# Local quality functions for graph clustering with non-negative matrix factorization

Twan van Laarhoven[*] and Elena Marchiori[†]
*Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands*

(Received 22 July 2014; published 29 December 2014)

Many graph clustering quality functions suffer from a resolution limit, namely the inability to find small clusters in large graphs. So-called resolution-limit-free quality functions do not have this limit. This property was previously introduced for hard clustering, that is, graph partitioning. We investigate the resolution-limit-free property in the context of non-negative matrix factorization (NMF) for hard and soft graph clustering. To use NMF in the hard clustering setting, a common approach is to assign each node to its highest membership cluster. We show that in this case symmetric NMF is not resolution-limit free, but that it becomes so when hardness constraints are used as part of the optimization. The resulting function is strongly linked to the constant Potts model. In soft clustering, nodes can belong to more than one cluster, with varying degrees of membership. In this setting resolution-limit free turns out to be too strong a property. Therefore we introduce *locality*, which roughly states that changing one part of the graph does not affect the clustering of other parts of the graph. We argue that this is a desirable property, provide conditions under which NMF quality functions are local, and propose a novel class of local probabilistic NMF quality functions for soft graph clustering.

DOI: 10.1103/PhysRevE.90.062813 PACS number(s): 89.75.Hc, 05.10.−a, 02.70.Rr

## I. INTRODUCTION

Graph clustering, also known as network community detection, is an important problem with real-life applications in diverse disciplines such as life and social sciences [1,2]. Graph clustering is often performed by optimizing a quality function, which is a function that assigns a score to a clustering. During the past few decades, many such functions (and algorithms to optimize them) have been proposed. However, relatively little effort has been devoted to the theoretical foundation of graph clustering quality functions, e.g., Ref. [3]. In this paper we try to provide a contribution in this direction by studying desirable locality properties of quality functions for hard and soft graph clustering.

We focus on the resolution-limit-free property, a property of hard graph clustering, recently introduced by Traag, Van Dooren, and Nesterov [4]. Resolution-limit freeness is essentially a locality property. Informally this property states that a subset of an optimal clustering in the original graph should also be an optimal clustering in the induced subgraph containing only the nodes in the subset of clusters. As the name suggests, resolution-limit-free quality functions do not suffer from the so-called resolution limit, that is, the inability to find small clusters in large graphs. In the seminal work by Fortunato and Barthélemy [5], it was shown that modularity [6], a popular quality function used for network community detection, has a resolution limit, in the sense that it may not detect clusters smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the clusters.

Our goal is to investigate resolution-limit freeness and other locality properties of non-negative matrix factorization (NMF) graph clustering quality functions. NMF [7,8] is a popular machine learning method initially used to learn the parts of objects, like human faces and text documents. It finds two non-negative matrices whose product provides a good approximation to the input matrix. The non-negative constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. Recently, NMF formulations have been proposed as quality functions for graph clustering, see, for instance, the surveys Wang *et al.* [9] and Li and Ding [10].

We consider symmetric and asymmetric NMF formulations based on Euclidean loss and a Bayesian NMF quality function recently proposed by Psorakis *et al.* [11], which can automatically determine the number of clusters.

The resolution-limit-free property is stated in the setting of hard clustering, where a clustering is a partition of the nodes. In contrast, NMF produces a soft clustering. Nodes have varying degrees of memberships of each clusters, and the clusters can overlap. To use NMF in the hard clustering setting, a common approach is to assign each node to its highest membership cluster.

In Sec. III we show that hard clustering based on NMF in this way is, in general, not resolution-limit free. For symmetric NMF we show that resolution-limit freeness can be obtained by using orthogonality constraints as part of the optimization and that the resulting function is strongly linked to the constant Potts model (CPM). CPM was introduced by Traag *et al.* as the simplest formulation of a (nontrivial) resolution-limit-free method. It is a variant of the Potts model by Reichardt and Bornholdt [12].

We argue in Sec. IV that in the soft clustering setting, resolution-limit freeness is a too-strong property and propose an alternative desirable locality property for soft graph clustering. We characterize an interesting class of local quality functions and show that symmetric and asymmetric NMF belong to this class. We show that Bayesian NMF is not local in general and that it suffers from a resolution limit. In Sec. V we

[*] tvanlaarhoven@cs.ru.nl
[†] elenam@cs.ru.nl

introduce a novel class of probabilistic NMF quality functions that are local and hence do not suffer from a resolution limit.

### A. Related work

The notion of resolution limit was introduced in Fortunato and Barthélemy [5]. They found a limitation of modularity, considered a state-of-the-art method for community detection. Van Laarhoven and Marchiori [13] showed empirically that the resolution limit is the most important difference between quality functions in graph clustering optimized using a fast local search algorithm, the Louvain method [14]. Traag *et al.* [4] introduced the notion of resolution-limit-free objective functions, which provides the motivation of this study.

Other local properties of quality functions for clustering have been considered in theoretical studies, but mainly in the hard setting, for distance-based clustering [15] and for graph clustering [16]. Locality as defined in Ackerman *et al.* [15] is a property of clustering functions, therein defined as functions mapping a data set and a positive integer $k$ to a partition of the data into $k$ clusters. This notion of locality was used together with other properties to characterize linkage-based clustering. The locality property considered in van Laarhoven and Marchiori [16] is part of an axiomatic study of quality functions for hard graph clustering. It states that local changes to a graph should have only local consequences to a clustering. It is slightly weaker than the locality property considered in this study, which corresponds more closely to the property there called strong locality.

### B. Definitions and notation

A (weighted) *graph* is a pair $(V,A)$ of a finite set $V$ of nodes and a function $A : V \times V \rightarrow \mathbb{R}_{\geqslant 0}$ of edge weights. For compactness we view $A$ as an adjacency matrix and write $a_{ij} = A(i,j)$. Edges with larger weights represent stronger connections, so $a_{ij} = 0$ means that there is no edge between nodes $i$ and $j$. A graph $G' = (V',A')$ is a *subgraph* of $G = (V,A)$ if $V' \subseteq V$ and $a'_{ij} = a_{ij}$ for all $i,j \in V'$.

Different clustering methods use different notions of a "cluster" and of a "clustering." For instance, in symmetric NMF a clustering is a matrix of membership coefficients; while in nonsymmetric NMF there are two such matrices. Some methods also have additional parameters for each cluster. In this paper we allow different types of "cluster" for different methods, but we use a common definition of "clustering."

Formally, each of these types of clusters can be specified by an injective function $\mathcal{C}$ from sets of nodes to sets of things which we call clusters. For a set of nodes $s$, for every cluster $c \in \mathcal{C}(s)$ we call $s$ the *support* of $c$, written as $\mathrm{supp}(c) = s$. The set of all clusters with support on a subset of $V$ is $\mathcal{C}^*(V) = \bigcup_{s \subseteq V} \mathcal{C}(s)$. In this paper we consider four types of clusters, which will be introduced in the next section.

A *clustering* of $V$ is a multiset of clusters with support on a subset of $V$. Note that we use multisets instead of sets to allow a clustering to contain two identical copies of the same cluster. For brevity, we also say that $C$ is a clustering of a graph $G$ if $C$ is a clustering of the nodes of $G$. If, in a slight abuse of notation, we define the support of a clustering as the union of the support of all clusters in that clustering, then the clusterings of $V$ are those multisets of clusters for which the support is a subset of $V$.

