



HAL
open science

Doubly Stochastic Scaling Unifies Community Detection

Luce Le Gorrec, Sandrine Mouysset, Daniel Ruiz

► **To cite this version:**

Luce Le Gorrec, Sandrine Mouysset, Daniel Ruiz. Doubly Stochastic Scaling Unifies Community Detection. *Neurocomputing*, 2022, 10.1016/j.neucom.2022.06.090 . hal-03633062v2

HAL Id: hal-03633062

<https://hal.science/hal-03633062v2>

Submitted on 22 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Doubly Stochastic Scaling Unifies Community Detection

Luce le Gorrec^{a,*}, Sandrine Mouysset^b, Daniel Ruiz^b

^a*Department of Mathematics and Statistics, University of Strathclyde, Glasgow, United Kingdom*

^b*Université de Toulouse, Toulouse, France*

Abstract

Graph partitioning, or community detection, has been widely investigated in network science. Yet, the correct community structure on a given network is essentially data-driven. Thus, instead of a formal definition, diverse measures have been conceived to capture intuitive desirable properties shared by most of the community structures. In this work, we propose a preprocessing based on a doubly stochastic scaling of network adjacency matrices, to highlight these desirable properties. By investigating a range of community detection measures, and carefully generalising them to doubly stochastic graphs, we show that such a scaling unifies a whole category of these measures—namely, the so-called linear criteria—onto two unique measures to set up. Finally, to help practitioners setting up these measures, we provide an extensive numerical comparison of the capacity of these measures to uncover community structures within stochastic block models, using the Louvain algorithm.

Keywords: Network Analysis, Community Detection, Graph Partitioning Measures, Doubly Stochastic Scaling.

1. Introduction

By mapping local-level elementary interactions between data, networks provide a powerful template that enables one to analyse emergent behaviours in complex systems, such as biological systems, social networks, etc. [1, Chap.5].
5 Hence, these last decades, analysis of complex networks has been at the core of several research works [2]. One aspect has gained a lot of attention: the problem of graph partitioning, also called community detection [3, 4, 1, Chap.21].
Defining a network as a set of entities (called nodes or vertices) connected by interactions (called links or edges), the aim of community detection is to partition
10 the set of the nodes into groups of nodes that are similar or strongly related.

*Corresponding Author.

Email address: `luce.le-gorrec@strath.ac.uk` (Luce le Gorrec)

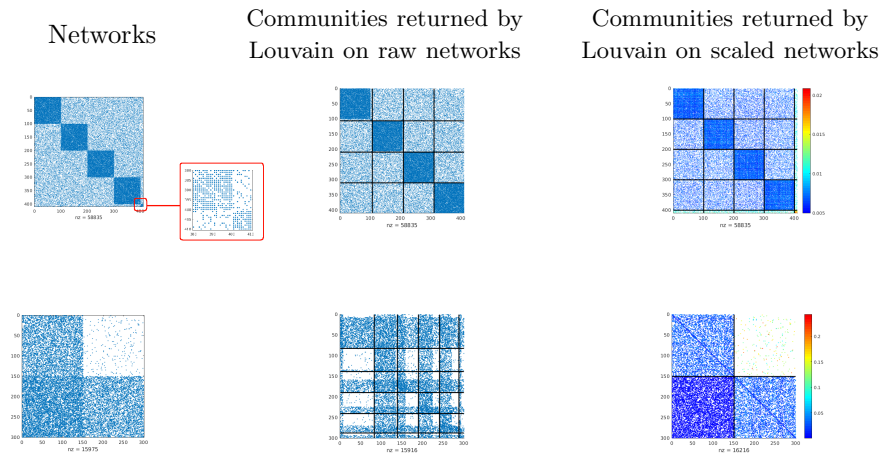


Figure 1: **Left:** Adjacency matrices of networks with community structures. **Middle:** The Louvain algorithm cannot detect the smallest community (top matrix, the smallest community highlighted in the red square); and it is unable to detect the two communities connected in an imbalanced fashion (bottom). **Right:** After scaling, Louvain can detect small communities in presence of larger ones (top); and it can detect the community structure when there is an imbalance in the flows of edges (bottom).

In real-world applications, the rightful community structure depends on the network. For this reason, there exists no formal definition of a community structure since it is always possible to find a community structure that contradicts the definition. However, it is generally admitted that community structures share similar properties: a community should be a group of nodes densely connected, and sparsely connected to the rest of the graph—see Table 1.1 from [5]. Thus, a number of measures that capture these properties has been designed to assess the quality of a community structure proposed on a network, e.g. [6, 7, 8]. Optimising such measures is generally a NP-complete problem [9, 3, 5], thus approximation algorithms have been proposed that perform community detection by approximating the “best” community structure. The most famous is probably the Louvain algorithm [10] that aims to maximise the so-called Newman-Girvan modularity [6]. Because of its simplicity, its accuracy in detecting communities, and its efficiency in terms of computational cost [11], it has been one of the most widely-used community detection algorithms for more than 10 years. But there are communities, very intuitive and yet poorly detected by algorithms in general, that even Louvain is unable to detect: 1) Small communities in large networks are generally missed—this is typically the so-called resolution limit [12]. 2) In directed networks, flow-based communities are usually not detected in presence of an imbalance of the edges leaving and entering these communities. Points 1) and 2) are illustrated in the middle panels of fig. 1, where the results of Louvain algorithm applied to two toy networks exhibiting such community structures are displayed.

The aim of this study is to investigate the potential of matrix scaling as a

35 preprocessing for community detection. Our contributions are three-folds:

- We propose a preprocessing based on the so-called doubly stochastic scaling, to increase the detectability of communities, in particular those usually hardly detectable as illustrated in fig. 1.
- By extending several graph partitioning measures to weighted graphs, in particular doubly stochastic graphs, we show that the proposed preprocessing unifies these measures onto two unique measures to set up.
- We conduct extensive comparisons of the capacity of these measures to uncover community structures within stochastic block models (SBMs), which provides guidance for customising them.

45 The paper is organised as follows: Section 2 lists the definitions and notations to be used through the paper. Section 3 gives an overview of related work. Section 4 presents the method: we introduce the doubly stochastic scaling (section 4.1) and detail the proposed preprocessing (section 4.2), showing its potential on toy examples and a real-world network (section 4.3). In section 5, we discuss the generalisation of six graph partitioning measures to weighted graphs, in particular doubly stochastic ones. Section 6 compares these measures, first theoretically in section 6.1, then experimentally in section 6.2. We finally conclude the study and discuss future work in section 7.

2. Definitions and Notations

55 In this section, we present some definitions and notations to be used through the study. Basic mathematical objects are listed in table 1.

Object	Typoface	Examples
Unweighted graph	2-element tuple	$G = (V, E)$
Weighted graph	3-element tuple	$G = (V, E, \Omega)$
Edge in a directed graph	Tuple of nodes	(u, v)
Edge in an undirected graph	Curly brackets of nodes	$\{u, v\}$
Matrix	Bold capital letter	A , S
Matrix entry	Letter with subscripts	$a_{i,j}$
Matrix of 1s	J	J
Identity matrix	I	I
Transpose of a matrix	\cdot^T	A ^T
Vector	Bold minuscule letter	u , x
Vector entry	Parentheses on a vector	u (i)
Vector of 1s	e	e
Diagonal matrix from a vector	$\mathcal{D}(\cdot)$	$\mathcal{D}(\mathbf{u})$, $\mathcal{D}(\mathbf{e}) = \mathbf{I}$
Cardinality of a set	$ \cdot $	$ S $

Table 1: Typography of mathematical objects.

Graphs. In this study, we investigate networks (also called graphs) that can be weighted or not. Except when stated otherwise, networks are undirected. For a network $G = (V, E, \Omega)$, V is the set of nodes, $E \subset V \times V$ the set of edges, and the function

$$\Omega : \begin{array}{ccc} E & \rightarrow & \mathbb{R}_+ \\ \{u, v\} & \mapsto & \omega(\{u, v\}) \end{array}$$

provides the weights of edges. We limit the study to graphs that are positively weighted. To simplify notations, we assume that graphs have integer nodes, i.e. $V = \{1, \dots, n\}$.

60 When there is no possible confusion about the network, letters n and m denote the number of nodes and the total weight of edges respectively, that is $n = |V|$ and $m = \sum_{\{u, v\} \in E} \omega(\{u, v\})$. The degree of a node u is defined as $d_u = \sum_{v: \{u, v\} \in E} \omega(\{u, v\})$. If $\exists \delta \in \mathbb{R} : \forall u \in V, d_u = \delta$, the graph is said to be δ -regular. We denote by simple graphs the unweighted undirected networks without self-loop—i.e. $\forall u \in V, \{u, u\} \notin E$.

Adjacency Matrices. A (directed) graph $G = (V, E, \Omega)$ can be represented by its adjacency matrix, that is a matrix $\mathbf{A} \in \mathbb{R}_+^{n \times n}$ where

$$a_{i,j} = \begin{cases} \omega((i, j)) & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases}.$$

Conversely, given a matrix $\mathbf{A} \in \mathbb{R}_+^{n \times n}$, we call the adjacency graph of \mathbf{A} the graph whose \mathbf{A} is the adjacency matrix.

For undirected graphs, when the adjacency graph of \mathbf{A} has no self-loop, then $2m = \sum_{i=1}^n \sum_{j=1}^n a_{i,j} = \mathbf{e}^T \mathbf{A} \mathbf{e}$. When the adjacency graph of \mathbf{A} is unweighted, we define the complementary of \mathbf{A} (denoted $\overline{\mathbf{A}}$) as the matrix in $\mathbb{R}_+^{n \times n}$ such that

$$\overline{a}_{i,j} = \begin{cases} 1 & \text{if } \{i, j\} \notin E \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

that is $\overline{\mathbf{A}} = \mathbf{J} - \mathbf{A}$.

Community Structures. Given a graph $G = (V, E)$, a community structure is a partitioning of the set of nodes V , that is a set of subsets of V : $\mathcal{C} = \{C_t\}_{t=1..k}$ such that $\bigcup_{t=1}^k C_t = V$ and $\forall t \neq s, C_t \cap C_s = \emptyset$. This community structure can be represented as an equivalence relation \mathcal{X} on $V \times V$ such that

$$u\mathcal{X}v \iff \exists t \in \{1, \dots, k\} : u, v \in C_t.$$

It can also be represented as a matrix $\mathbf{X} \in \{0, 1\}^{n \times n}$ such that

$$x_{i,j} = \begin{cases} 1 & \text{if } i\mathcal{X}j \\ 0 & \text{otherwise.} \end{cases}$$

A matrix $\mathbf{X} \in \{0, 1\}^{n \times n}$ encodes an equivalence relation \mathcal{X} (and hence a community structure) if and only if:

$$\begin{aligned} \forall i \in \{1, \dots, n\}, & \quad x_{i,i} = 1 & (a) \\ \forall i, j \in \{1, \dots, n\}, & \quad x_{i,j} = x_{j,i} & (b) \\ \forall i, j, k \in \{1, \dots, n\}, & \quad x_{i,k} + x_{j,k} - x_{i,j} \leq 1 & (c) \end{aligned}$$

where (a), (b), (c) indicate respectively the reflexivity, the symmetry and the transitivity of the equivalence relation represented by \mathbf{X} [13]. We denote by $Eq(n)$ the set of the equivalence relations on a set V such that $|V| = n$. That is, we write $\mathbf{X} \in Eq(n)$ when a matrix $\mathbf{X} \in \{0, 1\}^{n \times n}$ verifies (a), (b), (c), and $\mathcal{X} \in Eq(n)$ for an equivalence relation defined on the set V . For any $\mathbf{X} \in Eq(n)$, its complementary is defined by $\overline{\mathbf{X}} = \mathbf{J} - \mathbf{X}$.

Double Stochasticity. In the following, we specifically focus on networks that are doubly stochastic, that is such that their adjacency matrices have their row and column sums equal to 1. Formally, a (directed) network $G = (V, E, \Omega)$ is said to be doubly stochastic if its adjacency matrix $\mathbf{S} \in \mathbb{R}_+^{n \times n}$ is doubly stochastic, that is

$$\begin{cases} \mathbf{S}\mathbf{e} = \mathbf{e} \\ \mathbf{S}^T\mathbf{e} = \mathbf{e}. \end{cases} \quad (2)$$

We remark that doubly stochastic graphs are 1-regular graphs.

In this study, we preprocess graphs so that they (or equivalently their adjacency matrices) are doubly stochastic. Transforming a matrix $\mathbf{A} \in \mathbb{R}_+^{n \times n}$ onto a doubly stochastic matrix is an operation called “scaling \mathbf{A} onto its doubly stochastic form”. One achieves this by finding two vectors $\mathbf{r}, \mathbf{c} \in \mathbb{R}_+^n$ such that

$$\begin{cases} \mathcal{D}(\mathbf{r})\mathbf{A}\mathcal{D}(\mathbf{c})\mathbf{e} & = \mathbf{e} \\ \mathcal{D}(\mathbf{c})\mathbf{A}^T\mathcal{D}(\mathbf{r})\mathbf{e} & = \mathbf{e}. \end{cases} \quad (3)$$

The matrix $\mathbf{S} = \mathcal{D}(\mathbf{r})\mathbf{A}\mathcal{D}(\mathbf{c})$ is called the doubly stochastic scaling of \mathbf{A} , and vectors \mathbf{r} and \mathbf{c} are called the scaling factors. The existence of a doubly stochastic scaling is non-straightforward and is detailed in section 4.1.

3. Related Work

Doubly Stochastic Scaling for Community Detection. In this study, we design a preprocessing for community detection, based on doubly stochastic scaling. This scaling has already been used in the context of community detection. It is the first stage of the algorithm from [14] that partitions migration networks. However, the rationales for scaling in [14] (invariance of relative odds and approximation of maximum entropy) greatly differ from ours. Also in [15], doubly stochastic scaling is used as a preprocessing step for a spectral algorithm. Furthermore, in [16], authors aim to partition a dataset by finding the doubly stochastic matrix that best approximates the dataset similarity matrix. Finally,

in [17], authors use doubly stochastic scaling to perform co-clustering, as scal-
ing factors should approximate the joint densities between the random variables
90 inferring data, and the random variables inferring partitions. We remark that
all these studies use the doubly stochastic scaling for a very specific purpose:
achieving uniform marginals in the flow table [14], obtaining staircase-like singu-
lar vectors [15] or scaling factors [17], or approximating a similarity matrix [16].
95 On the other hand, the method proposed here is a wider-purpose preprocessing
that can be used prior to any community detection method.

Community Detection Measures. This study investigates a bunch of measures
designed to assess the quality of a community structure on a network, based
on the network structural properties. Nodes within a community are supposed
100 to be densely connected, while being loosely connected to nodes outside their
community. Different manners to define “densely” and/or “loosely” lead to dif-
ferent measures. They can be subdivided into three families. Measures based
on density, such as Newman-Girvan modularity [6] or coverage [18], define a
community as a group of nodes with a high density of edges. Measures based
105 on sparsity also exist, that consider that the amount of edges between two com-
munities must be low (e.g. conductance, expansion [19], or normalised cut [7]).
Some measures are a mixture of density and sparsity, such as LambdaCC [5] or
Balanced modularity [13]. Given a dynamic process defined on the graph edges
(e.g. a random walk), measures from the third kind consider a community as
110 a group of nodes from which the process struggle to escape (such as the Map
equation [8], the Markov stability [20], or the community distance from [21]).
Remark that some measures based on *a priori* hypotheses about the ground
truth community structure exist, such as the likelihood from SBM-based tech-
niques [22], the likelihood of preserving node neighbourhoods in *node2vec* [23],
115 or the cross-entropy error over nodes with known label in Graph Convolutional
Networks [24]. But they are beyond the scope of this study.

Most of the measures investigated here are listed in [25] to be used in the
Louvain algorithm. This requires them to be defined for graphs with integer
weights, and the measures are thus extended to such graphs when needed. Since
120 generalisation is not the purpose of [25], this is done straightforwardly and does
not always fit the philosophy of the initial measures, as shown in section 5.

4. Doubly Stochastic Scaling Preprocessing

In this section we describe and discuss the preprocessing that we propose
for community detection, that relies on a doubly stochastic scaling of the graph
125 adjacency matrix. Not every square matrix is amenable to a doubly stochastic
matrix. We thus first provide the conditions for such a scaling to exist, and
discuss the relations with graph connectivity. We then present the proposed
preprocessing, and discuss its impact on some community structures.

4.1. *Doubly Stochastic Scaling and Graph Connectivity*

130 *The Sinkhorn-Knopp Theorem.* Given a square matrix $\mathbf{A} \in \mathbb{R}_+^{n \times n}$, it is not always possible to find two vectors $\mathbf{r}, \mathbf{c} \in \mathbb{R}_+^n$ such that eq. (3) is verified. In order for such a scaling to exist, the pattern of \mathbf{A} —i.e. the positions of its nonzero entries—must respect certain conditions, which are provided by the so-called Sinkhorn-Knopp theorem [26]. In order to introduce this theorem, we
 135 first provide two definitions about the pattern of a matrix on which it relies. These definitions can be found in [27].

