The New England Journal of Statistics in Data Science Volume 2, 368–379 (2024)
DOI: https://doi.org/10.51387/23-NEJSDS32

Constrained Community Detection in Social Networks

Weston D. VILES* and A. James O'MALLEY

Abstract

Community detection in networks is the process by which unusually well-connected sub-networks are identified–a central component of many applied network analyses. The paradigm of *modularity quality function optimization* stipulates a partition of the network's vertexes that maximizes the difference between the fraction of edges within communities and the corresponding expected fraction if edges were randomly allocated among all vertex pairs while conserving the degree distribution. The modularity quality function incorporates exclusively the network's topology and has been extensively studied whereas the integration of constraints or external information on community composition has largely remained unexplored. We define a greedy, recursive-backtracking search procedure to identify the constitution of high-quality network communities that satisfy the global constraint that each community be comprised of at least one vertex among a set of so-called *special vertexes* and apply our methodology to identifying health care communities (HCCs) within a network of hospitals such that each HCC consists of at least one hospital wherein at least a minimum number of cardiac defibrillator surgeries were performed. This restriction permits meaningful comparisons in cardiac care among the resulting health care communities by standardizing the distribution of cardiac care across the hospital network.

KEYWORDS AND PHRASES: Modularity, Discrete and constrained optimization, Patient-sharing network, Health care network communities.

1. INTRODUCTION

Networks are collections of interconnected entities, e.g. social networks of communicating actors, ecological networks of flora and fauna commensalism, and computer networks.[31, 26] Network-graphs, or simply graphs, are mathematical objects consisting of a vertex set, e.g., one vertex per network entity, and an edge set, e.g., a set of pairs of vertexes involved in a network connection, that represent the arrangements of pairwise relationships in the network.[9]

Methodology developed in the fields of social networks, network science, and graph theory provides for the analysis of relational data generated from a variety of measurements across scientific disciplines.[22, 8] Community detection is the process of identifying exceptionally dense subnetworks of mutually well-connected network entities, known as communities, that often have functional meaning in the network.[12] Notable approaches to community detection include the clique percolation method [24], spectral partitioning [5], degree-corrected stochastic block models [18], modularity optimization [21], and multi-slice network community detection [20]. These approaches are designed for the unsupervised partitioning of the vertex set of a graph into unusually cohesive subsets of vertexes, and with varied applications in sociology [33], computer architecture [13], and biology [2], illustrate that myriad methodology in community detection procedures are applied broadly in scientific research.

Community detection procedures commonly integrate network connectivity exclusively, without regard for other quantities of interest, e.g., auxiliary measurements on vertexes.[10]. We refer the reader to recent expositions on the state of the science of community detection in networks including [16, 30, 34, 27] that detail existing methods with respect to certain applications. It is essential to note that constraints are not commonly imposed on the community structure in networks by existing methods.

We are motivated to partition a nation-wide network of hospitals into subnetworks of hospitals that (i) exhibit a high level of within-group patient sharing as quantified by the network modularity quality function and (ii) consist of at least one hospital that has hosted a minimum number of *implantable cardioverter defibrillator* (ICD) surgeries to define as *health care communities* (HCCs) that provide a comparable level of cardiac care. We utilize data acquired from health insurance claims made to the Medicare national social insurance program during the period 2006-2011 in addition to the quantity of ICD surgeries at the major cardiovascular referral centers known as *cardiac care facilities* (CCFs). The preeminent work in this domain is the *Dartmouth Atlas* in which the then Center for the Evaluative Clinical Sciences, Dartmouth Medical School [28] assigned hospitals to one of 306 *health referral regions* (HRRs) representing markets for tertiary medical care. The significant contributions made by the Dartmouth Atlas to health services research motivates our work but our methodology, which is based on network-graph topology, is a departure from the HRRs

*Corresponding author.

strict adherence to geographic proximity that continues to be pursued by some efforts to modernize the designations. [29] Related work in defining regions according to network topology for the purposes of measuring health care variation across regions are nonetheless fundamentally based on geography. [14, 17, 15] Note that geographic proximity could be an alternative or additional constraint to our network-based method but is not necessary. In particular, if the analysis health care services provided by telemedicine or remote monitoring is desired then geographic information may be ignored intentionally in the discovery of network communities.

We develop in the following a recursive backtracking procedure for greedy search of high modularity communities that are each constrained by the requirement to contain at least one of the so-called *special vertexes*, which in the present application are the network-graph representatives of the hospital network cardiac care facilities at which a minimum number of ICD surgeries were performed and apply the method to approximate the optimal health care community assignment for each hospital in the network.

The inclusion of constraints in community detection, particularly a constraint of the nature under present consideration, is pertinent across many domains and applications. The community structure in any social network consisting of entities possessing distinct classes or attributes, e.g., standard entities versus noteworthy entities, may be more accurately characterized by our methodology. For example, a researcher seeking to partition the individuals of social communication networks that are the result of correspondences among senior and junior members of an organization, e.g., the Enron corpus [19], may enforce through our methodology the constraint that each community of organization members in the network consist of at least one senior member. (A reason for doing this is that the researcher knows how the organization structures its workforce but does not know which senior employee is grouped with which junior employees). This example is retrospective in that latent groups are being rediscovered by the researcher. The same example could apply in a prospective form. For example, given a network of professional relationships among its employees of different ranks, a company could use the algorithm to form optimal groups with at least one senior employee per group. Analogous examples might also arise in peer mentoring situations – a teacher in a classroom might use a friendship network among the students to form workgroups such that each group consists of at least one student who is performing very well (on course for the top possible grade) who can help the other students understand the material. On the other hand, a researcher investigating the community structure in a trophic network [11] may desire a partition of species in which each community contains at least one member of the Canidae taxonomic family. We emphasize that applications of our methodology abound in networks consisting of heterogeneous entities.

2. BACKGROUND

We represent the nation-wide hospital network with the weighted, undirected graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ which designates one vertex $v \in \mathbf{V}$ for each hospital and a positively weighted edge $\{u, v, w_{uv}\} \in \mathbf{E}$ for each pair of vertexes $\{u, v\} \in \mathbf{V}^2$ that represents interacting entities in the network, e.g., patient-sharing hospitals, where $\mathbf{V}^2 = \mathbf{V} \times \mathbf{V}$ is the set of vertex pairs. Note that if $w_{uv} = 0$ then $\{u, v, 0\} \notin \mathbf{E}$. The weight w_{uv} of edge $\{v, u, w_{uv}\}$ reflects the quantity of shared patient visits recorded in Medicare claims data between the hospitals represented by vertexes $u, v \in \mathbf{V}$.

Suppose that the vertex subset $\mathbf{V}' \subseteq \mathbf{V}$ contains the vertexes that are called *special vertexes* and represent network entities of a distinguishing nature. In the present scenario, $\mathbf{V}'$ consists of the vertexes that represent the cardiac care facilities at which at least τ ICD surgeries, for some $\tau \geq 0$,[1] were performed and define $\mathbf{V}' = \mathbf{V}'_\tau$ accordingly. The information contained in the labeling of these special vertexes in $\mathbf{V}'_\tau$ is information not encoded in the graph topology induced by the edge set $\mathbf{E}$ itself and, therefore, its integration into a standard community detection procedure must be made explicitly.

