An Exact No Free Lunch Theorem for Community Detection

Arya D. McCarthy(✉), Tongfei Chen, and Seth Ebner

Johns Hopkins University, Baltimore, USA
arya@jhu.edu

Abstract. A precondition for a No Free Lunch theorem is evaluation with a loss function which does not assume *a priori* superiority of some outputs over others. A previous result for community detection by [12] relies on a mismatch between the loss function and the problem domain. The loss function computes an expectation over only a subset of the universe of possible outputs; thus, it is only *asymptotically* appropriate with respect to the problem size. By using the correct random model for the problem domain, we provide a stronger, exact No Free Lunch theorem for community detection. The claim generalizes to other set-partitioning tasks including core–periphery separation, k-clustering, and graph partitioning. Finally, we review the literature of proposed evaluation functions and identify functions which (perhaps with slight modifications) are compatible with an exact No Free Lunch theorem.

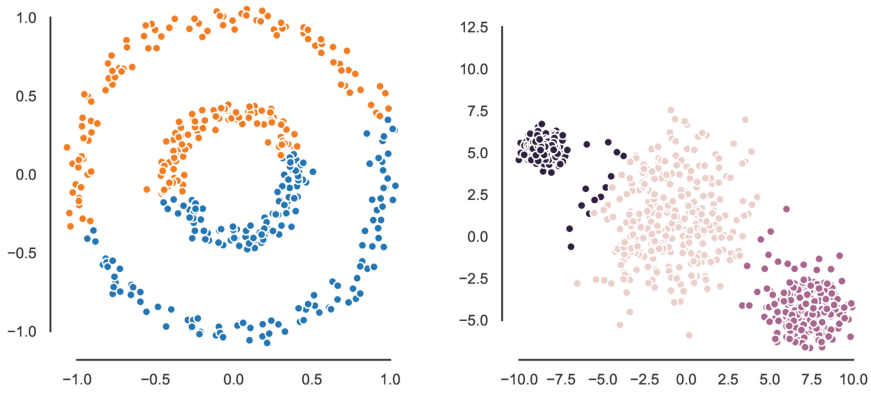
1 Introduction

A myriad of tasks in machine learning and network science involve discovering structure in data. Especially as we process graphs with millions of nodes, analysis of individual nodes is untenable, while global properties of the graph ignore local details. It becomes critical to find an intermediate level of complexity, whether it be communities, cores and peripheries, or other structures. Points in metric space and nodes of graphs can be clustered, and hubs identified, using algorithms from network science. A longstanding theoretical question in machine learning has been whether an "ultimate" clustering algorithm is a possibility or merely a fool's errand.

Largely, the question was addressed by [18] as a **No Free Lunch theorem**, a claim about the limitations of algorithms with respect to their problem domain. When an appropriate function is chosen to quantify the error (or **loss**), no algorithm can be superior to any other: an improvement across one subset of the problem domain is balanced by diminished performance on another subset. This is jarring at first. Are we not striving to find the best algorithms for our tasks? Yes—but by making specific assumptions about the subset of problems we expect to encounter, we can be comfortable tailoring our algorithms to those problems and sacrificing performance on remote cases.

As an example, the k-means algorithm for k-clustering is widely used for its simplicity and strength, but it assumes spherical clusters, equal variance in those

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2019, SCI 881, pp. 176–187, 2020.
https://doi.org/10.1007/978-3-030-36687-2_15



(a) Non-spherical clusters (b) Unequal variances

Fig. 1. k-means clustering when certain assumptions are violated. Although these are toy examples, the message is relevant to community detection, where algorithms' success is likewise predicated on assumptions about the problem.

clusters, and similar cluster sizes (equivalent to a homoscedastic Gaussian prior). Figure 1 shows the degraded performance on problems where these assumptions are violated.

To prove a No Free Lunch theorem for a particular task demands an appropriate loss function. A No Free Lunch theorem was argued for community detection [12], using the adjusted mutual information function [17].[1] However, the theorem is inexact. A No Free Lunch theorem relies on a loss function which imparts **generalizer-independence** (formally defined in Sect. 2.4): one which does not assume *a priori* that some prediction is superior to another. The loss function used in the proof is only *asymptotically* independent in the size of the input. We present a correction: by substituting an appropriate loss function, we are able to claim an exact version of the No Free Lunch theorem for community detection. The result generalizes to other set-partitioning tasks when evaluated with this loss function, including clustering, k-clustering, and graph partitioning.

