On the information-theoretic formulation of network participation

IOP Publishing J. Phys. Complex. 5 (2024) 015021 https://doi.org/10.1088/2632-072X/ad32da

Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
17 October 2023

REVISED
7 March 2024

ACCEPTED FOR PUBLICATION
12 March 2024

PUBLISHED
27 March 2024

PAPER

On the information-theoretic formulation of network participation

Pavle Cajic[1], Dominic Agius[1], Oliver M Cliff[1,2], James M Shine[2,3], Joseph T Lizier[2,4,*] and Ben D Fulcher[1,2,*]

[1] School of Physics, Faculty of Science, The University of Sydney, Sydney NSW 2006, Australia
[2] Centre for Complex Systems, The University of Sydney, Sydney NSW 2006, Australia
[3] Brain and Mind Centre, Faculty of Medicine, The University of Sydney, Sydney NSW 2006, Australia
[4] School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney NSW 2006, Australia
[*] Authors to whom any correspondence should be addressed.

E-mail: joseph.lizier@sydney.edu.au and ben.fulcher@sydney.edu.au

Keywords: participation, information theory, complex networks

Abstract

The participation coefficient is a widely used metric of the diversity of a node's connections with respect to a modular partition of a network. An information-theoretic formulation of this concept of connection diversity, referred to here as participation entropy, has been introduced as the Shannon entropy of the distribution of module labels across a node's connected neighbors. While diversity metrics have been studied theoretically in other literatures, including to index species diversity in ecology, many of these results have not previously been applied to networks. Here we show that the participation coefficient is a first-order approximation to participation entropy and use the desirable additive properties of entropy to develop new metrics of connection diversity with respect to multiple labelings of nodes in a network, as joint and conditional participation entropies. The information-theoretic formalism developed here allows new and more subtle types of nodal connection patterns in complex networks to be studied.

Many real-world networks exhibit modular structure, in which nodes form densely interconnected modules with relatively sparse connectivity between modules. Such modularity is observed in social networks, food webs, metabolic networks, protein–protein interaction networks, air-traffic networks, and brain networks [1]. Within such modular networks, individual nodes can vary substantially in their degree of within- versus across-module connectivity. These differences can provide important insights into a node's functional role within a network, such as facilitating local information processing (consistent with strong within-module connectivity) versus distributed/integrative communication (strong cross-module connectivity).

To measure the extent to which a given node's connections are distributed within or across modules, the participation coefficient was introduced by Guimerà and Amaral [1, 2]. It has been used widely to analyze networks across domains, including the Internet, metabolic, air transportation, protein-interaction, and neural networks [3, 4]. For example, the participation coefficient of nodes in macroscopic brain networks has been used to distinguish levels of consciousness caused by brain injury [5] and to identify emerging new research directions from scientific publication citation networks [6]. This concept of nodal connection diversity across modules was also formulated as a Shannon entropy by Rubinov and Sporns [7]. Quantifying diversity is a general problem studied across many fields, with a prominent application to species diversity in ecology for which the Shannon entropy and Gini–Simpson index (the measure underlying the participation coefficient [7]) formulations have been used for decades, among a host of alternative indices [8, 9]. Mathematical relationships between different formulations of diversity indices have been uncovered. For example, the Gini–Simpson index and Shannon entropy have each been shown to be special cases of 'generalised entropies' [10–12]. Zhang and Grabchak [13] have further shown that the Gini–Simpson index can be expressed as a first-order Taylor approximation to the Shannon entropy formulation of diversity.

Despite the wide variety of diversity indices used in ecology, the participation coefficient has remained the dominant measure of node participation in network theory since it was introduced in 2005 [1, 2]. Here we connect the problem of quantifying nodal connection diversity in networks with a large and existing literature on diversity indices, and in particular explain the relationship between the participation coefficient

and the corresponding Shannon entropy-based formulation of connection diversity [7], which we call 'participation entropy' here. We argue that participation entropy is a better-motivated measure of node participation diversity, primarily due to its additive behavior with respect to chaining probability distributions, an operation which can arise naturally when the nodes in a network have labels in multiple module sets. Taking advantage of this behavior, we define novel measures of connection diversity—'joint' and 'conditional' participation entropy—for quantifying more nuanced types of connection patterns in complex networks.