A useful mathematical representation of the weighted, undirected network-graph $\mathcal{G}$ in a variety of approaches to network science applications is the non-negative, symmetric $p \times p$ matrix $\mathbf{W}$ in which $W_{ij} = w_{v_i v_j}$, for $v_i, v_j \in \mathbf{V}^2$, i.e. $i, j \in \mathbb{Z}$ such that $1 \leq i < j \leq p$ corresponding to some ordering of the vertexes $v_k \in \mathbf{V}$. The *network modularity* quality function

$$Q(\mathbf{s}|\mathbf{W}) = \frac{1}{2m} \sum_{i,j}^{p} \left(W_{ij} - \frac{d_i d_j}{2m} \right) 1\{s_i = s_j\}, \tag{2.1}$$

where $\mathbf{s} \in \text{Ł}_p = \{1, 2, \ldots, p\}^p$ is a community assignment vector consisting of at most p unique community membership labels, $d_j = \sum_{i=1}^{p} W_{ij}$ is the degree of vertex v_j, and $2m = \sum_{j=1}^{p} d_j$ is the total degree of the network-graph $\mathcal{G}$. Note that W_{ij} is the observed (true) weight of the edge connecting vertexes v_i and v_j whereas $E_{ij} = \frac{d_i d_j}{2m}$ is the expected weight of the edge connecting these vertexes under randomization via the configuration model. [3]

The process of unconstrained modularity optimization involves the identifying of a partition $\{\mathbf{V}_1, \ldots, \mathbf{V}_n\} \subseteq \mathbf{V}$ of size $n \leq p$ corresponding to the community assignment vector $\mathbf{s}$ such that $s_k = r$ if $v_k \in \mathbf{V}_r$, for some $r \in \{1, 2, \ldots, n\}$. For a partition of size $n \leq p$, the *Stirling number of the second kind* [1]

$$S(p, n) = \frac{1}{n!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (n-k)^p$$

counts the number of unique community assignments. Note that if the $p' = |\mathbf{V}'_\tau|$ special vertexes are partitioned into n

[1] The case that $\tau = 0$ corresponds to unconstrained community detection, in which $p = p'$, and we include it here as a special case.

vertex sets then the number of feasible community assignments is

$$C(p, p', n) = n^{p-p'} S(p', n), \tag{2.2}$$

with $n \leq p'$, which displays asymptotics similar to the number of unconstrained partitions. The special case when $n = p'$, for which $C(p, p', n) = n^{p-n} \in O(n^p)$, amounts to the unconstrained optimization of the modularity quality function over the set of non-special vertexes $v \in \mathbf{V} \setminus \mathbf{V}'$ with each of the special vertexes $v \in \mathbf{V}'$ each belonging to a unique community. This observation accordingly supports a bottom-up approach that is a part of our forthcoming procedure.

On the extreme end of the spectrum, if $\tau = \omega$, where ω equals the maximum number of ICD surgeries performed at any cardiac care facility in the network, then p' achieves its minimum tenable value. In particular, if the maximum number ω of ICD surgeries performed at any hospital in the network is unique then $p' = 1$. Moreover, if $\tau > \omega$ then no feasible solution to the constrained optimization problem exists. In general, p' is a non-increasing function of $\tau \geq 0$.

The community assignment vector that optimizes the modularity quality function

$$\mathbf{s}^{\tau}_{opt} = \underset{\mathbf{s} \in \text{Ł}_p}{\arg\max} \;\; Q(\mathbf{s}|\mathbf{W}), \tag{2.3}$$

where $\text{Ł}_p = \{1, 2, \ldots, p\}^p$, is the community assignment vector which, on average, labels similarly vertexes that are more well-connected than expected to the same community among the $n \leq p$ different communities.

Suppose that $R(\mathbf{s}|\mathbf{V}'_{\tau})$ is the boolean function that returns true if the community assignment vector $\mathbf{s}$ satisfies the desired property that each community contain at least one special vertex $v' \in \mathbf{V}'_{\tau}$ and define the *restricted* community assignment vector that optimizes the modularity quality function

$$\mathbf{s}^{\tau}_{R} = \underset{\mathbf{s} \in \text{Ł}_p}{\arg\max} \;\; \left\{ Q(\mathbf{s}|\mathbf{W}) : R(\mathbf{s}|\mathbf{V}'_{\tau}) \right\}, \tag{2.4}$$

where, in the present scenario,

$$R(\mathbf{s}|\mathbf{V}'_{\tau}) = \begin{cases} \text{True} & \text{if } \mathcal{U}(\mathbf{s}|\mathbf{V}'_{\tau}) = \mathcal{U}(\mathbf{s}|\mathbf{V}) \\ \text{False} & \text{otherwise,} \end{cases}$$

where $\mathcal{U}(\mathbf{s}|\mathbf{V}'_{\tau})$ is, for example, the set of *unique* community labels for those vertexes $v \in \mathbf{V}'_{\tau}$. Defined as such, $R(\mathbf{s}|\mathbf{V}'_{\tau})$ returns True when each community is constituted by at least one special vertex $v' \in \mathbf{V}'_{\tau}$.

We define a *health care community assignment* as the designation of hospitals to communities in a community assignment vector that optimizes the network modularity quality function $Q(\mathbf{s}|\mathbf{W})$ in Equation (2.1). The quantity $Q(\mathbf{s}|\mathbf{W})$ is proportional to the difference between the fraction of (weighted) edges within communities and the expected fraction if edges were randomly distributed according to the configuration model, i.e. edge randomization with degree distribution conserved, subject to the constraint that to each community belongs at least one special vertex, e.g. a vertex representing a cardiac care facility where at least $\tau \geq 0$ ICD implantations occurred. Specifically, we seek to maximize modularity while requiring that the number of cardiac care facilities per health care community is at least $\tau \geq 0$: a restriction we encode with Boolean variable $R(\mathbf{s}|\mathbf{V}'_{\tau})$ indicating the feasibility of a community assignment.

Existing work in this realm considers incorporating additional information in the form of individual entity labels and pairwise constraints, i.e., that two vertexes must be labeled similarly or differently. [7] A restriction of the type we consider here has, to our knowledge, remained unstudied. In the context of the hospital network, the HRRs defined previously associated local health care markets to the tertiary care facilities where the plurality of the residents were referred for major cardiac procedures. An HRR is a reflection of its regional health care market and, because necessarily within each is a hospital specialized in cardiac surgery, a comparison across regions is facilitated. In an effort to standardize cardiac care among health care subnetworks, we present an initial undertaking towards developing a paradigm of constrained community detection. In particular, because constraint satisfaction in optimization problems provides context, the health care communities identified by our constrained community detection approach have real-world utility.

The organization of this article is as follows. In Section 3, we mathematically formulate the constrained optimization problem. In Section 4, we define a greedy, recursively-backtracking procedure for identifying high-quality, constrained communities. Utilizing our procedure in Section 5, we estimate the health care communities of the nation-wide hospital network and illustrate in Section 6 our method on a network-graph of well-known community structure.

3. PROBLEM SPECIFICATION

Suppose that $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, with $p = |\mathbf{V}|$, is a network-graph and that $\mathbf{V}' \subset \mathbf{V}$ is a subset of vertexes. Moreover, suppose that $\mathbf{s} \in \{1, 2, \ldots, n\}^p$, for some $n \in \mathbb{Z}_+$, is a vector of labels such at each label is applied at least once. We define the collection of $\binom{p}{2}$ binary variables $x_{ij} = 1\{s_i = s_j\}$, for $i, j \in \{1, 2, \ldots, p\}$, along with the corresponding collection of edge values $b_{ij} = w_{ij} - d_i d_j / 2m$ and the vertex values $u_k = 1\{v_k \in \mathbf{V}'\}$, for $k \in \{1, 2, \ldots, p\}$. Let $\mathbf{B} = [b_{ij}]$.

