Finding Better Web Communities in Digraphs via Max-Flow Min-Cut

Chung Chan, Ali Al-Bashabsheh, Da Sun Handason Tam and Chao Zhao

***Abstract*—We consider the web community detection problem by providing a cost function that, not only penalizes external connections, but also rewards the internal ones. Our formulation addresses limitations of cut-clustering and extends web communities to digraphs. The formulation is parametric, resulting in a hierarchy of communities that is representable in linear storage and computable in a linear number of maxflow computations. Experiments on synthetic and real-world datasets show that the proposed method can find better web communities and more densest subgraphs than previous formulations. Simple examples also show that it can return different and more meaningful communities compared to other formulations that are based on graph conductance, map equation and modularity score.**

I. Introduction

In this work, we consider the problem of graph clustering, see e.g., [2], where communities of highly related nodes are identified based on their interconnections. This has significant applications in studying social and biological systems where related entities tend to interact more with each other compared to unrelated ones. Although the notion of community is intuitive, the challenge, however, is to find a precise mathematical definition that efficiently identifies the communities of interest.

The most widely accepted definition of a community, or graph cluster, is perhaps that of [3–5]. Roughly speaking, a community is a group of nodes that share stronger connections among themselves compared with the rest of the graph. Such a definition naturally inspires graph cut [6, 7], modularity [5, 8], random walk [9] based methods, to name a few, for identifying communities. These methods have to refine the original definition of a community for efficient computation and storage. In particular, web communities defined in [3, 6] are NP-hard to compute, and so a cut-clustering algorithm was proposed in [6] that finds approximate web communities by minimizing a weighted sum of the external connections and the size of the community. It was shown that the computation reduces to finding the mincuts of some modified graphs, and the set of all such communities, or cut-clusters, form a hierarchy, which can be represented efficiently by a dendrogram.

However, the hierarchical structure of cut-clusters rely on the graph being undirected. Furthermore, we find that cut-clustering often fails to return dense subgraphs and instead returns the less interconnected peripherals of the dense subgraphs. This behaviour is not surprising since, for a given size, a set of nodes having few external connections does not necessarily imply they have many internal connections. In other words, a dense subgraph may have many external connections with the rest of the graph compared with the external connections of a sparser subgraph. In this work, we primarily focus on improving the cut-clustering algorithm. Our proposed solution not only resolves the problem with no additional cost in computation and storage, but also extends to cluster digraphs or more general non-graphical networks with a submodular cost function.

This paper is organized as follows. In Section II, we introduce the notion of web communities for digraphs and the approximate solution by cut-clustering for undirected graphs. In Section III, we improve the formulation of cut-clustering and extend it to digraphs. In Sections IV and V, we explain the hierarchical structure of the communities and its polynomial-time computation. In Section VI, we give test results comparing the proposed algorithm to cut-clustering on two small example graphs. Detailed comparisons with cut-clustering and other algorithms can be found in [1, Appendices A and B], which include more experiments on both synthetic and real-world data sets. The proofs and detailed calculations of some of the examples are given in [1, Appendices C and D].

II. Preliminaries

We consider a digraph with non-negative real-valued edge weights and a finite set V of $|V|>1$ vertices represented by the adjacency matrix $\boldsymbol{A}:=[a_{ij}]\in\mathbb{R}_+^{|V|\times|V|}$. For convenience, we write

$$w(B,C):=\sum_{i\in B}\sum_{j\in C}a_{ij}\quad\text{for } B,C\subseteq V,$$

$w(i,C):=\sum_{j\in C}a_{ij}$ for $i\in V$, and $w(B,j):=\sum_{i\in B}a_{ij}$ for $j\in V$. The weight a_{ij} of edge $(i,j)\in V^2$ is taken to mean the influence of i on j, which needs not be equal to the influence a_{ji} of j on i for $i\neq j$. In the following definition, we generalizes the idea of web communities in [3, 6] from graphs to digraphs.