1. Participation coefficient and participation entropy

We consider a binary, undirected network partitioned into M non-overlapping modules, with each node labeled as belonging to a module, from the set $\mathcal{M} = \{m_1, m_2, \ldots, m_M\}$. Note that this modular partition is most commonly obtained as the result of a community-detection algorithm operating on the network [14], but could in general represent any assignment of categorical labels to nodes in a network. Given $\mathcal{M}$, the participation coefficient, $\mathcal{P}_i$, of node i is defined as

$$\mathcal{P}_i(\mathcal{M}) = 1 - \sum_{j=1}^{M} \left(\frac{\kappa_{ij}}{k_i}\right)^2 , \tag{1}$$

where κ_{ij} is the number of edges between node i and a node in module m_j, and k_i is the degree of node i (the total number of connections made to all other nodes in the network) [1, 2]. For simplicity, we focus on undirected networks here, but note that this formulation extends straightforwardly to weighted networks (substituting κ_{ij} and k_i for weighted versions that sum edge weights) and directed networks (e.g. by defining κ_{ij} and k_i as counting connections outward from, or arriving to node i, as the in-degree or out-degree). Equation (1) exhibits the desired behavior of a connection diversity metric, taking a minimal value for a node with connections entirely within a single module ($\mathcal{P}_i = 0$) and a maximal value for a node that connects equally across all M modules ($\mathcal{P}_i = 1 - 1/M$).

1.1. A probabilistic formulation

An alternative interpretation of equation (1) can be considered by identifying κ_{ij}/k_i as the probability, $p_i(m_j)$, that a randomly selected connected neighbor of node i is assigned to module m_j. An example is depicted in figure 1(a), which depicts the connected neighbors of node i across each of three modules, $\mathcal{M} = \{m_1, m_2, m_3\}$. This, or any other pattern of connectivity, can be represented as a probability distribution, $\{p_i(m)\}_{m\in\mathcal{M}}$, plotted for this simple example in figure 1(b). In this probabilistic formulation, $\mathcal{P}_i$ can be expressed as a function of $p_i(m)$ by rewriting equation (1) as $\mathcal{P}_i(\mathcal{M}) = 1 - \sum_{m\in\mathcal{M}} p_i(m)^2$.

This formulation allows us to clearly see that the participation coefficient is an implementation of the Gini–Simpson index of diversity [15], as observed previously [7]. This is an important measure used in many other contexts, including quantifying biodiversity [8, 9]. Following the interpretation that motivated Simpson's original formulation [15], $\mathcal{P}_i$ can be interpreted as the probability that two randomly selected nodes connected to node i (with replacement) lie in different modules.

1.2. Participation entropy

The Shannon entropy [16, 17] of $p_i(m)$ is a natural measure of the connection diversity of node i across the label set, $\mathcal{M}$:

$$\mathcal{E}_i(\mathcal{M}) = H[p_i(m)] = -\sum_{m\in\mathcal{M}} p_i(m) \log p_i(m) . \tag{2}$$

We term $\mathcal{E}_i(\mathcal{M})$ the 'participation entropy' of node i, which measures the uncertainty (or average surprise) in the module labels (from $\mathcal{M}$) of its connected neighbors. This matches a previous formulation of nodal connection diversity introduced by Rubinov and Sporns [7] (named the 'diversity coefficient' in its implementation in code in the *Brain Connectivity Toolbox* [18]). Participation entropy exhibits the same desired qualitative behavior as the participation coefficient, $\mathcal{P}_i$; that is, $\mathcal{E}_i = 0$ is minimal when all connected neighbors of node i are in the same module (minimum uncertainty about the module label of node i's neighbors) and $\mathcal{E}_i = \log(M)$ is maximal when connected neighbors are equally distributed across all of the modules (maximum uncertainty about the module label of node i's neighbors). Note that both $\mathcal{E}_i$ and $\mathcal{P}_i$ may be normalized by dividing by their maximum value for a given number of modules M, if desired (as 'normalized connection diversity' [7], which has the effect of setting its range to the unit interval).