The constrained optimization problem under consideration is described as follows.

$$\text{Maximize: } f(\mathbf{s}, \mathbf{B}) = \sum_{ij}^{p} b_{ij} x_{ij} \text{ over } \mathbf{s} \in \{1, 2, \ldots, n\}^p$$

$$\text{and } n \in \mathbb{Z}_+$$

$$
\begin{aligned}
\text{Subject to: } & x_{ij} = 1\{s_i = s_j\} \text{ for } i, j \in \{1, 2, \ldots, p\} \\
& \sum_{\{h : s_h = r\}} u_h \geq \eta, \;\text{ for } r \in \{1, 2, \ldots, n\} \\
& \eta \geq 0 \\
& u_h = 1\{v_h \in \mathbf{V}'\} \text{ for } h \in \{1, 2, \ldots, p\}.
\end{aligned}
$$

This discrete optimization problem is, in general, at least as difficult as the corresponding, unconstrained NP-hard decision problem which seeks an answer to whether or not a partition of the vertex set exists with a modularity quality function value of at least some minimum value. In particular, the boundary value $\eta = 0$ corresponds to the special case of an unconstrained optimization problem in the description above.

A greedy unconstrained method may readily agglomerate communities toward achieving a local maximum of the network modularity function $f(\mathbf{s}, \mathbf{B})$. On the other hand, satisfying the constraint requires the consideration of many permutations of such mergers and any recursive backtracking procedure for identifying high-quality, with respect to the network modularity function, has potentially intractably-many paths in the search space to traverse. Note that if the graph $\mathcal{G}$ happens to be unweighted then the adjacency matrix $\mathbf{W}$ is binary and, up to quantities involving a $2m$ denominator, Equations (4.1) and (4.2) involve integers on similar scales. Accordingly, the various sequences of community assignment label updates through the recursive backtracking procedure are reduced. In the application of our method toward the defining of health care communities in the nationwide social network of cardiac-related Medicare referrals between hospitals we make this binary transformation $\mathbf{A} = 1\{g(\mathbf{W}) > \zeta\}$, where $g : \mathbb{R}^{p \times p} \mapsto \mathbb{R}^{p \times p}$ is defined in Section 5.1, the indicator function $1\{P\} = 1$ if proposition P is true and $1\{P\} = 0$ otherwise, and $\zeta > 0$ is determined from the data, to facilitate tractable computations.

4. GREEDY, RECURSIVE-BACKTRACKING PROCEDURE

We begin this section by presenting the existing work on unconstrained network modularity quality function optimization upon which our generalized procedure, which seeks to maximize the network modularity quality function over the space of feasible community assignment vectors, is based. Subsequently we outline our method for constrained optimization of the network modularity quality function and provide pseudocode for pertinent procedures.

4.1 Louvain Method

The Louvain method [4] is a greedy modularity optimization procedure that proceeds in two fundamental and repeating steps: (i) the local optimization of the modularity quality function and (ii) the folding of communities into super-vertexes to create a super-graph in which the newly-created super-vertexes are representatives in the super-graph of the communities in the former graph and the newly-created edge weights between super-vertexes are the sums of the edge weights between pairs of communities in the original graph. This process is continued until the super-graph resulting from repeated local optimizations and folds no longer warrants a subsequent fold, at which point, the folding process is reversed to recover the vertex-level community assignment vector, see [4] for details.

4.2 Modifications to Base Louvain Procedures

We modify the local optimization procedure of the Louvain method to cycle through the vertex set $\mathbf{V}$ in turn, with the exception of vertexes $v \in \mathbf{U}$, and assigning its community label s_k to

$$s_k \leftarrow \underset{s_k \in \{1,\ldots,n\}}{\arg\max} \; Q(s_k | \mathbf{s}_{-k}, \mathbf{W}).$$

This process is continued until no community label has been modified in one complete pass through the vertex set. The modification to restrict the updating of vertex community assignment labels to vertexes $v \in \mathbf{V} \setminus \mathbf{U}$ permits, in the present context, the community labels of special vertexes $v' \in \mathbf{V}'_\tau$ to be updated while the community labels of non-special vertexes $v \in \mathbf{V} \setminus \mathbf{V}'_\tau$ are held constant and *vice versa*. Moreover, we modify the folding procedure to trace a subset of vertexes $\mathbf{U} \subseteq \mathbf{V}$ through to the super-vertexes which ultimately represent them as part of the community folding process. This permits, in the present context, super-vertexes to inherit the *special* designation.

4.3 Constraint Corrected Louvain Method

Perhaps the most straightforward approach toward constrained greedy optimization of the modularity quality function leads through the unconstrained greedy optimization procedure of the Louvain method. That is, one may consider first estimating the unconstrained community assignment vector and then, subsequently, modify the unconstrained solution to a feasible state. We provide pseudocode in Procedure 1 for this recursive-backtracking procedure that modifies an infeasible community assignment vector in an agglomerative approach, i.e. by merging communities, in a manner that least reduces modularity until a constraint-satisfying solution is achieved. Of course, given enough community mergers, this procedure is guaranteed to eventually halt.

Note that within Procedure 1, the `Correct` local function is simplified by the fact that the merging of two communities that are respectively constituted by vertexes with indices in $\mathbf{I}_1$ and $\mathbf{I}_2$ results in the modularity gain proportional to

$$\Delta_{comm} Q \propto \sum_{i_1 \in \mathbf{I}_1} \sum_{i_2 \in \mathbf{I}_2} W_{i_1 i_2} - \frac{1}{2m} \left(\sum_{i_1 \in \mathbf{I}_1} d_{i_1} \right) \left(\sum_{i_2 \in \mathbf{I}_2} d_{i_2} \right). \tag{4.1}$$

```
Procedure 1: Recur.
Input:
$\mathbf{W}$: a $p \times p$ symmetric, weighted adjacency matrix
$\mathbf{V}'_\tau$: a set of size $p'$ of special vertexes
$\mathbf{s}^{(0)}$: a length $p$ community assignment vector,
    $\mathbf{s}^{(0)} \in \{1, 2, \ldots, p\}^p$ // Default: $\mathbf{s}^{(0)} = (1, 2, \ldots, p)$
Output:
$\mathbf{s}^{(1)}$: a locally optimized length $p$ community assignment
    vector // the community assignment vector with
    the greatest modularity among the rows of $\mathbf{Z}$
    (see below)
// Local Functions:
// Translate: removes one a special vertex from a
    community with a surplus of special vertexes
    and places it into a community in violation in
    constraint in a manner that least reduces
    modularity
// Merge_Correct: merges two communities in the
    existing community labeling assignment in a
    manner that least reduces modularity and
    applies Correct (see below) to the result
// Correct: applies Procedure 2 with $\mathbf{U} = \mathbf{V}'_\tau$ and,
    subsequently, applies Procedure 2 to the
    one-time folded graph resulting from Procedure
    3.
// $R(\cdot|\mathbf{V}'_\tau)$: Boolean constraint-satisfaction
    function
// Global Variables:
// $\mathbf{H} \leftarrow \mathbb{Z}_+^{0\times p}$: storage of community assignment
    vectors discovered by the procedure
// $\mathbf{Z} \leftarrow \mathbb{Z}_+^{0\times p}$: storage of community assignment
    vectors discovered by the procedure and
    satisfy the contraint
// modmax: maximum modularity community
    assignment vector discovered by the procedure
    at the present iteration

if $R(\mathbf{s}^{(0)}|\mathbf{W})$ then
    // Row append $\mathbf{s}^{(0)}$ to $\mathbf{Z}$
    modmax $\leftarrow \max\{$modmax$, Q(\mathbf{s}^{(0)}|\mathbf{W})\}$
else
    // Row append $\mathbf{s}^{(0)}$ to $\mathbf{H}$
    $\mathbf{s}_{tr} \leftarrow$ Translate$(\mathbf{s}^{(0)}, \ldots)$
    if $Q(\mathbf{s}_{tr}|\mathbf{W}) >$ modmax and $\mathbf{s}_{tr} \notin \mathbf{H}$ then
        Recur$(\mathbf{W}, \mathbf{V}'_\tau, \mathbf{s}_{tr})$
    $\mathbf{s}_{mc} \leftarrow$ Merge_Correct$(\mathbf{s}^{(0)}, \ldots)$
    if $Q(\mathbf{s}_{mc}|\mathbf{W}) >$ modmax and $\mathbf{s}_{mc} \notin \mathbf{H}$ then
        Recur$(\mathbf{W}, \mathbf{V}'_\tau, \mathbf{s}_{mc})$
return $\mathbf{s}^{(1)}$
```

```
Procedure 2: Local Optimization.
Input:
$\mathbf{W}$: a $p \times p$ symmetric, weighted adjacency matrix
$\mathbf{U}$: the set of vertex indices to exclude in local
    optimization procedure // Default: $\mathbf{U} = \emptyset$
$\mathbf{s}^{(0)}$: a length $p$ community assignment vector,
    $\mathbf{s}^{(0)} \in \{1, 2, \ldots, p\}^p$ // Default: $\mathbf{s}^{(0)} = (1, 2, \ldots, p)$
Output:
$\mathbf{s}^{(1)}$: a locally optimized length $p$ community assignment
vector

Change $\leftarrow$ True
while Change do
    Change $\leftarrow$ False
    for $k \in \{1, 2, \ldots, p\} \setminus \mathbf{U}$ do
        $s_k^{(1)} \leftarrow \arg\max_{s_k \in \{1,\ldots,n\}} Q(s_k|\mathbf{s}_{-k}, \mathbf{W})$
        if $Q(s_k^{(1)}|\mathbf{s}_{-k}, \mathbf{W}) > Q(s_k|\mathbf{s}_{-k}, \mathbf{W})$ then
            $s_k \leftarrow s_k^{(1)}$
            Change $\leftarrow$ True
return s
```