Definition 1 A *web community* is a non-empty subset $C\subseteq V$ of the vertices that satisfies

$$w(i,C\setminus\{i\})>w(V\setminus C,i),\ \forall i\in C,\tag{1}$$

i.e., the total external influence on any member of a web community is strictly smaller than the influence of that member on the remaining members of the community. □

C. Chan (corresponding author, email: chung.chan@cityu.edu.hk), D.S.H. Tam, and C. Zhao are with the Department of Computer Science, City University of Hong Kong. Their work is supported by a grant from the University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. 21203318).

A. Al-Bashabsheh is with the Big Data and Brain Computing (BDBC) center at Beihang University, Beijing, China.

978-1-5386-9291-2/19/$31.00 ©2019 IEEE

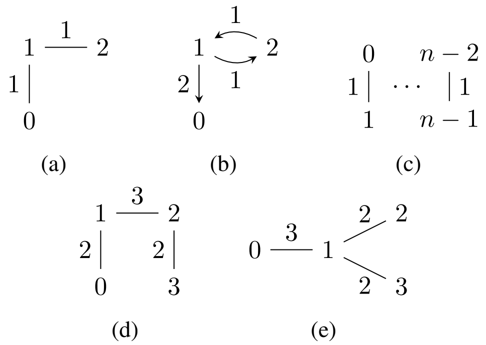


Fig. 1: The weighted graphs for Examples 1–4

Note that singleton sets $\{i\}$ for $i \in V$ are not web communities due to the strict inequality in (1). If $\boldsymbol{A}$ is symmetric, which corresponds to a graph, the above definition reduces to the original definition in [6]. The following example shows that our extension is non-trivial as the web communities can depend on the directions of the edges.

Example 1 Consider the undirected graph in Fig. 1a where $w(i,j) = 1$ for $\{i,j\} \in \{\{0,1\},\{1,2\}\}$, and the digraph in Fig. 1b where $w(1,0) = 2$ and $w(1,2) = w(2,1) = 1$. The two graphs differ only in the directions of the edges, i.e., they have the same values of $w(i,j) + w(j,i)$ for all $i,j \in V$. However, the two graphs have different web communities: $\{0,1,2\}$ for the undirected graph and $\{1,2\}$ for the digraph. $\{0,1,2\}$ is not a web community in the digraph in Fig. 1b since node 0 does not have any influence on other nodes, violating the condition (1). □

An issue with the above definition of communities is that there can be exponentially many of them as illustrated below.

Example 2 For the graph in Fig. 1c where the edges form a perfect matching, the web communities are all the non-empty subsets of the matched pairs. The number of communities is therefore $2^{|V|/2} - 1$, which is exponential in $|V|$. □

Since there can be exponentially many web communities, it is impractical to enumerate all of them for a large graph. Even for graphs of moderate sizes, returning many communities without any form of organization or measures of quality is not helpful, especially if the goal is to study a large graphical network by breaking it down into subnetworks.

In the case when the graph is undirected, [6] proposed the cut-clustering algorithm below:

1) For any parameter $\alpha \in \mathbb{R}$, add a new node s to the graph.
2) For each node $t \in V$, add an edge between s and t with weight α.
3) Construct a Gomory-Hu tree of the graph. Remove s from the tree and return the resulting connected components, i.e., a partition of V. If the Gomory-Hu tree is not unique, construct the tree that gives the finest partition.

The above procedure is called cut-clustering since, by the property of the Gomory-Hu tree, each of the resulting connected components corresponds to an s–t mincut of the graph augmented with s. We will refer to the non-singleton components as cut-clusters at threshold α. It was shown that α serves as a parameter that measures the quality of the returned clusters in terms of graph expansion [6, (3.3)]. However, the extension to digraphs is unclear since Gomory-Hu trees are defined for undirected graphs. Another limitation is that a cut-cluster may not be a web community as one of the nodes in a cut-cluster can fail to satisfy (1) [6, Lemma 3.1].

III. A better formulation

We propose the following more sophisticated definition of communities parameterized by different quality requirements. The communities will be shown to have a meaningful hierarchical structure that can be computed and represented in polynomial-time using maxflow algorithms.

