# Objective function-based clustering via near-Boolean optimization*

Giovanni Rossi
Department of Computer Science and Engineering DISI, University of Bologna
Mura Anteo Zamboni 7, Bologna 40126, Italy; e-mail: *giovanni.rossi6@unibo.it*

September 6, 2017

**Abstract**

Objective function-based clustering is here looked at as a maximum-weight set partitioning combinatorial optimization problem, with the instance given by a pseudo-Boolean (set) function assigning real-valued cluster scores (or costs, in case of minimization) to data subsets, while on every partition of data the global objective function takes the value given by the sum over clusters (or blocks) of their individual score. The instance may thus maximally consist of $2^n$ reals, where $n$ is the number of data, although in most cases the scores of singletons and pairs also fully determine the scores of larger clusters, in which case the pseudo-Boolean function is quadratic. This work proposes to quantify the cluster score of fuzzy data subsets by means of the polynomial MLE (multilinear extension) of pseudo-Boolean functions, thereby translating the original discrete optimization problem into a continuous framework. After analyzing in these terms the well-known modularity maximization in complex networks, two further examples of quadratic cluster score functions for graph clustering are proposed, while also considering alternative greedy search strategies.

**Keywords:** fuzzy clustering, pseudo-Boolean function, multilinear extension, similarity matrix, graph clustering, modularity, local search.

## 1 Introduction

Clustering methods are essential tools in a wide variety of disciplines and applications. Conceptually, the purpose of clustering a finite set of objects relies on some quantification of similarity (or, dually, of dissimalirity) *within* any cluster or subset of objects, as well as *between* any two disjoint clusters. The purpose of (hard) clustering is indeed to group the objects into disjoint clusters in order to have high similarity (or low dissimilarity) within each cluster, and high dissimilarity (or low similarity) between any two clusters. Here these objects shall be generic data, possibly points $X_1, \dots, X_n \in \mathbb{R}^m$ in a Euclidean space. For notational convenience, simply consider the set $N = \{1, \dots, n\}$ of their indices.

*Presented at the 4th International Conference on Big Data Analysis and Data Mining, 7-8 September 2017, Paris, France.

In objective function-based clustering, an explicit measure of (dis)similarity identifies clusters by means of optimization. Formally, for $2^N = \{A : A \subseteq N\}$ denoting the $2^n$-set of subsets or clusters, the idea is to construct a function $w : 2^N \to \mathbb{R}$ such that $w(A)$ measures the (dis)similarity within $A \in 2^N$. Hence if $A, B \in 2^N$ are such that $w(A) < w(B)$, then $A$ is a (better) worst cluster than $B$. Combinatorially speaking [2], a hard clustering is a partition $P = \{A_1, \ldots, A_{|P|}\} \subset 2^N$ of $N$, i.e. $A_l \cap A_k = \emptyset$ for $1 \leq l < k \leq |P|$ and $A_1 \cup \cdots \cup A_{|P|} = N$. The global score (cost) of a partition $P$, to be maximized (minimized), is the sum over its constituents blocks or clusters $A \in P$ of their own individual score (cost), thereby yielding a partition function $W(P) = \sum_{A \in P} w(A)$ sometimes called "*additive*" [12, p. 63] or "*additively separable*" [13, 14]. In the well-known $k$-means method [16] for example, applying to points $X_i = (X_i^1, \ldots, X_i^m), i \in N$ in a Euclidean space, every partition $P$ has an associated cost $C(P) = \sum_{A \in P} c(A)$ to be minimized, where $c(A) = \sum_{i \in A} d(X_i, \bar{X}_A)$ sums the $|A|$ distances $d(X_i, \bar{X}_A)$ of cluster members $i \in A$ from cluster centroid $\bar{X}_A \in \mathbb{R}^m$, i.e. $\bar{X}_A^l = \sum_{i \in A} X_i^l / |A|$ for $1 \leq l \leq m$.

In fuzzy clustering [22], an extension $\hat{W}$ of the global objective function $W$ takes values on collections $(q^1, \ldots, q^H) \subset [0,1]^n$ of fuzzy clusters over which every $i \in N$ distributes a unit membership, i.e. $\sum_{1 \leq h \leq H} q_i^h = 1$. Formally, $\hat{W}(q^1, \ldots, q^H) = \sum_{1 \leq h \leq H} \hat{w}(q^h)$, where $\hat{w} : [0,1]^n \to \mathbb{R}$ quantifies the cluster score (cost) of fuzzy data subsets. Considering again the fuzzy $k$-means method, a fuzzy cluster $q = (q_1, \ldots, q_n) \in [0,1]^n$ has centroid $\bar{X}_q$ weighted by memberships $q_i, i \in N$, i.e. $\bar{X}_q^l = \sum_{i \in N} \left[ q_i / \left( \sum_{j \in N} q_j \right) \right] X_i^l$ for $1 \leq l \leq m$, and similarly the cost is the weighted sum of distances, hence $\hat{c}(q) = \sum_{i \in N} q_i d(X_i, \bar{X}_q)$.

In this work, for given score (cost) set function $w$, fuzzy clusters $q$ are evaluated in a seemingly novel manner, namely by means of the polynomial MLE (multilinear extension [4]) $f^w : [0,1]^n \to \mathbb{R}$ of $w$. The first finding, along this route, might appear somehow discouraging, since the fuzzy model yields no better global score (cost) than the hard one. Yet, optimal partitions can be searched for as collections of pair-wise disjoint vertices (in a sense detailed shortly) of the $n$-cube $[0,1]^n$, thereby providing a useful geometric perspective.

A feature common to most objective function-based clustering methods is that, without imposing suitable conditions, the optimization problem yields a trivial solution. This is immediately seen for the $k$-means method, where the finest partition $P_\perp = \{\{1\}, \ldots, \{n\}\}$ (consisting of $n$ singleton blocks) is always optimal, as $\sum_{i \in N} c(\{i\}) = 0$ (in particular, this is the unique optimum as long as $X_i \neq X_j$ for all $i \in N, j \in N \backslash i$). The focus has thus been placed on determining an "optimal" number of clusters for the given data [7, 8, 17, 20, 40]. One approach is to validate fuzzy clusterings $(q^1, \ldots, q^H)$ obtained through optimization at different (constrained) values of $H$ by means of an index [38], and select next the value of $H$ for the output where the index is highest. Alternatively, an optimal number of clusters is often assessed by applying spectral methods to graph clustering [6, 33, 36, 37, 39]. Indeed, over the last decades graphs or networks have been used for modeling and increasing number of complex social, biological and technogical systems [21], from typical friendship/influence among humans to protein-to-protein functional relations [3, 35], across the Internet and financial time series [11]. In these settings, the objects to be clustered are vertices $v_1, \ldots, v_n$ of a simple graph $G = (N, E)$, where the edge set $E \subseteq N_2$ is

a subset of the $\binom{n}{2}$-set $N_2 = \{A \in 2^N : |A| = 2\}$ of unordered pairs $\{i, j\}$ for $1 \leq i < j \leq n$. More generally, weights $w : N_2 \to [0, 1]$ may quantify the ($[0, 1]$-normalized) similarities within pairs. Such a framework suits best those many clustering tasks dealing with protein structures, text documents, surveys or biological signals, where a vector space (such as $\mathbb{R}^m$ above) is not available.

The target of graph clustering is to partition the vertices in order to have blocks spanning each a densely connected subgraph which, when contracted into a single vertex, remains the endpoint of few (if any) edges. With this aim, spectral methods rely on the eigenvalues and eigenvectors of the adjacency $\mathcal{A} = (a_{ij})_{1 \leq i,j \leq n}$ and/or Laplacian $\mathcal{L} = (\ell_{ij})_{1 \leq i,j \leq n}$ matrices, where

$$a_{ij} = \begin{cases} 1 \text{ if } \{i, j\} \in E \\ 0 \text{ otherwise} \end{cases} \quad \text{as well as} \quad \ell_{ij} = \begin{cases} -a_{ij} \text{ if } i \neq j \\ d_i \text{ if } i = j \end{cases}$$

while $d_i = \sum_{j \in N} a_{ij}$ denotes the degree of vertex $i \in N$ in graph $G = (N, E)$. Spectral methods have also been applied to the so-called *modularity matrix* [26], whose $n \times n$ entries $(b_{ij})_{1 \leq i,j \leq n} = (a_{ij} - d_i d_j / (2|E|))_{1 \leq i,j \leq n}$ determine the additive partition function $\mathcal{Q}$ well-known as *modularity*, namely

$$\begin{aligned} \mathcal{Q}(P) &= \frac{1}{2|E|} \sum_{1 \leq i,j \leq n} \left( a_{ij} - \frac{d_i d_j}{2|E|} \right) \delta_P(i, j), \\ \text{where } \delta_P(i, j) &= \begin{cases} 1 \text{ if } i, j \in A \text{ for a block } A \in P, \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

Hence in particular $\delta_P(i, i) = 1$ for all $i \in N$ and all partitions $P$. Therefore,

$$\mathcal{Q}(P) = \sum_{A \in P} \left[ \sum_{i \in A} \left( -\frac{d_i^2}{4|E|^2} \right) + \sum_{\{i,j\} \subseteq A} \left( \frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2} \right) \right] = \sum_{A \in P} w(A).$$

This is an objective function, to be maximized: good clusterings $P$ will have high scores $\mathcal{Q}(P)$, because $w(A)$ is precisely a measure of cluster score. In fact, apart from constant terms, $w(A)$ is essentially determined by the difference between the fraction $\sum_{\{i,j\} \subseteq A} a_{ij} / |E| = |E(A)| / |E|$ of edges whose endpoints are both in $A$, and the expectation of such a fraction in the random graph with same degree sequence $d_i, i \in N$ or *configuration model* [24, p. 200], i.e. $\sum_{\{i,j\} \subseteq A} d_i d_j / (2|E|^2)$. Note that $\mathcal{Q}$ is commonly defined by the former expression above, namely in terms of a sum over *ordered* pairs of vertices, while here in the second expression the sum is over unordered ones [26, Section III].

The target of spectral graph clustering is of course to exploit the information given by the eigenvalues and eigenvectors of the chosen matrix. As outlined above, in optimization methods such an information is primarily used for selecting a range for the number of clusters [39]. Perhaps the most immediate example comes from the simplest conceivable graph clustering problem, where the components $G(A_1), \ldots, G(A_{|P|})$ of the given graph $G = (N, E)$ are each a clique or maximal complete subgraph. That is, $G(A_l) = K_{A_l}$ for $1 \leq l \leq |P|$, where $K_A$ is the complete graph on vertex set $A \in P$, hence $G = K_{A_1} \cup \cdots \cup K_{A_{|P|}}$. The adjacency matrix of this "partition-like" graph has eigenvalues $(-1)^{|A_l|-1}$ and $(|A_l| - 1)^1$, while its Laplacian matrix has eigenvalues $0^{|P|}$ and $|A_l|^{|A_l|-1}$ (for $1 \leq l \leq |P|$), where multiplicities are indicated as exponents. Note that

$r(P) = \sum_l (|A_l| - 1) = n - |P|$ is the rank function for the geometric lattice $(\mathcal{P}^N, \wedge, \vee)$ of partitions of $N$ [2], while $\sum_i d_i = 2|E| = \sum_l |A_l|(|A_l| - 1)$. In general, the multiplicity of 0 as an eigenvalue of $\mathcal{L}$ counts the number of components of $G$ (and the associated eigenvectors are linear combinations of the characteristic functions given by these components' vertex subsets, see below).