Definition 1. Bi-Irreducibility. A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is called bi-irreducible if there is no pair of permutation matrices \mathbf{R}, \mathbf{Q} such that

$$\mathbf{RAQ} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{1,2} \\ 0 & \mathbf{A}_2 \end{bmatrix}$$

with $\mathbf{A}_1, \mathbf{A}_2$ two square and non empty matrices.

This definition implies that \mathbf{A} is not amenable to a block triangular matrix by independent permutations of its rows and its columns.

Definition 2. Total Support. A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to have a total support if every nonzero entry lies on a strictly positive diagonal. One characterisation of this definition proposed in [27] is that there are two permutation matrices \mathbf{R}, \mathbf{Q} such that

$$\mathbf{RAQ} = \begin{bmatrix} \mathbf{A}_1 & & \\ & \ddots & \\ & & \mathbf{A}_k \end{bmatrix}$$

140 with $\mathbf{A}_1, \dots, \mathbf{A}_k$ bi-irreducible matrices.

We can now enunciate the Sinkhorn-Knopp theorem [26].

Theorem 1. Sinkhorn-Knopp. Given a matrix $\mathbf{A} \in \mathbb{R}_+^{n \times n}$, a necessary and sufficient condition that there exists a doubly stochastic matrix $\mathbf{S} = \mathcal{D}(\mathbf{r})\mathbf{A}\mathcal{D}(\mathbf{c})$ with $\mathbf{r}, \mathbf{c} \in \mathbb{R}_+^{*n}$, is that \mathbf{A} has a total support. If \mathbf{S} exists then it is unique.
 145 Vectors \mathbf{r} and \mathbf{c} are also unique up to a scalar multiple if and only if \mathbf{A} is bi-irreducible.

Relations with the Connectivity of the Adjacency Graph. We now introduce the definition of irreducibility, that draws a link between the connectivity of a network and the pattern of its adjacency matrix.

Definition 3. Irreducibility. A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is called irreducible if there is no permutation matrix \mathbf{Q} such that

$$\mathbf{QAQ}^T = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{1,2} \\ 0 & \mathbf{A}_2 \end{bmatrix}$$

150 with $\mathbf{A}_1, \mathbf{A}_2$ square and non empty. A characterisation of irreducible matrices from [28] is that they are the adjacency matrices of strongly connected graph.

Algorithm 1: Preprocessing Undirected Graphs

Data: A symmetric matrix $\mathbf{A} \in \mathbb{R}_+^{n \times n}$.

Result: A doubly stochastic matrix $\mathbf{S} \in \mathbb{R}_+^{n \times n}$.

1 $\varepsilon \leftarrow 10^{-8} \times \min_{i,j:a_{i,j} \neq 0} (a_{i,j});$

2 $\mathbf{S} = \text{symscalone}(\mathbf{A} + \varepsilon \mathbf{I});$

Algorithm 2: Preprocessing Directed Graphs

Data: A non-symmetric matrix $\mathbf{A} \in \mathbb{R}_+^{n \times n}$.

Result: A doubly stochastic matrix $\mathbf{S} \in \mathbb{R}_+^{p \times p}$, with $p \leq n$.

1 $\mathbf{B} \leftarrow$ largest block returned by $\text{dmperm}(\mathbf{A} + \mathbf{I});$

2 $\varepsilon \leftarrow 10^{-8} \times \min_{i,j:b_{i,j} \neq 0} (b_{i,j});$

3 $\mathbf{S} = \text{RAS}(\mathbf{B} + \varepsilon \mathbf{I});$

Every bi-irreducible matrix is also irreducible. Reciprocally, if a matrix is irreducible with its diagonal zero-free— $a_{i,i} \neq 0, \forall i \in \{1, \dots, n\}$ —, then this matrix is bi-irreducible (easily proven by applying the algorithm from [29] to such a matrix). Since definition 3 states that irreducible matrices are adjacency matrices of strongly connected graphs, then the adjacency matrix of every strongly connected graph can be made bi-irreducible (thus scalable) by ensuring that its diagonal is strictly positive (e.g. by adding a positive diagonal matrix to the adjacency matrix, which is equivalent to adding self-loop to the graph).

160 **Remark 1.** *For an undirected graph, adding a diagonal matrix to its adjacency matrix is sufficient to make it scalable onto a doubly stochastic graph, whatever its connectivity. Indeed, every symmetric matrix with zero-free diagonal has a total support (Lemma 3.3 from [30]).*

165 *For a directed graph, each strongly connected component must be scaled and partitioned apart¹. These components can be found by applying the Dulmage-Mendelsohn decomposition on the graph adjacency matrix whose diagonal has been made zero-free [29].*

4.2. The Preprocessing

170 We propose to apply a doubly stochastic scaling on networks as a preprocessing for community detection. As discussed in section 4.1, some requirements have to be fulfilled to ensure that the network can be scaled, which depend on whether the graph is directed. The steps to follow to scale a matrix $\mathbf{A} \in \mathbb{R}_+^{n \times n}$ are described in algo. 1 if \mathbf{A} is the adjacency matrix of an undirected graph, respectively in algo. 2 if the adjacency graph of \mathbf{A} is directed.

¹This implies that nodes from different strongly connected components cannot end within a same community.

175 In algo. 1, *symscalone* is the method from [31] that can compute a doubly stochastic scaling of a general square matrix with total support. It is well designed for symmetric matrices in particular, because it preserves matrix symmetry. In algo. 2, *dmperm* is the Dulmage-Mendelsohn decomposition, evoked in remark 1. When applied to $\mathbf{A} + \mathbf{I}$, it returns the strongly connected components of the adjacency graph of \mathbf{A} . The largest component is then scaled using 180 the so-called *RAS* or Sinkhorn-Knopp Algorithm [26, 30]. For both directed and undirected networks, the adjacency graph of the doubly stochastic matrix \mathbf{S} returned by the algorithm is the one on which communities are then detected. For directed network, it means that only the largest strongly connected component is partitioned. However, this is straightforward to extend to the whole 185 graph, by scaling and partitioning each component in turn.

For both algorithms, it is necessary to add entries in the diagonal of the matrix to scale, to ensure that conditions from theorem 1 are verified. We remark that adding diagonal elements leaves the community structure intact, 190 as the community structure of a graph is linked to the diagonal block structure of its adjacency matrix, which is not impacted by its diagonal entries. In both versions of the preprocessing, we choose to add very small entries (10^{-8} times the matrix smallest entry) to impact as little as possible the numerical values in the final scaling. This is an empirical choice which is not theoretically justified, 195 and it would be interesting to analyse how these diagonal entries impact the final scaling. We leave this analysis to further work.

4.3. Impact of the Preprocessing on Synthetic and Real-World Data

Rationales on Toy Examples. Our intuition that the doubly stochastic scaling may improve community detection comes from the two toy examples from fig. 1, 200 used in section 1 to illustrate the difficulty of detecting some community structures. First, doubly stochastic scaling leverages the weight of the edges in small and large communities. One may think at a trivial example, where a simple graph is composed of two disjoint communities of different size $n_1 > n_2$, such that the probability for two nodes in a same community to be linked is equal 205 to p_{in} , for both communities. Then, in average, a node in the large community shares more links with nodes from its community than a node in the small community ($p_{in} \times n_1 > p_{in} \times n_2$). This is not true anymore if we look at the doubly stochastic scaling of the adjacency matrix. In this case, every node in both communities shares strictly the same amount of edges with nodes from its 210 community, that is 1 by definition of the doubly stochastic scaling.

Secondly, doubly stochastic scaling can rationally be expected to mitigate against an existing imbalance in the direction of edges, because of its so-called vanishing effect. To understand it, we explain the behaviour of doubly stochastic scaling on $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. This matrix has no total support. Thus, according 215 to theorem 1, it is not amenable to a doubly stochastic form. Nevertheless, doubly stochastic scaling algorithms provide scaling factors \mathbf{r} and \mathbf{c} that tend towards $(0, +\infty)^T$ and $(+\infty, 0)^T$ respectively, so that the doubly stochastic

scaling of \mathbf{A} tends towards $\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, in which the off-diagonal element had vanished [30, 26].

220 As a matter of fact, the proposed preprocessing indeed improves the detectability of community structures of the toy examples from fig. 1: Louvain algorithm applied directly on the graphs fails to detect their structures; on the other hand, when applied on the preprocessed graphs, it returns their ground truth community structures, as shown in the right panels of fig. 1.

225 *Food Web of Florida Bay.* Here we observe the impact of the proposed preprocessing on the network of trophic dynamics within Florida Bay. In this directed network, a node is a compartment and an edge indicates carbon exchanges—roughly, an edge from node a to node b means that species in compartment a are eaten by species in compartment b . The network contains 128 compartments, that can be divided onto 9 types according to [32], namely Phytoplankton producers, Seagrass and seagrass roots, Microfauna, Macroinvertebrates, Fishes, Birds, Reptiles, Mammals, and Detritus. According to [33], this partitioning into types corresponds to the network underlying community structure. The largest strongly connected component contains 103 compartments: 11 are Microfauna, 22 Macroinvertebrates, 47 Fishes, 16 Birds, 3 Reptiles, 2 Mammals, and 2 are Detritus.

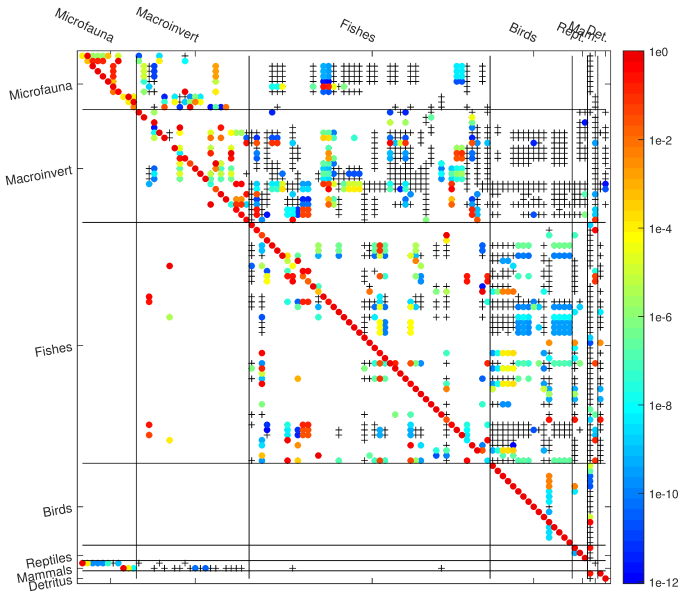


Figure 2: Output of algo. 2 on the Florida Bay network, and its ground truth partitioning. Black '+'s indicate nonzero entries with numerical values below 10^{-12} .

The matrix $\mathbf{S} \in \mathbb{R}_+^{p \times p}$ returned by algo. 2 is illustrated in fig. 2. The ground truth community structure is indicated by the black lines. Because numeri-

cal values range from 1 to 10^{-82} , only entries higher than 10^{-12} are plotted. Nonzero entries below this threshold are shown by black '+'s. From fig. 2 we observe that the preprocessing clearly tends to make vanish the edges between communities. This is highlighted by the high density of black '+'s in the off-diagonal blocks, meaning that numerous entries in \mathbf{S} off-diagonal blocks have a value that falls below 10^{-12} . To assess the extent to which the preprocessing indeed sharpens the network community structure, we need to compare the consistency of these communities on both the raw and the scaled networks. We base this comparison on the concepts defined below.

Definition 4. Community Structure Consistency. Assuming a matrix $\mathbf{M} \in \mathbb{R}_+^{p \times p}$ and \mathcal{C} its ground truth community structure. The level to which a node $u \in \{1, \dots, p\}$ belongs to a community $C \in \mathcal{C}$ is assessed² by $\varphi(u, C) = \frac{\sum_{i \in C} m(u, i)}{\sum_{j=1}^p m(u, j)}$, that is the ratio between the amount of edges that node u shares with nodes in C and the degree of u . Thus, the average level to which nodes from community $C \in \mathcal{C}$ belong to community $K \in \mathcal{C}$ is

$$\Phi(C, K) = \frac{1}{|C|} \sum_{u \in C} \varphi(u, K), \quad (4)$$

and matrix Φ provides a view of the community structure consistency.

Clearly, the higher the reflexive values of Φ , the more consistent the community structure. We compute the values of Φ for two matrices that are symmetrisations of the raw and preprocessed directed networks, namely $\mathbf{B} + \mathbf{B}^T$, where \mathbf{B} is the adjacency matrix of the raw network largest strongly connected component; and $\tilde{\mathbf{S}} + \tilde{\mathbf{S}}^T$, where $\tilde{\mathbf{S}}$ is the matrix \mathbf{S} with its diagonal values put to 0. We remove the diagonal because most of \mathbf{S} diagonal entries are scaled close to 1 (whereas they are initially very small). Thus, keeping the diagonal provides spuriously high values for $\Phi(C, C), \forall C \in \mathcal{C}$, whatever the community structure.

These values of Φ are displayed in fig. 3. The three last communities that contain no more than 3 nodes are missed by both the raw and the preprocessed matrices. And looking at the structure of these communities restricted to the analysed component in fig. 2, it is indeed not possible to consider them as standalone communities, without having been told so. The community corresponding to the Birds tends to be merged with Fishes by both raw and preprocessed networks. This is also in line with what can be observed from fig. 2. Finally, the three non trivial communities corresponding to Microfauna, Macroinvertebrates and Fishes, are assessed as fairly consistent in the preprocessed network (lowest reflexive value of Φ is 0.58, highest non reflexive value is 0.26). On the other hand, in the raw component, Microfauna and Macroinvertebrates are missed and

² φ is actually the opposite of the so-called mixing parameter introduced in section 6.2.

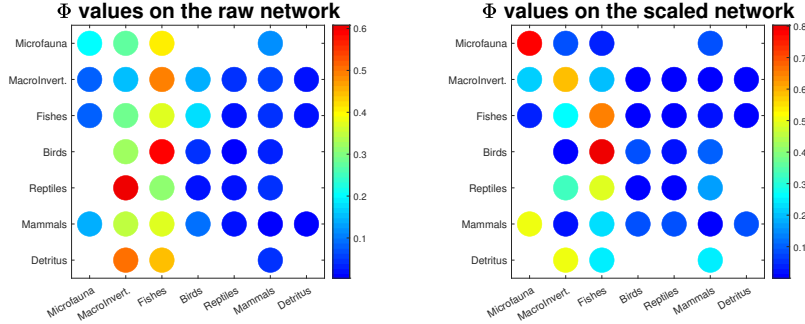


Figure 3: Φ values for the raw network (left) and the scaled network without self-loop (right). The three first communities are much more consistent in the scaled network (diagonal Φ values are dominant) than in the raw one.

merged with Fishes. We also remark that the preprocessing has more impact on the consistency of smaller communities—reflexive Φ values are 3.74 times higher in the preprocessed network than in the raw one for Microfauna and Macroin-
 270 vertebrates, 1.73 for Fishes. These observations illustrate the potential of the proposed preprocessing to increase the detectability of community structures within networks with an imbalance in edge direction between communities, as well as small-size communities.

275 5. Generalisation of some Graph Partitioning Measures to Weighted Networks

In this section, we investigate six measures—or criteria—that assess the quality of a community structure on a graph, namely: the Newman-Girvan modularity [6], Balanced modularity [13], the Deviation to Uniformity criterion [34, Chap.5.2.6], the Deviation to Indetermination criterion [13], the Zahn
 280 criterion [35] and the Correlation Clustering criterion [36].

This list of graph partitioning measures is not exhaustive. These measures are actually the linear criteria from [34]. Formally, denoting by F a criterion that assesses the quality of a community structure on a graph represented by its adjacency matrix, F is a linear criterion [25] if it can be written as

$$\begin{aligned}
 F : \mathbb{R}^{n \times n} \times Eq(n) &\longrightarrow \mathbb{R} \\
 (\mathbf{A}, \mathbf{X}) &\mapsto \sum_{i=1}^n \sum_{j=1}^n \varphi(a_{i,j}) x_{i,j} + K, \quad (5)
 \end{aligned}$$

where \mathbf{A} and \mathbf{X} are respectively a graph adjacency matrix and a community structure, $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is a function and K is some constant scalar.

For each criterion, we address three points:

- 285 • We explain quickly the measure background, that is how it works and why it assesses community structures, as well as its formulation.