More specifically, if vertex v_j currently belongs to the same community as vertexes with indices in $\mathbf{I}_0$ and vertex v_j is to be relabeled to belong to the same community as vertexes with indices in $\mathbf{I}_1$ then the change in modularity is proportional to

$$\Delta_{vert} Q \propto \left(W_{jj} - \frac{d_j^2}{2m}\right) + \left(\sum_{i_1 \in \mathbf{I}_1} W_{ji_1} - \sum_{i_0 \in \mathbf{I}_0} W_{ji_0}\right) - \frac{d_j}{2m}\left(\sum_{i_1 \in \mathbf{I}_1} d_{i_1} - \sum_{i_0 \in \mathbf{I}_0} d_{i_0}\right). \tag{4.2}$$

These two formulas provide, in general, substantial computational savings over direct $O(p^2)$ computation of the network modularity quality function in Equation (2.1), see Appendix A for derivations.

4.4 Modified Core Procedures

Many discrete optimization problems, including the present one, are computationally difficult and frequently intractable. That is, the size of the solution space is, in general, prohibitively large for a complete search and the objective function is, in general, not monotonic in a mathematically-useful manner. For these reasons, an exact solution to the modularity optimization problem is not often sought but, instead, a satisfactory solution that is encountered by an algorithmic procedure is frequently reported as is the case with, for example, the Louvain method described above.

Prior to proposing our approach to the constrained optimization problem described in Section 3, we present our modified versions of the greedy, local optimization (Procedure 2) and community folding (Procedure 3) procedures. In particular, we provide for the flexibility to exclude a subset of vertexes from the local optimization process in Procedure 2 so that, in the present context, the community

Procedure 3: Community Folding.

```
Input:
$\mathbf{W}^{(0)}$: a $p_0 \times p_0$ symmetric, weighted adjacency matrix
$\mathbf{U}^{(0)}$: the set of vertex indices to *track* in folding process
  $\mathbf{s}$: a length $p_0$ community assignment vector
    // $\mathbf{s} \in \{1, 2, \ldots, n\}^{p_0}$
Output:
$\mathbf{W}^{(1)}$: a $p_1 \times p_1$ symmetric, weighted adjacency matrix
    for super-graph // $p_1 = |\mathcal{U}(\mathbf{s})|$,  $\mathcal{U}(\mathbf{s})$ = unique
    elements in s
$\mathbf{U}^{(1)}$: the set of super-vertexes representing at least one
vertex $v \in \mathbf{U}^{(0)}$

$\mathbf{W}^{(1)} \leftarrow$ a $p_1 \times p_1$ matrix of zeros
for $u_1, u_2 \in \mathcal{U}(\mathbf{s})$ do
    $\mathbf{W}^{(1)}_{u_1,u_2} \leftarrow \sum_{\{i:s_i=u_1\}} \sum_{\{j:s_j=u_2\}} \mathbf{W}^{(0)}_{i,j}$
return $\mathbf{W}^{(1)}$, $\mathbf{U}^{(1)}$
```

labels corresponding to special vertexes $v' \in \mathbf{V}'_\tau$ may be held fixed while the community labels corresponding to *regular*, i.e. non-special, vertexes $v \in \mathbf{V} \setminus \mathbf{V}'_\tau$ are locally and greedily optimized. Moreover, in the process of folding, we provide the adapted Procedure 3 which traces a set of vertexes through the folding process and reports which super-vertexes, i.e. the vertexes resulting from folded communities, are representative of any vertexes from that original set. In the present context, this permits the tracing of special vertexes through the folding process and reporting on the status of the super-vertexes that result from the folding process. Pseudocode for Procedures 2 and 3 are provided in the following.

4.5 Initialization Generalizations

The default community assignment vector $\mathbf{s}^{(0)} = (1, 2, \ldots, p)$ in Procedure 2 is consistent with the initialization of the Louvain method and represents a bottom-up approach in the greedy optimization process. Experimentally, we have found that a top-down approach is often more effective in the context of the constraint that each community must consist of at least one special vertex.. We provide a summary of this procedure in the following pseudocode for Procedure 4.

4.6 CMOP

We finally present the pseudocode for the main Constrained Modularity Optimization Procedure (CMOP) in Procedure 5.

Among the three constrained, high-modularity community assignment vector $\mathbf{s}_{td}$, $\mathbf{s}_{mr}$, and $\mathbf{s}_{null}$, as computed in Algorithm 5, the vector corresponding to the greatest modularity value is ultimately returned.

Procedure 4: Initialization.

```
Input:
$\mathbf{W}$: a $p \times p$ symmetric, weighted adjacency matrix
$\mathbf{V}'_\tau$: special vertex set of size $p'$
top_down: Boolean, return the community assignment
vector computed by the Louvain method
merge_regular: Boolean, assign to the same community
regular (non-special) vertexes that belong to the same
connected component.
Output:
s: a community assignment vector

if top_down then
    s ← Louvain_Method($\mathbf{W}$)
else
    if merge_regular then
        // Assign in the community assignment
           vector s a unique community label to
           each $v' \in \mathbf{V}'_\tau$
        // Then label together all regular
           (non-special) vertexes $v \in \mathbf{V} \setminus \mathbf{V}'_\tau$ that
           belong to the same connected component
    else
        // Assign to each vertex $v \in \mathbf{V}$ a unique
           community label in the community
           assignment vector s
return s
```

Procedure 5: CMOP: Constrained Modularity Optimization Procedure.

```
Input:
$\mathbf{W}$: a $p \times p$ weighted, symmetric adjacency matrix
$\mathbf{V}'_\tau$: a set of size $p'$ consisting of *special vertexes*
Output:
$\hat{\mathbf{s}}^*_R$: the maximum modularity community assignment
vector discovered by the procedure that satisfies the
constraint that to each community belongs at least one
special vertex $v' \in \mathbf{V}'_\tau$

// Use Initialization in Procedure 4 to
   generate...
// $\mathbf{s}_{td}$: using the top_down flag
// $\mathbf{s}_{mr}$:: using the merge_regular flag
// $\mathbf{s}_{null}$: using neither flag
// Apply Recur in Procedure 1 to each $\mathbf{s}_{td}$, $\mathbf{s}_{mr}$,
   and $\mathbf{s}_{null}$
return  $\arg\max_{\mathbf{s} \in \{\mathbf{s}_{td}, \mathbf{s}_{mr}, \mathbf{s}_{null}\}} Q(\mathbf{s}|\mathbf{W})$
```

5. APPLICATION: DEFINITION OF HEALTH CARE COMMUNITIES

The nationwide hospital network we consider consists of $p = 4734$ hospitals, as depicted in Figure 1a, among which 1388 (29.3%) are CCFs hosting at least $\tau = 1$ ICD surgeries and correspond to special vertexes in $v' \in \mathbf{V}'$. We begin this analysis by processing the data.