2 Background

2.1 Community Detection

A number of tasks on graphs seek a partition of the graph's nodes that maximizes a score function. Situated between the microscopic node-level and the macroscopic graph-level, these partitions form a **mesoscopic structure**—be it a core–periphery separation, a graph coloring, or our focus: **community detection** (CD). Community detection has been historically ill-defined [13,19], though

[1] Throughout this work, we assume that we evaluate against a known ground truth, as opposed to some intrinsic measure of partition properties like modularity [10].

the intuition is to collect nodes with high interconnectivity (or edge density) into communities with low edge density between them. The task is analogous to clustering, in which points near one another in a metric space are grouped together.

To assess whether the formulation of community detection matches one's needs, one performs extrinsic evaluation against a known **ground truth** clustering. This ground truth can come from domain knowledge of real-world graphs or can be planted into a graph as a synthetic benchmark. After running community detection on the graph, some similarity or error measure between the computed community structure and the correct one can be computed.

No Bijection Between True Structure and Graph. Unfortunately, ground truth communities do not imply a single graph—and vice versa. [12] go as far as to claim, "Searching for the ground truth partition without knowing the exact generative mechanism is an impossible task".

We can imagine the following steps for how problem instances are created, given that we have $N = |V|$ nodes:

1. Sample (true) partition $\mathcal{T} \in \Omega$;
2. Generate graph G from $\mathcal{T}$ by adding edges according to the edge-generating process g.

where Ω is our **universe**: the space of all partitions of $N = |V|$ objects. Given a graph $G = (V, E)$, we can imagine multiple truths $\mathcal{T}_i \in \Omega$ that could define its edge set E by different generative processes $g_i : \Omega \to \Gamma$, where Γ is the set of all graphs with N nodes. [12] give a proof that extends from this simple example: Imagine that $\mathcal{T}_1$ partitions the N nodes into N components (the N-partition), and $\mathcal{T}_2$ partitions them into 1 component (the 1-partition). Let g_1 exactly specify the number of edges between each pair of communities, such that $g_1(\mathcal{T}_1)$ is G with probability 1. Similarly, let g_2 be an Erdős–Rényi model such that $g_2(\mathcal{T}_2)$ is G with nonzero probability. ([12] note that this is easily extended to graphs with more nodes). We thus have two different ways to create a single graph; how can a method discern the correct one, without knowledge of g?

Community detection is then an ill-posed inverse problem: Use a function $f : \Gamma \to \Omega$ to produce a clustering $\mathcal{C} = f(G)$, which is hopefully representative of $\mathcal{T}$ [12, Appendix C].[2] The function f is not a bijection, so there isn't a unique $\mathcal{T}$ represented in the given graph. Our algorithm f must encode our prior beliefs about the generative process g to select from among candidates. For this reason, we must hope that the benchmark graphs that we use are representative of the generative process for graphs in our real-world applications. That is, we hope that our benchmark domain matches our practical domain.

Other Set-Partitioning Tasks. While the remainder of this work focuses on community detection, our claims are relevant to other set-partitioning tasks. Notable examples are clustering (the vector space analogue to community detection), graph k-partitioning, and k-clustering. Metadata about the nodes and edges,

[2] That is, the objective is to find $f = g^{-1}$. In general, this does not exist.

such as vector coordinates, are used to guide the identification of such structure, but the tasks are all fundamentally set-partitioning problems. They can also have different universes Ω—the latter tasks have a smaller universe than does community detection, for a given graph G: They consider only partitions with a fixed number of clusters.

2.2 No Free Lunch Theorems

The **No Free Lunch theorem** in machine learning is a claim about the universal (in)effectiveness of learning algorithms. Every algorithm performs equally well when averaging over all possible input–output pairs. Formally, for any learning method f, the error (or **loss**) $\mathcal{L}$ of the method f, summed over all possible problems $(g, \mathcal{T})$ equals a loss-specific constant $\Lambda(\mathcal{L})$:

$$\sum_{(g,\mathcal{T})} \mathcal{L}\left(\mathcal{T}, f\left(g\left(\mathcal{T}\right)\right)\right) = \Lambda(\mathcal{L}), \tag{1}$$

defining the edge-generative process g and partition $\mathcal{T}$ as above. This loss is thus *generalizer-independent*. To reduce loss on a particular set of problems means sacrificing performance on others—"*there is no free lunch*" [16,18]. Judiciously choosing which set to improve involves making assumptions about the distribution of the data: as we've mentioned, k-means is a method for k-clustering which works well on data with spherical covariance, similar cluster sizes, and roughly equal class sizes. When these assumptions are violated, performance suffers and overall balance is achieved.