- Most of these measures are initially designed for unweighted networks, and some have been generalised to weighted graphs afterward. When such a generalisation exists, we may either use it or derive another one that we find more suitable for the community structure detection on doubly stochastic graphs. We hence discuss the measure generalisation to weighted graphs and especially to doubly stochastic ones.
- We provide a reduced form for the problem of finding the best community structure on a graph represented by its adjacency matrix \mathbf{A} using a criterion F . Namely, this problem is expressed as

$$\mathbf{X}^* = \underset{\mathbf{X} \in Eq(n)}{\operatorname{argmax}} \left(F(\mathbf{A}, \mathbf{X}) = \sum_{i,j} (\phi(a_{i,j}) - \bar{\phi}(a_{i,j})) x_{i,j} \right), \quad (6)$$

where ϕ and $\bar{\phi}$ are two functions in \mathbb{R}_+ respectively called the positive and negative agreements, as in [34]. This reduced form allows us to compare the criteria in section 6.

5.1. Newman-Girvan modularity

Principle. The Newman-Girvan modularity introduced in [6] is the most famous graph partitioning measure. The idea behind this criterion is that a community structure in a network actually characterises the property of assortative mixing in this network [37]. The assortative mixing is the tendency of similar nodes to draw connection amongst themselves instead than with dissimilar nodes: as an example, in a social network, people who speak the same language or have similar sociological background have more chance to be friends. Hence, given an assortative network, a good community structure is one such that the fraction of edges that connect nodes in a same community is high.

However, this notion cannot be used as a standalone. Indeed, the trivial structure that brings all the nodes in a same community always maximises this fraction of edges. Thus, to derive the modularity, Newman and Girvan also assume that random graphs do not exhibit a community structure [6]. The modularity is hence designed to compare the fraction of intra-community edges in a network with the expected fraction of intra-community edges in random graphs with the same degree sequence than the initial graph (i.e. generated by the configuration model). In the configuration model with degree sequence $\{d_1, \dots, d_n\}$, the probability of an edge between two nodes i and j can be approximated by $d_i d_j / 2m$. The modularity is thus defined as

$$F_{NG}(\mathbf{A}, \mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{i \in C} \sum_{j \in C} \left(\frac{a_{i,j}}{2m} - \frac{d_i d_j / 2m}{2m} \right),$$

with \mathbf{A} the adjacency matrix of the network, and \mathcal{C} a community structure. In turn, this can be re-written (as in [38])

$$F_{NG}(\mathbf{A}, \mathbf{X}) = \frac{1}{2m} \sum_{i,j} \left(a_{i,j} - \frac{d_i d_j}{2m} \right) x_{i,j}, \quad (7)$$

with $\mathbf{X} \in Eq(n)$ the matrix representation of the community structure \mathcal{C} .

Generalisation. The initial Newman-Girvan modularity from [6] is designed for unweighted graphs only. In [38], Newman proposes two steps to generalise modularity to weighted graphs. First, he investigates multi-graphs, that are simple networks in which two vertices can share more than one simple edge, as in fig. 4. Newman generalises some basics from simple networks to multi-

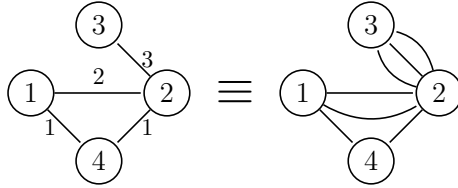


Figure 4: A weighted graph with positive integer edges (left) and the corresponding multi-graph (right).

graphs to derive an adapted modularity. Namely, let $\mathbf{A} \in \mathbb{N}^{n \times n}$ be a multi-graph adjacency matrix: 1) The degree d_i of a vertex i in the multi-graph is the number of simple edges adjacent to i : $d_i = \sum_k a_{i,k}$. 2) The constant $2m$ becomes the sum over the degrees, that is $2m = \sum_i d_i$. With these simple adaptations of degrees and number of edges, Newman generalises the modularity by simply applying eq. (7) to multi-graphs, with $a_{i,j}$, d_i and $2m$ as defined above.

Secondly, modularity is extended from multi-graphs to positively weighted graphs with the following remark: Given a graph whose adjacency matrix can be written as $\mathbf{A} = \alpha \mathbf{N}$, with $\alpha \in \mathbb{R}_+$ and $\mathbf{N} \in \mathbb{N}^{n \times n}$, and considering $d_i = \sum_k a_{i,k}$ and $2m = \sum_i d_i$, then for any $\mathbf{X} \in Eq(n)$, the results of the formula from eq. (7) applied to \mathbf{A} and to \mathbf{N} are equal. Hence, the modularity as defined in eq. (7) can be extended to graphs for which it exists a unit flow—i.e. an α —allowing to consider them as multi-graphs.

We show in [39, Property 1] that for every square matrix whose entries are rational, a unit flow can be found, but that this is not true for any weighted matrix. However, we also provide [39, Property 2] a proof that eq. (7) can be extended to any undirected positively weighted graph³. We thus apply directly eq. (7) to doubly stochastic matrices in the following.

Reduced Form. Given an adjacency matrix \mathbf{A} , finding the best community structure in the sense of the Newman-Girvan modularity provided in eq. (7) is equivalent to maximising the function

$$\mathbf{X} \mapsto F_{NG}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \frac{d_i d_j}{2m} \right) x_{i,j}. \quad (8)$$

³Since [38], modularity has been widely applied to any positively weighted graph. However, as far as we know, [39] is the first proof that eq. (7) can be consistently applied to these graphs.

330 This provides the reduced form of eq. (6), with positive and negative agreements equal to respectively $\phi(a_{i,j}) = a_{i,j}$ and $\bar{\phi}(a_{i,j}) = d_i d_j / 2m$.

Moreover, in a doubly stochastic graph, $\forall i, d_i = 1$ and $2m = n$. Thus, for a doubly stochastic matrix \mathbf{S} , we can simplify the Newman-Girvan modularity as

$$\mathbf{X} \mapsto F_{NG}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} (s_{i,j} - \frac{1}{n}) x_{i,j}, \quad (9)$$

and the negative agreement $\bar{\phi}(s_{i,j}) = 1/n$ does not depend on i, j .

5.2. Balanced Modularity

335 *Principle.* This criterion is proposed in [13] to complete the Newman-Girvan modularity. Recall from section 5.1 that, given a simple graph $G = (V, E)$ and a community structure, the Newman-Girvan modularity compares the ratio of edges within communities—i.e. intra-community edges—with the expected ratio of intra-community edges within a random graph with the same degree sequence than G . Then, the idea behind the Balanced modularity is to also take into account the ratio of inter-community edges. In other words, the Newman-Girvan modularity considers that a good community structure on G should have a ratio of intra-community edges “higher than by chance”, whereas the Balanced modularity considers that a good community structure should have a ratio of inter-community edges lower than by chance as well.

340 To take into account the ratio of inter-community edges, the Balanced modularity focuses on the complementaries of the graph and the community structure. We can state its concept as follows. Let us denote by $\Phi : \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ the function such that

$$\Phi(\mathbf{A}, \mathbf{B}) = \sum_{i,j} \left(a_{i,j} - \frac{\left(\sum_{k=1}^n a_{i,k} \right) \left(\sum_{l=1}^n a_{j,l} \right)}{\sum_{k=1}^n \sum_{l=1}^n a_{k,l}} \right) b_{i,j},$$

which is equivalent to the Newman-Girvan modularity from eq. (8) when \mathbf{A} is an adjacency matrix and $\mathbf{B} \in Eq(n)$. Thus, given \mathbf{A} the adjacency matrix of a simple graph and \mathbf{X} a community structure, the Balanced modularity is defined as

$$F_{BM}(\mathbf{A}, \mathbf{X}) = \Phi(\mathbf{A}, \mathbf{X}) + \Phi(\bar{\mathbf{A}}, \bar{\mathbf{X}}). \quad (10)$$

An explicit formula can be derived from eq. (10) by expressing the degrees and number of edges in the complementary of a simple graph through those from the graph. It can be indeed observed from fig. 5 that

$$\begin{cases} \forall i \in \{1, \dots, n\}, \bar{d}_i = \sum_{k=1}^n \bar{a}_{i,k} = n - d_i \\ \sum_{k=1}^n \bar{d}_k = n^2 - \sum_{k=1}^n d_k = n^2 - 2m \end{cases} .$$

Hence, we can write

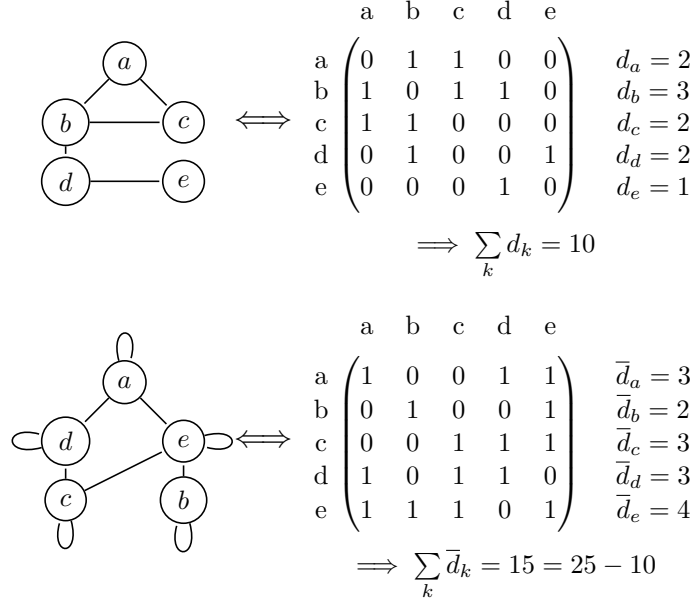


Figure 5: **Top:** A simple graph and its adjacency matrix. **Bottom:** the corresponding complementary graph and its adjacency matrix. Degree of each node is given next to the corresponding row, the sum of degrees lies below the matrices.

$$\begin{aligned}
 F_{BM}(\mathbf{A}, \mathbf{X}) = & \sum_{i,j} \left(a_{i,j} - \frac{d_i d_j}{2m} \right) x_{i,j} \\
 & + \sum_{i,j} \left(\bar{a}_{i,j} - \frac{(n - d_i)(n - d_j)}{n^2 - 2m} \right) \bar{x}_{i,j},
 \end{aligned} \tag{11}$$

345 which is the formula of the Balanced modularity provided in [13].

Generalisation. The Balanced modularity is built on the complementary of the graph, which stands for simple graphs only. However, a generalisation of this criterion to weighted graphs is proposed in [25]. It consists in stating that $\bar{a}_{i,j} = \max_{k,l} (a_{k,l}) - a_{i,j} = a_{max} - a_{i,j}$ in eq. (11). But this generalisation does not fit with the spirit of this criterion as stated in eq. (10), because it does not update \bar{d}_i and $\sum_k \bar{d}_k$ according to the new definition of $\bar{\mathbf{A}} = a_{max} \mathbf{J} - \mathbf{A}$ in the second sum of eq. (11). That is, it does not inject the weighted generalisation of $\bar{\mathbf{A}}$ in eq. (10). Hence, we propose another generalisation. Considering that, for a weighted graph defined by its adjacency matrix \mathbf{A} , the complementary of \mathbf{A} can be expressed as $\bar{\mathbf{A}} = \alpha \times \mathbf{J} - \mathbf{A}$, with α a scalar (that may depends on

A). Thus, the degrees of nodes in the complementary graph are

$$\begin{cases} \forall i \in \{1, \dots, n\}, \bar{d}_i = \sum_{k=1}^n \bar{a}_{i,k} = \alpha \times n - d_i \\ \sum_{k=1}^n \bar{d}_k = \alpha \times n^2 - \sum_{k=1}^n d_k = \alpha \times n^2 - 2m \end{cases} .$$

By injecting $\bar{\mathbf{A}}$ in eq. (10), the Balanced modularity becomes

$$\begin{aligned} F_{BM}(\mathbf{A}, \mathbf{X}) &= \sum_{i,j} \left(a_{i,j} - \frac{d_i d_j}{2m} \right) x_{i,j} \\ &+ \sum_{i,j} \left((\alpha - a_{i,j}) - \frac{(\alpha \times n - d_i)(\alpha \times n - d_j)}{\alpha \times n^2 - 2m} \right) \bar{x}_{i,j}. \end{aligned} \quad (12)$$

It remains to discuss the value of α . First, we remark that, for \mathbf{A} the adjacency matrix of any simple graph, the graph associated with $\mathbf{A} + \bar{\mathbf{A}}$ is the complete graph with self-loop: it is not possible to add any edge in this graph, that is, all edges are saturated. In a general case, given $\mathbf{A} \in \mathbb{R}_+^{n \times n}$ the adjacency matrix of some positively weighted graph, without any other
350 knowledge on the graph, we can assume that an edge is saturated if its value is a_{max} , with a_{max} as defined above. In this case, we cant state $\bar{\mathbf{A}} = a_{max} \mathbf{J} - \mathbf{A}$ as in [25]. This generalised Balanced modularity is provided by setting $\alpha = a_{max}$ in eq. (12), which is slightly different than changing $\bar{a}_{i,j}$ for $a_{max} - a_{i,j}$ in eq (11),
355 as proposed in [25]. For doubly stochastic graphs, there is an upper-bound on the weight of an edge, that is 1. Indeed, as a doubly stochastic graph is a 1-regular, positively weighted graph, no edge can have a weight above 1. Hence, 1 is the value that saturates an edge, and we can state $\alpha = 1$ in eq. (12) if the matrix is doubly stochastic.

Reduced Form. We derive the reduced form for the formula given in eq. (12), as this formula can be used for weighted and simple graphs as well (by setting $\alpha = 1$, it becomes equal to eq. (11) when \mathbf{A} represents a simple graph). Recalling that $\bar{\mathbf{X}} = \mathbf{J} - \mathbf{X}$ —or equivalently, $\forall i, j, \bar{x}_{i,j} = 1 - x_{i,j}$ —, maximising eq. (12) is equivalent to maximising

$$\begin{aligned} F_{BM}(\mathbf{A}, \mathbf{X}) &= \\ \sum_{i,j} &\left(a_{i,j} + \frac{(\alpha n - d_i)(\alpha n - d_j)}{2\alpha n^2 - 4m} - \frac{\alpha 2m + d_i d_j}{4m} \right) x_{i,j}, \end{aligned} \quad (13)$$

360 and the positive and negative agreements for the Balanced modularity in the general case can be stated as respectively $\phi(a_{i,j}) = a_{i,j} + \frac{(\alpha n - d_i)(\alpha n - d_j)}{2\alpha n^2 - 4m}$ and $\bar{\phi}(a_{i,j}) = \frac{\alpha 2m + d_i d_j}{4m}$.

However, for a doubly stochastic matrix \mathbf{S} , the formula of eq. (13) can be

greatly simplified. With $\alpha = 1$, by remarking that $\forall i, d_i = 1$ and $2m = n$

$$\begin{aligned}
F_{BM}(\mathbf{S}, \mathbf{X}) &= \sum_{i,j} \left(s_{i,j} + \frac{(n-d_i)(n-d_j)}{2n^2-4m} - \frac{2m+d_i d_j}{4m} \right) x_{i,j} \\
&= \sum_{i,j} \left(s_{i,j} + \frac{(n-1)^2}{2n^2-2n} - \frac{n+1}{2n} \right) x_{i,j} \\
&= \sum_{i,j} \left(s_{i,j} + \frac{n-1}{2n} - \frac{n+1}{2n} \right) x_{i,j} \\
&= \sum_{i,j} \left(s_{i,j} - \frac{1}{n} \right) x_{i,j},
\end{aligned}$$

which allows us to simplify the positive and negative agreements as $\phi(s_{i,j}) = s_{i,j}$ and $\bar{\phi}(s_{i,j}) = \frac{1}{n}$, with the latter one that does not depends on i, j .

365 5.3. Deviation to Uniformity

Principle. This criterion, proposed in [34, Chap.2.5.6], is based on a principle very similar to Newman-Girvan's one. The conceptual difference between these two criteria is that, given a graph and a community structure, the Deviation to Uniformity criterion compares the ratio of intra-community edges within the graph with the expected ratio of intra-community edges within δ -regular random graphs, by stating δ as the average degree in the initial graph—whereas the random model in Newman-Girvan modularity has the same degree sequence than the initial graph. Such a random model corresponds to graphs where edges are uniformly distributed among nodes. Thus the probability that there is an

edge between two nodes i and j is equal to $\frac{\sum_k d_k}{n^2}$, where d_k s are the degrees of the nodes in the initial graph. Hence, given $\mathbf{A} \in \mathbb{R}_+^{n \times n}$ the adjacency matrix of some positively weighted graph, and $\mathbf{X} \in Eq(n)$ a community structure, the Deviation to Uniformity can be written as

$$F_{DU}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \frac{\sum_k d_k}{n^2} \right) x_{i,j}. \quad (14)$$

This criterion is defined for weighted graphs such that those that fall into the scope of this study, so we do not discuss its generalisation.