Compared to $\mathcal{P}_i$, quantifying connection diversity as an entropy, $\mathcal{E}_i$, provides an interpretable measure of diversity as the uncertainty in a target node's label given the underlying probability distribution of these

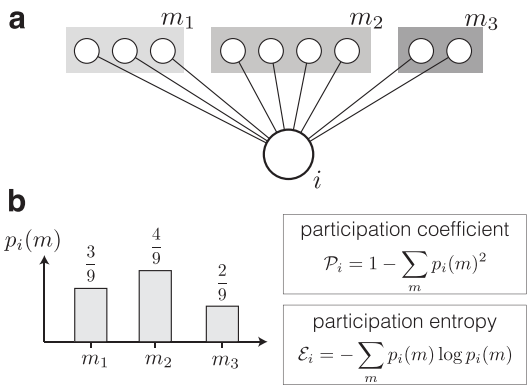

Figure 1. A probabilistic formulation of a node's connection diversity with respect to a set of labels. (a) We plot the connected neighbors of a given node, *i*, which span three labeled modules: m_1 (3 edges), m_2 (4 edges), and m_3 (2 edges). (b) This pattern can be represented as a probability distribution, $p_i(m)$, that captures the probability of node *i*'s connected neighbors being in each of the modules. The participation coefficient, $\mathcal{P}_i$, and participation entropy, are then computed from $p_i(m)$.

labels, $p_i(m)$. Moreover, $\mathcal{E}_i$ is the unique formulation that satisfies three key advantageous axioms simultaneously (see [16, 19], noting that there are multiple slightly different sets of axioms which lead to the same conclusion [17, 20]). First, it is continuous with respect to changes in $p_i(m)$. Second, it increases monotonically with the number of modules, M, when $p_i(m) = 1/M, \forall m$. Third, and most importantly, $\mathcal{E}_i$ can be decomposed consistently across multiple labeling sets for nodes [16, 19] which directly and uniquely leads to a chain rule relating univariate and multivariate measures of nodal participation. The ability to chain these entropy measures opens new ways of quantifying and interpreting nodal connection patterns in networks, as we develop later (in section 2). As per the original formulation of $\mathcal{P}_i$, it also generalizes straightforwardly to weighted and directed networks.

1.3. Connecting the two formulations

The mathematical relationship between the Gini–Simpson index and Shannon entropy is well-known [10–12] and has been demonstrated in the context of species diversity indices [13]. But the connection has not been reported for the corresponding measures of nodal connection diversity in networks, $\mathcal{P}$ and $\mathcal{E}$. The relationship can be seen through the series expansion of participation entropy via the logarithm in equation (2):

$$\mathcal{E}_i(\mathcal{M}) = -\sum_{m\in\mathcal{M}} p_i(m) \sum_{n=1}^{\infty} \frac{-[1-p_i(m)]^n}{n}. \tag{3}$$

This quantity converges for $0 < p_i(m) \leqslant 1$, and we take $0\log 0 \to 0$ by convention, so there is no contribution from any $p_i(m) = 0$. Limiting the expansion to the leading term, $n = 1$, yields

$$\begin{aligned} \mathcal{E}_i(\mathcal{M}) &\approx \sum_{m\in\mathcal{M}} p_i(m) - p_i(m)^2, \\ &= 1 - \sum_{m\in\mathcal{M}} p_i(m)^2, \qquad (4) \\ &= \mathcal{P}_i(\mathcal{M}). \qquad (5) \end{aligned}$$

We thus recapitulate the participation coefficient as a first-order approximation to participation entropy (as per the Gini–Simpson index and Shannon entropy [13]).