Note that this general definition implies that for certain clusterings the clusters can overlap, and some nodes can be in no cluster at all. We believe that this is a reasonable definition, because if we allow nodes to be in more than one cluster, there is little reason to not also allow them to be in less than one cluster.

Additionally, if $C$ and $D$ are clusterings of $G$, then their multiset sum $C \uplus D$ is also a clustering of $G$ [17], as is any subclustering (submultiset) of $C$. And if $G$ is a subgraph of $G'$, then $C$ and $D$ are also clusterings of $G'$. The symmetric difference of two clusterings is denoted $C \triangle D$ and is defined as the symmetric difference of multisets, that is, $C \triangle D = (C \backslash D) \cup (D \backslash C)$.

Graph clustering can be cast as an optimization problem. The objective that is being optimized is the *clustering quality function*, which is a function from graphs $G$ and clusterings of $G$ to real numbers. In this paper we take the convention that the quality is maximized.

Given a clustering quality function $q$, and a clustering $C$ of some graph $G$. We say that $C$ is $q$ *optimal* if $q(G,C) \geqslant q(G,C')$ for all clusterings $C'$ of $G$.

## II. NON-NEGATIVE MATRIX FACTORIZATION

At its core, non-negative matrix factorization decomposes a matrix $A$ as a product $A \approx WH^T$, where all entries in $W$ and $H$ are non-negative. For graph clustering the matrix $A$ is the adjacency matrix of a graph. For undirected graphs the adjacency matrix is symmetric, in which case it makes sense to decompose it as $A \approx HH^T$. Note that such a symmetric factorization has to be enforced explicitly, since the optimal nonsymmetric factorization of a symmetric matrix does not necessarily have $W = H$ [18].

The columns of $W$ and $H$ can be interpreted as clusters. To fit with the definitions of the previous paragraph we need to take a slightly different view. In the case of symmetric NMF, a cluster with support $s$ is a function that assigns a positive real number to each node in $s$, so $\mathcal{C}_{\mathrm{SymNMF}}(s) = \mathbb{R}^s_{>0}$. Equivalently, for a fixed set of nodes, we can represent a cluster as a vector of non-negative numbers with an entry for each node in $V$, such that the entries for the nodes not in $s$ are zero, that is, $\mathcal{C}^*_{\mathrm{SymNMF}}(V) \approx \mathbb{R}^V_{\geqslant 0}$. For a cluster $c$ we denote this vector as $h_c$, and a multiset of such vectors can be seen as a matrix $H$. The support of $c$ then coincides with the standard notion of support of the vector $h_c$, that is, the set $s$ of nodes for which the entry is nonzero. This representation of clusters in terms of a non-negative vector $h_c$ is more standard and more convenient than the one in terms of a function from $s$ to positive real numbers, and we use it in the rest of the paper.

For nonsymmetric NMF, a cluster is a tuple $c = (w_c,h_c)$ of two such vectors. That is, $\mathcal{C}^*_{\mathrm{AsymNMF}}(V) = \mathbb{R}^V_{\geqslant 0} \times \mathbb{R}^V_{\geqslant 0}$, with $\mathrm{supp}((w_c,h_c)) = \mathrm{supp}(w_c) \cup \mathrm{supp}(h_c)$. For Bayesian NMF [11] each cluster also contains a $\beta_c$ parameter, that is, $\mathcal{C}^*_{\mathrm{BayNMF}}(V) = \mathbb{R}^V_{\geqslant 0} \times \mathbb{R}^V_{\geqslant 0} \times \mathbb{R}_{>0}$.

A common notion to all NMF methods is that they predict a value for each edge. For symmetric NMF with per cluster membership vector $h_c$ this prediction can be written as

$\hat{a}_{ij} = \sum_{c\in C} h_{ci}h_{cj}$. For asymmetric NMF with cluster memberships $w_c$ and $h_c$ we can write $\hat{a}_{ij} = \sum_{c\in C} w_{ci}h_{cj}$.

The optimization problem then tries to ensure that $\hat{a}_{ij} \approx a_{ij}$. Different methods can have different interpretations of the "≈" symbol, and they impose different regularizations and possibly additional constraints. Perhaps the simplest NMF quality function for undirected graphs uses Euclidean distance and no additional regularization,

$$q_{\mathrm{SymNMF}}(G,C) = -\frac{1}{2}\sum_{i,j\in V}(a_{ij} - \hat{a}_{ij})^2.$$

## III. RESOLUTION-LIMIT-FREE FUNCTIONS FOR HARD CLUSTERING

Before we investigate the resolution limits of NMF, we will first look at traditional "hard" clustering, where each node belongs to exactly one cluster. In this setting a cluster is simply a subset of the nodes, and its support is the cluster itself, that is, $\mathcal{C}_{\mathrm{hard}}(s) = s$. There is the additional nonoverlapping or orthogonality constraint on clusters: In a valid hard clustering $C$ of $V$, each node $i \in V$ is in exactly one cluster $c_i \in C$. For symmetric NMF we may formulate these constraints as

$$\sum_{i\in V} h_{ci}h_{di} = 0 \quad \text{for all} \quad c,d \in C, c \neq d, \quad \text{and}$$

$$\sum_{c\in C} h_{ci} = 1 \quad \text{for all} \quad i \in V.$$

Traag *et al.* [4] introduced a locality property of clustering quality functions and called the functions that satisfy this property *resolution-limit free*. Their definition is as follows.

*Definition 1 (Resolution-limit free).* Let $C$ be a $q$-optimal clustering of a graph $G_1$. Then the quality function $q$ is called *resolution-limit free* if for each subgraph $G_2$ induced by $D \subset C$ the partition $D$ is a $q$-optimal clustering of $G_2$.

Thus in the setting of hard clustering, a quality function is resolution-limit free if any subset of clusters from an optimal clustering is also an optimal clustering on the graph that contains only the nodes and edges in those clusters.

NMF has been extended with a postprocessing step to yield a hard clustering. This is done by assigning each node to the cluster with the largest membership coefficient.

We can now ask if NMF with this postprocessing is resolution-limit free. In Fig. 1 we give a counterexample that answers this question negatively for the NMF-based methods of Psorakis *et al.* [11] and Ding *et al.* [19].

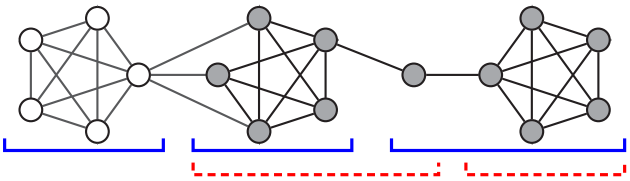

FIG. 1. (Color online) A counterexample that shows that NMF quality functions are not resolution limit free. When considering the entire graph, the first (solid blue) clustering is optimal. When considering only the gray nodes, the second (dashed red) clustering is optimal. The membership of the middle node is very unclear; it belongs to two clusters to almost the same degree. When another part of a cluster changes this can tip the balance one way or the other.

This counterexample consists of two cliques and one almost-clique. Additionally, there is a node with unclear membership. When the entire graph is considered, its membership of one cluster is slightly higher; when one clique and its incident edges are removed, its membership of another cluster is slightly higher. This difference is very small. For example, with Ding *et al.*'s method in the optimal clustering of the large graph, the disputed node belongs to the second and third clusters with membership coefficients 0.2306 and 0.2311, respectively; while in the smaller subgraph the membership coefficients are 0.2284 and 0.2607.