Reduced Form. The reduced form

$$F_{DU}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} (\phi(a_{i,j}) - \bar{\phi}(a_{i,j})) x_{i,j}$$

is directly derived from eq. (14) by stating the positive and negative agreements as respectively $\phi(a_{i,j}) = a_{i,j}$ and $\bar{\phi}(a_{i,j}) = \frac{\sum_k d_k}{n^2}$. For a doubly stochastic

matrix \mathbf{S} , then $\sum_k d_k = n$ and the criterion from eq. (14) can be simplified as

$$F_{DU}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{1}{n} \right) x_{i,j}. \quad (15)$$

The negative agreement thus becomes $\bar{\phi}(s_{i,j}) = \frac{1}{n}$.

5.4. Deviation to Indetermination

Principle. This criterion introduced in [13], is based on the principle of indetermination between two categorical variables as explained below. Given a set \mathcal{S} of M objects, and P, Q two categorical variables on \mathcal{S} . A categorical variable indicates the category taken by an object from the set. For instance, the objects can be human beings, and the categories are mother tongues, or first names, as long as we can consider that each human being has only one mother tongue and only one first name. Formally, we state

$$\begin{array}{ccc} P : \mathcal{S} & \rightarrow & \{p_1, \dots, p_\pi\} \\ u & \mapsto & P(u) \end{array} \quad \text{and} \quad \begin{array}{ccc} Q : \mathcal{S} & \rightarrow & \{q_1, \dots, q_\sigma\} \\ u & \mapsto & Q(u) \end{array}$$

where $\{p_1, \dots, p_\pi\}$ are the categories of variable P —e.g., languages if $P(u)$ is the mother tongue of individual u —, respectively $\{q_1, \dots, q_\sigma\}$ the categories of variable Q . We remark that, as a unique category is attributed to each object by a variable, P and Q also represent equivalence relations—two individuals named *Morgan* are in relation according to the Q that represents first names. We remark that P and Q can be represented by two matrices $\mathbf{P} \in \{0, 1\}^{M \times \pi}$, respectively $\mathbf{Q} \in \{0, 1\}^{M \times \sigma}$, such that

$$\mathbf{P}(u, i) = \begin{cases} 1 & \text{if } P(u) = p_i \\ 0 & \text{otherwise} \end{cases}, \quad \mathbf{Q}(u, i) = \begin{cases} 1 & \text{if } Q(u) = q_i \\ 0 & \text{otherwise} \end{cases},$$

which allows us to write the equivalence relations defined by the variables P and Q as $\mathbf{C}^{(p)} = \mathbf{P}\mathbf{P}^T \in Eq(M)$, respectively $\mathbf{C}^{(q)} = \mathbf{Q}\mathbf{Q}^T \in Eq(M)$. We can also derive their contingency table $\mathbf{N} = \mathbf{P}^T\mathbf{Q}$, with

$$n_{i,j} = |\{u \in \mathcal{S} : P(u) = p_i \text{ and } Q(u) = q_j\}|$$

370 the number of objects with both category p_i from P and category q_j from Q .

Given these matrix notations, we explain below the indetermination between categorical variables. Considering two categorical variables as two equivalence relations, an interesting problem is to measure their association [40]. This is done by comparing the agreements and disagreements between the two variables—these notions are illustrated in table 2. Indetermination is a special case of association. Strictly speaking, one says that two variables are indetermined if their number of agreements is equal to their number of disagreements, that is

$$\sum_{u,v \in \mathcal{S}} \left(c_{u,v}^{(p)} \times c_{u,v}^{(q)} + \overline{c_{u,v}^{(p)}} \times \overline{c_{u,v}^{(q)}} \right) = \sum_{u,v \in \mathcal{S}} \left(\overline{c_{u,v}^{(p)}} \times \overline{c_{u,v}^{(q)}} + c_{u,v}^{(p)} \times c_{u,v}^{(q)} \right).$$

$P \setminus Q$	$Q(u) = Q(v)$	$Q(u) \neq Q(v)$
$P(u) = P(v)$	agreement	disagreement
$P(u) \neq P(v)$	disagreement	agreement

Table 2: All possible agreement/disagreement relations between two objects u and v according to two categorical variables P and Q .

The notion of indetermination can be generalised to allow one to weight positive and negative cases differently. Indeed, it might worth to give more weight to objects that are related than to those that are not [41]. Recall that π (respectively σ) is the number of categories for variable P (respectively Q), an interesting generalisation of indetermination is to weight positive cases with $\pi - 1$ and negative cases with 1 in P , respectively $\sigma - 1$ for positive and 1 for negative cases in Q . This provides the following equality for indetermination

$$\begin{aligned}
& (\pi - 1)(\sigma - 1) \sum_{u,v \in \mathcal{S}} c_{u,v}^{(p)} \times c_{u,v}^{(q)} + \sum_{u,v \in \mathcal{S}} \overline{c_{u,v}^{(p)}} \times \overline{c_{u,v}^{(q)}} = \\
& (\pi - 1) \sum_{u,v \in \mathcal{S}} c_{u,v}^{(p)} \times \overline{c_{u,v}^{(q)}} + (\sigma - 1) \sum_{u,v \in \mathcal{S}} \overline{c_{u,v}^{(p)}} \times c_{u,v}^{(q)}.
\end{aligned} \tag{16}$$

This choice of weights is special because two categorical variables that verify eq. (16) verify also other properties, e.g. they make vanish the so-called Jansen-Vegelius criterion, one of the most famous association criteria. Besides, eq. (16) is strongly related to another special case of association, called the geometrical independence—see [41] for comparisons and discussions about the different notions of independence and indetermination. From here, we use the term indetermination to speak about the generalised indetermination weighted as in eq. (16).

It is shown in [41] that eq. (16) can be rewritten using the contingency table \mathbf{N} as

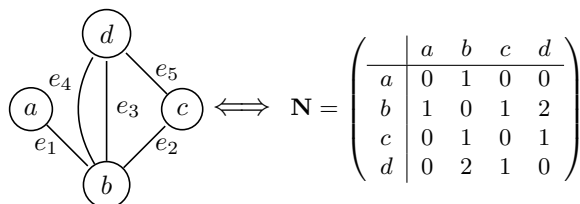
$$\forall i, j : n_{i,j} - \left(\frac{\sum_t n_{i,t}}{\sigma} + \frac{\sum_s n_{s,j}}{\pi} - \frac{M}{\pi \times \sigma} \right) = 0.$$

Thus, for any contingency table $\mathbf{N} \in \mathbb{N}^{p \times q}$, the deviation to indetermination is measured by

$$\sum_{i,j} \left(n_{i,j} - \frac{\sum_t n_{i,t}}{q} - \frac{\sum_s n_{s,j}}{p} + \frac{\sum_{t,s} n_{t,s}}{p \times q} \right). \tag{17}$$

The Deviation to Indetermination criterion is based on eq. (17), and can be understood as follows. Let $\mathbf{N} \in \mathbb{N}^{p \times p}$ be a contingency table built on two variables with the same categories $\mathbf{P}, \mathbf{Q} \in \{0, 1\}^{M \times p}$, thus, an equivalence relation $\mathbf{X} \in Eq(p)$ that maximises

$$\sum_{i,j} \left(n_{i,j} - \frac{\sum_t n_{i,t}}{p} - \frac{\sum_s n_{s,j}}{p} + \frac{\sum_{t,s} n_{t,s}}{p^2} \right) x_{i,j} \tag{18}$$



$$\mathbf{N} = \begin{pmatrix} & a & b & c & d \\ a & 0 & 1 & 0 & 0 \\ b & 1 & 0 & 1 & 2 \\ c & 0 & 1 & 0 & 1 \\ d & 0 & 2 & 1 & 0 \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} & a & b & c & d \\ e_1(s) & 1 & 0 & 0 & 0 \\ e_1(t) & 0 & 1 & 0 & 0 \\ e_2(s) & 0 & 1 & 0 & 0 \\ e_2(t) & 0 & 0 & 1 & 0 \\ e_3(s) & 0 & 1 & 0 & 0 \\ e_3(t) & 0 & 0 & 0 & 1 \\ e_4(s) & 0 & 1 & 0 & 0 \\ e_4(t) & 0 & 0 & 0 & 1 \\ e_5(s) & 0 & 0 & 1 & 0 \\ e_5(t) & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} & a & b & c & d \\ e_1(s) & 0 & 1 & 0 & 0 \\ e_1(t) & 1 & 0 & 0 & 0 \\ e_2(s) & 0 & 0 & 1 & 0 \\ e_2(t) & 0 & 1 & 0 & 0 \\ e_3(s) & 0 & 0 & 0 & 1 \\ e_3(t) & 0 & 1 & 0 & 0 \\ e_4(s) & 0 & 0 & 0 & 1 \\ e_4(t) & 0 & 1 & 0 & 0 \\ e_5(s) & 0 & 0 & 0 & 1 \\ e_5(t) & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 6: **Top:** A multi-graph (left) is the contingency table (right) of two categorical variables. **Bottom:** In these categorical variables, variables are edges and categories are end nodes. For undirected graphs, source and target end nodes can be swapped (right versus left).

groups together categories such that \mathbf{P} and \mathbf{Q} are highly determined (or far
 380 from the indetermination) when restricted to categories from a same group. The parallel with community structures is done by remarking that a multi-graph can be seen as the contingency table of two categorical variables, whose categories are nodes and which are defined on a set \mathcal{S} consisting in the end nodes of edges. An example is provided in fig. 6. In this figure, the edges of a
 385 multi-graph are named e_1, \dots, e_5 , and we define two categorical variables on their end nodes. Namely, each edge can be written $e = (u, v)$, where u and v are the end nodes of e , with u the source node ($e(s)$) and v the target node ($e(t)$). As the direction of an edge is immaterial in a undirected graph, the two categorical variables are created by swapping the role of end nodes—e.g. in fig. 6, \mathbf{P} sees
 390 $e_1 = (a, b)$, whereas \mathbf{Q} states $e_1 = (b, a)$.

Thus, considering $\mathbf{A} \in \mathbb{N}^{n \times n}$ the adjacency matrix of some multi-graph as the contingency table of two such variables, one can look for the community structure that groups together the nodes such that these two categorical variables are highly determined when restricted to these nodes. Roughly, given such a group of nodes, it means that most of the edges have either both or none of their end nodes in this group. By adapting eq. (18) to the specific case of multi-graphs, we remark that finding such a community structure is equivalent

to finding $\mathbf{X} \in Eq(n)$ that maximises

$$F_{DI}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \frac{d_i}{n} - \frac{d_j}{n} + \frac{\sum_k d_k}{n^2} \right) x_{i,j}. \quad (19)$$

Generalisation. This criterion is naturally defined on mutli-graphs, since we can write them as contingency tables. Using the same trick than for Newman-Girvan modularity, this criterion can be directly applied to undirected positively weighted graphs.

Reduced Form. The Deviation to Indetermination criterion from eq. (19) can be rewritten as in eq.(6) by choosing the positive and negative agreements as respectively

$$\phi(a_{i,j}) = a_{i,j} + \frac{\sum_k d_k}{n^2} \quad \text{and} \quad \bar{\phi}(a_{i,j}) = \frac{d_i + d_j}{n}.$$

For a doubly stochastic matrix \mathbf{S} , simplifications can be done. Indeed, since $\forall i, d_i = 1$ and $\sum_k d_k = n$, eq. (19) can be simplified as

$$F_{DI}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{1}{n} \right) x_{i,j}. \quad (20)$$

³⁹⁵ The positive and negative agreements become respectively $\phi(s_{i,j}) = s_{i,j}$ and $\bar{\phi}(s_{i,j}) = 1/n$, the latter not depending on i, j .

5.5. Zahn Criterion

Principle. Strictly speaking, the Zahn criterion does not assess the consistency of a community structure on a given network. However, it can be straightforwardly extended to such a purpose. The Zahn criterion is designed to compare two relations over a set of objects [35]. More precisely, given a finite set V and a symmetric relation \mathcal{R} over this set (that is $\forall (u, v) \in V \times V$, \mathcal{R} verifies $u\mathcal{R}v \iff v\mathcal{R}u$), Zahn wants to find the equivalence relation \mathcal{X} which is the closest to \mathcal{R} . To this aim, Zahn designs a distance between two relations by considering both relations as subsets of the cardinal set $V \times V$, and counting the number of pairs that belong to only one subset. An example is provided in the top panel of fig. 7, where the symmetric relation \mathcal{R} and the equivalence relation \mathcal{X} are defined on a set $V = \{a, b, c, d, e\}$. They are represented as subsets of $V \times V$ by grids, where a coloured cell means that the two corresponding objects are related. For instance, by looking at the row of object a in the grids, we see that $a\mathcal{R}b$ and $a\mathcal{R}c$ for \mathcal{R} , and $a\mathcal{X}a$, $a\mathcal{X}b$, and $a\mathcal{X}c$ for \mathcal{X} . This can be rewritten $(a, b), (a, c) \in \mathcal{R}$ and $(a, a), (a, b), (a, c) \in \mathcal{X}$. In both grids, the dark coloured cells correspond to pairs of objects that belong to both relations and the light ones are pairs that lie in only one subset. With this mapping between

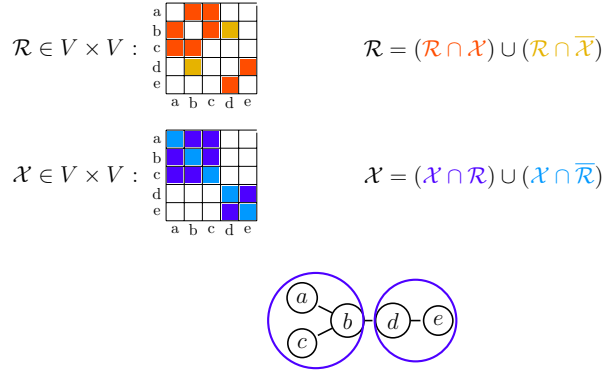


Figure 7: **Top:** The Zahn distance defined between two relations is based on the set representation of these relations. Coloured cells indicate pairs of related objects. **Bottom:** A symmetric relation can be seen as a simple graph. An equivalence relation corresponds to a community structure.

the relations defined on a set V and the subsets of $V \times V$, the distance defined by Zahn is

$$d_Z(\mathcal{R}, \mathcal{X}) = |\bar{\mathcal{X}} \cap \mathcal{R}| + |\bar{\mathcal{R}} \cap \mathcal{X}|, \quad (21)$$

with \mathcal{R} a symmetric relation, and \mathcal{X} an equivalence relation.

In [34], it is proposed to use this criterion to assess community structures on simple graphs, by remarking that a simple graph can be characterised by a symmetric relation over the set of its nodes, and a community structure on this graph is an equivalence relation over the graph nodes as well. The bottom panel of fig. 7 illustrates the relations \mathcal{R} and \mathcal{X} as respectively a graph and a community structure. Zahn distance is also rewritten in [34] to get a matrix-oriented formulation of this criterion. Denoting \mathbf{A} the adjacency matrix of the simple graph associated with the symmetric relation \mathcal{R} , respectively \mathbf{X} the matrix representation of the equivalence relation \mathcal{X} ,

$$\begin{aligned} \mathcal{R} \cap \bar{\mathcal{X}} &= \{(i, j) \in V \times V : a_{i,j}(1 - x_{i,j}) \neq 0\} \\ \mathcal{X} \cap \bar{\mathcal{R}} &= \{(i, j) \in V \times V : x_{i,j}(1 - a_{i,j}) \neq 0\} \end{aligned}$$

Hence, Zahn distance can be rewritten as

$$d_Z(\mathbf{A}, \mathbf{X}) = \frac{1}{2} \sum_{i,j} (a_{i,j} \bar{x}_{i,j} + x_{i,j} \bar{a}_{i,j}). \quad (22)$$

When used for community detection, the Zahn criterion is often stated as
 400 equivalent to the so-called Condorcet criterion [34]. However, in [39, Chap.1.1.3], we show that the Condorcet criterion cannot be extended to the problem of finding the best community structure given a graph.

Finally, we observe that the criterion of eq. (22) defines a distance in the formal mathematical sense (i.e., it is positive, symmetric, separable and verifies
 405 the triangle inequality). This does not hold anymore for any generalisation presented below.

Generalisation. Zahn distance, originally designed for comparing relations over a finite set, is straightforwardly extended to simple graphs and community structures. On the other hand, its generalisation to weighted graphs is not as straightforward since there is no trivial matching between a weighted graph and a symmetric relation. However, a generalisation of Zahn criterion to weighted graphs is proposed in [25]. It is directly derived from eq. (22) by defining the complementary of the real-valued matrix \mathbf{A} as $\bar{\mathbf{A}} = a_{max}\mathbf{J} - \mathbf{A}$, with $a_{max} = \max_{i,j} a_{i,j}$.