In this work, the key feature of additive partition functions such as modularity $\mathcal{Q}$ is that the $2^n$ values $w(A), A \in 2^N$ of the underlying cluster score function $w$ are fully determined by the $1 + n + \binom{n}{2} = 1 + \binom{n+1}{2}$ values $w(\emptyset)$, $(w(\{i\}))_{i \in N}$ and $(w(\{i,j\}))_{\{i,j\} \in N_2}$. For $\mathcal{Q}$ these values are $w(\emptyset) = 0$ for the empty set (obviously), $w(\{i\}) = -[d_i/(2|E|)]^2$ for singletons $\{i\}$, and

$$w(\{i,j\}) = \frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2} - \frac{d_i^2}{4|E|^2} - \frac{d_j^2}{4|E|^2}$$

for pairs $\{i,j\}$. When considered in their MLE $f^w$, these set functions $w$ provide a global cluster score taking the form of a polynomial of degree 2. Such an objective function is defined on $n$-tuples of membership distributions, each being a point in a $2^{n-1} - 1$-dimensional unit simplex, as $|\{A : i \in A \in 2^N\}| = 2^{n-1}$ $(i \in N)$. Partitions are special cases of these $n$-tuples of membership distributions. In fact, as already mentioned, the extremizers of the objective function always include some partitions. But this is actually good news, since it guarrantees that by searching for optimality in the continuous setting provided by generic membership distributions one ends up finding solutions that also fit the original discrete optimization problem. The framework allows to design objective function-based clustering in terms of iterative improvements of global score starting from alternative membership distributions, with different possible choices for the set function $w$ assigning scores to clusters. Also, the number of clusters shall be autonomously determined through optimization, rather than being required as an input. The paper is organized as follows: Section 2 provides background material on lattices and pseudo-Boolean functions; Section 3 is devoted to fuzzy clustering, by firstly introducing fuzzy covers with associated $n$-tuples of membership distributions, and by showing next that partitions are among the extremizers of the resulting objective function; Section 4 proposes two cluster score functions with so-called *quadratic* MLE, focusing respectively on (i) $[0,1]$-valued similarities within pairs of data modeled as weighted networks, and (ii) transitivity in simple (i.e. non-weighted) spanned subgraphs for detecting communities in social networks; Section 5 addresses clustering via near-Boolean optimization with the input consisting of both: (i) a cluster score function with quadratic MLE, and (ii) a fuzzy clustering to start from, while also exemplifying the method through a well-known greedy agglomerative algorithm which has been tested and theoretically analyzed for clustering via modularity maximization; Section 6 contains some concluding remarks.

## 2 Lattices and pseudo-Boolean functions

Clusters $A, B \in 2^N$ and clusterings $P, Q \in \mathcal{P}^N$ are elements of two fundamental posets (partially ordered sets), respectively the Boolean lattice $(2^N, \cap, \cup)$ of subsets of $N$ ordered by inclusion $\supseteq$ and the geometric lattice $(\mathcal{P}^N, \wedge, \vee)$ of partitions of $N$ ordered by coarsening $\geqslant$, i.e. $P \geqslant Q$ if for every $B \in Q$ there

is $A \in P$ such that $A \supseteq B$, where $\wedge$ and $\vee$ respectively denote the "coarsest-finer-than" or *meet* and the "finest-coarser-than" or *join* operators. Since these posets are finite, lattice functions $w : 2^N \to \mathbb{R}$ and $W : \mathcal{P}^N \to \mathbb{R}$ may be dealt with as points $w \in \mathbb{R}^{2^n}$ and $W \in \mathbb{R}^{\mathcal{B}_n}$ in vector spaces[1]. A fundamental basis of these spaces (apart from the canonical one) is provided by the so-called zeta function $\zeta$, which works as follows: for every $A \in 2^N$ and every $P \in \mathcal{P}^N$, define

$$\zeta_A : 2^N \to \{0,1\} \quad \text{by} \quad \zeta_A(B) = \begin{cases} 1 \text{ if } B \supseteq A \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } B \in 2^N,$$

$$\zeta_P : \mathcal{P}^N \to \{0,1\} \quad \text{by} \quad \zeta_P(Q) = \begin{cases} 1 \text{ if } Q \geqslant P \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } Q \in \mathcal{P}^N.$$

Then, $\{\zeta_A : A \in 2^N\}$ is a basis of $\mathbb{R}^{2^n}$ and $\{\zeta_P : P \in \mathcal{P}^N\}$ is a basis of $\mathbb{R}^{\mathcal{B}_n}$ (with axes indexed respectively by subsets $A$ and partitions $P$). Set functions $w$ and partition functions $W$ are linear combinations of the elements of these bases, with coefficients $\mu^w(A), A \in 2^N$ and $\mu^W(P), P \in \mathcal{P}^N$ respectively:

$$w(\cdot) = \sum_{A \in 2^N} \zeta_A(\cdot)\mu^w(A) \quad \text{and} \quad W(\cdot) = \sum_{P \in \mathcal{P}^N} \zeta_P(\cdot)\mu^W(P),$$

$$w(B) = \sum_{A \subseteq B} \mu^w(A) \ \ (\text{all } B \in 2^N) \quad \text{and} \quad W(Q) = \sum_{P \leqslant Q} \mu^W(P) \ (\text{all } Q \in \mathcal{P}^N).$$

Set function $\mu^w : 2^N \to \mathbb{R}$ and partition function $\mu^W : \mathcal{P}^N \to \mathbb{R}$ are the *Möbius inversions* [2, 32] respectively of $w$ and $W$, obeying recursion

$$\mu^w(A) = w(A) - \sum_{B \subset A} \mu^w(B) \text{ as well as } \mu^W(P) = W(P) - \sum_{Q < P} \mu^W(Q),$$

where $Q < P$ denotes *strict coarsening*, i.e. there exist at least two blocks $B, B' \in Q$ and a corresponding block $A \in P$ such that $A \supseteq (B \cup B')$.

If a partition function $W$ is additive or additively separable, meaning that there is a set function $w$ such that $W(P) = \sum_{A \in P} w(A)$ for all $P \in \mathcal{P}^N$, then of course the Möbius inversions $\mu^w$ and $\mu^W$ must be related. In fact, in this case $\mu^W$ takes value 0 on all partitions apart (possibly) from those where the number of non-singleton blocks is $\leq 1$ [13, 14]. Such partitions are the $2^n - n$ *modular elements* [2, 34] of geometric lattice $(\mathcal{P}^N, \wedge, \vee)$, namely the bottom $P_\perp$ (see above) and top $P^\top = \{N\}$, together with all those obtained for $1 < |A| < n$ with form $P_\perp^A = \{A, \{i_1\}, \ldots, \{i_{n-|A|}\}\}$, where $\{i_1, \ldots, i_{n-|A|}\} = N \backslash A = A^c$. The values taken by Möbius inversion $\mu^W$ on these modular elements[2] are determined through recursion [30] as follows:

(a) $\mu^W(P_\perp) = \sum_{i \in N} w(\{i\})$,

(b) $\mu^W(P_\perp^A) = \mu^w(A)$ for $1 < |A| < n$,

(c) $\mu^W(P^\top) = \mu^w(N)$.

[1] $\mathcal{B}_k = \sum_{1 \leq l \leq k} \mathcal{S}_{k,l}$ is the (Bell) number of partitions of a $k$-set, while $\mathcal{S}_{k,l}$ is the Stirling number of the second kind, i.e. the number of partitions of $k$-set into $l$ blocks [2, 15, 31].

[2] Modularity $\mathcal{Q} : \mathcal{P}^N \to \mathbb{R}$ in Section 1 is intended to identify the "modular structure" of complex networks [19], while the modular elements of $(\mathcal{P}^N, \wedge, \vee)$ are those partitions $\hat{P}$ realizing equality $r(\hat{P} \wedge Q) + r(\hat{P} \vee Q) = r(\hat{P}) + r(Q)$ for all $Q \in \mathcal{P}^N$ (where $r$ is the rank).

This means that if $W$ is additively separable, then the continuum of additively separating set functions $w'$ consists of all those satisfying

$$\mu^{w'}(A) = \mu^w(A) \text{ for all } A \in 2^N, |A| > 1 \text{ and } \sum_{i \in N} w'(\{i\}) = \sum_{i \in N} w(\{i\}).$$

Note that $w(\emptyset) = \mu^w(\emptyset)$ and $w(\{i\}) = \mu^w(\emptyset) + \mu^w(\{i\})$ for all $i \in N$. Since set functions $w$ are generally conceived here to quantify the cluster score of data/vertex subsets, attention shall be placed exclusively on the case where the empty set has no score: $w(\emptyset) = 0$, entailing $w(\{i\}) = \mu^w(\{i\})$ for all $i \in N$ as well as $\mu^w(\{i,j\}) = w(\{i,j\}) - w(\{i\}) - w(\{j\})$ for all $\{i,j\} \in N_2$.

The Boolean lattice $(2^N, \cap, \cup)$ whose elements are the $2^n$ subsets of $N$ is commonly looked at as the set $\{0,1\}^n$ of extreme points (or vertices) of the unit $n$-dimensional hypercube $[0,1]^n$, in that every $A \in 2^N$ bijectively corresponds to the extreme point $\chi_A = (\chi_A(1), \ldots, \chi_A(n)) \in \{0,1\}^n$ identified by its *characteristic function* $\chi_A : N \to \{0,1\}$, where $\chi_A(i) = 1$ for all $i \in A$ and $\chi_A(j) = 0$ for all $j \in A^c$. In this view [4], set functions $w : 2^N \to \mathbb{R}$ are pseudo-Boolean functions $f^w : \{0,1\}^n \to \mathbb{R}$, and their multilinear extension or MLE $f^w : [0,1]^n \to \mathbb{R}$ is defined over the whole $n$-cube by

$$f^w(q) = \sum_{A \in 2^N} \left( \prod_{i \in A} q_i \right) \mu^w(A) \text{ for all } q = (q_1, \ldots, q_n) \in [0,1]^n,$$

hence indeed $f^w(\chi_A) = \sum_{B \subseteq A} \mu^w(B) = w(A)$ for all $A \in 2^N$. (Conventionally, $\prod_{i \in \emptyset} q_i := 1$.) As $f^w(q)$ is a polynomial (in $n$ variables $q_1, \ldots, q_n$), its degree is $\max\{|A| : \mu^w(A) \neq 0\}$ and its coefficients are the non-zero values of Möbius inversion $\mu^w(A) \neq 0$. In particular, if $\mu^w(A) = 0$ for all $A \in 2^N, |A| > 1$, then $f^w$ is linear (and $w$ is a *valuation* [2] of Boolean lattice $(2^N, \cap, \cup)$, i.e. $w(A \cap B) + w(A \cup B) = w(A) + w(B)$ for all $A, B \in 2^N$). Similarly, if $\mu^w(A) = 0$ for all $A \in 2^N, |A| > 2$, then $f^w$ is quadratic. Since $w(\emptyset) = 0$, if $f^w$ is linear then $w(A) = \sum_{i \in A} w(\{i\})$, while if $f^w$ is quadratic then

$$w(A) = \sum_{i \in A} w(\{i\}) + \sum_{\{i,j\} \subseteq A} \mu^w(\{i,j\}).$$

The MLE $f^w$ of cluster score function $w$ for modularity $\mathcal{Q}(P) = \sum_{A \in P} w(A)$ in Section 1 is quadratic, with $\mu^w(\{i\}) = w(\{i\}) = -[d_i/(2|E|)]^2$ for singletons $i \in N$ and $\mu^w(\{i,j\}) = [a_{ij} - d_i d_j/(2|E|)]/|E|$ for pairs $\{i,j\} \in N_2$. As for conditions $(a) - (c)$ above, let $\hat{w}$ be an alternative cluster score function with $\hat{w}(\{i\}) = \sum_{j \in N} w(\{j\})/n$ for all $i \in N$. This means that in $w$ every vertex $i$ has its own score when considered as a singleton cluster, while in $\hat{w}$ all vertices score the same when considered as singleton clusters. However, condition $(a)$ is satisfied since $\sum_{i \in N} \hat{w}(\{i\}) = \sum_{i \in N} -d_i^2/(4|E|^2) = \sum_{i \in N} w(\{i\})$ or equivalently $\mu^{\hat{W}}(P_\perp) = \sum_{i \in N} -d_i^2/(4|E|^2) = \mu^W(P_\perp)$ (where $\hat{W}$ is the partition function additively separated by $\hat{w}$). Now, by setting $\mu^{\hat{w}}(\{i,j\}) = \mu^w(\{i,j\})$ for pairs and $\mu^{\hat{w}}(A) = 0 = \mu^w(A)$ for larger subsets $A \in 2^N, |A| > 2$, conditions $(b)$ and $(c)$ are satisfied too. In fact, the (signed) *net added* score of pairs over the two corresponding singletons is the same, where $\mu^{\hat{w}}(\{i,j\}) = \mu^w(\{i,j\})$ means

$$\hat{w}(\{i,j\}) - \hat{w}(\{i\}) - \hat{w}(\{j\}) = \hat{w}(\{i,j\}) - \frac{2}{n} \sum_{k \in N} - \left( \frac{d_k}{2|E|} \right)^2 = \frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2}$$

or equivalently

$$\hat{w}(\{i,j\}) = \frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2} + \frac{1}{2n}\sum_{k\in N}\left(\frac{d_k}{|E|}\right)^2.$$