In order to investigate the discrepancy between $\mathcal{E}_i$ and its first-order approximation, $\mathcal{P}_i$, we sampled from possible distributions, $p_i(m)$ for $M = 2, \ldots, 5$, and plotted the resulting accessible regions of $\mathcal{P}_i$–$\mathcal{E}_i$ space in figure 2. We used 5×10^4 random samples for $M = 3$ and $M = 4$, and 10^6 samples for $M = 5$, and the `boundary` function in Matlab, which was sufficient to obtain smooth boundaries in each case. Our numerical results match analytic expressions for these regions for the underlying measures on $p_i(m)$ derived by Vajda and Zvárová [12]. We find that $\mathcal{P}_i$ varies monotonically with $\mathcal{E}_i$ for $M = 2$, but for $M > 2$, allowed values of $\mathcal{P}_i$ and $\mathcal{E}_i$ are constrained to specific regions of the space. This accessible region expands with the addition of each new module; figure 2 annotates the additional accessible region with each increment of M. The results indicate that there can be a substantial discrepancy between an analysis using $\mathcal{P}_i$ versus $\mathcal{E}_i$, with

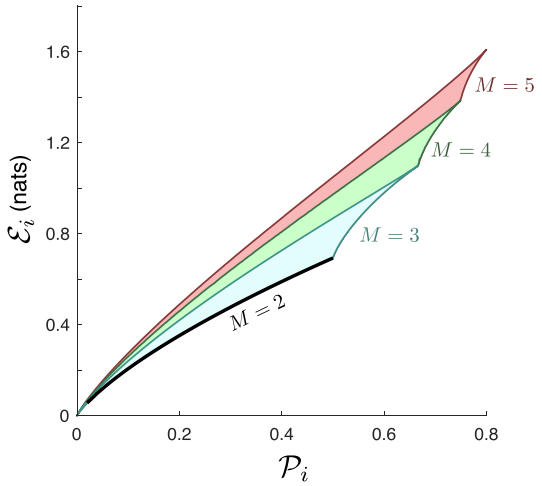

Figure 2. Constraints on the relationship between $\mathcal{P}_i$ and $\mathcal{E}_i$ for networks containing $M = 2, \ldots, 5$ modules. The shaded region for each M indicates the additional allowed region, in addition to that accessible for lower M values.

greater potential for differences at moderate-to-high values of $\mathcal{P}_i$ and with increasing M. Published results using $\mathcal{P}_i$ to quantify nodal diversity (or extract a list of 'high-participation nodes' [4, 21]), may thus obtain different results when using $\mathcal{E}_i$ instead of its first-order approximation, $\mathcal{P}_i$.

2. Joint and conditional participation entropy

A major advantage of formulating $\mathcal{E}_i$ as an entropy is the ability to capture more subtle types of connection-pattern diversity in networks. Here we demonstrate this capability by developing entropy-based network participation measures for the case that each node is annotated with *multiple* labels. Specifically, we consider L different module sets, $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_L$, that each define a labeling of network nodes. In a social network, this could correspond to individuals being labeled by both gender, $\mathcal{M}_g$, and friendship group, $\mathcal{M}_f$. Or, in a brain network, it could correspond brain regions being labeled by both their hemisphere, $\mathcal{M}_h$ (left or right) and their functional network module, $\mathcal{M}_f$ (e.g. auditory, visual, association, etc). There is no clear way of extending $\mathcal{P}_i$ to such a setting, but it can be incorporated naturally in the information-theoretic formulation of $\mathcal{E}_i$.