This leads to the criterion

$$d_{Z,1}^{\omega}(\mathbf{A}, \mathbf{X}) = \frac{1}{2} \sum_{i,j} (a_{i,j}(1 - x_{i,j}) + x_{i,j}(a_{max} - a_{i,j})). \quad (23)$$

Nevertheless, we propose other generalisations, as this one does not always seem suitable. Indeed, the purpose of generalising criteria to weighted graphs is to enable them to assess community structures on preprocessed doubly stochastic graphs. Assuming that a graph before preprocessing is simple and thus associated with a symmetric relation \mathcal{R} over its set of nodes. Calling \mathbf{S} the adjacency matrix of the preprocessed graph, and \mathbf{X} some community structure. Then, with Zahn criterion as defined in eq. (23):

- There is an imbalance between the impact on the criterion of pairs in $\mathcal{R} \cap \bar{\mathcal{X}}$ and in $\bar{\mathcal{R}} \cap \mathcal{X}$. Any pair in $\bar{\mathcal{R}} \cap \mathcal{X}$ results in a penalisation of the criterion equal to s_{max} , whereas a pair in $\mathcal{R} \cap \bar{\mathcal{X}}$ results in a penalisation equal to $s_{i,j} \leq s_{max}$. Hence, each pair in $\bar{\mathcal{R}} \cap \mathcal{X}$ penalises the criterion equally to the highest penalisation that can be reached by a pair in $\mathcal{R} \cap \bar{\mathcal{X}}$.
- Except for pairs (i, j) such that $s_{i,j} = s_{max}$, every pair that lies in $\mathcal{X} \cap \bar{\mathcal{R}}$ penalises the criterion.

We believe that these points are non desirable aspects of the previous generalisation of Zahn criterion. For this reason, we propose other generalisations. As authors of [25], given a positively weighted matrix \mathbf{A} , we choose generalisations that simply redefine the complementary of \mathbf{A} in eq. (22). That is, we define the generalised criterion as

$$d_Z^{\omega}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} (a_{i,j}(1 - x_{i,j}) + x_{i,j}(\alpha - a_{i,j})), \quad (24)$$

with α some constant to set up. We find that choosing α as the mean element of the matrix \mathbf{A} , that is, $\alpha = a_{mean}$, is a good trade-off to mitigate against the two drawbacks listed above:

- Each element in $\bar{\mathcal{R}} \cap \mathcal{X}$ penalises the criterion with the mean value of the adjacency matrix.
- An element in $\mathcal{X} \cap \bar{\mathcal{R}}$ penalises the criterion only if its value is lower than the adjacency matrix mean value. Otherwise it even favours the criterion.

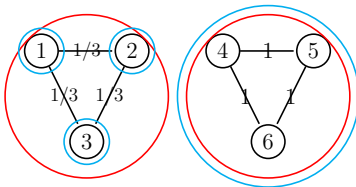


Figure 8: A weighted graph and two community structures \mathbf{X}_r (in red) and \mathbf{X}_b (in blue), for which $d_{Z,1}^\omega(\mathbf{A}, \mathbf{X}_r) = 5 \geq d_{Z,1}^\omega(\mathbf{A}, \mathbf{X}_b) = 4$, $d_{Z,2}^\omega(\mathbf{A}, \mathbf{X}_r) = -2 \leq d_{Z,2}^\omega(\mathbf{A}, \mathbf{X}_b) \sim -0.7$, and $d_{Z,3}^\omega(\mathbf{A}, \mathbf{X}_r) = d_{Z,3}^\omega(\mathbf{A}, \mathbf{X}_b) = 2$

We consider two possible definitions of $\alpha = a_{mean}$, namely

$$a_{mean} = \frac{1}{n^2} \sum_{i,j} a_{i,j}, \quad \text{and} \quad (25)$$

$$a_{mean} = \frac{1}{nnz} \sum_{i,j} a_{i,j}. \quad (26)$$

We denote by $d_{Z,2}^\omega$ the criterion from eq. (24) obtained with $\alpha = a_{mean}$ from eq. (25), respectively $d_{Z,3}^\omega$ the one obtained using $\alpha = a_{mean}$ from eq. (26).
 430 Different behaviours of $d_{Z,1}^\omega$, $d_{Z,2}^\omega$ and $d_{Z,3}^\omega$ are illustrated on a toy example in fig. 8. This figure shows a weighted graph with two disjoint components, where each component is a clique with its unique own edge value. In red and in blue, two community structures are proposed, that we denote respectively \mathbf{X}_r and \mathbf{X}_b . The criterion from eq (23) considers that \mathbf{X}_b is a community structure
 435 that better approximates the ground truth structure of the graph than \mathbf{X}_r . On the other hand, the criterion $d_{Z,2}^\omega$ states that \mathbf{X}_r is better than \mathbf{X}_b , which may be a more desirable situation. Finally, $d_{Z,3}^\omega$ considers the two structures as equivalent.

Remark 2. With both generalisations from eq (24) using $\alpha = a_{mean}$, negative
 440 values are possible, and the symmetry is not preserved ($d_Z(\mathbf{A}, \mathbf{X}) \neq d_Z(\mathbf{X}, \mathbf{A})$), whereas eq (23) ensures the positivity of the results and preserves the symmetry, since $\forall \mathbf{X} \in Eq(n), x_{max} = 1$.

Reduced Form. We now aim to find formulations of eqs. (22), (23) and (24) that fit with the reduced form from eq. (6). We first remark that, given \mathbf{A} the adjacency matrix of some graph, the Zahn criterion $d_Z(\mathbf{A}, \cdot)$ is an objective function that one aims to minimise, whereas in eq. (6), the criterion must be a function to maximise. Hence, we express the opposite of d_Z and remark that minimising the function given in eq (22) is equivalent to maximising

$$\mathbf{X} \mapsto F_Z(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \frac{1}{2} \right) x_{i,j}. \quad (27)$$

For unweighted graphs, the positive and negative agreements of the Zahn criterion are thus respectively $\phi(a_{i,j}) = a_{i,j}$ and $\bar{\phi}(a_{i,j}) = 1/2$.

For weighted graphs, the opposite of d_Z^ω from eq. (24) produces the following reduced form for Zahn criterion

$$F_Z^\omega(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \frac{\alpha}{2} \right) x_{i,j}, \quad (28)$$

445 with $\alpha = \underset{i,j}{max}(a_{i,j})$ for the generalisation of eq. (23) and $\alpha = a_{mean}$ for the generalisation proposed here. Moreover, this second generalisation on a doubly stochastic matrix \mathbf{S} implies that $\alpha = 1/n$ using eq. (25), respectively $\alpha = n/nmz$ using eq. (26). Both negative agreements $\bar{\phi}(s_{i,j}) = 1/2n$ and $\bar{\phi}(s_{i,j}) = n/2nmz$ do not depend on i, j .

450 5.6. Correlation Clustering Criterion

Principle. The Correlation Clustering is first introduced by Bansal et al. in [36]. Their problem can be stated as follows. Given a set of objects such that, for each pair of objects, one knows if the objects are similar or dissimilar, the aim is to find a clustering that “maximises agreements”, or equivalently “minimises
455 disagreements”. They model the set of objects as a complete graph such that each pair of nodes (objects)—or equivalently, each edge—has a label “+” if objects are similar, and a label “-” if objects are dissimilar (see fig. 9), and give a formal definition of maximising agreements/minimising disagreements.

- Maximising agreements means finding a clustering with both as many edges labelled “+” having end nodes in a same cluster as possible, and as many edges labelled “-” with end nodes in different clusters as possible. With notations from fig. 9, it means solving

$$\underset{\mathbf{X} \in Eq(n)}{argmax} \left(\sum_{\{i,j\} \in E^+} x_{i,j} + \sum_{\{i,j\} \in E^-} \bar{x}_{i,j} \right).$$

- Minimising disagreements means finding a clustering with both as few edges labelled “+” with end nodes in different clusters as possible, and as few edges labelled “-” having end nodes in a same cluster as possible. With notations from fig. 9, it means solving

$$\underset{\mathbf{X} \in Eq(n)}{argmin} \left(\sum_{\{i,j\} \in E^+} \bar{x}_{i,j} + \sum_{\{i,j\} \in E^-} x_{i,j} \right).$$

460 One of the authors’ rationales for formalising a clustering problem as a Correlation Clustering problem is that, on the contrary of other clustering methods that used to exist, the Correlation Clustering problem can be solved without

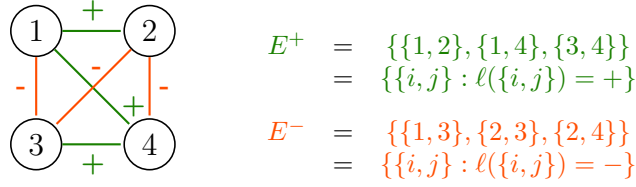


Figure 9: In the graph to the left, the green edges are similarities between the nodes they link. The set of these edges is denoted by E^+ . The orange edges link nodes that are dissimilar. The set of these edges is denoted E^- .

setting the number of clusters in advance. This makes this technique particularly suitable for community detection, where the number of communities is generally not known [36].

465 Demaine et al. extend the Correlation Clustering problem to general weighted graphs in [42]⁴. Given a weighted, labelled graph $G = (V, E, \Omega, \ell)$ where

- $\Omega : \begin{array}{l} E \longrightarrow \mathbb{R}_+ \\ \{i, j\} \mapsto \omega(\{i, j\}) \end{array}$ indicates edge weights,
- $\ell : \begin{array}{l} E \longrightarrow \{+, -\} \\ \{i, j\} \mapsto \ell(\{i, j\}) \end{array}$ indicates edge labels,

they focus on a generalised formulation of the “minimising disagreements” problem by looking for

$$\underset{\mathbf{X} \in E_q(n)}{\operatorname{argmin}} \left(\sum_{\substack{\{i, j\} \in E \\ \ell(\{i, j\}) = -}} \omega(\{i, j\}) x_{i, j} + \sum_{\substack{\{i, j\} \in E \\ \ell(\{i, j\}) = +}} \omega(\{i, j\}) \bar{x}_{i, j} \right) \quad (29)$$

In [34], it is proposed to separate positive and negative labels in the weight indicator Ω , that can be expressed as creating two functions Ω^+ and Ω^- such that

$$\Omega^+ : \begin{array}{l} E \longrightarrow \mathbb{R}_+ \\ \{i, j\} \mapsto \omega^+(\{i, j\}) = \begin{cases} \omega(\{i, j\}) & \text{if } \ell(\{i, j\}) = +, \\ 0 & \text{otherwise.} \end{cases} \end{array}$$

and

$$\Omega^- : \begin{array}{l} E \longrightarrow \mathbb{R}_+ \\ \{i, j\} \mapsto \omega^-(\{i, j\}) = \begin{cases} \omega(\{i, j\}) & \text{if } \ell(\{i, j\}) = -, \\ 0 & \text{otherwise.} \end{cases} \end{array}$$

⁴Bansal et al. generalise it for complete weighted graphs whose weights lie in $[-1, 1]$ in [36].

This allows to simplify eq. (29) as

$$\underset{\mathbf{X} \in Eq(n)}{\operatorname{argmin}} \sum_{i,j} (\omega^+(\{i,j\})\bar{x}_{i,j} + \omega^-(\{i,j\})x_{i,j}). \quad (30)$$

By denoting

$$g_{CC}(G, \mathbf{X}) = \sum_{i,j} (\omega^+(\{i,j\})\bar{x}_{i,j} + \omega^-(\{i,j\})x_{i,j}),$$

we remark that minimising $\mathbf{X} \mapsto g_{CC}(G, \cdot)$ is equivalent to minimising a function d_{CC} defined by

$$\mathbf{X} \mapsto d_{CC}(G, \mathbf{X}) = \sum_{i,j} (\omega^-(\{i,j\}) - \omega^+(\{i,j\})) x_{i,j}. \quad (31)$$

Generalisation. The case of graphs with positive and negative edges is beyond the scope of this study. However, in positively weighted networks, it is natural to assume that an edge indicates that its two end nodes are similar. In turn, one can assume that dissimilarities are indicated by an absence of edge. We use this idea to generalise the Correlation Clustering criterion to positively weighted graphs. For this purpose, we define the pattern of a matrix as the following function

$$\begin{aligned} \mathcal{P} : \mathbb{R}_+^{n \times n} &\longrightarrow \{0, 1\}^{n \times n} \\ \mathbf{M} &\mapsto \mathcal{P}(\mathbf{M}) = \mathbf{P}^{\mathbf{M}}, \end{aligned}$$

such that $p_{i,j}^{\mathbf{M}} = \begin{cases} 1 & \text{if } m_{i,j} \neq 0 \\ 0 & \text{otherwise.} \end{cases}$ Given $G = (V, E, \Omega)$ some positively weighted

graph and $\mathbf{A} \in \mathbb{R}^{n \times n}$ its adjacency matrix, the absence of edge in G is characterised by $\mathbf{J} - \mathbf{P}^{\mathbf{A}}$. Thus, denoting by $\lambda > 0$ the penalisation for clustering together nodes that are dissimilar, the Correlation Clustering from eq. (31) becomes

$$d_{CC}^\lambda(\mathbf{A}, \mathbf{X}) = \sum_{i,j} (\lambda \times (1 - p_{i,j}^{\mathbf{A}}) - a_{i,j}) x_{i,j}, \quad (32)$$

where \mathbf{A} is the adjacency matrix of some positively weighted graph, and \mathbf{X} is a community structure on this graph. The proposed generalised Correlation Clustering hence depends on some parameter $\lambda > 0$ to set up.

Remark 3. When focusing on simple networks, this generalised Correlation Clustering criterion is close to the LAMBDA_{CC} function proposed in [43].

Remark 4. Another way to generalise the Correlation Clustering criterion may be to consider that a positively weighted graph is actually a complete graph, where an edge whose weight is equal to 0 is the strongest case of dissimilarity. In this case, one can shift the weights so that the graph has positive and negative values. Given \mathbf{A} the adjacency matrix, the most straightforward way to do so

is to consider that an edge is a dissimilarity if it is below the mean value of \mathbf{A} , that is $\frac{\sum_k d_k}{n^2}$. In this case, the criterion from eq. (31) becomes

$$d_{CC}^\omega(\mathbf{A}, \mathbf{X}) = - \sum_{i,j} \left(a_{i,j} - \frac{\sum_k d_k}{n^2} \right) x_{i,j},$$

which is equivalent to the formula of the Deviation to Uniformity criterion developed in Section section 5.3.

Reduced Form. We aim to reduce the formula from eq. (32) to make it fit with eq. (6). As the Correlation Clustering criterion defined at eq. (32) is a criterion to minimise to obtain the best community structure, we look at its opposite. Minimising d_{CC}^λ is equivalent to maximising

$$F_{CC}^\lambda(\mathbf{A}, \mathbf{X}) = \sum_{i,j} (a_{i,j} - \lambda \times (1 - p_{i,j}^\mathbf{A})) x_{i,j}. \quad (33)$$

The positive and negative agreements for this generalised criterion are respectively $\phi(a_{i,j}) = a_{i,j}$ and $\bar{\phi}(a_{i,j}) = \lambda \times (1 - p_{i,j}^\mathbf{A})$.

6. Comparison of the Criteria

In this section, we compare the criteria from section 5. In table 3, we recall the reduced formulations of these criteria when applied on simple or doubly stochastic graphs.

6.1. Homogenisation on Doubly Stochastic Graphs

The first key result directly observed from table 3 is that, when applied to doubly stochastic graphs, many criteria become equivalent, as stated in theorem 2.

Theorem 2. *Given $\mathbf{S} \in \mathbb{R}^{n \times n}$ the adjacency matrix of some doubly stochastic graph, and $\mathbf{X} \in Eq(n)$ a community structure, thus*

$$F_{NG}^\omega(\mathbf{S}, \mathbf{X}) = F_{BM}^\omega(\mathbf{S}, \mathbf{X}) = F_{DU}^\omega(\mathbf{S}, \mathbf{X}) = F_{DI}^\omega(\mathbf{S}, \mathbf{X}).$$

Theorem 2 extends theorem 6.1 from [34] that states that these criteria are equivalent in the case of k -regular simple graphs. Furthermore, the doubly stochastic Zahn modularities, while not strictly equivalent to these four criteria, have very similar formulations. Actually, one can draw a parallel between Zahn formulations and the parametrised Newman-Girvan modularity, used to mitigate against the so-called resolution limit of the Newman-Girvan modularity—that is, its inability to highlight small communities [12]. This function is defined in [44] as

$$F_{NG}^\gamma(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \gamma \frac{d_i d_j}{2m} \right) x_{i,j},$$

Criteria	If \mathbf{A} represents a Simple Graph	If \mathbf{S} is Doubly Stochastic
Newman-Girvan Modularity	$F_{NG}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \frac{d_i d_j}{2m} \right) x_{i,j}$	$F_{NG}^{\omega}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{1}{n} \right) x_{i,j}$
Balanced Modularity	$F_{BM}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} + \frac{(n-d_i)(n-d_j)}{2n^2 - 4m} - \frac{2m + d_i d_j}{4m} \right) x_{i,j}$	$F_{BM}^{\omega}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{1}{n} \right) x_{i,j}$
Deviation to Uniformity	$F_{DU}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \frac{2m}{n^2} \right) x_{i,j}$	$F_{DU}^{\omega}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{1}{n} \right) x_{i,j}$
Deviation to Indetermination	$F_{DI}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} + \frac{2m}{n^2} - \frac{d_i + d_j}{n} \right) x_{i,j}$	$F_{DI}^{\omega}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{1}{n} \right) x_{i,j}$
Zahn Modularity	$F_Z(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \frac{1}{2} \right) x_{i,j}$	$F_{Z,1}^{\omega}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{s_{max}}{2} \right) x_{i,j}$ $F_{Z,2}^{\omega}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{1}{2n} \right) x_{i,j}$ $F_{Z,3}^{\omega}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{n}{2mnz} \right) x_{i,j}$
Correlation Clustering	$F_{CC}^{\lambda}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \lambda \times (1 - p_{i,j}^{\mathbf{A}}) \right) x_{i,j}$	$F_{CC}^{\omega,\lambda}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \lambda \times (1 - p_{i,j}^{\mathbf{S}}) \right) x_{i,j}$
Parametrised Modularity	$F_{NG}^{\gamma}(\mathbf{A}, \mathbf{X}) = \sum_{i,j} \left(a_{i,j} - \gamma \frac{d_i d_j}{2m} \right) x_{i,j}$	$F_{NG}^{\omega,\gamma}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{\gamma}{n} \right) x_{i,j}$

Table 3: Reduced formulations of the different criteria when applied on simple (respectively doubly stochastic) graphs.

with \mathbf{A} the adjacency matrix of some simple graph, $\mathbf{X} \in Eq(n)$, and $\gamma > 0$ the parameter. In definition 5, we also define a parametrised criterion for the doubly stochastic version of Newman-Girvan modularity.