2.3 Community Detection as Supervised Learning

We follow [12] in framing the task of community detection (CD) as a learning problem. While recent algorithms, e.g. [1], have introduced learnable parameters to community detection algorithms the CD literature's algorithms are by and large untrained. These untrained algorithms encode knowledge of the problem domain in prior beliefs. We note that our work and [12] straightforwardly handle both of these cases.

In general supervised machine learning problems, we seek to learn the function that maps an input space $\mathbf{X}$ to an output space $\mathbf{Y}$. We consider problem instances as sampled from random variables over each, so our goal is to learn the conditional distribution $p(Y \mid X)$. In the process of training on a dataset $\mathcal{D}$, we develop a distribution over hypotheses q which are estimates of the distribution p.

In the case of most community detection algorithms, our input space is the set of graphs on N nodes Γ, and the output space is Ω. There is no training data: $\mathcal{D} = \varnothing$. All of our prior beliefs about p must be encoded in the prior distribution $\Pr(q)$. That is, the model itself must contain our beliefs about the definition of community structure. Only from the encoded $\Pr(q)$ and an observed $x \in \mathbf{X}$ (our graph G) do we form our point estimate of the true distribution p [12]. However, in the case of trainable CD algorithms, we encode our beliefs in the posterior distribution $\Pr(q \mid \mathcal{D})$.

2.4 Loss Functions and a Priori Superiority

How should we evaluate an algorithm's predictions? Classification accuracy won't cut it: When comparing to the ground truth, there are no specific labels (e.g. no notion of a specific "Cluster 2")—only unlabeled groups of like entities. We settle for a measure of similarity in the groupings, quantifying how much the computed partition tells us about the ground truth.

A popular choice of measure is the *normalized mutual information* (NMI) [5] between the prediction and the ground truth. While this measure has a long history in community detection, its flaws have been well-noted [8,9,12,17]. It imposes a "geometric" structure upon the universe Ω,[3] so something as simple as guessing the trivial all-singletons clustering outperforms methods that try at all to find a mesoscopic-level structure [9]. The property which NMI lacks is *generalizer-independence*.

The property of generalizer-independence is defined by the generalization error function, an expectation of the loss $\mathbb{E}[L \mid p, q, \mathcal{D}]$. To satisfy this property, the generalization error must be independent of the particular true value $\mathcal{T}$. This is best expressed by Eq. 1.

The adjusted mutual information (AMI, defined in Sect. 3) [17] is a proposed replacement for NMI which does not impose a geometric structure upon the space. Unfortunately, this benefit is not fully realized when the expectation is computed over a space $\Psi \subset \Omega$. For the Ψ used in [12], the expected AMI across all problems is only *asymptotically* generalizer-independent as the graph size grows—it is within some diminishing amount of error $\varepsilon(N)$ of generalizer-independence, as proven by [12].

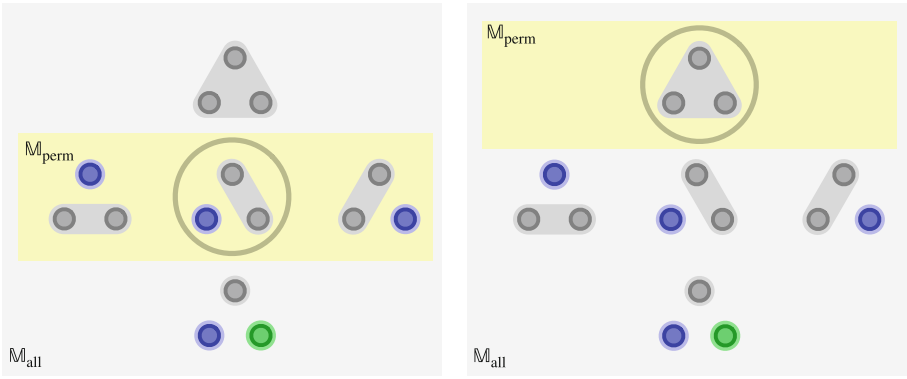
3 Previous Result: Approximate No Free Lunch Theorem