Extending participation entropy with respect to any single labeling of nodes, $\mathcal{M}$, we now consider the diversity of connections involving node i across multiple label sets jointly. Writing the L sets as $\underline{\mathcal{M}} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_L)$, and a combination of labels from $\underline{\mathcal{M}}$ for a given node as $\underline{m} = (m^{(1)}, m^{(2)}, \ldots, m^{(L)})$, we define the joint probability distribution $p_i(\underline{m})$ for the connected neighbors of node i. We can then define the *joint participation entropy* of node i as:

$$\begin{aligned} \mathcal{E}_i(\underline{\mathcal{M}}) &= H[p_i(\underline{m})] \\ &= -\sum_{\underline{m}} p_i(\underline{m}) \log p_i(\underline{m}) \,. \end{aligned} \tag{6}$$

This tells us the total diversity of connections across these multiple module sets, $\underline{\mathcal{M}}$.

Similarly, we can define the *conditional participation entropy* as the entropy of modular assignments $\underline{m}$ from sets $\underline{\mathcal{M}}$ of the connected neighbors of node i, given knowledge of the modular assignments $\underline{n}$ from other sets $\underline{\mathcal{N}}$:

$$\begin{aligned} \mathcal{E}_i(\underline{\mathcal{M}}|\underline{\mathcal{N}}) &= H[p_i(\underline{m}|\underline{n})] \,, \\ &= \mathcal{E}_i(\underline{\mathcal{M}}, \underline{\mathcal{N}}) - \mathcal{E}_i(\underline{\mathcal{N}}) \,. \end{aligned} \tag{7}$$

This quantifies the remaining uncertainty in the distributions of connections across the modules of sets $\underline{\mathcal{M}}$, given that we already know their distributions across sets $\underline{\mathcal{N}}$.

The joint participation entropy, $\mathcal{E}_i(\underline{\mathcal{M}})$, and conditional participation entropy, $\mathcal{E}_i(\underline{\mathcal{M}}|\underline{\mathcal{N}})$, are related via the chain rule for entropies [17] (vis-à-vis equation (7)), which means that we can consistently decompose and re-compose the diversity of connections over multiple module sets, regardless of which order we