Definition 5. *Given $\mathbf{S} \in \mathbb{R}^{n \times n}$ the adjacency matrix of some doubly stochastic graph, $\mathbf{X} \in Eq(n)$, and $\gamma > 0$ a scalar, the parametrised doubly stochastic Newman-Girvan modularity is defined as*

$$F_{NG}^{\omega, \gamma}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{\gamma}{n} \right) x_{i,j}.$$

These parametrised versions of the Newman-Girvan modularity are added
 490 to the list of criteria, as the last row of table 3, for both simple and doubly stochastic graphs. The doubly stochastic versions of the Zahn criterion can be expressed using definition 5, as stated in property 1.

Property 1. *Given $\mathbf{S} \in \mathbb{R}^{n \times n}$ the adjacency matrix of some doubly stochastic graph and $\mathbf{X} \in Eq(n)$. The doubly stochastic Zahn modularities can be ex-
 495 pressed as parametrised doubly stochastic Newman-Girvan modularities, using the following values for the γ parameter:*

- $F_{Z,1}^{\omega}(\mathbf{S}, \mathbf{X})$ is obtained with $\gamma = \frac{n \times s_{max}}{2}$,
- $F_{Z,2}^{\omega}(\mathbf{S}, \mathbf{X})$ is obtained with $\gamma = \frac{1}{2}$,
- $F_{Z,3}^{\omega}(\mathbf{S}, \mathbf{X})$ is obtained with $\gamma = \frac{n^2}{2 \times nnz(\mathbf{S})}$.

500 Thus, the Correlation Clustering criterion is the unique doubly stochastic criterion from table 3 that cannot be expressed as a parametrised doubly stochastic Newman-Girvan modularity. We now provide the main result of this study in result 1.

Result 1. *Generalising the criteria to doubly stochastic graphs unifies those
 505 criteria. Namely, there are two families of parametrised criteria:*

1. *The Newman-Girvan-like ones*

$$F_{NG}^{\omega, \gamma}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \frac{\gamma}{n} \right) x_{i,j}$$

2. *The Correlation Clustering-like ones*

$$F_{CC}^{\omega, \lambda}(\mathbf{S}, \mathbf{X}) = \sum_{i,j} \left(s_{i,j} - \lambda \times (1 - p_{i,j}^{\mathbf{S}}) \right) x_{i,j}$$

Each criterion is obtained from one of these parametrised criteria, using a specific parameter.

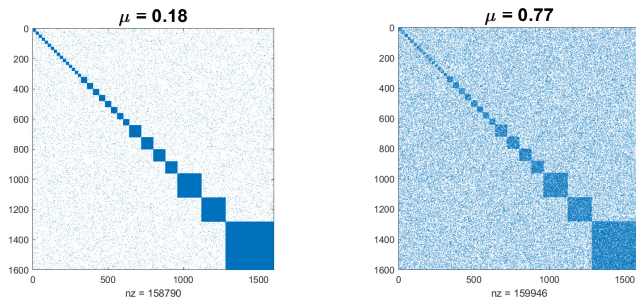


Figure 10: Two instances from the benchmark, built using the p_{in} and p_{out} from the first column (left) and the last column (right) of table 4.

6.2. Numerical Comparisons

In this section we compare the behaviours of the different criteria to un-
 510 cover community structures, applied on modular simple graphs on one hand,
 and their doubly stochastic preprocessing on the other hand. To that purpose,
 we optimise those criteria using the optimisation framework proposed by the
 Louvain algorithm [25].

Benchmark. For these numerical experiments, we build a range of random mod-
 515 ular networks, using eight Stochastic Block Models (SBMs). In brief, SBMs are
 random models for generating networks with some block structure, with pre-
 scribed probabilities of edges within and between the blocks. Models in which
 intra-block probabilities are higher than inter-block probabilities produce net-
 works with community structures [45]. Each SBM is used to generate 10 graphs
 520 of 1600 nodes, with an average degree equal to 100 and 31 blocks: 16 blocks
 of 20 nodes, 8 blocks of 40 nodes, 4 blocks of 80 nodes, 2 blocks of 160 nodes
 and one block of 320 nodes. They all have one unique probability of intra-block
 edge and one unique probability of inter-block edge, denoted respectively p_{in}
 and p_{out} . These SBMs differ in the values of parameters p_{in} and p_{out} , which
 525 are chosen so that the community structures of the random graphs become less
 and less sharp. The sharpness of the community structure is assessed by the so-
 called network mixing parameter [4]. The nodal mixing parameter measures the
 strength of a node’s community membership by computing the ratio between its
 links outside the community and its degree. The greater the mixing param-
 530 eter for each node, the weaker the community structure. The network mixing
 parameter μ is the mean value of the nodal mixing parameters [11]. Two in-
 stances from the benchmark are illustrated in fig. 10. These are two modular
 networks generated by the SBMs with highest and lowest mixing parameters. Finally, 80
 benchmark graphs are built using SBMs from the `NetworkX` library⁵, and pre-
 535 processed using algo. 1. The pairs of intra- and inter-edge probabilities p_{in} and

⁵<https://networkx.org/1>

p_{in}	7.32	6.58	5.83	5.09	4.34	3.60	2.86	2.11
p_{out}	0.06	0.13	0.19	0.25	0.31	0.38	0.44	0.5
μ_{theo}	1.79	3.02	3.98	4.81	5.56	6.27	6.97	7.67
$\widetilde{\mu}_{bin}$	1.80	3.04	4.00	4.84	5.58	6.30	6.99	7.69
μ_{stoch}	1.70	2.91	3.96	4.69	5.43	6.15	6.83	7.53

Table 4: Edge probabilities in each SBMs (p_{in} and p_{out}), theoretical mixing parameters (μ_{theo}), and the observed average mixing parameters on the simple ($\widetilde{\mu}_{bin}$) and preprocessed (μ_{stoch}) graphs.

p_{out} used in the SBMs are showed in table 4, along with the corresponding theoretical mixing parameters, and the average mixing parameters observed in the simple graphs, and in the doubly stochastic scaling of these graphs, respectively. All numbers are multiplied by 10 to improve readability. We observe that the mixing parameters of the preprocessed graphs tend to be slightly below those of the simple graphs, and of the theoretical value as well.

Scores. To assess the quality of the community structures returned by Louvain, we compare them to the ground truth by adapting the definitions of Precision, Recall and F1-score to community detection. Namely, assume we have $\mathbf{X}^* \in Eq(n)$ the ground truth, and $\widetilde{\mathbf{X}} \in Eq(n)$ the community structure returned by Louvain. We define the number of true positives as the number of pairs of different elements that are put together by both community structures, that is $TP = \sum_{i < j} \widetilde{x}_{i,j} \times x_{i,j}^*$. The number of false positives is the number of pairs that are put together by $\widetilde{\mathbf{X}}$ but not by \mathbf{X}^* : $FP = \sum_{i < j} \widetilde{x}_{i,j} \times (1 - x_{i,j}^*)$. And the number of false negatives is the number of pairs that are put together by \mathbf{X}^* but not $\widetilde{\mathbf{X}}$, namely $FN = \sum_{i < j} (1 - \widetilde{x}_{i,j}) \times x_{i,j}^*$. Now, we can derive Precision, Recall and F1-score of $\widetilde{\mathbf{X}}$ as usual.

- Precision: $Prec(\widetilde{\mathbf{X}}) = \frac{TP}{TP + FP}$
- Recall: $Rec(\widetilde{\mathbf{X}}) = \frac{TP}{TP + FN}$
- F1-score: $F1(\widetilde{\mathbf{X}}) = 2 \times \frac{Prec(\widetilde{\mathbf{X}}) \times Rec(\widetilde{\mathbf{X}})}{Prec(\widetilde{\mathbf{X}}) + Rec(\widetilde{\mathbf{X}})}$.

Furthermore, since the Louvain algorithm is sensitive to node labelling, we apply it four times to each network in the benchmark, using a random labelling. Thus, in the following figures, the points on the curves are the average score of the 40 returned community structures (10 networks and 4 runs of Louvain). For each point, these 40 community structures are summarised by box plots, that

indicate the median (white circle with black point), 25th and 75th percentiles (edges of the box), and extreme values (extrema points of vertical segments).

Finally, the number of communities returned by Louvain often helps to explain some observations done on the scores. Indeed, as the number of communities is not constrained, Louvain algorithm may find either more or less communities than expected, with different impacts on Precision and Recall. Thus, we also compare the number of communities returned by Louvain with the expected number (31). Given n_c the number of communities in $\tilde{\mathbf{X}}$, we compute

$$r(\tilde{\mathbf{X}}) = \begin{cases} n_c/31 & \text{if } n_c \geq 31, \\ 31/n_c & \text{otherwise.} \end{cases}$$

This allows a fairer comparison between the criteria that over- or under-partition the graphs. These ratios are displayed in table 5. Structures with more (respectively less) communities than expected are highlighted by a “+” (respectively a “-”) exponent. Also, it may happen that some of the 40 structures have more communities than expected, while other have less. Such cases are indicated by the exponent “*”.

Parametrised Newman-Girvan Modularities. We first focus on the behaviours of the parametrised Newman-Girvan modularities, when varying the parameter γ . The $F1$ -scores (y-axis) over γ parameters (x-axis) of the Newman-Girvan modularities applied on simple and preprocessed graphs are provided in fig. 11 and 12 respectively. As stated in the legends, each curve corresponds to one mixing parameter value $\widetilde{\mu}_{bin}$ from table 4.

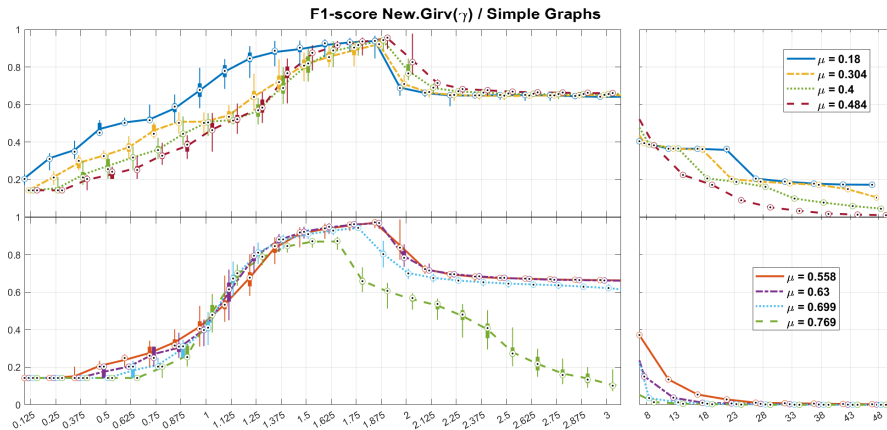


Figure 11: The $F1$ -score (y-axis) over γ (x-axis) of the parametrised Newman-Girvan modularity on simple graphs.

On these figures, we observe that both modularities are able to provide community structures close to the ground truth for some $\gamma \in [1.25, 2]$. In the right panels, one can also observe that, for very large γ 's, $F1$ -score tends to

0. This means that, whatever the sharpness of the ground truth community structure, it exists some γ beyond which Louvain returns the structure with one community per node. We also observe that the fundamental difference between simple and doubly stochastic criteria is that the parametrised Newman-Girvan modularity is much more sensitive to γ variations when applied on simple graphs. Indeed, in fig. 11, the $F1$ -score curves are quite sharp. For each μ , there is a peak at the γ value that maximises the $F1$ -score. Moreover, this peak is not located at the same γ across the μ 's (1.625 for $\mu = 0.769$, 1.75 for $\mu = 0.699$ and 1.875 for the other values of μ). On the other hand, in fig. 12, the $F1$ -score curves are much smoother and the maxima lie along a plateau, whose length depends on the mixing parameter μ . Thus, there is much more chance to pick a γ that provides a sound community structure for preprocessed graphs than for simple ones.

Correlation Clustering Criteria. We now focus on the behaviours of the Correlation Clustering criteria, when varying the parameter λ . As previously, $F1$ -scores over λ parameters are provided in fig. 13 and 14, for Louvain algorithm applied on simple and preprocessed graphs, respectively. This time, criteria on both simple and doubly stochastic graphs highlight plateaus at their maxima. However, one can observe that these plateaus do not appear for the same parameter values. Indeed, there is a factor 100 between the x-axes of the two figures (on the left panel of fig. 13, x-axis limits are $10/n$ and $500/n$, while these are $1/10n$ and $50/10n$ for fig. 14). This observation is consistent with property 2.

Property 2. Given \mathbf{A} the adjacency matrix of a simple graph, and $\lambda > nnz(\mathbf{A})/2$. Assuming that $\mathbf{X}^* = \underset{\mathbf{X} \in Eq(n)}{\operatorname{argmax}} F_{CC}^\lambda(\mathbf{A}, \mathbf{X})$. Thus

$$\forall i \neq j, a_{i,j} = 0 \implies x_{i,j}^* = 0.$$

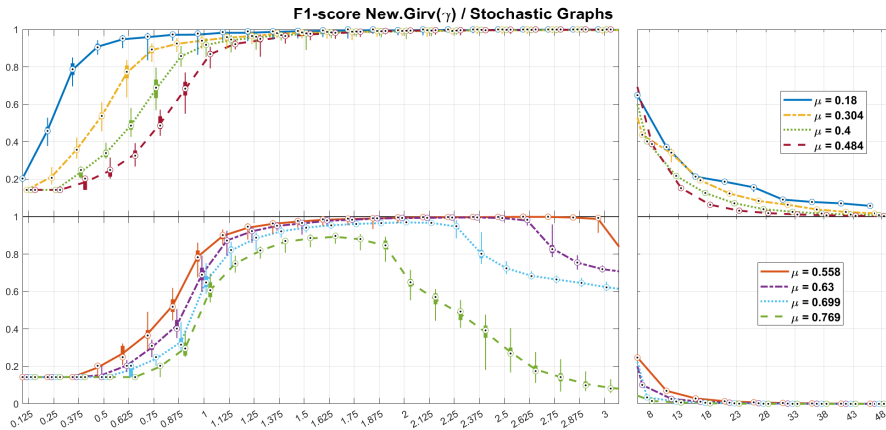


Figure 12: The $F1$ -score (y-axis) over γ (x-axis) of the parametrised Newman-Girvan modularity on doubly stochastic graphs.

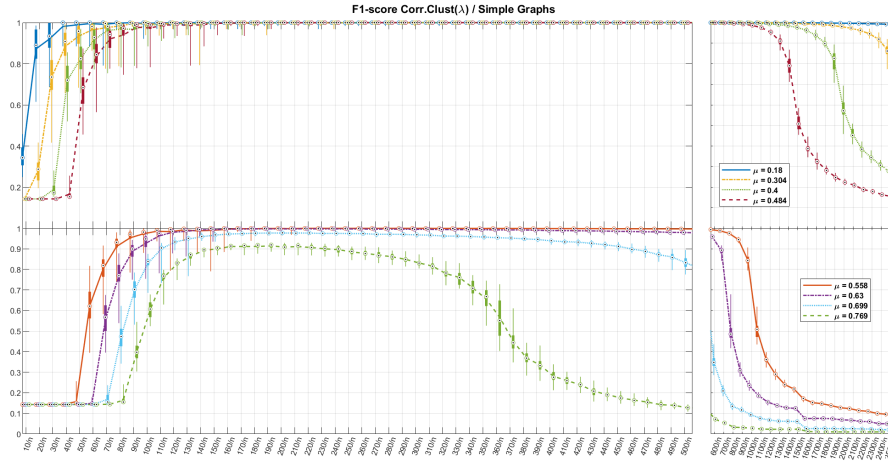


Figure 13: The $F1$ -score (y-axis) over λ (x-axis) of the Correlation Clustering criterion on simple graphs.