[12] frame community detection in the style of learning algorithms, letting them prove a No Free Lunch theorem for community detection. They note that the claim holds for "an appropriate choice of $\ldots\mathcal{L}$"—specifically a loss function $\mathcal{L}$ that is generalizer-independent—but their chosen loss function is not fully generalizer-independent. They also consider a stricter property than generalizer-independence: **homogeneity**. With a homogeneous loss function, the *distribution* of the error (not just its expectation) is identical, regardless of the ground truth. A measure which deviates from homogeneity may have this deviation bounded by a function of the number of vertices (the graph order).

Lemma 1 ([12]). *Adjusted mutual information (AMI) is a homogeneous loss function over the interior of the space of partitions of N objects, i.e., excluding the 1-partition and the N-partition. Including these, AMI is homogeneous within $\frac{1}{\mathcal{B}_N}$.*[4]

[3] To take the example of [12], L^2 loss (squared Euclidian distance) imposes a geometric structure: In the task of guessing points in the unit circle, guessing the center will garner a higher reward, on average, than any other point.

[4] $\mathcal{B}_N$ is the N-th Bell number, i.e., the number of partitions of a set of N nodes.



(a) Ground truth has cluster size pattern $\{2, 1\}$.

(b) Ground truth has cluster size pattern $\{3\}$.

Fig. 2. $\mathbb{M}_{\text{all}}$ and $\mathbb{M}_{\text{perm}}$ when clustering three nodes, for two different ground truths (circled). The top and bottom clusterings—the 1 and N clusterings—are the boundary partitions. All other partitions form the interior. $\mathbb{M}_{\text{perm}}$ changes based on the ground truth, but $\mathbb{M}_{\text{all}}$ stays the same.

[18] gives a generalized No Free Lunch theorem, which assumes a homogeneous loss.

Theorem 1 ([18]). *For homogeneous loss $\mathcal{L}$, the uniform average over all distributions p of* $\Pr(\ell \mid p, \mathcal{D})$ *equals* $\frac{\Lambda(\ell)}{|\mathbf{Y}|}$. (Plainly, "There is no free lunch").

[12] then use Wolpert's result with their inexactly homogeneous measure to claim a No Free Lunch result.

Theorem 2 ([12]). *By Lemma 1 and Theorem 1, for the community detection problem with a loss function of AMI, the uniform average over all distributions p of* $\Pr(\ell \mid p, \mathcal{D})$ *equals* $\frac{\Lambda(\ell)}{|\mathbf{Y}|}$.

But this choice of measure (AMI) is not, in fact, homogeneous over the *entire* universe Ω (Lemma 1). A strategy that guesses either of the non-interior (i.e., boundary) partitions—the 1-partition or N-partition—will yield a higher-than-average reward. There is indeed a negligible amount of free lunch—a free morsel, if you will.

4 Diagnosis: Random Models

[12] use AMI out of the box, as proposed by [17], which involves subtracting an expected value from a raw score. Unfortunately, AMI as given takes its expectation over the wrong distribution. Because of the mismatch, [12]'s claim of homogeneity is accurate only to within $\frac{1}{\mathcal{B}_N}$ when considering the trivial partitions into either one community or N communities.

Correcting this is arguably a pedantic demand, for two reasons:

1. The fraction $\frac{1}{\mathcal{B}_N}$ converges to 0 superexponentially as N increases.
2. The deficiency is only present when $\mathcal{T}$ is one of the trivial partitions. Otherwise, AMI as used is exactly homogeneous. But the trivial partitions reflect a lack of any mesoscopic community structure.

Nevertheless, we'd like to see a tight claim of generalizer independence. To do this, we must select the proper **random model**, a sample space for a distribution.