Traag *et al.* [4] showed that the CPM is the simplest formulation of any (nontrivial) resolution-limit-free method. The CPM quality function $q_{\mathrm{cpm}}(G,C)$ can be formulated as

$$q_{\mathrm{cpm}}(G,C) = \sum_{i,j\in V}(a_{ij} - \gamma)\mathbf{1}[c_i = c_j],$$

where $\mathbf{1}[c_i = c_j]$ is 1 if nodes $i$ and $j$ belong to the same cluster and 0 otherwise.

Symmetric NMF and CPM are closely related. This can be shown with a technique similar to that used by Ding *et al.* [19] to link symmetric NMF and spectral clustering.

*Theorem 2.* Symmetric NMF is an instance of CPM with $\gamma = 1/2$ and orthogonality constraints relaxed.

*Proof.* Recall that in symmetric NMF, $\hat{a}$ is defined as $\hat{a}_{ij} = \sum_{c\in C} h_{ci}h_{cj}$. With orthogonality constraints, any two nodes $i$ and $j$ are either in the same cluster, in which case $\hat{a}_{ij} = 1$, or they are in different clusters, in which case $\hat{a}_{ij} = 0$. So $\hat{a}_{ij} = \hat{a}_{ij}^2 = \mathbf{1}[c_i = c_j]$.

Symmetric NMF is given by the optimization problem

$$\underset{C}{\operatorname{argmax}}\, q_{\mathrm{SymNMF}}(G,C) = -\frac{1}{2}\sum_{i,j\in V}(a_{ij} - \hat{a}_{ij})^2.$$

Expanding the square shows that this is equivalent to

$$\underset{C}{\operatorname{argmax}} \sum_{i,j\in V}\left(a_{ij}\hat{a}_{ij} - \frac{1}{2}\hat{a}_{ij}^2\right).$$

With orthogonality constraints this is equivalent to

$$\operatorname{argmax}_C \sum_{i,j\in V}\left(a_{ij} - \frac{1}{2}\right)\mathbf{1}[c_i = c_j],$$

which is the CPM objective with $\gamma = 1/2$. ■

The CPM *is* resolution-limit free. Therefore in order to perform hard clustering using symmetric NMF it is preferable to act on the quality function, for instance, by enforcing orthogonality as done in Refs. [19,20], instead of assigning each node to the cluster with the highest membership coefficient.

## IV. RESOLUTION-LIMIT-FREE FUNCTIONS FOR SOFT CLUSTERING

We could still try to directly adapt Definition 1 to the soft clustering setting by defining what a graph induced by a subclustering is. The obvious idea is to include all nodes in the support of the subclustering. So for a clustering $C$ of $G$, the graph $G'$ induced by $D \subseteq C$ would contain only the nodes

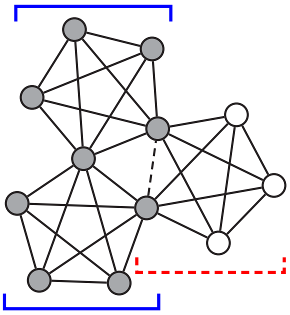

FIG. 2. (Color online) Three cliques sharing two nodes each. The obvious clustering consists of three overlapping clusters, with the three central nodes in two clusters each. The white nodes are not in the support of the solid blue clusters.

which are in at least one cluster in $D$, that is, $V' = \text{supp}(D)$, and all edges between these nodes from the original graph.

However, in contrast to the hard clustering case, an optimal soft clustering might have clusters in $C \setminus D$ that overlap with clusters in $D$. This makes the notion of resolution-limit free too restrictive, since it effectively disallows any interesting uses of overlapping clusters.

Consider the graph with three overlapping 5-cliques shown in Fig. 2. In an NMF-style method such as Ref. [19], the optimal clustering of this graph will have three overlapping clusters, corresponding to the three cliques. The subgraph introduced by the support of the solid blue clusters includes just the dark nodes, but neither cluster covers both nodes incident to the dashed edge. Therefore, with these two clusters the prediction $\hat{a}$ for this edge will be 0. But the optimal clustering of this subgraph would have a nonzero prediction for this edge. In other words, the optimal clustering for the induced subgraph is not the same as the solid blue clustering, and even the support of the clusters is different. Hence no NMF method is resolution-limit free in this sense.

An alternative approach is to only consider subclusterings with disjoint support in the definition of resolution-limit free, that is, with $\text{supp}(D) \cap \text{supp}(C \setminus D) = \emptyset$. Unfortunately this variant has the opposite problem: The condition almost never holds. So many quality functions would trivially satisfy this variant of resolution-limit freeness. For example, the optimal clusterings in NMF methods based on a Poisson likelihood will always have overlapping clusters covering every edge, so the disjointness condition only holds when the graph has multiple connected components.

Clearly we need a compromise.

### A. Locality

The resolution-limit-free property looks at the behavior of a clustering quality function on graphs of different sizes. Intuitively, a quality function suffers from a resolution limit if optimal clusterings at a small scale depend on the size of the entire graph.

As shown in the previous paragraph we cannot just zoom in to the scale of any subclustering $D$ by discarding the rest of the graph. But if we let go of only considering the optimal clustering, it does become possible to zoom in only partially, leaving the part of the graph covered by clusters that overlap

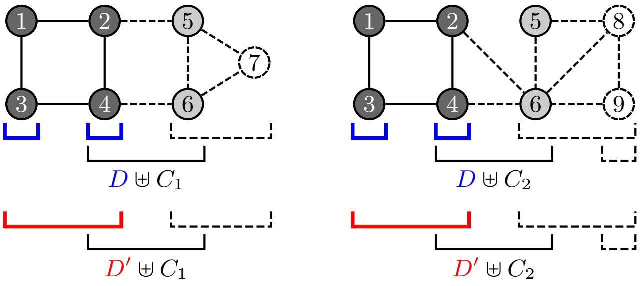

FIG. 3. (Color online) An example illustrating locality. Between the left and right sides, the dashed part of the clustering and the dashed part of the graph changes. The top and bottom clusterings differ only on the constant part (red or blue), and these differences do not overlap with changing clusters (dashed). Therefore if the top clustering has a higher quality than the bottom clustering on the left graph, then the same must hold on the right graph. Formally, the dark gray nodes are in the common subgraph $G_S$, and the light gray nodes are in $\text{supp}(C_1 \cap C_2)$. The thick blue clustering is $D$, the thick red clustering $D'$, the solid black clusters are in both $C_1$ and $C_2$, and the dashed clusters are in only one of $C_1$ and $C_2$. Since the dashed clusters do not cover the dark gray nodes, the black clusterings agree on the dark gray subgraph.

clusters in $D$ intact. If $D$ is an optimal clustering of the original graph, then it should be a "locally optimal" clustering of the smaller graph in some sense.

We take this to mean that if a clustering $D$ is better than some other clustering $D'$ on the original graph, then the same holds on the smaller graph, as long as $D$ and $D'$ induce the same zoomed-in graph.

It then makes sense to not only consider zooming in by discarding the rest of the graph but also consider arbitrary changes to the rest of the graph, as well as arbitrary changes to clusters not overlapping with $D$ or $D'$.