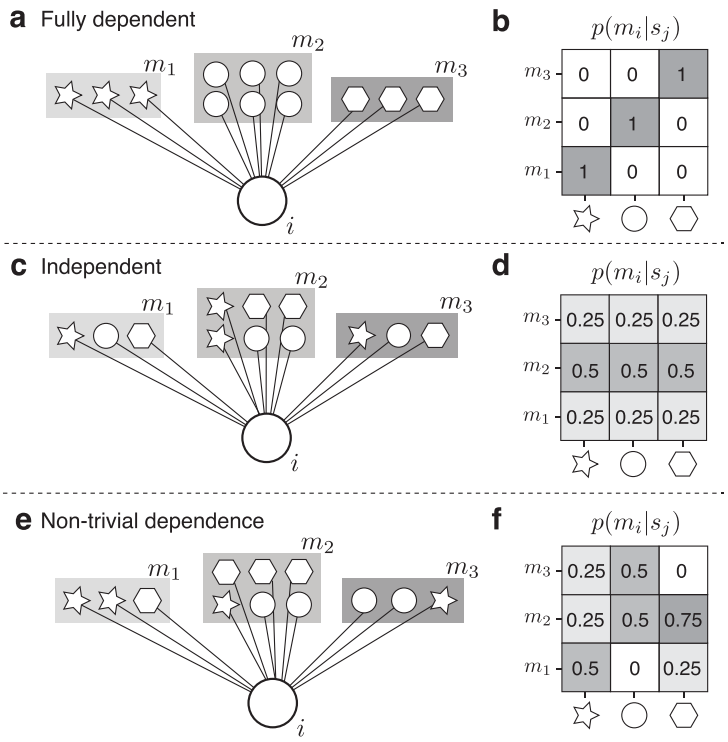

Figure 3. Participation entropy can be extended to the case where each node is labeled according to multiple module sets. Here we illustrate the case in which each node is labeled by two different module sets: $\mathcal{M} = \{m_1, m_2, m_3\}$ and $\mathcal{S} = \{\star, \bigcirc, \hexagon\}$. Three cases are illustrated: (a), (b), labelings $\mathcal{M}$ and $\mathcal{S}$ are redundant; (c), (d), labelings $\mathcal{M}$ and $\mathcal{S}$ are statistically independent; and (e), (f), labelings $\mathcal{M}$ and $\mathcal{S}$ exhibit non-trivial dependence. (a), (c), (e) show the pattern of connectivity from a target node, *i*, to a set of nodes labeled by $\mathcal{M}$ and $\mathcal{S}$. (b), (d), (f) show conditional probability matrices, $p(m_i|s_j)$ for connected neighbors of node *i* from both labelings.

chain our knowledge of the module labelings. This property is unique to the information-theoretic formulation [19].

To illustrate the calculation of conditional participation entropy, we show some illustrative examples in figure 3 for the simple case of two node labelings: $\mathcal{M} = \{m_1, m_2, m_3\}$ and $\mathcal{S} = \{\star, \bigcirc, \hexagon\}$. The three cases shown in figure 3 correspond to distinct types of connection patterns of node *i* with respect to $\mathcal{M}$ and $\mathcal{S}$. In figure 3(a), the labels assigned to node *i*'s connected neighbors are *redundant* with respect to $\mathcal{M}$ and $\mathcal{S}$. That is, for a given connected neighbor, knowledge of the label *s* leaves no uncertainty about the label *m* (and vice-versa), resulting in the symmetric $p(m_i|s_j)$ matrix shown in figure 3(b). For this case, the conditional participation entropy of node *i*, $\mathcal{E}_i(\mathcal{M}|\mathcal{S}) = 0$.

For the connection pattern shown in figure 3(c), the labelings *m* and *s* are statistically independent. That is, for a given connected neighbor, knowledge of the label *s* does not reduce our uncertainty about the label *m*, as reflected in the $p(m_i|s_j)$ matrix in figure 3(d). In this case, $\mathcal{E}_i(\mathcal{M}|\mathcal{S}) = \mathcal{E}_i(\mathcal{M})$.

In general, a node's connection pattern will involve non-trivial statistical dependencies between the combinations of labels. Such a case is shown in figure 3(e), where knowledge of the label *s* reduces our uncertainty about *m*. For example, as depicted in figure 3(f), if we learn that a node is labeled $s = \bigcirc$, then our uncertainty about its label, *m*, is reduced, from $\{p(m_1), p(m_2), p(m_3)\} = \{0.25, 0.5, 0.25\}$ to $\{0, 0.5, 0.5\}$. As such, $0 < \mathcal{E}_i(\mathcal{M}|\mathcal{S}) < \mathcal{E}_i(\mathcal{M})$ here.

The conditional participation entropy thus provides a new way to quantify a node's connection diversity across multiple labelings of network nodes. For example, in a structural brain network in which brain areas (nodes) are annotated by both by a functional annotation, $\mathcal{M}_f$ (e.g. visual, auditory, motor, etc) and their hemisphere, $\mathcal{M}_h$ (left or right), $\mathcal{E}(\mathcal{M}_h|\mathcal{M}_f)$ could be used to highlight nodes whose diversity of connectivity between left and right hemispheres depends on which functional module they connect to.

3. Conclusion

We have introduced an information-theoretic formulation of nodal connection diversity in complex networks, incorporating results from the broader literature on quantitative diversity indices that builds on a prior introduction of the Shannon entropy formulation of participation coefficient [7]. Quantifying

connection diversity as the average uncertainty in the module label of a connected neighboring node, termed participation entropy, $\mathcal{E}_i$, has mathematically favorable properties over the more commonly used participation coefficient. Using a probabilistic formulation of the two measures, we show that the participation coefficient is a first-order approximation to the participation entropy (as per the relationship of the underlying measures of diversity [13]). Using the additivity of participation entropy with respect to chaining probability distributions for multiple module sets, we introduce new ways of measuring connection diversity for cases where nodes are labeled from multiple label sets, defining joint and conditional participation entropy.