AMI adjusts NMI by subtracting the expected value from both the numerator and the denominator, shown in blue:

$$\mathrm{AMI}(\mathcal{C},\mathcal{T}) \triangleq \frac{I(\mathcal{C},\mathcal{T}) - \mathbb{E}_{\mathcal{C}',\mathcal{T}'}\left[I(\mathcal{C}',\mathcal{T}')\right]}{\max_{\mathcal{C}',\mathcal{T}'} I(\mathcal{C}',\mathcal{T}') - \mathbb{E}_{\mathcal{C}',\mathcal{T}'}\left[I(\mathcal{C}',\mathcal{T}')\right]}, \tag{2}$$

where I is the mutual information, maximized when the specific clustering $\mathcal{C}$ equals the ground truth $\mathcal{T}$. By inspecting Eq. 2, we see that AMI's value is 1 (the maximum) when $\mathcal{C} = \mathcal{T}$, 0 in expectation, and negative when the agreement between $\mathcal{C}$ and $\mathcal{T}$ is worse than chance.

Subtly hidden in this equation is the decision of which distribution to compute the expectation over. For decades, this distribution has been what [2] call $\mathbb{M}_{\mathrm{perm}}$: all partitions of the same **partition shape**[5] as $\mathcal{C}$ or $\mathcal{T}$. For example, if $\mathcal{C}$ partitioned 7 nodes into clusters of sizes 2, 2, and 3, then we would compute the expected mutual information over all clusterings where one had cluster sizes of 2, 2, and 3.

[9] argue that $\mathbb{M}_{\mathrm{perm}}$ is inappropriate. To use this random model assumes that we can only produce outputs within that restricted space, when in actuality Ω is the set of *all* partitions of N nodes. Furthermore, during evaluation, we hold our ground truth fixed, rather than marginalizing over possible ground truths. Were we to instead consider a distribution over $\mathcal{T}$s, we would add noise from other possible generative processes which yield the same graph from different underlying partitions. In our average, we might be including scores on ground truths that better align with our notions of, say, core–periphery partitioning. For this reason, we take a **one-sided expectation**—over $\mathcal{C}$, holding $\mathcal{T}$ fixed. The one-sided distribution over all partitions of N nodes is called $\mathbb{M}_{\mathrm{all}}^1$ [2]. This distribution is what we use for our AMI expectation, giving a measure denoted as $\mathrm{AMI}_{\mathrm{all}}^1$, which is recommended by [9]. It takes the form

$$\mathrm{AMI}_{\mathrm{all}}^1(\mathcal{C},\mathcal{T}) \triangleq \frac{I(\mathcal{C},\mathcal{T}) - \mathbb{E}_{\mathcal{C}'\sim\mathbb{M}_{\mathrm{all}}^1}\left[I(\mathcal{C}',\mathcal{T})\right]}{\max_{\mathcal{C}'} I(\mathcal{C}',\mathcal{T}) - \mathbb{E}_{\mathcal{C}'\sim\mathbb{M}_{\mathrm{all}}^1}\left[I(\mathcal{C}',\mathcal{T})\right]}. \tag{3}$$

The differences between $\mathbb{M}_{\mathrm{all}}$ and $\mathbb{M}_{\mathrm{perm}}$ are illustrated in Fig. 2 under $|V| = 3$. We will now show that substituting $\mathbb{M}_{\mathrm{all}}$ for $\mathbb{M}_{\mathrm{perm}}$, hence using $\mathrm{AMI}_{\mathrm{all}}^1$, allows for an exact No Free Lunch theorem.

[5] A multiset of cluster sizes, also called a decomposition pattern [3] or a group-size distribution [6]. It is equivalent to an integer partition of N.

5 An Exact No Free Lunch Theorem

We strengthen the No Free Lunch theorem for community detection given by [12] by using an improved loss function, $\mathrm{AMI}^1_{\mathrm{all}}$, for community detection. Our proof does not distinguish the "boundary" partitions (the two trivial partitions) from the "interior" partitions (the remainder). It is entirely agnostic toward the particular ground truth $\mathcal{T}$, which is exactly what we need. We improve the previous result by moving from $\mathbb{M}_{\mathrm{interior}}$ (which excludes the boundary partitions) to $\mathbb{M}_{\mathrm{all}}$.