More precisely, if one subclustering $D$ is better than another subclustering $D'$ on a subgraph $G_S$ of some graph $G_1$, and one changes the graph to $G_2$ in such a way that the changes to the graph and to the clustering are disjoint from this subgraph $G_S$, then $D$ will stay a better clustering than $D'$. This idea is illustrated in Fig. 3.

To formalize this idea we introduce the notion of agreement. We say that two clusterings $C_1$ of $G_1$ and $C_2$ of $G_2$ *agree* on a common subgraph $G_S = (V_S, A_S)$ of $G_1$ and $G_2$ if $\text{supp}(C_1 \triangle C_2) \cap V_S = \emptyset$. Note that this subgraph can be the smallest subgraph containing $\text{supp}(D)$ and $\text{supp}(D')$. This leads to the following definition.

*Definition 3 (Locality).* A clustering quality function $q$ is *local* if for all graphs $G_1$, $G_2$, and common subgraphs $G_S$ of $G_1$ and $G_2$, for all clusterings $C_1$ of $G_1$ and $C_2$ of $G_2$ that agree on $G_S$, and clusterings $D, D'$ of $G_S$, it is the case that $q(G_1, C_1 \uplus D) \geqslant q(G_1, C_1 \uplus D')$ if and only if $q(G_2, C_2 \uplus D) \geqslant q(G_2, C_2 \uplus D')$.

Locality as defined in Ackerman *et al.* [15] differs from our definition because it is a property of clustering functions. The locality property considered in van Laarhoven and Marchiori [16] differs from our definition because it also enforces that the graphs agree "on the neighborhood" of the common subgraph. Instead, we require agreement between overlapping *clusters*.

They also briefly discussed and dismissed a “strong locality” property, which is closer to our definition.

Even in the case of hard clustering locality and resolution-limit free are not equivalent. For hard clustering, locality implies resolution-limit freeness, but the converse is not true.

*Theorem 4.* If a hard clustering quality function is local, then it is resolution-limit free.

*Proof.* Let $q$ be a local hard cluster quality function and $C$ be a $q$-optimal clustering of a graph $G_1 = (V_1, A_1)$.

Consider the subgraph $G_2$ induced by $D \subset C$.

Let $C_1 = C \backslash D$ and $C_2 = \emptyset$, and let $G_S = G_2$. Because $C$ is a partition of $V_1$, we have that $\mathrm{supp}(C_1)$ is disjoint from $G_S$, and so $C_1$ and $C_2$ agree on $G_S$.

Then for each clustering $D'$ of $G_2$ we have $q(G_1, C_1 \uplus D) \geqslant q(G_1, C_1 \uplus D')$ because $C_1 \cup D = C$ is an optimal clustering of $G_1$. By locality it follows that $q(G_2, C_2 \uplus D) \geqslant q(G_2, C_2 \uplus D')$.

So $D$ is a $q$-optimal clustering of $G_2$. ■

*Theorem 5.* If a hard clustering quality function is resolution-limit free, then it is not necessarily local.

*Proof.* Consider the following quality function:

$$q(G,C) = \max_{c \in C} |c| + \min_{c \in C} |c|.$$

For each graph $G = (V,A)$, the clustering $C = \{V\}$ is the single $q$-optimal clustering, with quality $2|V|$. Since there are no strict subsets of $C$ the quality function is trivially resolution-limit free.

Now consider the graphs $G_1$ with nodes $\{1,2, \ldots ,7\}$ and $G_2$ with nodes $\{1,2, \ldots ,6\}$, both with no edges. These graphs have a common subgraph $G_S$ with nodes $\{1,2, \ldots ,6\}$. Take the clusterings $D = \{\{1,2,3,4\},\{5\},\{6\}\}$, $D' = \{\{1,2,3\},\{4,5,6\}\}$, $C_1 = \{\{7\}\}$, and $C_2 = \{\}$. Then $q(G_1, C_1 \uplus D) = 5 > 4 = q(G_1, C_1 \uplus D')$, while $q(G_2, C_2 \uplus D) = 5 < 6 = q(G_2, C_2 \uplus D')$.

So $q$ is not local.

This counterexample is illustrated in Fig. 4. ■

### B. Characterizing local quality functions

Many quality functions can be written as a sum with a term for each edge, characterizing a goodness of fit, a term for each node, controlling the amount of overlap, and a term for each cluster, indicating some kind of complexity penalty. There might also be a constant term not actually depending on the clustering and so not affecting the optimum. We call such quality functions additive.

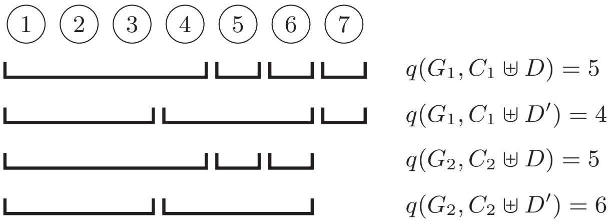
1
2
3
4
5
6
7
$q(G_1, C_1 \uplus D) = 5$
$q(G_1, C_1 \uplus D') = 4$
$q(G_2, C_2 \uplus D) = 5$
$q(G_2, C_2 \uplus D') = 6$

FIG. 4. The counterexample from the proof of Theorem 5.

*Definition 6.* A qualify function is *additive* if it can be written as

$$\begin{aligned} q(G,C) = {} & q_{\mathrm{graph}}(G) + \sum_{c \in C} q_{\mathrm{clus}}(c) \\ & + \sum_{i \in V} q_{\mathrm{node}}(\{c \in C \mid i \in \mathrm{supp}(c)\}) \\ & + \sum_{i \in V} \sum_{j \in V} q_{\mathrm{edge}}(a_{ij}, \{c \in C \mid i,j \in \mathrm{supp}(c)\}) \end{aligned}$$

for some functions $q_{\mathrm{graph}}$, $q_{\mathrm{clus}}$, $q_{\mathrm{node}}$, $q_{\mathrm{edge}}$.

Note that $q_{\mathrm{node}}$ can depend on all clusters that contain node $i$, and $q_{\mathrm{edge}}$ can depend on all clusters that contain the edge $ij$.

*Theorem 7.* If a quality function is additive, then it is local.

*Proof.* Let $q$ be an additive quality function. Let $G_1$ $G_2$ and $G_S = (V,A)$ be graphs such that $G_S$ is a subgraph of both $G_1$ and $G_2$.

Let $C_1$ be a clustering of $G_1$, $C_2$ a clustering of $G_2$ and, $D, D'$ clusterings of $G_S$

such that $C_1$ and $C_2$ agree on $G_S$.

Let $E = C_1 \cap C_2$. Then for every node $i \in \mathrm{supp}(C_1 \backslash C_2)$, we have $i \notin V$, which implies that $i \notin \mathrm{supp}(D)$ and $i \notin \mathrm{supp}(D')$. So $\{c \in C_1 \uplus D \mid i \in \mathrm{supp}(c)\} = \{c \in C_1 \uplus D' \mid i \in \mathrm{supp}(c)\} = \{c \in C_1 \mid i \in \mathrm{supp}(c)\}$.

Conversely, for every node $i \notin \mathrm{supp}(C_1 \backslash C_2)$, we have $\{c \in C_1 \uplus D \mid i \in \mathrm{supp}(c)\} = \{c \in E \uplus D \mid i \in \mathrm{supp}(c)\}$.