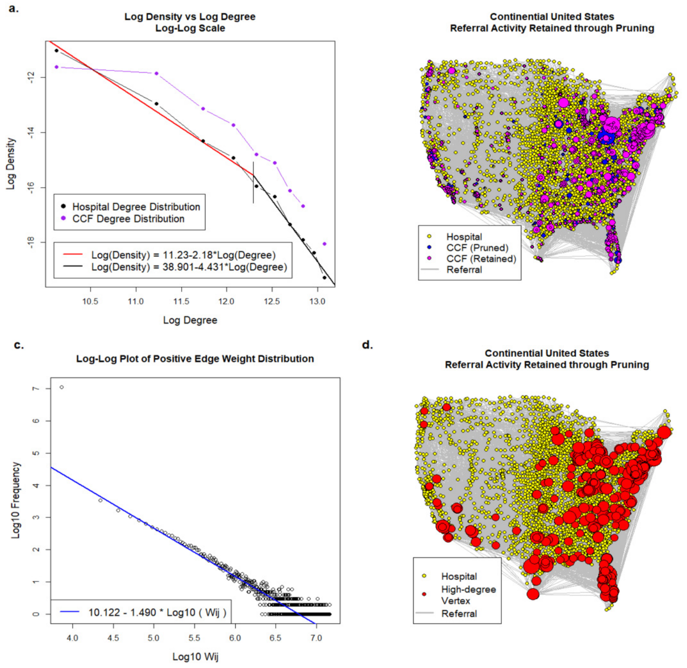

Figure 1: **Hospital Network-graph**: **(a)** plot of log histogram density vs log degree for all hospitals (black points) and only CCF hospitals (green points). Lines of best fits for the head (red line) and tail (blue line) of the log degree distribution of all hospitals. **(b, c)** The edges retained after pruning, see Appendix B. All hospitals are marked as small yellow points whereas, in (b) the vertexes with degrees in the tail of the degree distribution are marked with red points with radii proportional to their degree and in (c) the vertexes representing cardiac care facilities (CCFs) are marked in green points with radii proportional to the total number of ICD implantations that took place during the study period at that facility. **(d)** Using a database of all United States zip codes, along with their respective latitude and longitude coordinate, we indicated on the map the community of the nearest hospital to that zip code, with the understanding that a patient in that zip code is likely to travel to the nearest hospital in case of cardiac emergency.

5.1 Data Processing

The weighted degree distributions of all hospitals and the CCFs in Figure 1a illustrate the different orders of magnitude in their respective quantities. We compute the Pearson chi-square test statistic for independence

$$\chi^2_{ij} = \frac{(W_{ij} - E_{ij})^2}{E_{ij}},$$

where $E_{ij} = \frac{d_i d_j}{2m}$, for $i, j \in \{1, 2, \dots, p\}$, and note that in Figure 1c, a rather clear trend reflecting the exponential decay of the χ^2_{ij} quantities in the nationwide hospital network-graph. We set $\zeta = 266.4843$ as the 0.995 quantile threshold of the collection of $\binom{p}{2}$ chi-square quantities above and define the unweighted adjacency matrix $\mathbf{A}$ of the pruned nationwide hospital network as $\mathbf{A} = 1\{\boldsymbol{\chi}^2 > \zeta\}$, where $\boldsymbol{\chi}^2 \in \mathbb{R}^{p \times p}$ with elements χ^2_{ij} for $i, j \in \{1, 2, \dots, p\}$ and the indicator function $1\{\cdot\}$ is applied element-wise.

The argument $\mathbf{s}$ of the objective function $Q(\mathbf{s}|\mathbf{W})$ in Equation (2.1) is of principal concern, as opposed to the weighted network itself. Accordingly, we reduce the network dataset by pruning the weighted network edges that are relatively inconsequential in the evaluation of the modularity quality function in Equation (2.1). If, for instance, $W_{ij} - E_{ij} \approx 0$, i.e., the observed weight of the edge connecting vertexes v_i and v_j is approximately as expected under the configuration model, then whether vertexes v_i and v_j have the same or different community assignments results in a small marginal change in $Q(\mathbf{s}|\mathbf{W})$. Conversely, if the observed weight $W_{ij} \ll E_{ij}$ or $W_{ij} \gg E_{ij}$ then the community assignments of vertexes v_i and v_j has a relatively greater marginal impact on $Q(\mathbf{s}|\mathbf{W})$. Accordingly, a large value of $(W_{ij} - E_{ij})^2$ implies that the correspondence of the community assignments of vertexes v_i and v_j are important. The χ^2_{ij} value standardizes this value so that comparisons among the magnitudes of $(W_{ij} - E_{ij})^2$ across all vertex pairs $\{(v_i, v_j) : i, j \in \{1, 2, \dots, p\}\}$ are meaningful.

The United States map in Figure 1b indicates that the bulk of ICD surgeries occur in hospitals located in the Eastern States and have a higher frequency of shared patients with physicians associated with other hospitals compared to the entire population of hospitals in the network. Note that in Figure 1a the degree distributions of all hospitals and, separately, that of cardiac care facilities only are similar and, moreover, Figure 5c of the quantiles of respective degree distribution of regular (non-special) vertexes against special vertexes indicates that there is no significant difference between the two distributions, see Appendix A. The consequence of this fact is that the distribution of cardiac care facilities within communities is not expected to be an artifact of the network modularity quality function.

5.2 Unconstrained Hospital Network Communities

To estimate unconstrained communities in the hospital network we used the `cluster_louvain` function that belongs to the igraph R package. [25, 6] The resulting communities are mapped to local zip codes, and displayed in Figure 2, to reflect the hospital community into which a resident of each zip code would likely be entered upon a cardiac emergency.

The geographical proximity of unconstrained hospital communities, as quantified by the network modularity quality function and estimated via the Louvain method, is

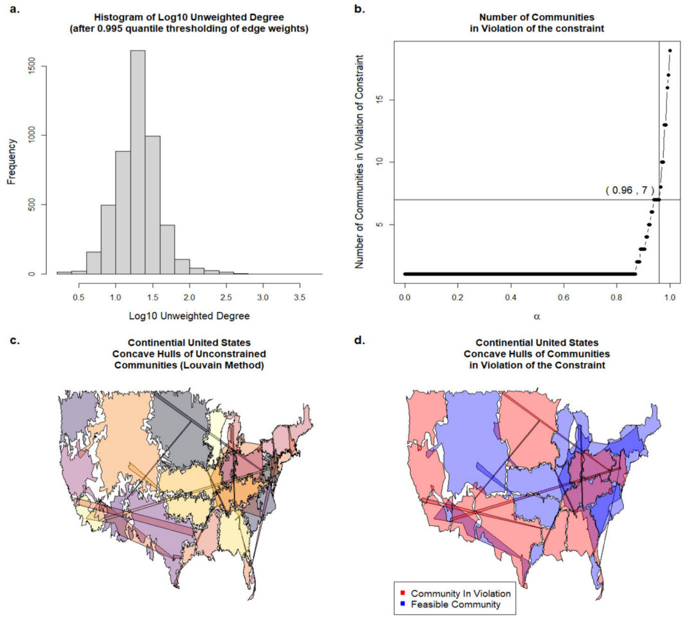

Figure 2: **Unconstrained Hospital Communities**: **(a)** Histogram of base ten logarithm of the pruned network degree distribution using 0.995 quantile threshold on original patient-sharing edge weights in the hospital network. **(b)** Selection of threshold $\tau_\alpha = 36$, where the subscript reflects the $\alpha = 0.265$ quantile of the distribution of ICD procedures performed at CCF hospitals, see Figure 5a, as the change point in the trend of number of communities in violation of the constraint, i.e., communities that do not possess a hospital where at least τ_α ICD implantations were performed, appears to change. Note that this trend and the corresponding location of interest depends on the graph topology. **(c)** The result of first employing the Louvain method on the nationwide network of hospital referrals for cardiac care and subsequently extending the communities to local zip codes via the $K = 1$ nearest neighbor criterion. The unconstrained communities exhibit some contiguity. However, because the communities are not constrained to be geographically contiguous, some of the communities have highly elongated shapes as they include at least one hospital far apart from the others. **(d)** Zip codes labeled as belonging to a community in violation of the constraint (red) or as belonging to a feasible community (blue).