5.1 Generalizer-Independence of $\mathbf{AMI}^1_{\mathbf{all}}$

Lemma 2. $\mathrm{AMI}^1_{\mathrm{all}}$ *is a generalizer-independent loss function over the* entire *space* $\mathbb{M}_{all}$ *of partitions of* N *objects.*

Proof. Like [12], we must show that the sum of scores is independent of $\mathcal{T}$:

$$\forall \mathcal{T}_1, \mathcal{T}_2, \quad \sum_{\mathcal{C} \in \Omega} \mathrm{AMI}^1_{\mathrm{all}}(\mathcal{C}, \mathcal{T}_1) = \sum_{\mathcal{C} \in \Omega} \mathrm{AMI}^1_{\mathrm{all}}(\mathcal{C}, \mathcal{T}_2), \tag{4}$$

where Ω is the space of all partitions of N objects. Unlike [12], we take the AMI expectation over all $\mathcal{B}_N$ clusterings in Ω using the random model $\mathbb{M}^1_{\mathrm{all}}$ [2].

To prove our claim about Eq. 4, we note that denominator of $\mathrm{AMI}^1_{\mathrm{all}}$ is a constant with respect to $\mathcal{C}$ (Eq. 3), so we can factor it out of the sum and restrict our attention to the numerator. This is because the max-term in the denominator is the constant $\log N$ [2] and the expectation term for a given $\mathcal{T}$ is independent of the particular $\mathcal{C}$. Having factored this out, we will now prove Eq. 4 by the stronger claim:

$$\sum_{\mathcal{C} \in \Omega} \left[I(\mathcal{C}, \mathcal{T}) - \mathbb{E}_{\mathcal{C}' \sim \mathbb{M}^1_{\mathrm{all}}} [I(\mathcal{C}', \mathcal{T})] \right] \stackrel{?}{=} 0 \quad \forall \mathcal{T} \tag{5}$$

To prove Eq. 5, we separate the summation's two terms:

$$\sum_{\mathcal{C} \in \Omega} [I(\mathcal{C}, \mathcal{T})] - \sum_{\mathcal{C} \in \Omega} \left[\mathbb{E}_{\mathcal{C}' \sim \mathbb{M}^1_{\mathrm{all}}} [I(\mathcal{C}', \mathcal{T})] \right] \tag{6}$$

The expectation is uniform over the universe Ω,[6] so we can apply the law of the unconscious statistician, then push the constant probability out, to get

$$\sum_{\mathcal{C} \in \Omega} [I(\mathcal{C}, \mathcal{T})] - \sum_{\mathcal{C} \in \Omega} \left[\frac{1}{|\Omega|} \sum_{\mathcal{C}' \in \Omega} [I(\mathcal{C}', \mathcal{T})] \right] \tag{7}$$

[6] Why do we assume uniformity over Ω? Because this is the highest-entropy (i.e., least informed) distribution—it places the fewest assumptions on the distribution.

Because the inner sum is independent of any particular $\mathcal{C}$, the outer sum is a sum of constants—one for each element in Ω. We can now express Eq. 5 as follows, where the reciprocals straightforwardly cancel out:

$$\sum_{\mathcal{C}\in\Omega} [I(\mathcal{C},\mathcal{T})] - |\Omega|\frac{1}{|\Omega|}\sum_{\mathcal{C}'\in\Omega} [I(\mathcal{C}',\mathcal{T})] \equiv 0. \tag{8}$$

This equivalence implies that Eq. 4 is true. □

The proof is valid without loss of generality vis-à-vis the distribution—that is, as long as the AMI expectation is computed uniformly over the problem universe Ω, AMI is a generalizer-independent measure. This stipulation is relevant to tasks which assume a fixed number of clusterings—using $\mathbb{M}_{\text{num}}$—like k-clustering and graph partitioning.

Having demonstrated the generalizer-independence of AMI, we can define our loss function as, say,

$$\mathcal{L}(\mathcal{C},\mathcal{T}) = 1 - \text{AMI}(\mathcal{C},\mathcal{T}). \tag{9}$$

The loss is zero when we exactly match the true clustering and positive otherwise.

Having proven the generalizer-independence of $\text{AMI}^1_{\text{all}}$, we now turn to a more general form of the No Free Lunch theorem, which admits not just a homogeneous loss function but any generalizer-independent loss.