Definition 2 Given $\beta \in [0,1]$, define for $\alpha \in \mathbb{R}$

$$\hat{f}(\alpha) := \min_{C \subseteq V : |C| \geq 1} f_\alpha(C) \quad \text{where} \tag{2a}$$

$$f_\alpha(C) := f(C) + \alpha \cdot |C| \tag{2b}$$

$$f(C) := (1-\beta) \cdot w(V \setminus C, C) - \beta \cdot w(C,C), \tag{2c}$$

where we have made the dependency on β implicit for notational simplicity. The set of communities is defined as

$$\mathcal{C} := \bigcup\nolimits_{\alpha \in \mathbb{R}} \mathcal{S}_\alpha \tag{3}$$

where $\mathcal{S}_\alpha$ is defined as the collection of $C \subseteq V$ such that $|C| > 1$, i.e., C is non-singleton, and

$$f_\alpha(C) = \hat{f}(\alpha) < \min_{B \subsetneq C : |B| \geq 1} f_\alpha(B), \tag{4}$$

i.e., C is an inclusion-wise minimal solution to (2a). For each community $C \in \mathcal{C}$, we define

$$\sigma(C) := \sup\{\alpha \in \mathbb{R} \mid C \in \mathcal{S}_\alpha\} \tag{5}$$

as a measure of the strength of the community. □

Subsequently, unless otherwise specified in the context, a community will refer to one according the definition above. In the definition, the parameters α and β allow for a tuning of the quality of the communities. Namely, the cost function $f(C)$ (2c) penalizes the *external influence* $w(V \setminus C, C)$ and rewards the *internal influence* $w(C,C)$. Since the entire set V trivially minimizes the external influence and maximizes the internal influence, we further penalize the size of C in the objective function $f_\alpha(C)$ (2b) with $\alpha \geq 0$ to obtain more compact communities. A simple connection to web communities is given by the following result, which provides a stronger guarantee on the quality of the communities compared to that of the cut-clusters in [6, Lemma 3.1].

Proposition 1 *Every $C \in \mathcal{S}_\alpha$ defined with* (4) *satisfies*

$$w(i,C) > w(V \setminus C, i) + (\alpha - \beta d_i) \quad \forall i \in C, \tag{6a}$$

$$w(V \setminus C, i) \geq w(i,C) - (\alpha - \beta d_i) \qquad \forall i \in V \setminus C, \tag{6b}$$

where $d_i := w(V \setminus \{i\}, i)$ is the in-degree of vertex i. □

Corollary 1 *A community $C \in \mathcal{C}$ defined in (3) is a web community if $\sigma(C) > \beta \max_{i \in C} d_i$.* □

Equation (6a) relates our communities and web communities by bounding the gap between the internal and external influences in terms of the community parameters. For instance, for $\beta = 0$, a community is always a web community for $\alpha \geq 0$, and, moreover, α provides a lower bound on the gap between the internal and external influences. Note that while the parameter β might appear to have an undesirable effect by diminishing the gap between the internal and external influences, and so leading to communities that are not web communities, it is one of the contributions we claim in this work. Indeed, as will be argued in a subsequent section, there can be meaningful communities that may or may not be web communities, and they cannot be identified unless $\beta > 0$.

PROOF (SKETCH) (6a) follows from that fact that $C \setminus \{i\}$ for any $i \in C$ and $C \in \mathcal{C}$ is a strictly suboptimal solution to (2), while (6b) follows from the fact that $C \cup \{i\}$ for any $i \in V \setminus C$ is a feasible but not necessarily optimal solution. The corollary follows from (6a) since $C \in \mathcal{S}_\alpha$ for some α arbitrarily close to $\sigma(C)$ by the definition (5) of $\sigma(C)$. ■

The following example illustrates the definition of the communities and its desired property.