relatively expected. In fact, if we consider the five most geographically-proximal hospitals to each zip code in the United States then we find that 72.1% of zip codes are closest to five hospitals all of the same unconstrained community and 24.5% are closest to five hospitals from two different unconstrained communities.

5.3 Defining Heath Care Communities

We now apply Procedure 1 to the nationwide unweighted hospital network-graph to identify the health care communities. As it turns out, the unconstrained community assignment vector $\mathbf{s}_{opt}$ results in a single network community that is in violation of the constraint that each community have at least one cardiac care facility (CCF) where *at least one* implantable cardioverter defibrillator (ICD) procedure was performed during the study period.

In order to more completely and accurately illustrate the utility of our methodology, we restrict the special vertex set to consist of vertexes corresponding to hospitals at which at least $\tau_{0.265} = 36$ ICD procedures were performed, where τ_α is the α quantile of the distribution of ICD procedure counts across all CCFs in the nationwide network, see Figure 2. This restriction tightens the constraint of the optimization problem and, practically, corresponds to the requirement that each health care community discovered by Procedure 1 has a greater lower-bound on the quantity of ICD procedures performed therein.

Our procedure considers each of the initial conditions of the recursive-backtracking Procedure 1 and ultimately selects $\mathbf{s}_{opt}$. Subsequently, two options are considered: (i) should a special vertex be relabeled according to the community in violation and then subsequently update the elements of the community assignment vector by applying an alternating sequence of Procedures 2 and 3 until the local optimum $\mathbf{s}_{hcc} = \mathbf{s}_R$ is identified via many computations of the forms in Equations (4.1) and (4.2). We consider the length $p = 4734$ vector $\mathbf{y}_0$ containing the number of ICD procedures taking place at each corresponding hospital, e.g. $y_k = 0$ if v_k does not represent a hospital where any ICD procedures were performed and otherwise $y_k > 0$. Define the vector $\mathbf{y}^{(\alpha)}$ with elements

$$y_k^{(\alpha)} = \begin{cases} y_k & \text{if } y_k > \tau_\alpha \\ 0 & \text{otherwise,} \end{cases}$$

for some $\tau_\alpha \geq 0$. Note that, as displayed in Figure 2b, with $\tau_{0.96} = \texttt{quantile}(\mathbf{y}_0, 0.96)$ and $\mathbf{V}'_\alpha = \{v_k : y_k > \tau_\alpha,\ \text{for } k \in \{1, 2, \ldots, p\}\}$ then, with this subset $\mathbf{V}'_\alpha \subset \mathbf{V}'$ of special vertexes, the number of communities in violation of the constraint has risen to seven. We subsequently executed Procedure 1 with $\mathbf{W} = \mathbf{A}$ and $\mathbf{V}' = \mathbf{V}'_\alpha$. Please see Appendix B for more details.

The role of τ_α in this application is that of τ in the preceding description of the general procedure. Note that this is the only parameter that modifies the contents of the auxiliary information contained in $\mathbf{V}' \subseteq \mathbf{V}$. Other parameters, including ζ, which control the topology of the unweighted network-graph $\mathbf{A}$, are pertinent to the general problem of modularity quality function optimization and are related to the base Louvain method.

5.4 Results Using a Subset of Special Vertices

By applying Procedure 1 to the unweighted adjacency matrix $\mathbf{A}$ and the reduced special vertex set $\mathbf{V}'_\alpha$, we approximate the maximizer community assignment vector $\mathbf{s}_{hcc} = \mathbf{s}_R$

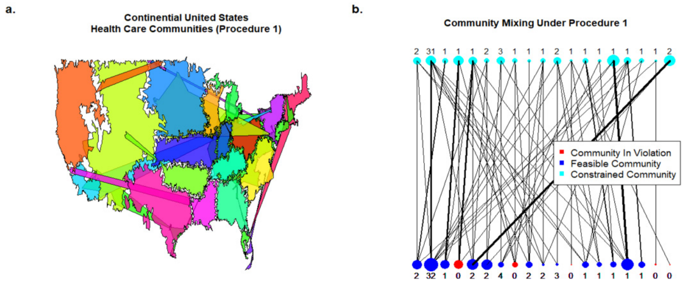

Figure 3: **Depiction of Health Care Communities** **(a)** The nationwide-network of hospitals is partitioned into a constraint-satisfying communities of hospitals such that each community contains at least one of the CCFs exceeding the defined ICD procedure threshold. Because the communities are not constrained to be geographically contiguous, some of the communities have highly elongated shapes as they include at least one hospital far apart from the others. **(b)** A bipartite graph reflecting the overlap between unconstrained communities (bottom row) and constrained communities (top row). The width of the line segment between vertexes, each of which is representing a community consisting of a number of vertexes that is proportional to its dot radius on the plot, reflects the relative overlap of the communities in each class. The number below or above each vertex equals the number of CCFs contained within each community.

of the network modularity quality function subject to the constraint that each community include at least one vertex $v' \in \mathbf{V}'_\alpha$ and note that it corresponds to a network modularity quality function value of $Q \approx 0.6814$, whereas the unconstrained modularity value of $Q \approx 0.6805$. It turns out that the marginal adjustments subsequent to the constraint-satisfaction elements of Procedure 1 may indeed identify a yet greater local maximum than the community assignment vector identified by the Louvain procedure. We do not consider this marginal improvement as worthwhile, however, due to the computational time requirements of such adjustments. We nevertheless note that, importantly, our strategy for identifying high-quality constrained communities is able to do so effectively. A plot of the health care communities identified by Procedure 1 is provided in Figure 3. We note that $\mathrm{nmi}(\mathbf{s}_{opt}, \mathbf{s}_R) \approx 0.9432$, implying that the relative mutual correspondence between the two vertex community assignments. Note that this procedure requires approximately 14.35 minutes to halt compared with the near instantaneous computation of unconstrained communities on the same network with the Louvain method.

It is imperative for the reader to recognize the Louvain method as agnostic to vertex attributes and, in particular, to the accumulation of vertexes with particular attributes in communities discovered by the method. The modularity quality function, which the Louvain method optimizes in a greedy manner, is exclusively a function of the edge set of a graph. Accordingly, modifying the designation of special vertexes, as in the present context, does not modify the composition of the resulting communities discovered by the method. Our proposed method, by contrast, is specifically devised to address the composition of discovered network communities, that is, the assignment of special vertexes to each community. It follows that, if the number of special vertices is diminished, as was demonstrated in the present application, then the constraint that each community contain at least one special vertex becomes more stringent and the community structure discovered by our procedure is modified. We have depicted, in the context described in this section, precisely how the community structure of the network is modified by including a strict constraint. Although the computational time required to compute the constraint-satisfying community structure exceeds that of the time required to compute an unconstrained community structure, the leveraging of additional information related to vertex attributes is worthwhile, in this context and in many other contexts as discussed in the Introduction section of the present article, to facilitate meaningful comparisons across communities that are standardized by the constraint.