Theorem 3 ([18]). *For generalizer-independent loss ℓ, the uniform average over all p, $\mathbb{E}[\ell \mid p, \mathcal{D}]$, equals $\frac{\Lambda(\ell)}{|\mathbf{Y}|}$. (Plainly, "There is no free lunch.")*

Proof. See [18]. □

Theorem 4 (No Free Lunch theorem for community detection and other set-partitioning tasks). *For a set-partitioning problem with a loss function of adjusted mutual information* using the appropriate random model for the task*, the uniform average over all p, $\mathbb{E}[\ell \mid p, \mathcal{D}]$, equals $\frac{\Lambda(\ell)}{|\mathbf{Y}|}$.*

Proof. Lemma 2 proves that AMI *using the appropriate random model* is generalizer-independent. Applying Theorem 3 completes the proof [12]. □

5.2 Other Measures

AMI stemmed from a series of efforts to improve normalized mutual information (NMI). We note that six other measures, when extended to $\mathbb{M}^1_{\text{all}}$ instead of $\mathbb{M}_{\text{perm}}$, are also generalizer-independent: the adjusted Rand index (ARI) [4], relative NMI (rNMI) [21], ratio of relative NMI (rrNMI) [20], Cohen's κ [7], corrected NMI (cNMI) [6], and standardized mutual information (SMI) [14]. We elide the proofs because they are similar to Lemma 2. Each of the six measures satisfies the precondition for the No Free Lunch theorem when the random model matches the problem domain.

Of late, a renewed push has advocated using the adjusted Rand index (ARI) [4] to evaluate community detection; in fact, ARI and AMI are specializations of the same underlying function which uses *generalized* information-theoretic measures [15]. Every claim in the proof works for ARI, by replacing every mutual information I term with the Rand index RI.

Another line of research, focusing on improving NMI, produced rNMI [21], rrNMI [20], and cNMI [6]. We note that rrNMI is identical to one-sided AMI when both are extended to $\mathbb{M}^1_{\text{all}}$. Consequently, our claim above works just as well for rrNMI. Further, because we were able to ignore the denominator of AMI in our proof of Lemma 2, we can do the same for rrNMI, which gives its unnormalized variant, rNMI. This means that rNMI is a generalizer-independent measure as well, when used in the appropriate one-sided random model. The practical benefit of normalizing rNMI into rrNMI is that the normalized measure gives a more interpretable notion of success.

Additionally, Lemma 2 holds true for standardized mutual information (which is equivalent to standardized variation of information and standardized V-measure) [14], the adjusted variation of information [17], and for Cohen's κ, advocated for CD by [7]. This is because each measure shares the form of AMI: an observed score minus an expectation.

Finally, to show whether cNMI is generalizer-independent under the correct random model, we must show how to specialize it into a one-sided variant, because there is room for interpretation about how this should be done, even restricting our focus to $\mathbb{M}^1_{\text{all}}$. The expression for cNMI

$$\text{cNMI}(\mathcal{C},\mathcal{T}) \triangleq \frac{2\,\text{NMI}(\mathcal{C},\mathcal{T}) - \mathbb{E}_{\mathcal{C}'}\left[\text{NMI}(\mathcal{C}',\mathcal{T})\right] - \mathbb{E}_{\mathcal{T}'}\left[\text{NMI}(\mathcal{C},\mathcal{T}')\right]}{2 - \mathbb{E}_{\mathcal{C}'}\left[\text{NMI}(\mathcal{C}',\mathcal{C})\right] - \mathbb{E}_{\mathcal{T}'}\left[\text{NMI}(\mathcal{T},\mathcal{T}')\right]} \tag{10}$$

depends on both $\mathcal{C}$ and $\mathcal{T}$ relative to the universes that contain them. Our specialization should remove dependence on the family of $\mathcal{T}$, so we arrive at the following expression after cancellation and noting that the NMI between a clustering and itself is 1:

$$\text{cNMI}(\mathcal{C},\mathcal{T}) = \frac{\text{NMI}(\mathcal{C},\mathcal{T}) - \mathbb{E}_{\mathcal{C}'}\left[\text{NMI}(\mathcal{C}',\mathcal{T})\right]}{1 - \mathbb{E}_{\mathcal{C}'}\left[\text{NMI}(\mathcal{C}',\mathcal{C})\right]} \tag{11}$$

As it turns out, this quasi-adjusted measure is also generalizer-independent.

In general, we now have a recipe for generalizer-independent loss functions: They can be created by subtracting the expected score from the observed score. This recipe works whenever a uniform expectation can be well defined.