Respectively, if $\mathbf{S} \in \mathbb{R}_+^{n \times n}$ is doubly stochastic, and $\lambda_\omega > n/2$, thus $\mathbf{X}^* \in Eq(n)$ that maximises $F_{CC}^{\omega, \lambda_\omega}(\mathbf{S}, \cdot)$ is such that

$$\forall i \neq j, s_{i,j} = 0 \implies x_{i,j}^* = 0.$$

600

Proof. Straightforward by adapting the proof from [39, Property 8]. \square

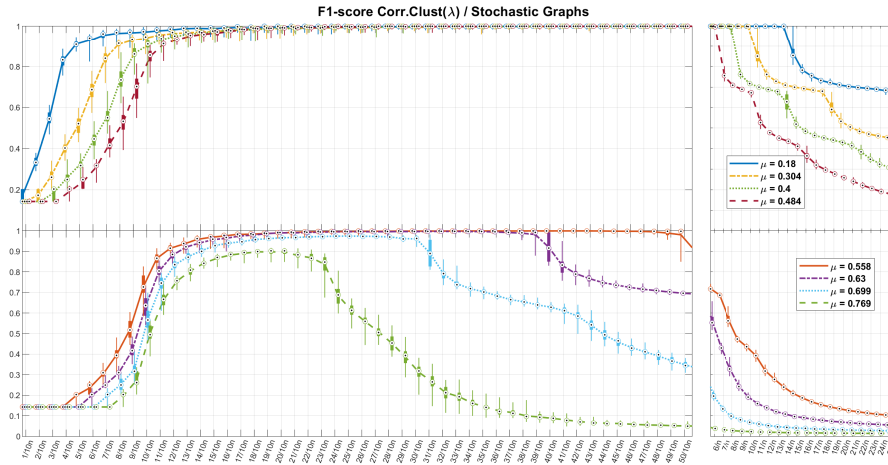


Figure 14: The $F1$ -score (y-axis) over λ (x-axis) of the Correlation Clustering criterion on doubly stochastic graphs.

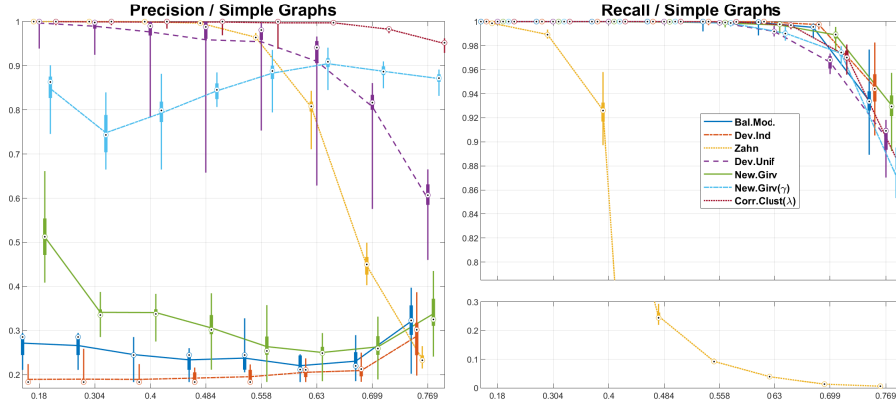


Figure 15: Precision and Recall (y-axes) of all the criteria applied on simple graphs, over the mixing parameters (x-axes).

This property states that, for large values of the λ parameter, in the community structure that optimises the Correlation Clustering criterion, each community must be a clique of the graph. In the benchmark analysed here, this strong constraint implies that, for such λ values, the optimal community structure does not fit the ground truth one. Thus, in property 2, λ and λ_ω provide an upper bound beyond which the Correlation Clustering criteria are not able to uncover the ground truth community structures of the graphs from the benchmark. Those graphs having an average degree equal to 100 implies that those upper bounds are such that $\lambda \approx 100 \times \lambda_\omega$, which is consistent with the differences of x-axes between fig. 13 and 14. Finally, we remark that, opposite to the observations made on Newman-Girvan modularities, maximum plateaus are smoother for simple graphs.

All Criteria on Simple Graphs. In this paragraph, we compare the different behaviours of all the criteria designed for simple graphs. Recall and Precision are displayed in fig. 15. The parameters for the Correlation Clustering and the parametrised Newman-Girvan criteria, respectively $\lambda = 210/n$ and $\gamma = 1.625$, are chosen so that the average $F1$ -score is maximised over all the mixing parameters. From the right panel, we observe that, except for Zahn criterion, all the measures return high scores of Recall (all above 0.8 even for the largest mixing parameter). This is consistent with the fact that, except when used with the Zahn criterion, the Louvain algorithm tends to return structures with less communities than expected, when applied on simple graphs, as it can be seen from ■-highlighted cells in table 5. Thus, some of the ground truth communities are merged into the returned ones. And a high value of Recall means that the returned communities tend to cover the ground truth ones. On the other hand, Louvain with Zahn criterion returns almost 5 times more communities than expected when $\mu = 0.484$, and this ratio keeps increasing with μ , which

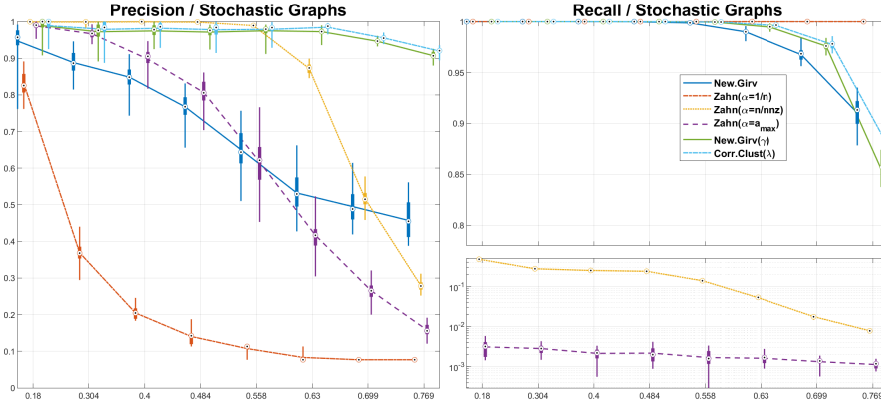


Figure 16: Precision and Recall (y-axes) of all the criteria applied on doubly stochastic graphs, over the mixing parameters (x-axes).

explains the slump of this criterion Recall curve. This is an expected result, since it is proven in [34] that the community structure that maximises the Zahn criterion is such that subgraphs induced by each community must be 1/2-dense. Looking at the values of p_{in} and p_{out} from table 4, ground truth communities are expected to respect this property up to $\mu = 0.484$, included. However, Louvain algorithm only approximates the best community structure for the criterion, which explains why the slump starts at $\mu = 0.484$ in the tests.

When looking at the left panel, we can roughly divide the remaining measures into two categories: the Balanced Modularity, the Deviation to the Indetermination and the Newman-Girvan modularity, that exhibit low Precision scores, and the Deviation to the Uniformity, the Correlation Clustering criterion and the parametrised Newman-Girvan modularity that exhibit much better Precision values. Once again, this is consistent with the ratio of the number of communities returned by Louvain, highlighted in table 5. Indeed, low Precision values are expected when ground truth communities are merged into the returned ones. And from ■-highlighted cells in table 5, one can remark that the tendency of Louvain algorithm to provide less communities than expected is emphasised for the criteria with lower Precision scores. Finally, one can focus on the somehow strange shape of the parametrised Newman-Girvan modularity, that achieves its minimum for the second smallest value of mixing parameter. As observed from fig. 11, the parameter value γ that maximises the $F1$ -score is not consistent over all the mixing parameters. Thus the choice of γ , which is a trade-off between the mixing parameters, clearly disadvantages graphs with lowest mixing parameters.

All Criteria on Doubly Stochastic Graphs. Here, we discuss the behaviours of the criteria applied to doubly stochastic graphs. As previously, the parameters for the Correlation Clustering criterion ($\lambda = 20/10n$) and for the parametrised

Newman-Girvan modularity ($\gamma = 1.75$), are chosen to maximise the average F1-score over mixing parameters. Recall and Precision are displayed in fig. 16. From the Recall curves on the right panel, we observe three different behaviours. First, we remark that two of the three versions of the Zahn criteria ($F_{Z,1}^\omega$ and $F_{Z,3}^\omega$) exhibit very low Recall values, even for the smallest mixing parameter. From ■-highlighted cells in table 5, it can be seen that, for $\mu = 0.18$, Louvain used with $F_{Z,1}^\omega$ returns in average almost 47 times more communities than expected. Recalling that there are 31 ground truth communities for 1600 nodes in the networks from the benchmark, this means that, in the community structure returned by this version of Louvain, a community contains in average about 1.1 node. On the other hand, Louvain used with $F_{Z,3}^\omega$ returns communities containing 5.35 nodes in average, when $\mu = 0.18$. While larger than for $F_{Z,1}^\omega$, this size is yet more than three times smaller than the smallest communities from the ground truth community structures, which contain 20 nodes. This explains the low Recall scores of $F_{Z,1}^\omega$ and $F_{Z,3}^\omega$. Opposite to this is the behaviour of Louvain used with the other version of the Zahn criterion, namely $F_{Z,2}^\omega$. Its Recall curve is constant, equal to 1, which means that it does not split any of the ground truth communities. However, when looking at ■-highlighted cells in table 5, we see that, from $\mu = 0.558$, the number of communities returned by Louvain with $F_{Z,2}^\omega$ is less than two, meaning that it returns some community structures where all the nodes belong to a unique community, and the Recall of this trivial community structure is 1. Thus, Louvain used with $F_{Z,2}^\omega$ does not split existing communities, but tends to merge them into one unique community as the mixing parameter increases. Finally, the parametrised and non-parametrised Newman-Girvan modularities, along with the Correlation Clustering criterion, have Recall values that remain equal to 1 for mixing parameters up to $\mu = 0.558$, and then start to decrease. The fact that parametrised Newman-Girvan and Correlation Clustering criteria have a lower Recall value than the Newman-Girvan modularity for $\mu = 0.769$ is explained by the fact that Louvain algorithm used with the two former criteria returns more communities than expected for this mixing parameter—see ■-highlighted cells in table 5.

Looking at the Precision curves on the left panel of fig. 16, we observe that the non-parametrised version of the Newman-Girvan modularity is not competitive with the parametrised Newman-Girvan and the Correlation Clustering criteria. On the other hand, these two latter ones exhibit an extremely close behaviour. About the two versions of the Zahn criteria with low Recall values, Louvain with $F_{Z,3}^\omega$ highlights the highest Precision up to $\mu = 0.558$, while Precision of Louvain with $F_{Z,1}^\omega$ decreases quickly. This is due to the fact that, as already discussed, the latter one returns essentially one community per node. Such communities account for 0 in the Recall. Finally, the version of the Zahn criterion $F_{Z,2}^\omega$ with the best possible Recall score, also exhibit the worst results in terms of Precision, with a slump of its Precision as soon as the second smallest mixing parameter.

Summary. Average F1-scores of all measures are provided in table 6, along with standard deviations. To improve readability, F1-scores are multiplied by

10, and standard deviations by 100. Parameters for the Correlation Clustering criteria and parametrised Newman-Girvan modularities are those that maximise the average F1-score, as explained in the previous paragraphs. We observe that the most accurate criteria are the parametrised ones. Indeed, the criterion
705 which provides the best F1-score overall is the Correlation Clustering criterion on simple graphs (F_{CC}^λ), closely followed by the Correlation Clustering and parametrised Newman-Girvan criteria on doubly stochastic graphs ($F_{CC}^{\omega,\lambda}$ and $F_{NG}^{\omega,\gamma}$). Last from this pool is the parametrised Newman-Girvan modularity on simple graphs (F_{NG}^γ). These four criteria exhibit average F1-scores above 0.85
710 for all the mixing parameters.

We now compare the four criteria unified by theorem 2, namely the Deviation to Indetermination (F_{DI}), Balanced Modularity (F_{BM}), Newman-Girvan modularity (F_{NG}) and Deviation to Uniformity (F_{DU}). First, we observe that the latter provides very high F1-scores compared to the other measures. Except
715 for the largest mixing parameter value, and its high standard deviations, the Deviation to Uniformity is almost competitive with the parametrised criteria. This is an artifact due to the benchmark, in which network community structures are typical deviations to regular graphs. On the other hand, the three others are not competitive with the doubly stochastic Newman-Girvan modularity (F_{NG}^ω) that generalises them all.
720

Our last observations concern the Zahn criteria. From table 6, it seems that none of the doubly stochastic versions of the Zahn criterion can compete with the one for simple graphs. However, it can be seen from table 5 that the number of communities returned by the Zahn criteria are quite different,
725 making them hard to compare based on their F1-score. To highlight this, in fig. 17, we plot the confusion matrices of a community structure with $\mu = 0.4$ returned using F_Z (left panel), and $F_{Z,3}^\omega$ (right panel) (for each criterion, the chosen community structure is the one that provides the maximum F1-score). We observe that their behaviours are opposite: the community structure found
730 on a simple graph correctly detects the largest communities, but split those of sizes 20 and 40: nodes from the ground truth 20-node communities are assigned to 36 communities by Louvain (16 are expected), and nodes from the 40-node communities are split into 22 communities (8 expected). On the other hand, Louvain used with $F_{Z,3}^\omega$ on a preprocessed graph perfectly detects communities
735 of size 20, 40 and 80. However, it splits the two communities of size 160 into 20 communities, and the 320-node community into 225 ones. This illustrates again the tendency of the proposed preprocessing to sharpen small-size communities, here at the expense of the larger ones. It also highlights that finding the more desirable partitioning remains application-dependant, and should not be chosen
740 on the basis of maximum F1-score only.

μ	Comparison # Uncovered Blocks VS # Expected Blocks							
	0.18	0.304	0.4	0.484	0.558	0.63		
F_{DI}	7.56 ⁻ (0.52)	7.6 ⁻ (0.47)	7.52 ⁻ (0.56)	7.48 ⁻ (0.6)	7.17 ⁻ (0.76)	6.78 ⁻ (0.84)	6.77 ⁻ (0.84)	6.43 ⁻ (0.85)
F_{BM}	4.72 ⁻ (0.48)	4.86 ⁻ (0.53)	5.29 ⁻ (0.5)	5.63 ⁻ (0.69)	5.66 ⁻ (0.69)	6.21 ⁻ (0.84)	6.05 ⁻ (0.74)	6.05 ⁻ (0.74)
F_{NG}	2.79 ⁻ (0.2)	3.95 ⁻ (0.27)	4.1 ⁻ (0.45)	4.53 ⁻ (0.58)	5.13 ⁻ (0.71)	5.47 ⁻ (0.68)	5.58 ⁻ (0.78)	5.76 ⁻ (0.78)
F_Z	1.16 ⁺ (0.06)	1.68 ⁺ (0.1)	2.9 ⁺ (0.23)	4.99 ⁺ (0.19)	7.97 ⁺ (0.29)	11.26 ⁺ (0.27)	13 ⁺ (0.14)	13.38 ⁺ (0.14)
F_{DU}	1.01 ⁻ (0.01)	1.02 ⁻ (0.03)	1.04 ⁻ (0.03)	1.06 ⁻ (0.05)	1.09 ⁻ (0.05)	1.2 ⁻ (0.07)	1.55 ⁻ (0.1)	1.94 ⁻ (0.14)
F_{NG}^*	1.79 ⁻ (0.07)	2.01 ⁻ (0.09)	2.07 ⁻ (0.13)	1.92 ⁻ (0.1)	1.73 ⁻ (0.12)	1.6 ⁻ (0.09)	1.54 ⁻ (0.08)	1.08 [*] (0.05)
F_{CC}^*	1 ⁻ (0.01)	1 ⁻ (0.01)	1.01 ⁻ (0.01)	1 [*] (0.01)	1.01 [*] (0.02)	1.04 [*] (0.03)	1.39 ⁺ (0.07)	1.94 ⁺ (0.07)
$F_{Z,1}^*$	46.9 ⁺ (1.39)	46.8 ⁺ (1)	47.2 ⁺ (1.22)	46.4 ⁺ (1.93)	45.8 ⁺ (2.04)	43 ⁺ (1.73)	40.5 ⁺ (2.23)	36.3 ⁺ (2.26)
$F_{Z,2}^*$	1.89 ⁻ (0.13)	3.95 ⁻ (0.32)	7.4 ⁻ (0.66)	11.67 ⁻ (2.8)	17.44 ⁻ (5.19)	28.28 ⁻ (5.96)	31 ⁻ (0)	31 ⁻ (0)
$F_{Z,3}^*$	9.72 ⁺ (0.24)	9.46 ⁺ (0.2)	8.94 ⁺ (0.18)	8.12 ⁺ (0.28)	7.38 ⁺ (0.22)	8.73 ⁺ (0.18)	10.93 ⁺ (0.09)	12.69 ⁺ (0.17)
F_{NG}^*	1.2 ⁻ (0.06)	1.54 ⁻ (0.07)	1.83 ⁻ (0.09)	2.21 ⁻ (0.12)	2.7 ⁻ (0.24)	3.36 ⁻ (0.32)	3.95 ⁻ (0.35)	4.47 ⁻ (0.39)
F_{NG}^*	1.02 ⁻ (0.03)	1.07 ⁻ (0.05)	1.1 ⁻ (0.05)	1.14 ⁻ (0.06)	1.16 ⁻ (0.05)	1.16 ⁻ (0.05)	1.18 ⁻ (0.05)	1.23 ⁺ (0.08)
F_{CC}^*	1.02 ⁻ (0.02)	1.04 ⁻ (0.04)	1.06 ⁻ (0.04)	1.1 ⁻ (0.05)	1.1 ⁻ (0.04)	1.08 ⁻ (0.05)	1.1 ⁻ (0.05)	1.28 ⁺ (0.09)

Table 5: Average ratio scores of the community structures returned by Louvain, with standard deviations between parentheses. Top : ■ : more communities than expected; ■ : number of communities less than half the expected number, in average. Bottom: ■ : less than 2 communities, in average; ■ : less than 6 nodes per community, for $\mu = 0.18$.