Therefore, $\sum_{A\in P}\hat{w}(A) = \mathcal{Q}(P) = \sum_{A\in P} w(A)$ for all partitions $P$.

# 3 Fuzzy clustering

Denote by $2^N_i = \{A : i \in A \in 2^N\}$ the $2^{n-1}$-set consisting of all subsets where every $i \in N$ is included, and by $\Delta_i$ the associated $2^{n-1}-1$-dimensional unit simplex whose extreme points are indexed by these subsets $A \in 2^N_i$. With some abuse of notation, let $q_i \in \Delta_i$ be a generic membership distribution, where $q_i^A \in [0,1]$ quantifies the membership of $i$ in cluster $A \in 2^N_i$. It must be stressed that $q_i \in [0,1]$ in Sections 1 and 2 denoted (as usual) $i$'s membership in a generic fuzzy cluster $q = (q_1,\ldots,q_n) \in [0,1]^n$. Throughout the remainder of this work, $q_i \in \Delta_i$ shall denote instead a generic membership distribution, i.e. $q_i : 2^N_i \to [0,1]$ with $q_i(A) = q_i^A$ and $\sum_{A\in 2^N_i} q_i^A = 1$.

**Definition 1** *A fuzzy cover is a collection $\boldsymbol{q} = \{q^A : A \in 2^N\}$ of $2^n$ fuzzy clusters $q^A = (q_1^A,\ldots,q_n^A) \in [0,1]^n$, where $q_i^A \in [0,1]$ if $i \in A$ and $q_j^A = 0$ if $j \in A^c$, while $\sum_{A\in 2^N_i} q_i^A = 1$ for all $i \in N$.*

Apart from zero entries, fuzzy covers $\mathbf{q}$ thus essentially correspond to $n$-tuples $(q_1,\ldots,q_n) \in \underset{i\in N}{\times}\Delta_i$ of membership distributions [18, 23]. Also, given a cluster score function $w$, fuzzy covers $\mathbf{q} = \{q^A : A \in 2^N\}$ attain additive global score $W(\mathbf{q})$ given by the sum of the $2^n$ values taken by the MLE $f^w$ of $w$, namely

$$W(\mathbf{q}) = \sum_{A\in 2^N} f^w(q^A) = \sum_{A\in 2^N}\sum_{B\supseteq A}\left(\prod_{i\in A} q_i^B\right)\mu^w(A). \tag{1}$$

In pseudo-Boolean optimization [4], the goal is to minimize or maximize a pseudo-Boolean function $f^w : \{0,1\}^n \to \mathbb{R}$, where $w : 2^N \to \mathbb{R}$ is a set function, and the MLE $f^w : [0,1]^n \to \mathbb{R}$ thus allows to turn several discrete optimization problems into a continuous setting. In near-Boolean optimization [30], the objective function has the form of $W(\mathbf{q})$ defined by expression (1), and the MLE allows to deal with discrete optimization problems involving additive partition functions (namely maximum-weight set partitioning/packing) into a continuous setting.

**Definition 2** *A fuzzy clustering is a fuzzy cover $\boldsymbol{q} = \{q^A : A \in 2^N\}$ satisfying $|\{i : q_i^A > 0\}| \in \{0,|A|\}$ for all $A \in 2^N$.*

In words, in a fuzzy clustering for every subset $A$ the number of those $i \in A$ with strictly positive membership $q_i^A > 0$ is either 0 or else $|A|$. As shown below, the set of values taken by $W$ on fuzzy covers coincides with the set of values taken (solely) on fuzzy clusterings.

**Proposition 3** *For any set function $w$, the range of $W$ defined by expression (1) is saturated by the values taken on fuzzy clusterings.*

**Proof:** In a fuzzy cover $\mathbf{q} = \{q^A : A \in 2^N\}$, let $A^+_{\mathbf{q}} = \{i : q^A_i > 0\}$ satisfy $\emptyset \subset A^+_{\mathbf{q}} \subset A$ for some ($\supseteq$-minimal) $A \in 2^N$, i.e. $0 < |A^+_{\mathbf{q}}| = \alpha < |A|$. Then,

$$W(\mathbf{q}) = \sum_{B \in 2^{A^+_{\mathbf{q}}}} f^w(q^B) + \sum_{A' \in 2^N \setminus 2^{A^+_{\mathbf{q}}}} f^w(q^{A'})$$

with, in particular,

$$f^w(q^A) = \sum_{B \in 2^{A^+_{\mathbf{q}}}} \left( \prod_{i \in A^+_{\mathbf{q}}} q^A_i \right) \mu^w(B).$$

Now consider another fuzzy cover $\hat{\mathbf{q}}$ such that $\hat{q}^{A'} = q^{A'}$ for all $A' \in 2^N \setminus 2^{A^+_{\mathbf{q}}}$, while $\hat{q}^A_i = 0$ for all $i \in A$, with group membership $q^A_A = \sum_{i \in A} q^A_i = \sum_{i \in A^+_{\mathbf{q}}} q^A_i$ redistributed over subsets $B \in 2^{A^+_{\mathbf{q}}}$ according to the following conditions:

$$\sum_{B \in (2^N_i \cap 2^{A^+_{\mathbf{q}}})} \hat{q}^B_i = q^A_i + \sum_{B \in (2^N_i \cap 2^{A^+_{\mathbf{q}}})} q^B_i \text{ for all } i \in A^+_{\mathbf{q}},$$

$$\prod_{i \in B} \hat{q}^B_i = \prod_{i \in B} q^B_i + \prod_{i \in B} q^A_i \text{ for all } B \in 2^{A^+_{\mathbf{q}}}, |B| > 1.$$

These $2^\alpha - 1$ equations with $\sum_{1 \leq k \leq \alpha} k\binom{\alpha}{k} > 2^\alpha$ variables $\hat{q}^B_i, \emptyset \neq B \in 2^{A^+_{\mathbf{q}}}$ admit a continuum of solutions, each providing a fuzzy cover $\hat{\mathbf{q}}$ where

$$\sum_{B \in 2^{A^+_{\mathbf{q}}}} f^w(\hat{q}^B) = f^w(q^A) + \sum_{B \in 2^{A^+_{\mathbf{q}}}} f^w(q^B) \Rightarrow W(\mathbf{q}) = W(\hat{\mathbf{q}}).$$

When reiterated for all (if any) $A' \in 2^N \setminus 2^{A^+_{\mathbf{q}}}$ where $0 < |\{i : q^{A'}_i > 0\}| < |A'|$, this procedure yields a final fuzzy clustering $\hat{\mathbf{q}}^*$ satisfying $W(\mathbf{q}) = W(\hat{\mathbf{q}}^*)$. ■

**Example 4** *Let* $A = \{1, 2, \ldots\} \supset A^+_{\boldsymbol{q}} = \{1, 2\}$*, hence*

$$f^w(q^A) = q^A_1 \mu^w(\{1\}) + q^A_2 \mu^w(\{2\}) + q^A_1 q^A_2 \mu^w(\{1, 2\}),$$

*with the three conditions for* $\hat{\boldsymbol{q}}$ *as follows*

- $\hat{q}^{\{1,2\}}_1 + \hat{q}^{\{1\}}_1 = q^{\{1,2\}}_1 + q^{\{1\}}_1 + q^A_1$,
- $\hat{q}^{\{1,2\}}_2 + \hat{q}^{\{2\}}_2 = q^{\{1,2\}}_2 + q^{\{2\}}_2 + q^A_2$,
- $\hat{q}^{\{1,2\}}_1 \hat{q}^{\{1,2\}}_2 = q^{\{1,2\}}_1 q^{\{1,2\}}_2 + q^A_1 q^A_2$,

*while the four variables are* $\hat{q}^{\{1\}}_1, \hat{q}^{\{1,2\}}_1, \hat{q}^{\{2\}}_2$ *and* $\hat{q}^{\{1,2\}}_2$*. One solution thus is*

- $\hat{q}^{\{1,2\}}_1 = \hat{q}^{\{1,2\}}_2 = \sqrt{q^{\{1,2\}}_1 q^{\{1,2\}}_2 + q^A_1 q^A_2} > 0$,
- $\hat{q}^{\{1\}}_1 = q^{\{1,2\}}_1 + q^{\{1\}}_1 + q^A_1 - \sqrt{q^{\{1,2\}}_1 q^{\{1,2\}}_2 + q^A_1 q^A_2} > 0$,
- $\hat{q}^{\{2\}}_2 = q^{\{1,2\}}_2 + q^{\{2\}}_2 + q^A_2 - \sqrt{q^{\{1,2\}}_1 q^{\{1,2\}}_2 + q^A_1 q^A_2} > 0$.