6 Conclusion

We now have a proof of the No Free Lunch theorem for community detection and clustering that is both complete and exact. We show that a corrected form

of AMI, namely $\text{AMI}^1_{\text{all}}$, computes its expectation in a way that does not advantage the boundary partitions (1 cluster and N singleton clusters). Indeed, this expectation is over the entire universe of partitions Ω, rather than any proper subset, such as the historically common $\mathbb{M}_{\text{perm}}$. We affirm the claim: "Any subset of problems for which an algorithm outperforms others is balanced by another subset for which the algorithm underperforms others. Thus, there is no single community detection algorithm that is best overall" [12].

It is still possible for an algorithm to perform better on a *subset* of community detection problems, so we can strive toward improved results on such a subset. To hope to perform well, we must note the assumptions about the subset of problems we expect to encounter. Some work has been done on estimating network properties to select the correct algorithm for the task at hand—a coarse way of checking assumptions [11,19]. Beyond this, though, we must clarify what the problem of community detection *is*; the formulation we choose will guide which subset of problem instances to prioritize and which to sacrifice.

References

1. Chen, Z., Li, L., Bruna, J.: Supervised community detection with line graph neural networks. In: International Conference on Learning Representations (2019)
2. Gates, A.J., Ahn, Y.Y.: The impact of random models on clustering similarity. J. Mach. Learn. Res. **18**(87), 1–28 (2017)
3. Hauer, B., Kondrak, G.: Decoding anagrammed texts written in an unknown language and script. Trans. Assoc. Comput. Linguist. **4**, 75–86 (2016)
4. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)
5. Kvalseth, T.O.: Entropy and correlation: some comments. IEEE Trans. Syst. Man Cybern. **17**(3), 517–519 (1987)
6. Lai, D., Nardini, C.: A corrected normalized mutual information for performance evaluation of community detection. J. Stat. Mech: Theory Exp. **2016**(9), 093403 (2016)
7. Liu, X., Cheng, H.M., Zhang, Z.Y.: Evaluation of community structures using kappa index and F-score instead of normalized mutual information. ArXiv e-prints, July 2018
8. McCarthy, A.D., Matula, D.W.: Normalized mutual information exaggerates community detection performance. In: SIAM Workshop on Network Science, SIAM NS 2018, pp. 78–79. SIAM, Portland, July 2018
9. McCarthy, A.D., Rudinger, R., Chen, T., Matula, D.W.: Metrics matter in community detection. In: Proceedings of the 8th International Conference on Complex Networks and Their Applications: Complex Networks, Lisbon, Portugal (2019)
10. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
11. Peel, L.: Estimating network parameters for selecting community detection algorithms. J. Adv. Inform. Fusion **6**, 119–130 (2011)
12. Peel, L., Larremore, D.B., Clauset, A.: The ground truth about metadata and community detection in networks. Sci. Adv. **3**(5), e1602548 (2017)
13. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. **101**(9), 2658–2663 (2004)

14. Romano, S., Bailey, J., Nguyen, V., Verspoor, K.: Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In: International Conference on Machine Learning, pp. 1143–1151 (2014)
15. Romano, S., Vinh, N.X., Bailey, J., Verspoor, K.: Adjusting for chance clustering comparison measures. J. Mach. Learn. Res. **17**(1), 4635–4666 (2016)
16. Schumacher, C., Vose, M.D., Whitley, L.D.: The no free lunch and problem description length. In: Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation, GECCO 2001, pp. 565–570. Morgan Kaufmann Publishers Inc., San Francisco (2001)
17. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, pp. 1073–1080. ACM, New York (2009)
18. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Comput. **8**(7), 1341–1390 (1996)
19. Yang, Z., Algesheimer, R., Tessone, C.J.: A comparative analysis of community detection algorithms on artificial networks. Sci. Rep. **6**, 30750 (2016)
20. Zhang, J., Chen, T., Hu, J.: On the relationship between gaussian stochastic blockmodels and label propagation algorithms. J. Stat. Mech: Theory Exp. **2015**(3), P03009 (2015)
21. Zhang, P.: Evaluating accuracy of community detection using the relative normalized mutual information. J. Stat. Mech: Theory Exp. **2015**(11), P11006 (2015)