Therefore,

$$\begin{aligned} & q(G_1, C_1 \uplus D) - q(G_1, C_1 \uplus D') \\ & \quad = \sum_{c \in D} q_{\mathrm{clus}}(c) - \sum_{c \in D'} q_{\mathrm{clus}}(c) \\ & \qquad + \sum_{i \in V} q_{\mathrm{node}}(\{c \in E \uplus D \mid i \in \mathrm{supp}(c)\}) \\ & \qquad - \sum_{i \in V} q_{\mathrm{node}}(\{c \in E \uplus D' \mid i \in \mathrm{supp}(c)\}) \\ & \qquad + \sum_{i,j \in V} q_{\mathrm{edge}}(a_{ij}, \{c \in E \uplus D \mid i,j \in \mathrm{supp}(c)\}) \\ & \qquad - \sum_{i,j \in V} q_{\mathrm{edge}}(a_{ij}, \{c \in E \uplus D' \mid i,j \in \mathrm{supp}(c)\}), \end{aligned}$$

and similarly for $G_2$ and $C_2$ in place of the $G_1$ and $C_1$.

Which implies that $q(G_1, C_1 \uplus D) - q(G_1, C_1 \uplus D') = q(G_2, C_2 \uplus D) - q(G_2, C_2 \uplus D')$.

And so $q(G_1, C_1 \uplus D) \geqslant q(G_1, C_1 \uplus D')$ if and only if $q(G_2, C_2 \uplus D) \geqslant q(G_2, C_2 \uplus D')$.

In other words, $q$ is local. ■

The converse of Theorem 7 does not hold; not all local quality functions are additive. For example, any monotonic function of a local quality function is also local.

Another example are quality functions that use higher-order interactions, that is, it includes terms not only for nodes and edges but also for triangles and larger structures. For instance, the clique percolation method [21] finds clusters which are cliques. That method is local, but it is not additive. We could imagine including higher-order terms in the definition of additivity,

$$q(G,C) = \cdots + \sum_{i,j,k \in V} q_{\text{triangle}} \times (a_{ij}, a_{ik}, a_{jk}, \{c \in C \mid i,j,k \in \text{supp}(c)\}),$$

and so on. But for most purposes the edge term is sufficient; and the local quality functions that we consider in this paper are all additive in the sense of Definition 6.

Additivity provides additional insight into how quality functions behave: the quality is composed of the goodness-of-fit of a the clustering to nodes and edges (and perhaps larger structures), together with a cost term for each cluster. By Theorem 7, it also gives us a convenient way to *prove* that a certain quality function is local, while locality can more convenient if we want to *reason* about the behavior of a quality function.

For symmetric NMF, $\hat{a}_{ij}$ can be written as a sum over clusters that contain nodes $i$ and $j$,

$$\hat{a}_{ij} = \sum_{c \in C \text{ s.t. } i,j \in \text{supp}(c)} h_{ci} h_{cj}.$$

As a consequence, NMF quality functions without regularization, such as $q_{\text{SymNMF}}$, are additive. Therefore these quality functions are local.

Many regularization terms can also be encoded in an additive quality function. For example the L2 term $\sum_{c \in C} \sum_{i \in V} h_{ci}^2$ is a sum over clusters and independent of the graph, and so it fits in $q_{\text{clus}}$.

### C. Fixed number of clusters

The question of automatically finding the right number of clusters is still not fully solved. Therefore in most NMF-based clustering methods the number of clusters $k$ is specified by the user.

For most quality functions, if they are optimized directly without taking this restriction into account, then the number of clusters will tend to infinity. So we somehow need to fix the number of clusters.

The most direct way to incorporate this restriction of a fixed number of clusters is by adding it as a constraint to the quality function. That is, use $q(G,C,k) = q(G,C) + \mathbf{1}[|C| = k]\infty$. Strictly speaking this is not a function to the real numbers. But we never need the fact that $q$ is such a function, all we need is that the quality of different clusterings can be compared. Unfortunately, encoding a fixed $k$ restriction in the quality function violates locality.

Take two clusterings $C$ and $D$ of a graph $G$, with a different number of clusters. Let $C'$, $D'$ and $G'$ be copies of $C$, $D$, and $G$ on a disjoint set of nodes, and let $k$ be $|C| + |D|$. Then the quality $q(G \cup G', D \uplus C', k)$ is finite, while $q(G \cup G', D \uplus D', k)$ is infinite. On the other hand, $q(G \cup G', C \uplus C', k)$ is infinite, while $q(G \cup G', C \uplus D', k)$ is finite. This contradicts locality.

Instead, we need to consider the restriction on the number of clusters as separate from the quality function. In that case the definition of locality can be used unchanged.

Equivalently, if we call a clustering consisting of $k$ clusters a $k$-clustering, then we can extend the definitions of locality to take the restricted number of clusters into account. This approach is also used by Ackerman and Ben-David [15].

If we call a function $q(G,C,k)$ for graphs $G$, clusterings $C$ and number of clusters $k$ a fixed-size quality function, then this leads to the following fixed-size variant of locality.

*Definition 8 (Fixed size locality).* A fixed-size quality function $q$ is *fixed-size local* if for all graphs $G_1$, $G_2$ and a common subgraph $G_S$, for all $k_1$-clusterings $C_1$ of $G_1$ and $k_2$ clusterings $C_2$ of $G_2$ that agree on $G_S$, and $m$-clustering $D$ of $G_S$ and $m'$-clusterings $D'$ of $G_S$, it is the case that $q(G_1, C_1 \uplus D, k_1 + m) \geqslant q(G_1, C_1 \uplus D', k_1 + m')$ if and only if $q(G_2, C_2 \uplus D, k_2 + m) \geqslant q(G_2, C_2 \uplus D', k_2 + m')$.

Every local quality function that does not depend on $k$ is fixed-size local when combined with a constraint that the number of clusters must be $k$. And so NMF with a fixed number of clusters is fixed-size local.

### D. Varying number of clusters

Psorakis *et al.* [11] formulated a Bayesian formulation of NMF for overlapping community detection that uses automatic relevance determination (ARD) [22] to determine the number of clusters. Their quality functions can be written as

$$\begin{aligned} q_{\text{BayNMF}}(G,C) &= -\sum_{i \in V} \sum_{j \in V} \left( a_{ij} \log \frac{a_{ij}}{\hat{a}_{ij}} + \hat{a}_{ij} \right) \\ &\quad - \frac{1}{2} \sum_{c \in C} \left( \sum_{i \in V} \beta_c w_{ci}^2 + \sum_{i \in V} \beta_c h_{ci}^2 - 2|V| \log \beta_c \right) \\ &\quad - \sum_{c \in C} (\beta_c b - (a-1) \log \beta_c) - \kappa, \end{aligned}$$

where each cluster is a triple $c = (w_c, h_c, \beta_c)$ of two vectors and a scalar and $\kappa$ is a constant. ARD works by fixing the number of clusters to some upper bound. In the optimal clustering many of these clusters $c$ will be empty, that is, have $\text{supp}(c) = \emptyset$.