A main advantage of fuzzy clusters over hard ones is that they may display non-empty pair-wise intesections while also maintaining a unit (cumulative) membership that every $i \in N$ distributes over $2_i^N$ [28, 41, 42]. In this view, if fuzzy clusterings are evaluated via MLE as in expression (1), then they cannot yield a better global score than hard ones or partitions $P = \{A_1, \ldots, A_{|P|}\}$, where these latter correspond to $2^n$-collections $\mathbf{p} = \{p^A : A \in 2^N\}$ defined by $p^A = \begin{cases} \chi_A \text{ if } A \in P \\ \mathbf{0} \text{ if } A \in 2^N \backslash P \end{cases}$, with $\mathbf{0} \in \{0,1\}^n$ denoting the all-zero $n$-vector. Hence, apart from zero entries, $\mathbf{p}$ coincides with the collection $(\chi_{A_1}, \ldots, \chi_{A_{|P|}})$ of the characteristic functions of $P$'s blocks, which are pair-wise disjoint extreme points of the $n$-cube, i.e. $\langle \chi_{A_l}, \chi_{A_k} \rangle = 0$ for all $1 \leq l < k \leq |P|$, satisfying $\chi_1 + \cdots + \chi_{A_{|P|}} = \chi_N = \mathbf{1}$, where $\langle \cdot, \cdot \rangle$ denotes scalar product and $\mathbf{1} \in \{0,1\}^n$ is the all-one $n$-vector. In terms of expression (1), interpreting partitions $P \in \mathcal{P}^N$ as these collections $\mathbf{p} \subset \{0,1\}^n$ of disjoint extreme points of the $n$-cube means

$$W(\mathbf{p}) = \sum_{A \in 2^N} f^w(p^A) = \sum_{A \in P} f^w(\chi_A) = \sum_{A \in P} \sum_{B \in 2^A} \mu^w(B) = \sum_{A \in P} w(A).$$

**Proposition 5** *For any fuzzy clustering $\boldsymbol{q}$ and set function $w$, there are partitions $P, P' \in \mathcal{P}^N$ such that expression (1) satisfies $W(\boldsymbol{p}) \geq W(\boldsymbol{q}) \geq W(\boldsymbol{p}')$.*

**Proof:** Consider isolating the contribution of membership $q_i, i \in N$ to objective function $W(\mathbf{q}) = W(q_i|\mathbf{q}_{-i})$ when all other $n-1$ memberships $q_j, j \neq i$ are given:

$$W(\mathbf{q}) = W_i(q_i|\mathbf{q}_{-i}) + W_{-i}(\mathbf{q}_{-i}), \tag{2}$$

where $W(\mathbf{q}) = \sum_{A \in 2_i^N} f^w(q^A) + \sum_{A' \in 2^N \backslash 2_i^N} f^w(q^{A'})$ and

$$W_i(q_i|\mathbf{q}_{-i}) = \sum_{A \in 2_i^N} q_i^A \left[ \sum_{B \subseteq A \backslash i} \left( \prod_{j \in B} q_j^A \right) \mu^w(B \cup i) \right] \text{ as well as } W_{-i}(\mathbf{q}_{-i}) =$$

$$= \sum_{A \in 2_i^N} \left[ \sum_{B \subseteq A \backslash i} \left( \prod_{j \in B} q_j^A \right) \mu^w(B) \right] + \sum_{A' \in 2^N \backslash 2_i^N} \left[ \sum_{B' \subseteq A'} \left( \prod_{j' \in B'} q_{j'}^{A'} \right) \mu^w(B') \right].$$

Define $w_{\mathbf{q}_{-i}} : 2_i^N \to \mathbb{R}$ by

$$w_{\mathbf{q}_{-i}}(A) = \sum_{B \subseteq A \backslash i} \left( \prod_{j \in B} q_j^A \right) \mu^w(B \cup i). \tag{3}$$

Let $\mathbb{A}^+_{\mathbf{q}_{-i}} = \arg\max w_{\mathbf{q}_{-i}}$ and $\mathbb{A}^-_{\mathbf{q}_{-i}} = \arg\min w_{\mathbf{q}_{-i}}$, with $\emptyset \subset \mathbb{A}^+_{\mathbf{q}_{-i}}, \mathbb{A}^-_{\mathbf{q}_{-i}} \subseteq 2_i^N$. Most importantly,

$$W_i(q_i|\mathbf{q}_{-i}) = \sum_{A \in 2_i^N} \left( q_i^A \cdot w_{\mathbf{q}_{-i}}(A) \right) = \langle q_i, w_{\mathbf{q}_{-i}} \rangle. \tag{4}$$

In words, for given membership distributions $q_j, j \neq i$, global score is affected by $i$'s membership distribution $q_i$ through a scalar product. In order to maximize (or minimize) $W$ by suitably choosing $q_i$ for given $\mathbf{q}_{-i}$, the whole of $i$'s

membership mass has to be placed over $\mathbb{A}^+_{\mathbf{q}_{-i}}$ (or $\mathbb{A}^-_{\mathbf{q}_{-i}}$), anyhow. Hence there are precisely $|\mathbb{A}^+_{\mathbf{q}_{-i}}| > 0$ (or $|\mathbb{A}^-_{\mathbf{q}_{-i}}| > 0$) available extreme points of $\Delta_i$. After reiteration for all $i \in N$, the outcome shall generally consist of two fuzzy covers $\overline{\mathbf{q}}$ and $\underline{\mathbf{q}}$ such that $W(\overline{\mathbf{q}}) \geq W(\mathbf{q}) \geq W(\underline{\mathbf{q}})$ as well as $\overline{q}_i, \underline{q}_i \in ex(\Delta_i)$, where $ex(\Delta_i)$ is the $2^{n-1}$-set of extreme points of simplex $\Delta_i$. When this is combined with Proposition 3, the desired conclusion follows. ■

These findings suggest to search for optimal partitions through reiterated improvements $W(\mathbf{q}(t+1)) > W(\mathbf{q}(t)), t = 0, 1, \ldots$ of the objective function, while only requiring the cluster score function $w$ and an initial fuzzy clustering $\mathbf{q}(0)$ as inputs. Before considering such a possibility, attention now turns on (further) cluster score functions $w$ with quadratic MLE $f^w$.

# 4 Quadratic cluster score functions

As outlined in sections 1 and 2, a main example of cluster score function $w$ with quadratic MLE $f^w$ is given by modularity $\mathcal{Q}$. This section proposes two further examples of these cluster score functions, based respectively on the notions of *similarity* and *transitivity* that can be associated with networks. More precisely, the former is concerned with $n \times n$ (symmetric) similarity matrices quantifying the $\binom{n}{2}$ similarities within data pairs. Such matrices basically are the adjacency matrices of weighted graphs, when weights on edges are $[0, 1]$-normalized, and thus comprehend Boolean adjacency matrices as those special cases where each pair $\{i, j\}$ has weight $a_{ij} \in \{0, 1\}$. On the other hand, the latter is concerned with the density of triangles (or complete graphs $K_3$ on 3 vertices) in spanned subgraphs $G(A), A \in 2^N$, when $G = (N, E)$ is a simple (non-weighted) graph.

***Quadratic cluster scores and similarity matrices***

Pair-wise similarities may be quantified by a weighted graph $G_{\mathcal{W}} = (N, \mathcal{W})$, where weights $w_{ij} = w_{ji}$ for $1 \leq i < j \leq n$ are the entries of a symmetric similarity matrix $\mathcal{W} = (w_{ij})_{1 \leq i,j \leq n} \in [0, 1]^{n \times n}$. The issue addressed hereafter is how to construct a quadratic cluster score function $w$ (thus $\mu^w(A) = 0$ if $|A| = 0$ or $|A| > 2$) relying exclusively on these $\binom{n}{2}$ entries $w_{ij}, 1 \leq i < j \leq n$. A preliminary observation is that none of the following two choices works:

(i) $w(\{i\}) = 1$ for all $i \in N$ and $w(\{i, j\}) = w_{ij}$ for all $\{i, j\} \in N_2$,

(ii) $w(\{i\}) = 0$ for all $i \in N$ and $w(\{i, j\}) = w_{ij}$ for all $\{i, j\} \in N_2$.

The former simply applies the idea that every $i \in N$ has full (i.e. equal to 1) similarity with itself, while the latter associates zero cluster score to singletons, independently from the network $G_{\mathcal{W}}$. The reason why these values do not work is that choice (i) yields a set function $w$ such that for any $A \in 2^N, |A| > 1$ strict inequality $w(A) < w(B) + w(A \backslash B)$ holds for all $\emptyset \subset B \subset A$. Therefore, as already observed for the $k$-means method, global score attains its maximum $W(\mathbf{p}_\perp) = n$ solely on the finest partition $P_\perp$. In terms of Möbius inversion, $\mu^w(\{i\}) = 1$ for all $i \in N$ and $\mu^w(\{i, j\}) < 0$ for all $\{i, j\} \in N_2$. Conversely, choice (ii) yields a set function $w$ such that for all $A, B \in 2^N$ with $A \cap B = \emptyset$ inequality $w(A \cup B) \geq w(A) + w(B)$ holds, entailing that the coarsest partition $P^\top$ satisfies $W(\mathbf{p}^\top) \geq w(\mathbf{p})$ for all partitions $P$ (i.e. $\mathbf{p}$). In terms of Möbius inversion, $\mu^w(\{i\}) = 0$ for all $i \in N$ and $\mu^w(\{i, j\}) \geq 0$ for all $\{i, j\} \in N_2$.

In fact, for any quadratic cluster score function $w$, if matrix $\mathcal{W}$ is used to set $w(\{i,j\}) = w_{ij}$ for all $\{i,j\} \in N_2$, then of course there only remains to be defined the cluster score $w(\{i\})$ of singletons $i \in N$, for this univocally determines the $\binom{n+1}{2}$ non-zero values of Möbius inversion, namely $\mu^w(\{i\}) = w(\{i\}), i \in N$ and $\mu^w(\{i,j\}) = w_{ij} - w(\{i\}) - w(\{j\}), \{i,j\} \in N_2$.

Assuming $w(\{i\}) \geq 0$ for all $i \in N$, the cluster score of singletons has to take greater values on those $i \in N$ such that $\sum_{j \in N \setminus i} w_{ij}$ is small (namely outliers), and smaller values on those $i \in N$ where conversely $\sum_{j \in N \setminus i} w_{ij}$ is great. One way to achieve this is by setting

$$w(\{i\}) = \sum_{j \in N \setminus i} \frac{1 - w_{ij}}{2(n-1)}. \tag{5}$$

The idea is that $(1 - w_{ij}) \in [0,1]$ measures dissimilarity between $i$ and $j$, which has to be equally distributed over the two of them [29]. Accordingly, $w(\{i\})$ is the arithmetic mean of these $n-1$ half dissimilarities, entailing $w(\{i\}) \in [0, \frac{1}{2}]$, where the upper bound attains on isolated vertices $i$ such that $\sum_{j \in N \setminus i} w_{ij} = 0$, while the lower bound attains on those $i$ such that $w_{ij} = 1$ for all $j \in N \setminus i$. The $\binom{n}{2}$ values taken by Möbius inversion on pairs thus are

$$\begin{aligned} \mu^w(\{i,j\}) &= w_{ij} - 2\frac{1 - w_{ij}}{2(n-1)} - \sum_{k \in N \setminus \{i,j\}} \frac{2 - w_{ik} - w_{jk}}{2(n-1)} = \\ &= \frac{n \cdot w_{ij} - 1}{n-1} - \frac{n-2}{n-1} + \sum_{k \in N \setminus \{i,j\}} \frac{w_{ik} + w_{jk}}{2(n-1)}. \end{aligned} \tag{6}$$

The seemingly simplest way to check the functioning of this $w$ defined by expressions (5) and (6) is by focusing on simple graphs $G = (N, E)$ or equivalently on weighted ones $G_{\mathcal{W}} = (N, \mathcal{W})$ with Boolean similarity matrix $\mathcal{W} \in \{0,1\}^{n \times n}$. Then, $w(\{i\}) = (n - 1 - d_i)/[2(n-1)]$, where $d_i = \sum_{j \in N \setminus i} a_{ij}$ and $a_{ij} \in \{0,1\}$ is the $ij$-th entry of the adjacency matrix $\mathcal{A}$ $(= \mathcal{W})$, while for pairs

$$\begin{aligned} \mu^w(\{i,j\}) &= \frac{n \cdot a_{ij} - 1}{n-1} - \frac{n-2}{n-1} + \sum_{k \in N \setminus \{i,j\}} \frac{a_{ik} + a_{jk}}{2(n-1)} = \\ &= \frac{n}{n-1} a_{ij} - 1 + \frac{d_i + d_j - 2a_{ij}}{2(n-1)} = a_{ij} - 1 + \frac{d_i + d_j}{2(n-1)}. \end{aligned}$$