μ	F1-scores									
	0.18	0.304	0.4	0.484	0.558	0.63	0.699	0.769		
F_{DI}	3.18 (1.7)	3.19 (2)	3.18 (1.6)	3.22 (1.7)	3.27 (2)	3.4 (2)	3.46 (2.5)	4.38 (5.7)		
F_{BM}	4.27 (2.7)	4.20 (3.1)	3.94 (2)	3.78 (2.4)	3.83 (3)	3.61 (2.5)	3.74 (3)	4.76 (5.1)		
F_{NG}	6.75 (5.4)	5.09 (1.9)	5.08 (2.4)	4.67 (4.3)	4.16 (4)	4 (2.9)	4.14 (4.1)	4.94 (5.2)		
F_Z	9.99 (0.03)	9.95 (0.1)	9.62 (0.7)	9.99 (2.2)	1.69 (0.6)	0.75 (0.5)	0.25 (0.2)	0.12 (0.07)		
F_{DU}	9.98 (0.6)	9.95 (0.8)	9.88 (2.1)	9.78 (4)	9.75 (3.6)	9.47 (4.6)	8.79 (3.2)	7.19 (3.5)		
F_{NG}^*	9.17 (2.5)	8.55 (3.3)	8.84 (2.6)	9.15 (1.3)	9.37 (1.6)	9.45 (1.2)	9.28 (0.1)	8.70 (1.7)		
F_{CC}^*	10 (0.1)	10 (0.1)	9.99 (0.2)	9.99 (0.3)	9.98 (0.6)	9.97 (0.1)	9.77 (0.4)	9.11 (1.4)		
$F_{Z,1}^*$	0.06 (0.3)	0.06 (0.2)	0.04 (0.1)	0.04 (0.2)	0.03 (0.1)	0.03 (0.1)	0.03 (0.1)	0.02 (0.1)		
$F_{Z,2}^*$	9.08 (2.1)	5.38 (3.2)	3.4 (2.2)	2.46 (3.3)	1.94 (2.1)	1.53 (2.3)	1.43 (0)	1.43 (0)		
$F_{Z,3}^*$	6.47 (0.3)	4.33 (0.5)	4.01 (0.2)	3.86 (0.3)	2.45 (0.7)	1 (0.5)	0.3 (0.3)	0.2 (0.1)		
F_{NG}^*	9.72 (2.6)	9.40 (1.8)	9.18 (2.1)	8.68 (2.6)	7.86 (4.7)	6.9 (4.8)	6.54 (4)	6.07 (4.4)		
F_{NG}^*	9.94 (1)	9.86 (1.3)	9.87 (1)	9.85 (0.8)	9.87 (0.8)	9.84 (0.6)	9.61 (0.5)	8.79 (1.9)		
F_{CC}^*	9.94 (0.9)	9.89 (1.3)	9.91 (0.9)	9.89 (0.9)	9.89 (0.9)	9.9 (0.4)	9.66 (0.6)	8.98 (1.5)		

Legend for the F1-scores
■ 0 ■ 2.5 ■ 5 ■ 7.5 ■ 10

Table 6: Average F1-scores ($\times 10$) of the community structures returned by Louvain with criteria from table 3. Standard deviations ($\times 10^2$) between parentheses.

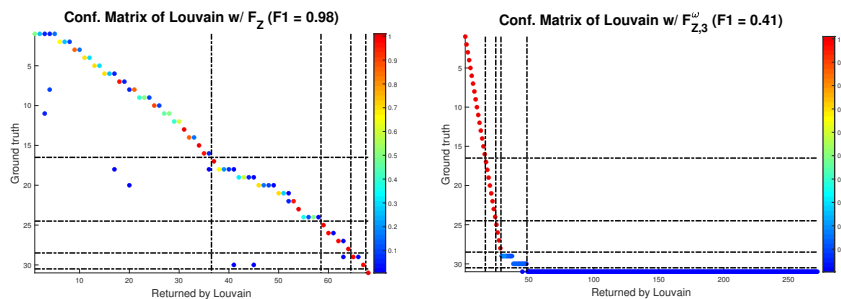


Figure 17: Confusion matrix of one community structure returned by Louvain used with F_Z (left), respectively $F_{Z,3}^\omega$ (right).

7. Conclusion and Future Perspectives

Broadly speaking, the aim of this study was to investigate the utility of doubly stochastic scaling as a preprocessing for community detection. In particular, its capacity to increase the detectability of small-size communities and communities with an imbalance in edge direction, such communities being in general poorly detected by community detection algorithms. The proposed preprocessing was presented in section 4, along with illustrations of its potential to sharpen those kinds of communities on toy examples and on a real-world network. In section 5, we have generalised a range of graph partitioning measures to weighted networks, with a particular focus on the case of doubly stochastic ones. Of utmost interest is the result that the doubly stochastic scaling unifies these measures, as stated in section 6.1. That is, all of the six measures defined for simple graphs can be expressed using only two parametrised measures for doubly stochastic graphs. Extensive comparisons of these measures have been conducted using SBMs in section 6.2, where we observed that the measures the most able to accurately uncover community structures are the parametrised ones, for both simple and preprocessed graphs, but foremost that a great care should be given to the choice of the measure to maximise, as different measures behave extremely differently.

In the future, we would like to investigate the impact of the diagonal added to ensure the convergence of the scaling in algo. 1 and 2, in terms of numerical values within the resulting preprocessed graph. This would provide us with theoretical basis to help making the right choice. Furthermore, to keep improving community detection methods, we would like to incorporate the knowledge obtained from scaling factors to the process of discovering communities. Indeed, after scaling, all nodes have the same degree. This may be seen as a non desirable feature, as it means that some initial information about node centrality (namely, the degree) is lost. And for real applications, the more central the node, the more harmful an error of assignation on this node. However, as stated in [30], another kind of information about node centrality, similar to hub and authority centralities from [46], is conveyed by the scaling factors, and should

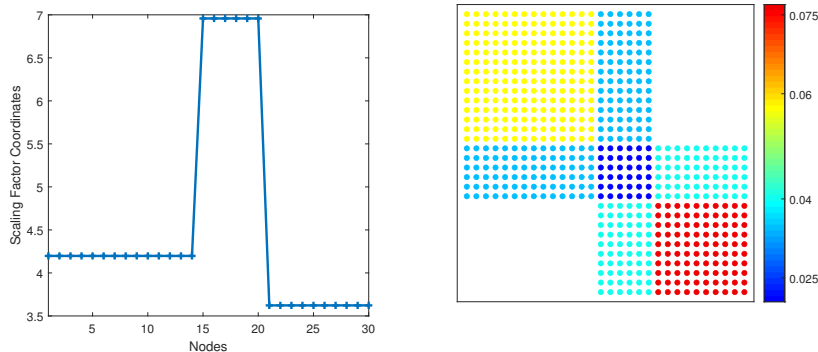


Figure 18: Doubly stochastic scaling of a toy example of overlapping communities. Left: Values of the scaling factor. Right: The scaling form of a simple graph exhibiting two overlapping communities.

be exploited to ensure that a greater care is taken to the correct assignation of nodes with high centrality. Finally, we would like to extend the proposed preprocessing to the detection of overlapping communities. Indeed, in many applications, one node can be involved in more than one community [47]. In a doubly stochastic scaling, a node belonging to many communities should produce high scaling factors (because of its high degree) and thus low numerical values in the doubly stochastic scaling, as illustrated in fig. 18. This may provide a framework to identify those nodes.

Funding

The study started when LIG was a PhD student at the Université Paul Sabatier, Toulouse, and ended while she hold a postdoctoral position at the University of Strathclyde. In this context, this project has been partially supported by an “Axes Thématiques Prioritaires” PhD funding from the Université Paul Sabatier, and by the Royal Academy of Engineering and the Office of the Chief Science Advisor for National Security under the UK Intelligence Community Postdoctoral Fellowship Programme.

Acknowledgements

We would like to thank Jean-Francois Marcotorchino for his feedback at an early stage of this work. We also would like to thank Mario Arioli and Renaud Lambiotte, for their constructive comments on this work when reviewing LIG’s PhD dissertation. We are also grateful to the reviewers from the Neurocomputing Journal for their helpful remarks and suggestion that have helped to improve the quality of this study.

795 **Conflict of interest**

The authors declare that they have no conflict of interest.

Availability of code and data

All the material needed to reproduce the experimental results from Section 6.2 can be found at github.com/luleg/StochasticMeasuresLouvain.

800 **References**

- [1] E. Estrada, P. A. Knight, A first course in network theory, Oxford University Press, USA, 2015.
- [2] A. R. Benson, D. F. Gleich, H. D. J., Higher-order network analysis takes off, fueled by old ideas and new data, SIAM News Blog, 21/1/2021.
- 805 [3] S. E. Schaeffer, Graph clustering, Computer science review 1 (1) (2007) 27–64.
- [4] S. Fortunato, D. Hric, Community detection in networks: A user guide, Physics reports 659 (2016) 1–44.
- [5] L. N. Veldt, Optimization frameworks for graph clustering, Ph.D. thesis, 810 Purdue University Graduate School (2019).
- [6] M. E. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical review E 69 (2) (2004) 026113.
- [7] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on pattern analysis and machine intelligence 22 (8) (2000) 888–905.
- 815 [8] M. Rosvall, D. Axelsson, C. T. Bergstrom, The map equation, The European Physical Journal Special Topics 178 (1) (2009) 13–23.
- [9] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hofer, Z. Nikoloski, D. Wagner, On modularity clustering, IEEE transactions on knowledge and data engineering 20 (2) (2007) 172–188.
- 820 [10] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of statistical mechanics: theory and experiment 2008 (10) (2008) P10008.
- [11] Z. Yang, R. Algesheimer, C. J. Tessone, A comparative analysis of community detection algorithms on artificial networks, Scientific reports 6 (1) 825 (2016) 1–18.
- [12] S. Fortunato, M. Barthelemy, Resolution limit in community detection, Proceedings of the national academy of sciences 104 (1) (2007) 36–41.

- [13] P. C. Céspedes, J.-F. Marcotorchino, Comparing different modularization criteria using relational metric, in: International Conference on Geometric Science of Information, Springer, 2013, pp. 180–187.
- 830 [14] P. B. Slater, Hubs and clusters in the evolving us internal migration network, arXiv preprint arXiv:0809.2768.
- [15] L. Le Gorrec, S. Mouysset, I. Duff, P. Knight, D. Ruiz, Uncovering hidden block structure for clustering, in: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer, 2019.
- 835 [16] F. Wang, P. Li, A. C. König, M. Wan, Improving clustering by learning a bi-stochastic data similarity matrix, Knowledge and information systems 32 (2) (2012) 351–382.
- 840 [17] C. Laclau, I. Redko, B. Matei, Y. Bennani, V. Brault, Co-clustering through optimal transport, in: International Conference on Machine Learning, PMLR, 2017, pp. 1955–1964.
- [18] S. Emmons, S. Kobourov, M. Gallant, K. Börner, Analysis of network clustering algorithms and cluster quality metrics at scale, PloS one 11 (7) (2016) e0159161.
- 845 [19] F. R. Chung, Lectures on spectral graph theory, CBMS Lectures, Fresno 6 (92) (1996) 17–21.
- [20] J.-C. Delvenne, S. N. Yaliraki, M. Barahona, Stability of graph communities across time scales, Proceedings of the national academy of sciences 107 (29) (2010) 12755–12760.
- 850 [21] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: Computer and Information Sciences - ISCIS 2005, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 284–293.
- [22] B. Karrer, M. E. Newman, Stochastic blockmodels and community structure in networks, Physical review E 83 (1) (2011) 016107.
- 855 [23] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–864.
- [24] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proceedings of the 5th International Conference on Learning Representations (ICLR-17), 2016.
- 860 [25] R. Campigotto, P. C. Céspedes, J.-L. Guillaume, A generalized and adaptive method for community detection, arXiv preprint arXiv:1406.2518.
- [26] R. Sinkhorn, P. Knopp, Concerning nonnegative matrices and doubly stochastic matrices, Pacific Journal of Mathematics 21 (2) (1967) 343–348.
- 865

- [27] R. A. Brualdi, *Combinatorial matrix classes*, Vol. 13, Cambridge University Press, 2006.
- [28] A. Berman, R. J. Plemmons, *Nonnegative matrices in the mathematical sciences*, SIAM, 1994.
- 870 [29] A. Pothén, C.-J. Fan, Computing the block triangular form of a sparse matrix, *ACM Transactions on Mathematical Software (TOMS)* 16 (4) (1990) 303–324.
- [30] P. A. Knight, The sinkhorn–knopp algorithm: convergence and applications, *SIAM Journal on Matrix Analysis and Applications* 30 (1) (2008) 261–275.
- 875 [31] P. A. Knight, D. Ruiz, B. Uçar, A symmetry preserving algorithm for matrix scaling, *SIAM journal on Matrix Analysis and Applications* 35 (3) (2014) 931–955.
- [32] R. Ulanowicz, C. Bondavalli, M. Egnotovitch, Network analysis of trophic dynamics in south florida ecosystem, fy 97: The florida bay ecosystem, Annual Report to the United States Geological Service Biological Resources Division. Ref. No. [UMCES]CBL.
- 880 [33] A. R. Benson, Tools for higher-order network analysis, Ph.D. thesis, Stanford University (2017).
- 885 [34] P. Conde Cespedes, Modélisations et extensions du formalisme de l’analyse relationnelle mathématique à la modularisation des grands graphes, Ph.D. thesis, Paris 6 (2013).
- [35] C. Zahn, Jr, Approximating symmetric relations by equivalence relations, *Journal of the Society for Industrial and Applied Mathematics* 12 (4) (1964) 840–847.
- 890 [36] N. Bansal, A. Blum, S. Chawla, Correlation clustering, *Machine learning* 56 (1) (2004) 89–113.
- [37] M. E. Newman, Mixing patterns in networks, *Physical review E* 67 (2) (2003) 026126.
- 895 [38] M. E. Newman, Analysis of weighted networks, *Physical review E* 70 (5) (2004) 056131.
- [39] L. Le Gorrec, Équilibrage bi-stochastique des matrices pour la détection de structures par blocs et applications, Ph.D. thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier (2019).
- 900 [40] L. Hubert, P. Arabie, Comparing partitions, *Journal of classification* 2 (1) (1985) 193–218.

- [41] J. Ah-Pine, J.-F. Marcotorchino, Statistical, geometrical and logical independences between categorical variables, in: 12th International Conference on Applied Stochastic Models and Data Analysis (ASMDA 2007), 2007.
- 905 [42] E. D. Demaine, D. Emanuel, A. Fiat, N. Immorlica, Correlation clustering in general weighted graphs, *Theoretical Computer Science* 361 (2-3) (2006) 172–187.
- [43] N. Veldt, D. F. Gleich, A. Wirth, A correlation clustering framework for community detection, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 439–448.
- 910 [44] J. Reichardt, S. Bornholdt, Statistical mechanics of community detection, *Physical review E* 74 (1) (2006) 016110.
- [45] E. Abbe, Community detection and stochastic block models: recent developments, *The Journal of Machine Learning Research* 18 (1) (2017) 6446–6531.
- 915 [46] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM* 46 (5) (1999) 604–632.
- [47] J. Xie, S. Kelley, B. K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, *Acm computing surveys (csur)* 45 (4) (2013) 1–35.
- 920