This quality function is *not* additive, for two reasons. First, there is the term $2|V| \log \beta_c$ for each cluster, which stems from the half-normal priors on $W$ and $H$. This term depends on the number of nodes. Second, the $\kappa$ term actually depends on the number of clusters and the number of nodes, since it contains the normalizing constants for the hyperprior on $\beta$, as well as constant factors for the half-normal priors. For a fixed graph and fixed number of clusters the $\kappa$ term can be ignored, however.

As a result, Psorakis *et al.*'s method is also not local, as the following counterexample shows:

*Theorem 9.* $q_{\text{BayNMF}}$ is not local.

*Proof.* Consider a graph $G_1$, consisting of a ring of $n = 10$ cliques, where each clique has $m = 5$ nodes, and two edges connecting it to the adjacent cliques.

We follow Psorakis *et al.*, and use hyperparameters $a = 5$ and $b = 2$. This choice is not essential, similar counterexamples exist for other hyperparameter values. As might be hoped, the $q_{\text{BayNMF}}$-optimal clustering $C_1$ of this graph then puts each clique in a separate cluster, with a small membership for the directly connected nodes in adjacent cliques.

This clustering is certainly better than the clustering $C_2$ with 5 clusters each consisting of two cliques, and 5 empty clusters.

However, on a larger graph with two disjoint copies of $G_1$, the clustering with two copies of $C_2$ is better than the clustering with two copies of $C_1$.

But by locality we would have $q_{\text{BayNMF}}(G_1 \cup G_1', C_1 \uplus C_1') \geqslant q_{\text{BayNMF}}(G_1 \cup G_1', C_2 \uplus C_1')$ as well as $q_{\text{BayNMF}}(G_1 \cup G_1', C_2 \uplus C_1') \geqslant q_{\text{BayNMF}}(G_1 \cup G_1', C_2 \uplus C_2')$, where the primed variables indicate copies with disjoint nodes. So $q_{\text{BayNMF}}$ is not local. ■

In the above counterexample things do not change if one uses a ring of 20 cliques instead of two disjoint rings of 10 cliques. This is closer to the original characterization of the resolution limit by Fortunato and Barthélemy [5]. In a ring of 20 cliques, the solution with 10 clusters is better than the solution with 20 clusters. But it is harder to show that this violates locality.

## V. NMF AS A PROBABILISTIC MODEL

NMF can be seen as a maximum likelihood fit of a generative probabilistic model. The quality function that is optimized is then the log likelihood of the model conditioned on the observed graph,

$$q(C,G) = \log P(C|G).$$

One assumes that there is some underlying hidden cluster structure, and the edges in the graph depend on this structure. The clustering structure in turn depends on the nodes under consideration. So, by Bayes rule, we may decompose $P(C|G)$ as

$$P(C|V,A) = P(A|C,V)P(C|V)P(V)/P(V,A).$$

The terms $P(V)$ and $P(V,A)$ are constant given the graph, so the quality function becomes

$$q(C,G) = \log P(A|C,V) + \log P(C|V) + \kappa,$$

where $\kappa = \log P(V) - \log P(V,A)$ is a constant. The first term is the likelihood of the edges given the clustering, and the second factor is the prior probability of a clustering for a certain set of nodes.

To make the above general formulation into an NMF model, one assumes that the edge weights are distributed independently, depending on the product of the membership matrices. Then a prior is imposed on the membership coefficients. Usually a conjugate prior is used, which for Gaussian likelihood has a half-normal distribution, and for Poisson likelihood has a gamma distribution. So the simplest symmetric Gaussian NMF method would be

$$\begin{aligned} a_{ij} &\sim \mathcal{N}(\hat{a}_{ij},1) \\ \hat{a}_{ij} &= \sum_c h_{ci} h_{cj} \\ h_{ci} &\sim \mathcal{HN}(0,\sigma). \end{aligned}$$

Which leads to the quality function

$$q(C,G) = -\frac{1}{2}\sum_{i,j\in V}(a_{ij}-\hat{a}_{ij})^2 - \frac{1}{2\sigma^2}\sum_{c\in C}\sum_{i\in V} h_{ci}^2 + |V|^2 \log\sqrt{2\pi} + |C||V|\log\sqrt{\pi\sigma^2/2}.$$

This is a regularized variant of symmetric NMF discussed previously.

Such a model implicitly assumes a fixed number of clusters; and the corresponding quality function will not be local if the number of clusters is not fixed. Intuitively, this happens because the model has to "pay" the normalizing constant of the prior distribution for each $h_{ci}$, the number of which is proportional to the number of clusters.

The method of Psorakis *et al.* also stems from a probabilistic model. They use a Poisson likelihood and a half-normal prior. Note that these are not conjugate. For finding the maximum likelihood solution conjugacy is not important. Using a conjugate prior becomes important only when doing variational Bayesian inference or Gibbs sampling [23].

To determine the number of clusters, Psorakis *et al.* put a gamma hyperprior on the inverse variance $\beta$. This allows a sharply peaked distribution on $w_c$ and $h_c$ when the support of a cluster is empty. The model is

$$\begin{aligned} a_{ij} &\sim \text{Poisson}(\hat{a}_{ij}) \\ \hat{a}_{ij} &= \sum_c h_{ci} w_{cj} \\ h_{ci} &\sim \mathcal{HN}(0,1/\sqrt{\beta_c}) \\ w_{ci} &\sim \mathcal{HN}(0,1/\sqrt{\beta_c}) \\ \beta_c &\sim \text{Gamma}(a,b). \end{aligned}$$

As shown in Sec. IV D, the corresponding quality function is not local. The problems stem from the priors on $W$, $H$, and $\beta$, which depend on the number of nodes and clusters. We will next try to find a different prior that is local.

### A. A local prior

To get a local quality function from a probabilistic model, that does not assume a fixed number of clusters, we clearly need a different prior. The approach we take will be to construct an additive quality function, which is local by Theorem 7.

First assume as above that the likelihoods of the edges are independent and depending on the product of membership degrees, that is, $P(A|C,V) = \prod_{ij} P(a_{ij}|\hat{a}_{ij})$. This fits nicely into the fourth term, $q_{\text{edge}}$, of an additive quality function.

Without loss of generality we can split the prior into two parts. First, the support of each cluster is determined, and based on this support the membership coefficients are chosen. If we define $S = \{\text{supp}(c)|c \in C\}$, then this means that

$$P(C|V) = P(C|V,S)P(S|V).$$

Just like $C$, $S$ should be seen as a multiset, since multiple clusters can have the same support. A reasonable choice for the first term $P(C|V,S)$ is to assume that the clusters are independent, and that the membership coefficients inside each cluster are also independent, so

$$C = \{C_s \mid s \in S\}$$

$$P(C_s|V,s) = \prod_{c\in C}\left(\prod_{i\in s} P(h_{ci}) \prod_{i\in V\setminus s} \delta(h_{ci},0)\right),$$

where $\delta$ is the Kronecker delta, which forces $h_{ci}$ to be zero for nodes not in $s$. The logarithm of $P(C|V,S)$ is a sum of terms that depend only on a single cluster, so it can be encoded in the $q_{\text{clus}}$ term of an additive quality function.