Future work may build on the theoretical foundations laid here, including applying the new measures to data. This will require developing statistical significance tests against appropriate null distributions. For example, analysis on the conditional participation entropy of a node, $\mathcal{E}_i(\mathcal{M}|\mathcal{N})$ (i.e. the diversity of connectivity across modules $\mathcal{M}$ given the labeling $\mathcal{N}$) requires comparison to an appropriate null hypothesis. One choice of null hypothesis is that node i connects randomly with respect to $\mathcal{M}$, while preserving the distribution of connections over $\mathcal{N}$ (which could be sampled from numerically). Future work could also explore alternative probabilistic formulations of connection diversity that may differently account for module size [21, 22]. In summary, the new theory introduced here enables practical new ways of understanding and quantifying more subtle types of nodal connection patterns in complex networks.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Joseph T Lizier https://orcid.org/0000-0002-9910-8972
Ben D Fulcher https://orcid.org/0000-0002-3003-4055

References

[1] Guimerà R and Amaral L A N 2005 Cartography of complex networks: modules and universal roles *J. Stat. Mech.* **2005** 02001
[2] Guimerà R and Amaral L A N 2005 Functional cartography of complex metabolic networks *Nature* **433** 895
[3] Guimerà R, Sales-Pardo M and Amaral L A 2007 Classes of complex networks defined by role-to-role connectivity profiles *Nat. Phys.* **3** 63
[4] Sporns O, Honey C J and Kötter R 2007 Identification and classification of hubs in brain networks *PLoS One* **2** e1049
[5] Rizkallah J *et al* 2019 Decreased integration of EEG source-space networks in disorders of consciousness *NeuroImage* **23** 101841
[6] Shibata N, Kajikawa Y, Takeda Y and Matsushima K 2008 Detecting emerging research fronts based on topological measures in citation networks of scientific publications *Technovation* **28** 758
[7] Rubinov M and Sporns O 2011 Weight-conserving characterization of complex functional brain networks *NeuroImage* **56** 2068
[8] Peet R 2003 The measurement of species diversity *Annu. Rev. Ecol. Syst.* **5** 285
[9] Daly A, Baetens J and Baets B D 2018 Ecological diversity: measuring the unmeasurable *Mathematics* **6** 119
[10] Havrda J and Charvát F 1967 Quantification method of classification processes. Concept of structural *a*-entropy *Kybernetika* **03** 30
[11] Keylock C J 2005 Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy *Oikos* **109** 203
[12] Vajda I and Zvárová J 2007 On generalized entropies, Bayesian decisions and statistical diversity *Kybernetika* **43** 675
[13] Zhang Z and Grabchak M 2014 Entropic representation and estimation of diversity indices *J. Nonparametric Stat.* **28** 563–575
[14] Fortunato S and Hric D 2016 Community detection in networks: a user guide *Phys. Rep.* **659** 1
[15] Simpson E H 1949 Measurement of diversity *Nature* **163** 688
[16] Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 379
[17] Cover T M and Thomas J A 2005 *Elements of Information Theory* (Wiley)
[18] Rubinov M and Sporns O 2010 Complex network measures of brain connectivity: uses and interpretations *NeuroImage* **52** 1059
[19] Ash R B 1965 *Information Theory* (Dover Publications Inc.)
[20] Khinchin A Y 2013 *Mathematical Foundations of Information Theory* (Courier Corporation)
[21] Pedersen M, Omidvarnia A, Shine J M, Jackson G D and Zalesky A 2020 Reducing the influence of intramodular connectivity in participation coefficient *Netw. Neurosci.* **4** 416
[22] Ruiz Vargas E and Wahl L M 2014 The gateway coefficient: a novel metric for identifying critical connections in modular networks *Eur. Phys. J.* B **87** 161