Therefore, if $a_{ij} = 1$ then $\mu^w(\{i,j\}) = (d_i + d_j)/[2(n-1)] \geq 0$, while if $a_{ij} = 0$ then $\mu^w(\{i,j\}) = -1 + (d_i + d_j)/[2(n-1)] \leq 0$. By letting $d_A = \sum_{i \in A} d_i$ denote the group degree for all $A \in 2^N$, cluster score thus is

$$\begin{aligned} w(A) &= \sum_{i \in A} \left( \frac{1}{2} - \frac{d_i}{2(n-1)} \right) + \sum_{\{i,j\} \subseteq A} \left( a_{ij} - 1 + \frac{d_i + d_j}{2(n-1)} \right) = \\ &= \frac{|A|}{2} - \frac{d_A}{2(n-1)} + \frac{d_A(|A|-1)}{2(n-1)} - \left( \binom{|A|}{2} - |E(A)| \right) = \\ &= \frac{|A|}{2} + \frac{d_A(|A|-2)}{2(n-1)} - \left( \binom{|A|}{2} - |E(A)| \right), \end{aligned} \tag{7}$$

where $E(A) = \{\{i,j\} : E \ni \{i,j\} \subseteq A\}$. Hence the score of a cluster $A$ obtains by summing half its cardinality $|A|$ and $(|A| - 2)/[2(n-1)]$ times its group

degree $d_A$, and next subtracting the number $\binom{|A|}{2} - |E(A)|$ of edges that the spanned subgraph $G(A)$ lacks with respect to the complete one $K_A$ (see above). Accordingly, if $A$ is such that $G(A) = (A, E(A)) = K_A$ is both complete and a component of $G$, then $|E(A)| = \binom{|A|}{2}$ and $d_A = |A|(|A| - 1) = 2|E(A)|$. Substituting into expression (7) yields $w(A) = \frac{|A|}{2} + \frac{|A|(|A|-1)(|A|-2)}{2(n-1)}$ or

$$w(A) = \frac{|A|}{2} + \binom{|A|}{2}\frac{|A| - 2}{n - 1}.$$

If $G = (N, E) = K_N$ is the complete graph, then $w(N) = \binom{n}{2}$. More generally, for any partition $P = \{A_1, \ldots, A_{|P|}\}$ of $N$ and corresponding partition-like graph $G = K_{A_1} \cup \cdots \cup K_{A_{|P|}}$ introduced above, global score $W$ attains its unique mamximum $W(\mathbf{p}) = \sum_{A \in P}\left(\frac{|A|}{2} + \binom{|A|}{2}\frac{|A|-2}{n-1}\right)$ on $\mathbf{p} = (\chi_{A_1}, \ldots, \chi_{A_{|P|}})$ (since $\binom{1}{2} = 0$, for the finest partition $P_\perp$ or $\mathbf{p}_\perp$ and associated empty graph $G = (N, \emptyset)$ such a maximum is indeed $W(\mathbf{p}_\perp) = \frac{n}{2}$).

***Quadratic cluster scores and transitivity in spanned subgraphs***

A further way to take into account a given network's topology when assigning a score $w(A)$ to every vertex subset $A$ obtains by focusing on the density of triangles (or complete subgraphs on three vertices) included in the spanned subgraph $G(A)$. In fact, if cluster scores are specifically intended to detect the community structure of (simple) social networks $G = (N, E)$, then such scores may incorporate the empirical evidence that transitivity is an essential property of such networks with respect to non-social ones. In fact, social networks display a "higher-than-expected" value for the *clustering coefficient* $cc(G)$, where

$$cc(G) = \frac{3 \times \text{ number of triples of vertices spanning a complete subgraph}}{\text{number of connected triples of vertices}} .$$

Quoting [27], *"a "connected triple" means a vertex connected directly to an unordered pair of others"* (hence every $\{i, j, k\} \in 2^N$ spanning a complete subgraph $G(\{i, j, k\}) = K_{\{i,j,k\}} = K_3$ counts for three such triples). In other terms, $cc(G)$ is the expectation that by randomly picking a vertex and two of its neighbors, these latter are also adajcent. Now consider the aim to assign scores $w(A)$ to clusters $A$ in a way such that higher values of the clustering coefficient $cc(G(A))$ for spanned subgraphs provide greater scores. From a combinatorial perspective, the most natural way to achieve this is certainly by means of a *cubic* $f^w$, i.e. such that $\mu^w(A) = 0$ if $|A| > 3$ (or $|A| = 0$). For instance, the $\binom{n+1}{2}$ values of the sought $w$ on singletons and pairs can remain those defined respectively by expressions (5) and (6) above (with Boolean weights $w_{ij} \in \{0, 1\}$), while on the $\binom{n}{3}$ triples $\{i, j, k\}$ Möbius inversion may be defined by $\mu^w(\{i, j, k\}) =$

$$= \begin{cases} 1 \text{ if } G(\{i,j,k\}) = K_{\{i,j,k\}} \text{ is complete (i.e. } G(\{i,j,k\}) = K_3), \\ 0 \text{ if } G(\{i,j,k\}) \text{ is connected but non-complete (i.e. } G(\{i,j,k\}) \neq K_3), \\ -1 \text{ if } G(\{i,j,k\}) \text{ is disconnected.} \end{cases}$$

The resulting cluster scores $w(A)$ clearly provide additional reward/penalty for proximity/distance to/from completeness of the spanned subgraph $G(A)$, as

$$w(A) = \frac{|A|}{2} + \frac{d_A(|A| - 2)}{2(n - 1)} - \left(\binom{|A|}{2} - |E(A)|\right) + \sum_{\{i,j,k\} \subseteq A} \mu^w(\{i, j, k\})$$

replaces expression (7) above, and again a spanned subgraph $G(A) = K_A$ which is both complete and a component of $G$ yields a cluster score

$$w(A) = \frac{|A|}{2} + \binom{|A|}{2}\frac{|A|-2}{n-1} + \binom{|A|}{3}.$$

However, a similar result also obtains with a quadratic $f^w$ where the count of both common and non-common neighbors (of cluster members) enters explicitly in the values taken by Möbius inversion $\mu^w$ on pairs. Formally, following [1, 41], let $N_i = \{i\} \cup \{j : \{i,j\} \in E\}$ be the set of vertices at distance $\leq 1$ from vertex $i$ in the given network $G$, i.e. $N_i$ contains $i$ and all its neighbors $j$, entailing $\{i,j\} \subseteq (N_i \cap N_j)$ for all edges $\{i,j\} \in E$, with $|N_i \cap N_j|$ counting the number of common neighbors, while symmetric difference $N_i \Delta N_j = (N_i \backslash N_j) \cup (N_j \backslash N_i)$ contains non-common neighbors. A simple quadratic cluster score function $w$ taking into account transitivity in spanned subgraphs can thus be defined by

$$\begin{aligned} \mu^w(\{i\}) &= \frac{1}{|N_i|} = w(\{i\}) \text{ on singletons and} \\ \mu^w(\{i,j\}) &= a_{ij} + \frac{|N_i \cap N_j| - |N_i \Delta N_j|}{|N_i \cup N_j|} \text{ on pairs.} \end{aligned}$$

The resulting cluster score takes form

$$w(A) = \sum_{i \in A} \frac{1}{1+d_i} + |E(A)| + \sum_{\{i,j\} \subseteq A} \frac{|N_i \cap N_j| - |N_i \Delta N_j|}{|N_i \cup N_j|}, \tag{8}$$

hence again a spanned subgraph $G(A) = K_A$ which is both complete and a component of $G$ yields score $w(A) = 1 + 2\binom{|A|}{2} = 1 + |A|(|A|-1)$.

The next section is focused on searching for partitions $\mathbf{p}$ that locally maximize objective function $W$ in expression (1), when the input consists of $\binom{n+1}{2}$ non-zero values of Möbius inversion $\mu^w$ and an initial fuzzy clustering $\mathbf{q}(0)$.

# 5 Greedy clustering

A seemingly interesting way to see how near-Boolean optimization may be employed in the present setting is through comparison with the so-called greedy agglomerative approach [5, 25], which starts from the finest partition and iteratively selects one union of two blocks that results in a maximal increase of global score, thereby yielding a sequence $P(t+1) \succcurlyeq P(t)$ of partitions as the search path, where $P(0) = P_\perp$ and $\succcurlyeq$ denotes the covering relation between partitions, i.e. both $|P(t+1)| = |P(t)| - 1$ and strict coarsening $P(t+1) > P(t)$ hold. If there are tails, meaning that different unions of two blocks of $P(t)$ yield the same maximal increase of global score, then the two blocks to be merged are randomly selected. The stopping criterion is the absence of any further improvement. The iterative procedure thus is as follows.

GreedyMerging$(w, P)$

*Initialize:* Set $t = 0$ and $P(0) = P_\perp$.

*Loop:* While $w(A \cup B) - w(A) - w(B) > 0$ for some $A, B \in P(t)$,

set $t = t + 1$ and

[1] randomly select $A, B \in P(t-1)$ satisfying, for all $A', B' \in P(t-1)$,

$$w(A \cup B) - w(A) - w(B) \geq w(A' \cup B') - w(A') - w(B'),$$

[2] define $P(t) = \{A \cup B\} \cup (P(t-1) \backslash \{A, B\})$

(hence $P(t)$ obtains from $P(t-1)$ by merging blocks $A$ and $B$).

*Output:* Set $P^* = P(t)$.

This algorithm has been tested [25] for maximizing modularity $\mathcal{Q}$, hence with $w$ and quadratic MLE $f^w$ as specified above. For notational convenience, also let $W$ be the corresponding objective function defined by expression (1). Therefore, $W(\mathbf{p}) = \mathcal{Q}(P)$ for all partitions $\mathbf{p}$ or $P$. However (and most importantly), $W$ is defined on fuzzy clusterings, while $\mathcal{Q}$ is defined solely on hard clusterings or partitions. If employed for modularity clustering, then *GreedyMerging* admits no finite approximation. In fact, for the class of $\frac{n}{2}$-regular graphs considered in [5, Theorem 5.1, p. 8], *GreedyMerging* provides a *worst-case* solution $\hat{P}$ with zero global cluster score $W(\hat{\mathbf{p}}) = 0$, while the optimal solution $P^*$ provides a strictly positive global score $W(\mathbf{p}^*) > 0$. In these graphs $G = (N, E)$, the number $n > 4$ of vertices is even and $N = N^1 \cup N^2$ includes two vertex subsets of equal size, i.e. $N^1 = \{i_1, \ldots, i_{\frac{n}{2}}\}$ as well as $N^2 = \{j_1, \ldots, j_{\frac{n}{2}}\}$. The edge set is $E = \{\{i, i'\} : \{i, i'\} \subset N^1\} \cup \{\{j, j'\} : \{j, j'\} \subset N^2\} \cup \{\{i_k, j_k\} : 1 \leq k \leq \frac{n}{2}\}$. In words, $G \supset K_{N^1}, K_{N^2}$ includes the two complete graphs on vertex sets $N^1$ and $N^2$, together with all the $\frac{n}{2}$ edges with endpoints $i_k \in N^1$ and $j_k \in N^2$ for $1 \leq k \leq \frac{n}{2}$. At $t = 0$, for each of the $\binom{n}{2}$ possible unions of two blocks of $P_\perp$, the corresponding variation of global score is $w(\{i, j\}) - w(\{i\} - w(\{j\} =$

$$= \mu^w(\{i, j\}) = \frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2} = \begin{cases} 2/n^2 \text{ if } \{i, j\} \in E, \\ -2/n^2 \text{ if } \{i, j\} \in N_2 \backslash E, \end{cases}$$

where $|E| = 2\binom{\frac{n}{2}}{2} + \frac{n}{2} = (\frac{n}{2})^2$ and $d_i = \frac{n}{2} = d_j$ for all $i, j \in N$. Hence the worst-case output of *GreedyMerging* is the partition $\hat{P} = \{\{i_1, j_1\}, \ldots, \{i_{\frac{n}{2}}, j_{\frac{n}{2}}\}\}$ obtained in $\frac{n}{2}$ iterations through unions $\{i_k\} \cup \{j_k\}$ for $1 \leq k \leq \frac{n}{2}$, where

$$W(\hat{\mathbf{p}}) = \sum_{i \in N} w(\{i\}) + \sum_{1 \leq k \leq \frac{n}{2}} \mu^w(\{i_k, j_k\}) = -\sum_{i \in N} \frac{d_i^2}{4|E|^2} + \frac{n}{2}\frac{2}{n^2} = -\frac{1}{n} + \frac{1}{n} = 0$$

is the resulting global score, while $n > 4$ entails that the unique maximum attains at $P^* = \{N^1, N^2\}$ where

$$W(\mathbf{p}^*) = \sum_{i \in N} w(\{i\}) + 2 \sum_{\{i, i'\} \subset N^1} \mu^w(\{i, i'\}) = -\frac{1}{n} + 2\binom{\frac{n}{2}}{2}\frac{2}{n^2} = \frac{n-4}{2n} > 0.$$