Now consider $P(S|V)$. If we know nothing about the nodes, then the two simplest aspects of $S$ we can look at are (1) how many clusters cover each node and (2) how many nodes are in each cluster. The only local choice for (1) is to take the number of clusters that cover node $i$, $n_i = \#\{s \in S \mid i \in s\}$, be independent and identically distributed according to some $f(n_i)$. While for (2), the probability of a cluster $s \in S$ must be independent of the other clusters. And since we have no information about the nodes, the only property of $s$ we can use is its size. This suggests a prior of the form

$$P(S|V) = \frac{1}{Z}\prod_{i\in V} f(n_i)\prod_{s\in S} g(|s|),$$

where $n_i = |\{s \in S \mid i \in s\}|$ is the number of clusters covering node $i$. The term $f(n_i)$ is local to each node and can be encoded in $q_{\text{node}}$. The term $g(|s|)$ is local to each cluster and can therefore be encoded in $q_{\text{clus}}$. The normalizing constant $Z$ depends only on $V$, and so it can be encoded in $q_{\text{graph}}$.

If we take $f(n_i) = \mathbf{1}[n_i = 1]$ and $g(|s|) = (|s| - 1)!$, then the prior on $S$ is exactly a Chinese restaurant process [24]. If we relax $f$, then we get a generalization where nodes can belong to multiple clusters. Another choice is $f(n_i) = \mathbf{1}[n_i = 1]$ and $g(|s|) = 1$. Then the prior on $S$ is the flat prior over partitions, which is commonly used for hard clustering.

Yet another choice is to put a Poisson prior on either the number of clusters per node or the number of nodes per cluster. That is, take $f(n_i) = \lambda^{n_i}/(n_i!)e^{-\lambda}$ for some constant $\lambda$ or do the same for $g$. This parameter allows the user to tune the number or size of clusters that are expected *a priori*.

To summarize, we obtain a local quality function of the form

$$\begin{aligned} q(G,C) = &\sum_{i\in V} \log f(|\{c \in C \mid i \in \text{supp}(c)\}|) \\ &+ \sum_{c\in C} \log g(|\text{supp}(c)|) + \sum_{c\in C}\sum_{i\in\text{supp}(c)} \log P(h_{ci}) \\ &+ \sum_{i,j\in V} \log P(a_{ij} \mid \hat{a}_{ij}) + \kappa, \end{aligned}$$

which has four independent parts: a score for a node being in a certain number of clusters, a score for the size of each cluster, a prior for each nonzero membership coefficient, and the likelihood of an edge $a_{ij}$ given the $\hat{a}_{ij}$.

The discrete nature of this quality function makes it harder to optimize. It is not clear if the multiplicative gradient algorithm that is commonly employed for NMF [25] can be adapted to deal with a prior on the support of clusters. On the other hand, it might become possible to use discrete optimization methods, such as the successful Louvain method used for modularity maximization.

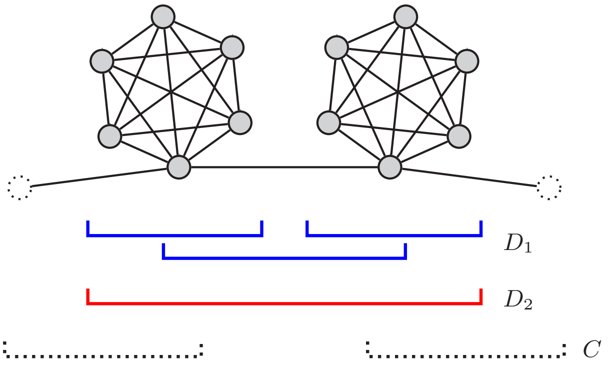

FIG. 5. (Color online) Two possible clusterings in a subgraph of a ring of cliques. In the first clustering ($D_1$, blue), the two cliques are in separate clusters, and there is a third cluster for the edge between them. In the second clustering ($D_2$, red) two cliques are put into a single cluster. A third possibility is to include the middle edge in a cluster together with one of the two cliques. A clustering of this entire subgraph will also include two clusters covering the connecting edges ($C$, dotted).

### B. Analysis of the quality functions on two types of graphs

We will now investigate the local quality function proposed in the previous section.

First consider the original resolution limit model [5], which consists of a ring of cliques. Two possible clusterings of a part of such a ring are illustrated in Fig. 5.

If a quality function is local, then we know that if $D_1 \uplus C$ is a better clustering than $D_2 \uplus C$ in this subgraph, then $D_1$ will also be better than $D_2$ as part of a larger graph. In other words, if the cliques are clustered correctly in a small ring, then this is true regardless of the number of cliques in the ring (unless a clustering with very large clusters is suddenly better).

We have performed experiments with the prior from the previous section to see what the optimal clustering will be in practice. We use a Poisson likelihood, a half normal prior on the supported membership coefficients (with precision $\beta = 1$), a Poisson prior on the number of clusters-per-node (with $\lambda = 1$), and a flat prior on the number of nodes per cluster. To find the optimal clustering we use a general purpose optimization method, combined with a search over the possible supports of the clusters.

Figure 6 shows that, as expected, the optimal solution is always to have one cluster per clique when using the local quality function. For comparison we also looked at the simpler nonlocal NMF method without a prior on the support. In that case the optimal solution depends strongly on the prior on membership coefficients $\beta$. If $\beta$ is small, then there is a penalty for every zero in the membership matrix and hence a penalty on the number of clusters that increases with the number of nodes. If $\beta$ is large enough, then the probability density $p(0) > 1$, and this penalty becomes a "bonus." In that case adding even an empty cluster would improve the quality, and the optimal clustering has an infinite number of clusters.

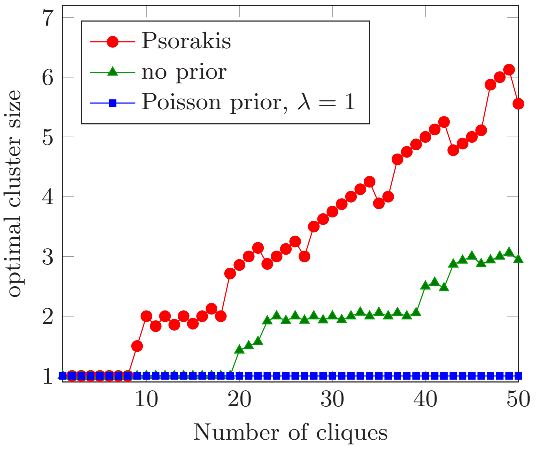

FIG. 6. (Color online) Optimal cluster size (average number of cliques per cluster) in a ring of $n$ 5-cliques, when varying the number $n$ of cliques.

The method of Psorakis *et al.* has the same resolution limit problem but to an even larger extent. To automatically determine the number of clusters, this method keeps the actual number of clusters fixed to a large upper bound, for which the authors take the number of nodes. This means that there are very many clusters which will be empty in the optimal solution. For these empty clusters, the parameter $\beta_c$ becomes very large. And as said in the previous paragraph, this results in a bonus for empty clusters. Hence the method will tend to maximize the number of empty clusters, which results in a few large clusters actually containing the nodes. For this experiment we used the prior $\beta_c \sim$ Gamma(5,2), as is also done in the code provided by Psorakis *et al.* Note that the jaggedness in the plot is due to the fact a ring of $n$ cliques cannot always be divided evenly into $m$ clusters of equal size. Between 24 and 50 cliques, the optimal number of clusters is always 8 or 9.