6. SIMULATION ON ZACHARY KARATE CLUB

In the following, we consider each k-tuple, for $k \in \{2,3,4,5,6\}$ as the set of special vertexes $\mathbf{V}'$, among the set of $p = 34$ vertexes in the unweighted Zachary Karate Club social network-graph. [32] We apply Procedure 1 to determine, for each of these tuples of special vertexes, to record the modularity of each constraint-satisfying community assignment vector returned by Procedure 1, the initial position selected by the procedure, and the relative length of the computational time for the procedure to halt. The results of the simulation study are presented in Figure 4.

The quantity $C(p, p', n)$ in Equation (2.2) counts the number of feasible community assignments for a given $p = |\mathbf{V}|$, $p' = |\mathbf{V}'|$, and number of communities n. While this number reflects the number of community assignments necessary to brute-force check and, therefore, guarantee that the optimum defined in Equation (2.4) has been obtained, our greedy procedure is guided by the network topology and halts in many fewer iterations. Since the number of communities n is automatically chosen by the procedure, the computational time required for our procedure to halt is a function of (i) the number p' of special vertexes and (ii) the distribution of the special vertexes within the network-graph. For example, if n_{opt} is the number of unique labels, i.e. communities, represented in $\mathbf{s}_{opt}$ in Equation (2.3) and $p' < n_{opt}$ or if $p' \geq n_{opt}$ but the special vertexes are frequently labeled similarly in $\mathbf{s}_{opt}$ then some work is necessary to compute $\mathbf{s}_R$ of in Equation (2.4).

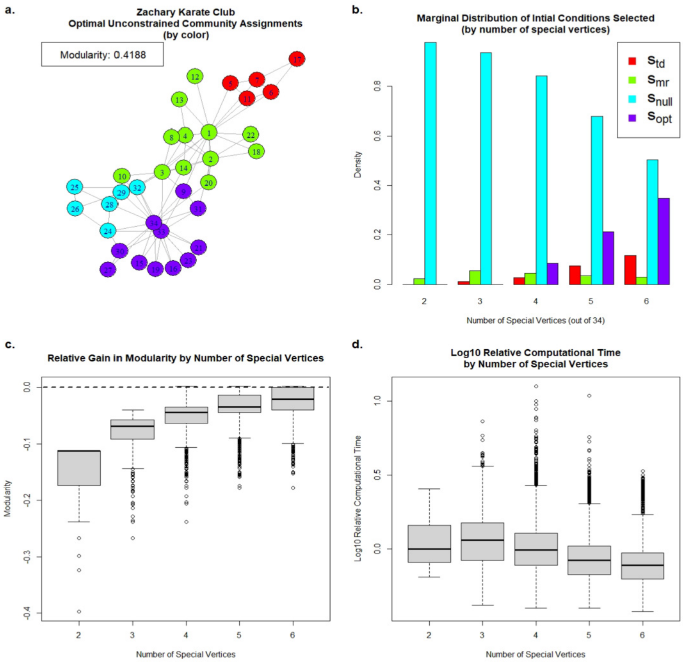

Figure 4: **Simulation: Zachary Karate Club (a)** The optimal community assignments in this social network of $p = 34$ vertexes. **(b)** The marginal relative frequency that each initial condition was selected by Procedure 5 over all $\binom{p}{d}$ choices of special vertexes, for $d = 2, 3, 4, 5, 6$. **(c)** The relative gain in modularity (modularity of constrained community assignment)/(modularity of unconstrained community assignment) - 1 by number of special vertexes in the network $d = 2, 3, 4, 5, 6$. **(d)** Logarithm base 10 of the relative (to the median computational time with two special vertexes in the network) amount of computational time for Procedure 5 to halt. We did not include instances when the unconstrained community assignment vector satisfied the constraint.

7. CONCLUSION

Our method for identifying network communities optimizes a quality function while adhering to constraints. The results in this paper establish the utility of penalized optimization in community detection. Our method is versatile and amenable to many types of constraints on the composition of communities. We note that our procedure is valid for any constraint which is an increasing function of the variable of interest, e.g., number of CCF hospitals belonging to a community, number of cardiac surgeons, quantity of cases involving improper medical procedures, etc. The key requirement of our constrained optimization procedure is that the merging of two communities must not be the basis for the resulting community to be in violation of the constraint. A constraint that imposes a maximum ICD volume is, for example, not of this type.

There exists a disconnect between network science and health services research due in part to the incongruence between mathematical elegance and real-world constraints. We have provided an illustration of the application of both a pure (unconstrained) method and one with constraints. We solved the practical problem of partitioning a network of hospitals with the constraint that the number of ICD surgical procedures that have taken place at at least one hospital belonging to each community exceeds some threshold. Though our method advances both the community detection and heath services literature, it is not complete from the perspective of a health care policy maker since many real-world constraints remain to be incorporated. Another type of constraint is, for example, given the geographic locations of hospitals, a requirement that communities not exceed a defined geographic maximum diameter or that they satisfy a geographic congruity constraint. On the other hand, one shouldn't necessarily seek to impose geographic contiguity constraints, for example, if one is interested in analyzing the effect of telemedicine or remote monitoring, for which the organization of health care does not need to conform as much to geography. This article is an initial exploration of a line of thinking that we anticipate will substantially advance the practical utility of community detection.

In terms of health policy, our future research involving an outcomes-based analysis of the communities discovered by our method, as constrained here by minimum ICD surgery volume and subsequently by other factors, will lead to enhanced acuity and potentially greater statistical power for studying variations in health care markets (based on patient referral patterns). Through standardizing the composition of the HCCs, our method provides the tools for such comparisons to be made meaningfully.

APPENDIX A. CHANGE IN MODULARITY

Let $\mathbf{I}_i, \mathbf{I}_2 \subseteq \mathbf{V}$ be two disjoint vertex subsets that are to be merged into $\mathbf{I} = \mathbf{I}_1 \cup \mathbf{I}_2 \subseteq \mathbf{V}$ and define $Q_{\mathbf{I}_1,\mathbf{I}_2}$ and $Q_{\mathbf{I}}$ respectively as the modularity of a vertex partition prior to and subsequent to the merger. The change in modularity

$$\begin{aligned}\Delta_{comm}Q &= Q_{\mathbf{I}} - Q_{\mathbf{I}_1,\mathbf{I}_2} \\ &= \frac{1}{2m}\sum_{i_1 \in \mathbf{I}_1}\left(\sum_{i_2 \in \mathbf{I}_2} W_{i_1 i_2} - \frac{1}{2m} d_{i_1} d_{i_2}\right)\end{aligned}$$

gives rise to Equation (4.1). On the other hand, suppose that $\mathbf{I}_0, \mathbf{I}_1 \subseteq \mathbf{V}$ and that the vertex $v_j \in \mathbf{I}_0$ is to be reassigned to $\mathbf{I}_1$ and define $\Delta Q^-_{\mathbf{I}_0}$ as the change in modularity resulting from removing vertex v_j from $\mathbf{I}_0$ and define $\Delta Q^+_{\mathbf{I}_1}$ as the change in modularity resulting from adding vertex v_j to $\mathbf{I}_1$. The change in modularity

$$\begin{aligned}\Delta_{vert}Q &= \Delta^+_{\mathbf{I}_1} - \Delta^-_{\mathbf{I}_0} \\ &\propto \sum_{i_1 \in \mathbf{I}_1}\left(W_{i_1 j} - \frac{d_{i_1} d_j}{2m}\right) - \sum_{\substack{i_0 \in \mathbf{I}_0 \\ i_0 \neq j}}\left(W_{i_0 j} - \frac{d_{i_0} d_j}{2m}\right)\end{aligned}$$