One immediate observation is that *GreedyMerging* may well fall in the same worst-case trap even when the input cluster score function $w$ is that defined by expression (7), in which case the $\binom{n}{2}$ possible unions of two blocks of $P_\perp$ result in a variation of global score $w(\{i, j\}) - w(\{i\} - w(\{j\} = \mu^w(\{i, j\}) =$

$$= a_{ij} - 1 + \frac{d_i + d_j}{2(n-1)} = \begin{cases} n/[2(n-1)] \text{ if } \{i, j\} \in E, \\ -(n-2)/[2(n-1)] \text{ if } \{i, j\} \in N_2 \backslash E. \end{cases}$$

Conversely, if the input cluster score function $w$ is defined by expression (8), then *GreedyMerging* surely finds the optimum $P^*$, as the $\binom{n}{2}$ unions of two blocks of $P_\perp$ result in a variation of global score $w(\{i,j\}) - w(\{i\} - w(\{j\} = \mu^w(\{i,j\}) =$

$$= a_{ij} + \frac{|N_i \cap N_j| - |N_i \Delta N_j|}{|N_i \cup N_j|} = \begin{cases} 2n/(n+4) \text{ if } \{i,j\} \subset N^1 \text{ or } \{i,j\} \subset N^2, \\ 4/n \text{ if } \{i,j\} \in E, i \in N^1, j \in N^2, \\ -(n-2)/n \text{ if } \{i,j\} \in N_2 \backslash E. \end{cases}$$

Coming to near-Boolean optimization, firstly note that a quadratic $f^w$, whatever its chosen form, reduces expression (1) to

$$W(\mathbf{q}) = \sum_{i \in N} w(\{i\}) + \sum_{\{i,j\} \in N_2} \left( \sum_{A \supseteq \{i,j\}} q_i^A q_j^A \right) \mu^w(\{i,j\}). \tag{9}$$

In terms of this objective function, *GreedyMerging* develops from the initial $n$-tuple of membership distributions $\mathbf{p}_\perp = (p_{1\perp}, \ldots, p_{n\perp})$ where $p_{i\perp}^{\{i\}} = 1$ for all $i \in N$, and at any partition $P(t+1)$ or $\mathbf{p}(t+1)$ obtained through the union of two blocks $A, B \in P(t)$, the corresponding membership distributions are $p_{i'}^{A'}(t+1) = 1 = p_{i'}^{A'}(t)$ for all $i' \in A'$ and all $A' \in P(t), A \neq A' \neq B$, while $p_i^A(t) = 1 = p_i^{A \cup B}(t+1)$ for all $i \in A$ and $p_j^B(t) = 1 = p_j^{A \cup B}(t+1)$ for all $j \in B$. Equivalently, apart from zero entries, $\mathbf{p}(t) = \left(\chi_A, \chi_B, \chi_{A'_1}, \ldots, \chi_{A'_{|P(t)|-2}}\right)$ and $\mathbf{p}(t+1) = \left(\chi_A + \chi_B, \chi_{A'_1}, \ldots, \chi_{A'_{|P(t)|-2}}\right)$. The corresponding change

$$W(\mathbf{p}(t+1)) - W(\mathbf{p}(t)) = \sum_{\substack{\{i,j\} \subseteq (A \cup B) \\ A \not\supseteq \{i,j\} \not\subseteq B}} \mu^w(\{i,j\})$$

of global score is a maximal one among the $\binom{|P(t)|}{2}$ available. The main advantage of expression (9) is that $W$ takes values on fuzzy clusterings, hence search paths may take the form of sequences $\mathbf{q}(t)$ such that $W(\mathbf{q}(t)) > W(\mathbf{q}(t-1))$. This requires to first formalize: (I) how $\mathbf{q}(t+1)$ obtains from the reached $\mathbf{q}(t)$, (II) the stopping criterion, and (III) the initial $\mathbf{q}(0)$. Before addressing these issues, it may be outlined that the search proposed below may be regarded as a local one, since it develops from an input $\mathbf{q}(0)$. However, the more these initial $n$ membership distributions $q_i(0), i \in N$ are each spread over $2_i^N$, the more the search tends to be global. As for local optimality, which essentially determines the stopping criterion, the neighborhood of $\mathbf{q}$ is $\mathcal{N}(\mathbf{q}) = \bigcup_{i \in N} \{\hat{q}_i | \mathbf{q}_{-i} : \hat{q}_i \in \Delta_i\}$, hence $\mathbf{q}^*$ is a local optimum if $W(\mathbf{q}^*) \geq W(\hat{\mathbf{q}})$ for all $\hat{\mathbf{q}} \in \mathcal{N}(\mathbf{q}^*)$. In words, the neighborhood of $\mathbf{q}$ consists of all $n$-tuples of membership distributions where $n-1$ distributions are as in $\mathbf{q}$ while only one may vary, and $\mathbf{q}^*$ is a local optimum if $W(\mathbf{q}^*)$ is the greatest value taken by $W$ when restricted to $\mathcal{N}(\mathbf{q}^*)$. It is shown below that for any partition $P$ or $\mathbf{p}$ a necessary and sufficient condition for local optimality, i.e. $W(\mathbf{p}) \geq W(\mathbf{q})$ for all $\mathbf{q} \in \mathcal{N}(\mathbf{p})$, is $w(A) \geq w(A \backslash i) + w(\{i\})$ for all $i \in A$ and all $A \in P$. A typical greedy local search would thus progress through a sequence $\mathbf{q}(t)$ such that $\mathbf{q}(t+1) \in \mathcal{N}(\mathbf{q}(t))$ and $W(\mathbf{q}(t+1)) - W(\mathbf{q}(t))$ is maximal, but none of these two conditions is here maintained. In fact, $\mathbf{q}(t+1) \notin \mathcal{N}(\mathbf{q}(t))$ as more than one of the $n$ membership distributions $(q_1(t), \ldots, q_n(t)) = \mathbf{q}(t)$ vary within the same $t$-th iteration (and

the same clearly charaterizes *GreedyMerging* too, apart from fuzziness). Also, rather than being applied directly to the increase $W(\mathbf{q}(t+1))-W(\mathbf{q}(t))$ of global score, greediness is applied to the "average derivative", formalized hereafter.

Concerning (I), recall that the (first order) $i$-th derivative [4, p. 157] of the MLE $f^w$ at $x=(x_1,\ldots,x_n)\in[0,1]^n$ is $f_i^w(x)=\frac{\partial f^w}{\partial x_i}(x)=$

$$\begin{aligned} &= f^w(x_1,\ldots,x_{i-1},1,x_{i+1},\ldots,x_n)-f^w(x_1,\ldots,x_{i-1},0,x_{i+1},\ldots,x_n)= \\ &= \sum_{A\in 2_i^N}\left(\prod_{j\in A\backslash i} x_j\right)\mu^w(A). \end{aligned}$$

At vertices $\chi_B, B\in 2^N$ of the $n$-cube it takes values $f_i^w(\chi_B)=w(B)-w(B\backslash i)$ if $B\in 2_i^N$ and $f_i^w(\chi_B)=w(B\cup i)-w(B)$ if $B\notin 2_i^N$. This derivative may be reproduced for objective function $W$ as follows. For all $i\in N$ and all $A\in 2_i^N$, define membership $q_{i_A}$ by $q_{i_A}^B=\left\{\begin{array}{l} 1 \text{ if } B=A \\ 0 \text{ otherwise}\end{array}\right.$ for all $B\in 2_i^N$. Also let $q_{i\emptyset}^B=0$ for *all* $B\in 2_i^N$, noting that $q_{i_\emptyset}$ is *not* a membership distribution, as it places no membership over $2_i^N$ at all. Now define

$$W_{i_A}(\mathbf{q})=\frac{\partial W}{\partial q_i^A}(\mathbf{q})=W(q_{i_A}|\mathbf{q}_{-i})-W(q_{i_\emptyset}|\mathbf{q}_{-i})=W_i(q_{i_A}|\mathbf{q}_{-i})=w_{\mathbf{q}_{-i}}(A)$$

to be the (first order) $i_A$-derivative of $W$ at $\mathbf{q}$, where the last two equalities obtain from expressions (2-3) in Section 3. If the $|A|-1$ membership distributions $q_j, j\in A\backslash i$ are $q_j^A=1$, then $W_{i_A}(\mathbf{q})=w(A)-w(A\backslash i)$, and $W_{i_{\{i\}}}(\mathbf{q})=w(\{i\})$ independently from $\mathbf{q}$. These $n2^{n-1}$ derivatives $(W_{i_A}(\mathbf{q}(t)))_{i\in N, A\in 2_i^N}$ inform about how to obtain $\mathbf{q}(t+1)$ from the reached $\mathbf{q}(t)$ in order to maximize the objective function. In particular, any greedy strategy requires first to make clear what "maximum distance" may separate $\mathbf{q}(t+1)$ from $\mathbf{q}(t)$. As already mentioned, $\mathbf{q}(t+1)\in\mathcal{N}(\mathbf{q}(t))$ does not suit. The rule maintained here is the same as for *GreedyMerging*, namely that precisely one block is formed when transforming $\mathbf{q}(t)$ into $\mathbf{q}(t+1)$. In other terms, there is exactly one $A$ such that $\sum_{i\in A} q_i^A(t)<|A|=\sum_{i\in A} q_i^A(t+1)$ or equivalently $q^A(t)\neq\chi_A=q^A(t+1)$. Given this constraint, greediness is applied to *average derivative*

$$\begin{aligned} \bar{W}_A(\mathbf{q}) &= \frac{1}{|A|}\sum_{i\in A} w_{\mathbf{q}_{-i}}(A)=\frac{1}{|A|}\sum_{i\in A}\left[\sum_{B\subseteq A\backslash i}\left(\prod_{j\in B} q_j^A\right)\mu^w(B)\right]= \\ &= \frac{1}{|A|}\sum_{B\subseteq A}\left[\sum_{i\in B}\left(\prod_{j\in B\backslash i} q_j^A\right)\right]\mu^w(B). \end{aligned}$$

Hence for a quadratic $w$

$$\bar{W}_A(\mathbf{q})=\frac{1}{|A|}\left[\sum_{i\in A} w(\{i\})+\sum_{\{i,j\}\subseteq A}\left(q_i^A+q_j^A\right)\mu^w(\{i,j\})\right].$$

That is to say, the chosen $A$ (at iteration $t$, to be a block of the output partition $\mathbf{p}^*$ being constructed) is one where $q^A(t)\neq\chi_A$ and $\bar{W}_A(\mathbf{q}(t))$ is maximal (the case of tails can be dealt with arbitrarily). Then, it remains to specify, for all

$j \in A^c$, how to reallocate membership $\sum_{B \in 2_j^N : B \cap A \neq \emptyset} q_j^B(t)$. Basically, this shall be redistributed over those $B \in 2_j^N$ such that $A \cap B = \emptyset$.

