 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (https://www.R-project.org/)
51. Paradis E, Claude J, Strimmer K. 2004 Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)
52. Pinheiro J, Bates D, DebRoy S, Sarkar D, team, RC. 2009 nlme: linear and nonlinear mixed effects models. R package version 3.1-123. (http://CRAN.R-project.org/package=nlme)
53. Legendre P, Galzin R, Harmelin-Vivien ML. 1997 Relating behaviour to habitat: solutions to the fourth-corner problem. *Ecology* **78**, 547–562.
54. Dray S, Choler P, Dolédec S, Peres-Neto PR, Thuiller W, Pavoine S, ter Braak CJF. 2014 Combining the fourth-corner and the rlq methods for assessing trait responses to environmental variation. *Ecology* **95**, 14–21. (doi:10.1890/13-0196.1)
55. Spitz J, Ridoux V, Brind'Amour A. 2014 Let's go beyond taxonomy in diet description: testing a trait-based approach to prey–predator relationships. *J. Anim. Ecol.* **83**, 1137–1148. (doi:10.1111/1365-2656.12218)
56. Dehling DM, Töpfer T, Schaefer HM, Jordano P, Böhning-Gaese K, Schleuning M. 2014 Functional relationships beyond species richness patterns: trait matching in plant-bird mutualisms across scales. *Glob. Ecol. Biogeogr.* **23**, 1085–1093. (doi:10.1111/geb.12193)
57. Ives AR, Godfray HCJ. 2006 Phylogenetic analysis of trophic associations. *Am. Nat.* **168**, E1–E14. (doi:10.1086/505157)