Figure 7 shows the influence of the parameter $\lambda$ of the Poisson prior that we put on the number of clusters per node. When $\lambda$ becomes smaller, it becomes *a priori* more likely for a node to be in only a single cluster or, in fact, to be in no cluster at all. It actually requires a quite strong prior to get two cliques to merge into one cluster, when using 5-cliques, we need $\lambda$ to be smaller than approximately $10^{-5}$.

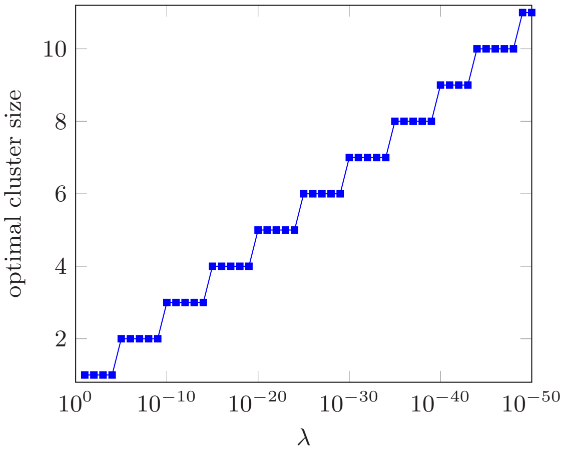

FIG. 7. (Color online) Optimal cluster size (average number of cliques per cluster) in a ring of 5-cliques, when varying the $\lambda$ parameter of the Poisson prior on the number of clusters per node. The number of cliques in the ring does not matter because of locality.

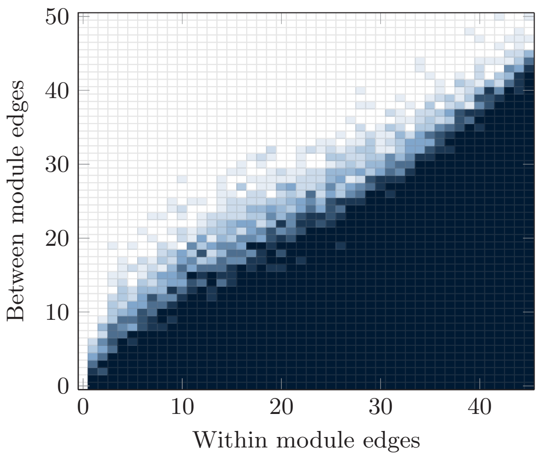

FIG. 8. (Color online) Varying the number of within and between module edges. The modules each have 10 nodes. A Poisson prior on the number of clusters per node ($\lambda = 1$) was used. We consider two possible clusterings: (a) A solution with three clusters, two clusters for the two modules and one cluster for the between module edges. And (b) the solution with a single cluster containing all nodes. The color in the plot indicates which clustering has a higher quality. In the dark region, the clustering (a) with three clusters is better. In the light region, the solution (b) with a single cluster is better. Results are the average over 10 random graphs with the given number of edges.

A ring of cliques is not a realistic model of real-world graphs, since on most graphs the clustering is not as clear-cut as it is there. The clustering problem can be made harder by removing edges inside the cliques, which are then no longer cliques, and better called modules, or by adding more edges between the modules.

We consider such a generalization, where there are two modules connected by zero or more edges. We then generated random modules and random between module edges. The two modules are either clustered together in one big cluster or separated. In Fig. 8 we show simulation results of such a more realistic situation. As we can see, as the number of between module edges increases, or the number of within module edges decreases, it becomes more likely to combine the two modules into one cluster. At the threshold between the two situations, the number of between module edges is roughly equal to the number of within module edges. This matches the notion of a *strong community*, which is defined by Radicchi *et al.* [26] as a set of nodes having more edges inside the cluster than edges leaving the cluster. A theoretical justification of these empirical results is beyond the scope of this work.

## VI. CONCLUSION

To our knowledge, this work is the first to investigate resolution-limit free and local NMF quality functions for graph clustering. We gave a characterization of a class of good (i.e., local) additive quality functions for graph clustering that provides a modular interpretation of NMF for graph clustering. The definitions of locality and of additive quality functions are general and can also be applied to other soft clustering methods. We proposed the class of local probabilistic NMF quality functions. The design and assessment of efficient algorithms for optimizing these quality functions remains to be investigated.

Results of this paper provide novel insights on NMF for hard clustering, on the resolution limit of Bayesian NMF for soft clustering, and on the beneficial role of a local prior in probabilistic formulations of NMF.

## ACKNOWLEDGMENTS

We thank the anonymous reviewer for their comments. This work has been partially funded by the Netherlands Organization for Scientific Research (NWO) within the NWO Projects No. 612.066.927 and No. 612.001.352.

[1] S. E. Schaeffer, Comput. Sci. Rev. **1**, 27 (2007).

[2] S. Fortunato, Phys. Rep. **486**, 75 (2010).

[3] M. Ackerman and S. Ben-David, in *NIPS*, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Curran Associates, Inc., Red Hook, NY, 2008), pp. 121–128.

[4] V. A. Traag, P. Van Dooren, and Y. E. Nesterov, Phys. Rev. E **84**, 016114 (2011).

[5] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).

[6] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).

[7] P. Paatero and U. Tapper, Environmetrics **5**, 111 (1994).

[8] D. D. Lee and H. S. Seung, Nature **401**, 788 (1999).

[9] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding, Data Min. Knowl. Discov. **22**, 493 (2011).

[10] T. Li and C. H. Q. Ding, in *Data Clustering: Algorithms and Applications* (Chapman & Hall/CRC, Boca Raton, Florida, 2013), pp. 149–176.

[11] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, Phys. Rev. E **83**, 066114 (2011).

[12] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).

[13] T. van Laarhoven and E. Marchiori, Phys. Rev. E **87**, 012812 (2013).

[14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech.: Theory Exp. (2008), P10008.

[15] M. Ackerman, S. Ben-David, and D. Loker, in *COLT 2010: The 23rd Conference on Learning Theory*, edited by A. T. Kalai and M. Mohri (Omnipress, New York, 2010), pp. 270–281.

[16] T. van Laarhoven and E. Marchiori, J. Mach. Learn. Res. **15**, 193 (2014).

[17] $C \uplus D$ denotes multiset sum, the multiplicity of $c$ in $C \uplus D$ is the sum of multiplicities of $c$ in $C$ and of $c$ in $D$.

[18] M. Catral, L. Han, M. Neumann, and R. J. Plemmons, Linear Algebra and its Applications **393**, 107 (2004).

[19] C. Ding, X. He, and H. D. Simon, *Proceedings of the SIAM International Conference on Data Mining (SDM'05)* (SIAM, Philadelphia, PA, 2005), pp. 606–610.

[20] C. Ding, T. Li, W. Peng, and H. Park, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'06 (ACM, New York, NY, 2006), pp. 126–135.

[21] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Nature **435**, 814 (2005).

[22] V. Y. F. Tan and C. Févotte, *Signal Processing with Adaptative Sparse Structured Representations (SPARS), Saint Malo, France* (2009).

[23] A. T. Cemgil, Comput. Intell. Neurosci. **2009**, 785152 (2009).

[24] J. Pitman, *Combinatorial Stochastic Processes* (Springer-Verlag, Berlin, 2006).

[25] D. D. Lee and H. S. Seung, in *14th Annual Conference on Neural Information Processing Systems*, NIPS 2000 (MIT Press, Cambridge, MA, 2000), pp. 556–562.

[26] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. USA **101**, 2658 (2004).