Coming to (II), the greedy procedure stops when for all $A \in 2^N$ either $q^A(t) = \mathbf{0}$ or $q^A(t) = \chi_A$, i.e. when $\sum_{i \in A} q_i^A(t) \in \{0, |A|\}$. That is, apart from zero entries, $\mathbf{q}(t) = \mathbf{p}^* = (\chi_{A_1}, \ldots, \chi_{A_{|P^*|}})$ for a partition $P^* = \{A_1, \ldots, A_{|P^*|}\}$. Next, this is checked to be a local optimum, i.e. $w(A) \geq w(A \backslash i) + w(\{i\})$ for all $i \in A$ and all $A \in P^*$. If this inequality is not satisfied, then the partition updates by splitting block $A$ in the two (new) blocks $A \backslash i$ and $\{i\}$.

Finally, as for (III), there surely exist many reasonable alternatives for the choice of input $\mathbf{q}(0)$, including the simplest one given by the $n$-tuple of uniform distributions $q_i^A(0) = 2^{1-n}$ for all $A \in 2_i^N$ and all $i \in N$. In general, the initial fuzzy clustering establishes the terms of trade between computational burden and search width. Specifically, the more the $n$ distributions $q_i(0), i \in N$ are each spread over $2_i^N$, the more computational demanding and wider becomes the search. In fact, if a family $\mathcal{F} = \{A_1, \ldots, A_k\} \subset 2^N$ satisfies $q_i^B(0) = 0$ for all $B \in 2^N \backslash (2^{A_1} \cup \cdots \cup 2^{A_k})$ and all $i \in N$ as well as $q^{A_1}(0), \ldots, q^{A_k}(0) \neq \mathbf{0}$, then the algorithm proposed hereafter only searches for optimal clusters (or blocks) among those $B \in (2^{A_1} \cup \cdots \cup 2^{A_k})$, and thus cannot output any partition $P$ such that $B \in P$ for some $B \in 2^N \backslash (2^{A_1} \cup \cdots \cup 2^{A_k})$. In particular, if the input $\mathbf{q}(0) = \mathbf{p}$ is a partition $P = \{A_1, \ldots, A_{|P|}\}$, then the algorithm only checks if local optimality (i.e. $w(A) \geq w(A \backslash i) + w(\{i\})$ for all $A \in P, i \in A$) holds, and for the limit case $\mathbf{q}(0) = \mathbf{p}_\perp$ the output $\mathbf{p}^* = \mathbf{p}_\perp$ coincides with the input. Now let $\hat{w}(A) = \frac{w(A)}{|A|}$ and consider choosing $\mathbf{q}(0)$ through an arbitrary threshold $\theta \geq 0$ as follows: if $\hat{w}(A) \leq \theta$ then $q^A(0) = \mathbf{0}$, while if $\hat{w}(A) > \theta$ then

$$q_i^A(0) = \hat{w}(A) \Big/ \sum_{B \in 2_i^N : \hat{w}(B) > \theta} \hat{w}(B) \text{ for all } i \in A, \tag{10}$$

entailing $\frac{q_i^A(0)}{q_i^B(0)} = \frac{\hat{w}(A)}{\hat{w}(B)}$ for all $i \in N$ and $A, B \in 2_i^N$ such that $\hat{w}(A) > \theta < \hat{w}(B)$.

The following greedy (local) search strategy can now be formalized.

GreedyClustering$(w, \mathbf{q})$

*Initialize:* Set $t = 0$ and $\mathbf{q}(0)$ as in expression (10).

*GreedyLoop:* While $0 < \sum_{i \in A} q_i^A(t) < |A|$ for some $A \in 2^N$, set $t = t + 1$ and

(a) select arbitrarily one such $A^*(t) \in 2^N$ where, in addition, the average derivative is maximal, i.e. for all $B$ such that $0 < \sum_{i \in B} q_j^B(t) < |B|$

$$W_{\bar{A}}(\mathbf{q}(t-1)) \geq W_{\bar{B}}(\mathbf{q}(t-1));$$

(b) for $i \in A^*(t)$ and $A \in 2_i^N$, define $q_i^A(t) = \begin{cases} 1 \text{ if } A = A^*(t), \\ 0 \text{ if } A \neq A^*(t); \end{cases}$

(c) for $j \in N \backslash A^*(t)$ and $A \in 2_j^N$ with $A \cap A^*(t) = \emptyset$, define

$$q_j^A(t) = q_j^A(t-1) + \left( \hat{w}(A) \sum_{\substack{B \in 2_j^N \\ B \cap A^*(t) \neq \emptyset}} q_j^B(t-1) \right) \left( \sum_{\substack{B' \in 2_j^N \\ B' \cap A^*(t) = \emptyset}} \hat{w}(B') \right)^{-1};$$

(d) for $j \in N \backslash A^*(t)$ and $A \in 2_j^N$ with $A \cap A^*(t) \neq \emptyset$, define $q_j^A(t) = 0$.

*CheckLoop:* While $q^A(t) = \chi_A, |A| > 1$ and $w(A) < w(\{i\}) + w(A\backslash i)$ for some $A \in 2^N, i \in A$, set $t = t + 1$ and define:

$$\begin{aligned} q_i^{\hat{A}}(t) &= \begin{cases} 1 \text{ if } |\hat{A}| = 1 \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } \hat{A} \in 2_i^N, \\ q_j^B(t) &= \begin{cases} 1 \text{ if } B = A\backslash i \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } j \in A\backslash i, B \in 2_j^N, \\ q_{j'}^{\hat{B}}(t) &= q_{j'}^{\hat{B}}(t-1) \text{ for all } j' \in A^c, \hat{B} \in 2_{j'}^N. \end{aligned}$$

*Output:* Set $\mathbf{p}^* = \mathbf{q}(t)$.

**Proposition 6** *The output $\boldsymbol{p}^*$ of GreedyClustering is a local optimum, i.e. $W(\boldsymbol{p}^*) \geq W(\boldsymbol{q})$ for all $\boldsymbol{q} \in \mathcal{N}(\boldsymbol{p}^*)$.*

**Proof:** For the partition $P^*$ corresponding to output $\mathbf{p}^*$, the case of singleton blocks (if any) is trivial, in that if $\{i\} \in P^*$ then $W(q_i|\mathbf{p}^*_{-i}) = W(\mathbf{p}^*)$ for all membership distributions $q_i \in \Delta_i$. Hence let $i \in A \in P^*$ with $|A| > 1$. By switching from $p_i^*$ to any different $q_i \in \Delta_i$, the change in global score is

$$\begin{aligned} W(q_i|\mathbf{p}^*_{-i}) - W(\mathbf{p}^*) &= w(\{i\}) - w(A) + \\ &+ \left( q_i^A \sum_{B \in 2^A \backslash 2^{A\backslash i} : |B|>1} \mu^w(B) + \sum_{B' \in 2^{A\backslash i}} \mu^w(B') \right) = \\ &= (q_i^A - 1) \sum_{B \in 2^A \backslash 2^{A\backslash i} : |B|>1} \mu^w(B), \end{aligned}$$

where the last equality is due to $w(A) - w(A\backslash i) = \sum_{B \in 2^A \backslash 2^{A\backslash i}} \mu^w(B)$. Now assume that $\mathbf{p}^*$ is not a local optimum, i.e. $W(q_i|\mathbf{p}^*_{-i}) - W(\mathbf{p}^*) > 0$. Since $q_i^A - 1 < 0$, it must also be

$$\sum_{B \in 2^A \backslash 2^{A\backslash i} : |B|>1} \mu^w(B) = w(A) - w(A\backslash i) - w(\{i\}) < 0,$$

but this is not possible in view of *CheckLoop*. ■

For the class of graphs detailed above where *GreedyMerging* provides a worst-case modularity score equal to zero, it is easy to check that for reasonable input $\mathbf{q}(0)$ *GreedyClustering* does not fall into the same trap and surely finds the unique optimum. In particular, consider for simplicity the initial $n$-tuple of uniform distributions $q_i^A(0) = 2^{1-n}$ for all $A \in 2_i^N, i \in N$. Then for every edge $\{i_k, j_k\} \in E$ where $i_k \in N^1$ and $j_k \in N^2$ the average derivative takes value

$$\bar{W}_{\{i_k,j_k\}}(\mathbf{q}(0)) = \frac{1}{2}\left[ -\frac{2}{n^2} + \frac{2}{2^{n-1}} \frac{2}{n^2} \right] = -\frac{1}{n^2}\left( 1 - \frac{1}{2^{n-2}} \right) < 0,$$

while for every subset $A \subseteq N^1$ (or $A \subseteq N^2$) its value is

$$\bar{W}_A(\mathbf{q}(0)) = \frac{1}{|A|}\left[ -\frac{|A|}{n^2} + \binom{|A|}{2} \frac{4}{n^2 2^{n-1}} \right] = -\frac{1}{n^2}\left( 1 - \frac{|A|-1}{2^{n-2}} \right) < 0.$$

Hence for $|A| = 2$ there is no difference, but $\bar{W}_A(\mathbf{q}(0))$ increases with $|A|$ and

$$\bar{W}_{N^1}(\mathbf{q}(0)) = -\frac{1}{n^2}\left(1 - \frac{\frac{n}{2}-1}{2^{n-2}}\right) = \bar{W}_{N^2}(\mathbf{q}(0)),$$

where $\frac{1}{2^{n-2}} < \frac{n-2}{2^{n-1}}$ as long as $n > 4$ indeed.

Apart from the obvious case of a partition $\mathbf{p} = \mathbf{q}(0)$ as input, *CheckLoop* is also strictly necessary in general. In fact, if at some iteration $t$ with $\mathbf{q}(t) = \mathbf{q}$ the greedily formed block was some $A, |A| > 1$ rather than $A\backslash i$ for any $i \in A$, then $\bar{W}_A(\mathbf{q}) - \bar{W}_{A\backslash i}(\mathbf{q}) \geq 0$ or

$$\begin{aligned} &\frac{1}{|A|}\left[\sum_{j\in A} w(\{j\}) + \sum_{\{j,j'\}\subseteq A}\left(q_j^A + q_{j'}^A\right)\mu^w(\{j,j'\})\right] \quad + \\ -\ &\frac{1}{|A|-1}\left[\sum_{j\in A\backslash i} w(\{j\}) + \sum_{\{j,j'\}\subseteq A\backslash i}\left(q_j^{A\backslash i} + q_{j'}^{A\backslash i}\right)\mu^w(\{j,j'\})\right] \quad \geq \quad 0. \end{aligned}$$

However, the sought condition $w(A) \geq w(A\backslash i) + w(\{i\})$ cannot be cast for generic $q^A, q^{A\backslash i}$, even when these latter are maintained at a certain ratio $q_j^A/q_j^{A\backslash i}$ for all $j \in A\backslash i$ according to both: (i) expression (10) defininining input $\mathbf{q}(0)$, and (ii) the updating rule (c) in *GreedyLoop*.