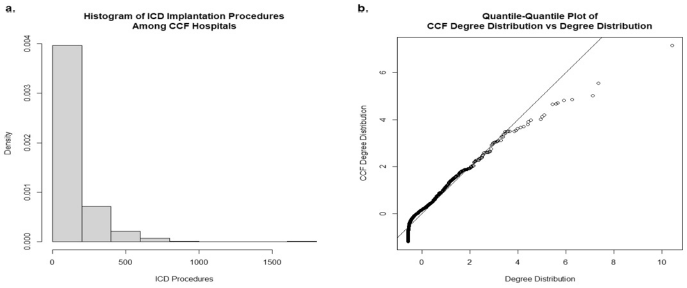

Figure 5: **Pruning the Special Vertex Set: (a.)** Histogram of number of ICD implantation procedures performed at CCF hospitals. **(b)** Quantiles of the degree distribution conditional on CCF status vs quantiles of the full vertex set degree distribution.

$$
\begin{aligned}
&\propto \sum_{i_1 \in \mathbf{I}_1} \left(W_{i_1 j} - \frac{d_{i_1} d_j}{2m} \right) - \sum_{i_0 \in \mathbf{I}_0} \left(W_{i_0 j} - \frac{d_{i_0} d_j}{2m} \right) \\
&+ \left(W_{jj} - \frac{d_j^2}{2m} \right),
\end{aligned}
$$

where the individual term $W_{jj} - d_j^2/2m$ must be added since $v_j \in \mathbf{I}_0$ at the outset, gives rise to Equation (4.2).

APPENDIX B. PRUNING

There are several hospitals where few ICD implantations occurred and, in order to isolate those where relatively many were performed, we first estimated the unconstrained network communities using the Louvain method. We subsequently counted the number of communities in violation of the constraint over a range of $\alpha \in [0, 1]$ and with the corresponding

$$
\tau_\alpha = \inf\{x \in \mathbb{Z}_+ : \hat{F}_{icd}(x) \geq \alpha\},
$$

that is, the $\alpha \cdot 100\%$ quantile of the distribution of ICD procedures across all hospitals where at least one such procedure was performed (Figure 5a.). We find that $\alpha = 0.96$, which corresponds to $\tau_{0.265} = 36$ (Figure 5b.)

FUNDING

Research for the paper was supported by NIH grants U01 AG046830 and P01 AG019783.

Accepted 5 March 2023

REFERENCES

[1] Abramowitz, M. and Stegun, I. A. (1964) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover. MR0415956

[2] Andrade, R. F. S., Rocha-Neto, I. C., Santos, L. B. L., de Santana, C. N., Diniz, M. V. C., Lobão, T. P., Góes-Neto, A., Pinho, S. T. R. and El-Hani, C. N. (2011). Detecting Network Communities: An Application to Phylogenetic Analysis. *PLoS Computational Biology* **7**. https://doi.org/10.1371/journal.pcbi.1001131. MR2821650

[3] Barabási, A. L. and Pósfai, M. (2016) *Network science*. Cambridge University Press, Cambridge.

[4] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10) 10008.

[5] Chung, F. R. K. (1997) *Spectral Graph Theory*. American Mathematical Society. MR1421568

[6] Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal* **Complex Systems** 1695.

[7] Eaton, E. and Mansbach, R. (2021). A Spin-Glass Model for Semi-Supervised Community Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* **26** 900–906.

[8] Estrada, E. (2011) *The Structure of Complex Networks: Theory and Applications*. Oxford University Press, Inc., USA.

[9] Estrada, E. and Knight, P. A. (2015) *A First Course in Network Theory*. Oxford University Press, United Kingdom.

[10] Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **486**(3-5) 75–174. https://doi.org/10.1016/j.physrep.2009.11.002. MR2580414

[11] Gauzens, B., Thébault, E., Lacroix, G. and Legendre, S. (2015). Trophic groups and modules: two levels of group detection in food webs. *Journal of the Royal Society Interface* **12**(106) 20141176.

[12] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**(12) 7821–7826. https://doi.org/10.1073/pnas.122653799. MR1908073

[13] Hendrickson, B. and Kolda, T. G. (2000). Graph Partitioning Models for Parallel Computing. *Parallel Computing* **26**(12) 1519–1534. https://doi.org/10.1016/S0167-8191(00)00048-X. MR1786938

[14] Hu, Y., Wang, F. and Xierali, I. M. (2018). Automated Delineation of Hospital Service Areas and Hospital Referral Regions by Modularity Optimization. *Health Services Research* **53**(1) 236–255.

[15] Hu, Y., Wang, F. and Xierali, I. M. (2018). Automated Delineation of Hospital Service Areas and Hospital Referral Regions by Modularity Optimization. *Health services research* **53** 236–255.

[16] Javed, M. A., Younis, M. S., Latif, S., Qadir, J. and Baig, A. (2018). Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications* **108** 87–111.

[17] Jia, P., Wang, F. and Xierali, I. M. (2020). Evaluating the effectiveness of the Hospital Referral Region (HRR) boundaries: a pilot study in Florida. *Annals of GIS* **26**(3) 251–260.

[18] Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* **83** 016107. https://doi.org/10.1103/PhysRevE.83.016107. MR2788206

[19] Leskovec, J., Lang, K. J., Dasgupta, A. and Mahoney, M. W. (2009). Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics* **6**(1) 29–123. MR2736090

[20] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328**(5980) 876–878. https://doi.org/10.1126/science.1184819. MR2662590

[21] Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the USA* **103**(23) 8577–8582.

[22] Newman, M. (2010) *Networks: An Introduction*. Oxford University Press, Inc., USA. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001. MR2676073

[23] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer,

D., Kaski, K., Kertész, J. and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* **104**(18) 7332–7336.

[24] Palla, G., Derényi, I., Farkas, I. and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435** 814–818.

[25] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[26] Schmid, J. S., Taubert, F., Wiegand, T., Sun, I.-F. and Huth, A. (2020). Network science applied to forest megaplots: tropical tree species coexist in small-world networks. *Scientific Reports* **10** 13198.

[27] Tandon, A., Albeshri, A., Thayananthan, V., Alhalabi, W., Radicchi, F. and Fortunato, S. (2021). Community detection in networks using graph embeddings. *Phys. Rev. E* **103** 022316.

[28] The Center for the Evaluative Clinical Sciences, Dartmouth Medical School (1996) *The Dartmouth atlas of health care*. American Hospital Publishing, Chicago.

[29] Wang, C. and Wang, F. (2022). GIS-automated delineation of hospital service areas in Florida: from Dartmouth method to network community detection methods. *Annals of GIS* **0**(0) 1–17.

[30] Wang, C., Tang, W., Sun, B., Fang, J. and Wang, Y. (2015). Review on community detection algorithms in social networks. In *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)* 551–555.

[31] Wasserman, S. and Faust, K. (1994) *Social network analysis: Methods and applications* **8**. Cambridge university press.

[32] Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research* 452–473.

[33] Zhang, Y., Friend, A. J., Traud, A. L., Porter, M. A., Fowler, J. H. and Mucha, P. J. (2008). Community Structure in Congressional Cosponsorship Networks. *Physica A-statistical Mechanics and Its Applications* **387** 1705–1712.

[34] (2020). A Guide for Choosing Community Detection Algorithms in Social Network Studies: The Question Alignment Approach. *American Journal Preventative Medicine* **59**(4) 597–605.

Weston D. Viles. Khoury College of Computer Science, Northeastern University, Roux Institute, Portland, ME USA.
E-mail address: w.viles@northeastern.edu

A. James O'Malley. Department of Biomedical Data Science and The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Dartmouth College, Hanover, NH USA.
E-mail address: james.omalley@dartmouth.edu