Although analyzing *GreedyMerging* in terms of near-Boolean optimization offers a useful perspective, still its comparison with *GreedyClustering* has a purely illustrative purpose. Indeed, together with the general similarities observed thus far, the two also share an autonomous (i.e. optimization-driven) determination of the number of clusters (rather than requiring it as an input). However, they differ crucially in terms of computational burden: at each iteration $t$ the former only explores $\binom{|P(t)|}{2}$ possible unions of two blocks (of the current partition $P(t)$), while the latter has to quantify (ideally) the average derivatives $\bar{W}_A(\mathbf{q}(t))$ for all clusters $A$ where $\mathbf{0} \neq q^A(t) \neq \chi_A$. Of course, the same general arguments towards a restricted search for *GreedyMerging* also apply to *GreedyClustering*. More precisely, if two blocks $A, B \in P(t)$ are such that in the given graph $G$ their union $A \cup B$ spans a disconnected subgraph $G(A \cup B)$, then $A \cup B$ can be excluded from the $\binom{|P(t)|}{2}$ possible unions of two blocks of $P(t)$. Similarly, although $\mathbf{0} \neq q^A(t) \neq \chi_A$, if $G(A)$ is disconnected, then average derivative $\bar{W}_A(\mathbf{q}(t)$ can be ignored. But even in view of these restrictions, the two greedy procedures remain far too diverse. In fact, apart from the computational demand, another difference is that *GreedyMerging* starts from the finest partition independently from the input cluster score function $w$, and thus is not local in any manner. Conversely, although not progressing from neighborhood to neighborhood, still the search conducted by *GreedyClustering* may well be regarded as a local one, since it develops from an arbitrary initial $\mathbf{q}(0)$, which in particular can be determined depending on $w$ as in expression (10).

One common way to employ local search methods is by means of several runs, for different initial candidate solutions, with the associated outputs thus providing a range for locally optimal values taken by the objective function. In such settings, meeting the computational demand of a single run is usually fast, and hence diversified initial candidate solutions enable to figure globally optimal values. The same approach may be applied to *GreedyClustering*, while

also *nesting* the sequence $\mathbf{q}(0_1), \ldots, \mathbf{q}(0_T)$ of different initial candidate solutions in the following manner. Define the very first input fuzzy clustering $\mathbf{q}(0_1)$ by

$$q_i^A(0_1) = \begin{cases} \frac{2}{(n-1)(n-2)} \text{ if } |A| = 2 \text{ (i.e. if } A = \{i, j\}, j \in N \backslash i) \\ 0 \text{ if } |A| \neq 2 \end{cases}$$

for all $A \in 2_i^N, i \in N$. The associated output $\mathbf{p}_1^*$ clearly is a partition $P_1^*$ each of whose blocks is either a singleton or a pair. At this point, the procedure may be reiterated by transforming the original $n$-set $N = N^0 = \{1, \ldots, n\}$ of indices into the updated $N^1 = \{1, \ldots, |P_1^*|\}$, and with input $\mathbf{q}(0_2)$ placing the uniform distribution over all $\binom{|P_1^*|}{2}$ clusters obtained as the union of two blocks of $P_1^*$. In other terms, the original set $2^N$ containing all $2^n$ clusters is replaced with the novel (restricted) set $2^{P_1^*}$, namely with the *field of subsets* (of $N$) generated by partition $P_1^*$. Hence, just like in the first iteration, the novel input $\mathbf{q}(0_2)$ distributes memberships uniformly over all and only those clusters obtained as the union of exactly two blocks of $P_1^*$. The stopping condition is reached at $T$, relying on criterion $P_T^* = P_{T-1}^*$. Embedded within such a larger loop, *GreedyClustering* appears quite more similar to *GreedyMerging*, in that: (i) it no longer relies on a local search but is conversely conceived to start exclusively from the finest partition, (ii) at each iteration $t$ (of the outer loop) it only explores the average derivative $\bar{W}_A(\mathbf{q}(t))$ for a limited (by $\binom{|P_t^*|}{2}$) number of clusters, and (iii) the two stopping criteria are evidently equivalent.

# 6 Conclusions

Developing from the evaluation of fuzzy clusters via MLE of pseudo-Boolean functions, this work proposes a general framework where to design and analyze objective function-based clustering. With respect to other optimization approaches based on fuzzy modeling, the main difference is that here fuzzy clusterings constitute the mean for exploring a larger (i.e. continuous rather than discrete) search space, and are not intended to also achieve better global values than hard ones, as they cannot. The general setting applies to any clustering problem, as it only requires to formalize a cluster score (set) function. In the simplest case (such as modularity maximization), this latter has a quadratic MLE. Indeed, the input of many clustering problems is a similarity matrix, which finds a seemingly natural translation into a cluster score function with quadratic MLE. A further example obtains by quantifying (through suitable values of Möbius inversion) the empirical evidence that in social networks "good clusters" or communities have members who, apart from being "densely adjacent", also display a greater-than-expected number of neighbors. It can be mentioned that these quadratic cluster score functions need not be mutually exclusive, as all of the above applies invariate to their linear combinations.

From a general perspective, local search methods such as the proposed *GreedyClustering* may also lead to a final observation focused on overlapping community detection in complex networks via objective function-based graph clustering. The fundamental issue is that, independently from the chosen formal definition of community [9, 10], in many environments vertices must be allowed to be members of different such communities (thus making these latter overlapping or with non-empty intersection). In this view, objective function-based

graph clustering methods such as modularity maximization are sometimes questioned on the basis that they provide optimal partitions, namely collections of disjoint (i.e. non-overlapping) clusters, and thus cannot suitably address the issue of finding "optimal" set systems of overlapping clusters. However, from such an overall perspective local search optimization methods may be useful for overlapping community detection, while also providing a formal definition of community. In fact, consider again modularity $\mathcal{Q}$ as the fundamental example and assume to have computational resources sufficient for many runs of a (fast) local search algorithm maximizing $\mathcal{Q}$. Accordingly, following a sequence of $T$ (non-nested) inputs, partitions $P_1^*, \ldots, P_T^*$ are the corresponding outputs. These latter shall be local optima with respect to the notion of neighborhood underlying the algorithm itself, but in any case there is a (non-empty) subset $\mathfrak{T} \subseteq \{1, \ldots, T\}$ such that for all $t \in \mathfrak{T}$ inequality $\mathcal{Q}(P_t^*) \geq \mathcal{Q}(P_{t'}^*)$ holds for $1 \leq t' \leq T$. Now let $\mathcal{C} = \bigcup_{t \in \mathfrak{T}} P_t^* = \{A : A \in P_t^* \text{ for some } t \in \mathfrak{T}\}$. In words, $\mathcal{C}$ is the set system consisting of all clusters that are blocks of at least one optimal output $P_t^*, t \in \mathfrak{T}$, and thus such clusters may well be overlapping. The resulting definition of community clearly is: $A$ is a community $\Leftrightarrow A \in \mathcal{C}$.

# References


[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466:761–764, 2010.

[2] M. Aigner. *Combinatorial Theory - Reprint of the 1979 ed.* Springer, 1997.

[3] S. Asur, D. Ucar, and S. Parthasarathy. An ensemble framework for clustering protein-protein interaction networks. *Bioinformatics*, 23:i29–i40, 2007.

[4] E. Boros and P. Hammer. Pseudo-Boolean optimization. *Discrete Applied Mathematics*, 123:155–225, 2002.

[5] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Trans. on Knowledge and Data Engineering*, 20(2):172–188, 2007.

[6] A. E. Brower and W. H. Haemers. *Spectra of Graphs*. Springer, 2011.

[7] M. Cottrell, B. Hammer, A. Hasenfuß, and T. Villmann. Batch and median neural gas. *Neural Networks*, 19(6-7):762–771, 2006.

[8] K.-L. Du. Clustering: a neural network approach. *Neural Networks*, 23:89–107, 2010.

[9] E. Estrada and N. Hatano. Communicability in complex networks. *Physical Review E*, 77:036111, 2008.

[10] E. Estrada and N. Hatano. Communicability graph and community structures in complex networks. *Applied Mathematics and Computation*, 214(2):500–511, 2009.

[11] D. J. Fenn, M. A. Porter, P. J. Mucha, M. McDonald, S. Williams, N. F. Johnson, and N. S. Jones. Dynamical clustering of exchange rates. *Quantitative Finance*, 12(10):1493–1520, 2012.

[12] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[13] I. Gilboa and E. Lehrer. Global games. *International Journal of Game Theory*, (20):120–147, 1990.

[14] I. Gilboa and E. Lehrer. The value of information - an axiomatic approach. *Journal of Mathematical Economics*, 20(5):443–459, 1991.

[15] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics - A Foundation for Computer Science (Second Edition)*. Addison-Wesley, 1994.

[16] A. K. Jain. Data clustering: 50 years beyond $k$-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[17] M. Junjie, M. K. Ng, Y.-M. Cheung, and J. Z. Huang. Agglomerative fuzzy $k$-means clustering algorithm with selection of number of clusters. *IEEE Trans. on Knowledge and Data Engineering*, 20(11):1519–1534, 2008.

[18] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[19] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physics Review E*, 78(4):046110, 2008.

[20] E. Lughofer. Extensions of vector quantization for incremental clustering. *Pattern Recognition*, 41:995–1011, 2008.

[21] M. E. J. Newman and A.-L. Barabási and D. J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[22] S. Miyamoto, H. Ichihashi, and K. Honda. *Algorithms for Fuzzy Clustering*. Springer, 2008.

[23] T. Nepusz, A. Petróczi, L. Négyessy, and F. Baszó. Fuzzy communities and the concept of bridgeness in complex networks. *Physics Review E*, 77(1):016107, 2008.

[24] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003.

[25] M. E. J. Newman. Fast algorithm for detecting communities in networks. *Physics Review E*, 69(6):066133, 2004.

[26] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[27] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003.

[28] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters*, 93(21):218701, 2004.

[29] G. Rossi. Multilinear objective function-based clustering. In *Proceedings of the 7th Int. J. Conf. on Computational Intelligence*, volume 2 Fuzzy Computation Theory and Applications, pages 141–149, 2015.

[30] G. Rossi. Near-Boolean optimization - a continuous approach to set packing and partitioning. In A. Fred, M. De Marsico, and G. Sanniti di Baja, editors, *LNCS 10163 Pattern Recognition Applications and Methods*, pages 60–87. Springer, 2017.

[31] G.-C. Rota. The number of partitions of a set. *American Mathematical Monthly*, 71:499–504, 1964.

[32] G.-C. Rota. On the foundations of combinatorial theory I: theory of Möbius functions. *Z. Wahrscheinlichkeitsrechnung u. verw. Geb.*, 2:340–368, 1964.

[33] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1:27–64, 2007.

[34] R. Stanley. Modular elements of geometric lattices. *Algebra Universalis*, 1:214–217, 1971.

[35] J. Vlasblom and S. J. Wodak. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, 10:99, 2009.

[36] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[37] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.

[38] W. Wang and Y. Zhang. On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, 158:2095–2117, 2007.

[39] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. In H. Kargupta, J. Srivastava, C. Kamath, and A. Goodman, editors, *Proceedings of the 2005 SIAM Conference on Data Mining*, pages 274–285, 2005.

[40] S. Wu and T. W. S. Chow. Clustering of the self-organizing map using a cluster validity index based on inter-cluster and intra-cluster density. *Pattern Recognition*, 37:175–188, 2004.

[41] J. Xie, S. Kelley, and B. K. Szymanski. Overlapping community detection in networks: the state of the art and a comparative study. *ACM Computing Surveys*, 45(43):1–35, 2012.

[42] S. Zhang, R.-S. Wang, and X.-S. Zhang. Identification of overlapping community structure in complex networks using fuzzy $c$-means clustering. *Phisica A*, 374:483–490, 2007.