Example 3 Consider Fig. 1c as in Example 2. Assuming $\beta = 0$, we have $f(C) = w(V \setminus C, C)$ by (2c). It is easy to see that $\hat{f}(0) = 0$ because the solution to (2a) when $\alpha = 0$ are the unions of the matched pairs $C_i := \{2i, 2i+1\}$ for $0 \leq i < n/2$. By (4), $\mathcal{S}_0$ is the set of matched pairs C_i's since they are inclusion-wise minimal solutions that are non-singleton. Similarly, it is straightforward to show that for

- $\alpha < 0$: $\hat{f}(\alpha) = \alpha|V|$ and $\mathcal{S}_\alpha = \{V\}$.
- $\alpha \in [0, 1)$: $\hat{f}(\alpha) = 2\alpha$ and $\mathcal{S}_\alpha = \{C_i \mid 0 \leq i < n/2\}$.
- $\alpha \geq 1$: $\hat{f}(\alpha) = 1 + \alpha$ and $\mathcal{S}_\alpha = \emptyset$.

By (3), the set $\mathcal{C}$ of communities consists of V and the matched pairs C_i's. By (5), $\sigma(V) = 0$ and $\sigma(C_i) = 1$. By Corollary 1, since $\sigma(C_i) > 0 = \beta \sum_{i \in C} d_i$, C_i's are web communities. □

In the above example, it can be seen that $\mathcal{C}$ captures the essential web communities for Fig. 1c, which form a meaningful hierarchy with respect to the quality measure σ.

IV. COMMUNITY HIERARCHY

The following result shows that, similar to hierarchical clustering methods, the set of communities forms a hierarchy and so can be represented by a dendrogram in linear storage.

Theorem 1 *For $\alpha_1 \geq \alpha_2 \geq 0$ and $C_i \in \mathcal{S}_{\alpha_i}$ for $i = 1, 2$, we have $C_1 \subseteq C_2$ or $C_1 \cap C_2 = \emptyset$. Furthermore, $C_1 \subsetneq C_2$ implies $\alpha_1 > \alpha_2$. Hence, the set $\mathcal{C}$ of communities can be represented by a dendrogram with the strength σ measuring the cophenetic similarity of the dendrogram.* □

The hierarchical structure can be observed from Example 3, where matched pairs C_i's are disjoint communities with the same strength $\sigma(C_i) = 1$. Since the trivial community V contains a matched pair, it has a strictly smaller strength $\sigma(V) = 0$ as expected. We remark that the proof of the theorem relies only on the submodularity (discussed in Section V) of f, and so the theorem extends to submodular functions that are not necessarily defined in terms of graph cut.

To understand the parameter α as a measure of similarity, we will strengthen Proposition 1 to give a more precise interpretation of α as a bound on the marginal change in the cost of a community.

Theorem 2 *Each $C \in \mathcal{S}_\alpha$ satisfies*

$$\alpha < \min_{B \subsetneq C : |B| \geq 1} \frac{f(B) - f(C)}{|C \setminus B|}, \tag{7a}$$

$$\alpha \geq \max_{A \subseteq V : A \supsetneq C} \frac{f(C) - f(A)}{|A \setminus C|}. \tag{7b}$$

(By convention, we set the r.h.s. of (7b) to $-\infty$ if $C = V$.) □

In other words, α is both a lower bound (7a) on the marginal increase in the cost f when the community shrinks, and an upper bound (7b) on the marginal decrease in the cost f when the community expands. The above theorem can be viewed as a generalization of Proposition 1 because (6a) and (6b) are the special cases when we further impose $B = C \setminus \{i\}$ for $i \in C$ in (7a), and $A = C \cup \{i\}$ for $i \in V \setminus C$ in (7b) respectively. As the following corollary shows, the result also ties back to the notion of graph expansion used to measure cluster quality.

Corollary 2 *Each $C \in \mathcal{S}_\alpha$ satisfies for $\beta = 0$*

$$\frac{w(V \setminus C, C)}{|V \setminus C|} \overset{\text{(i)}}{\leq} \alpha \overset{\text{(ii)}}{<} \min_{B \subsetneq C : |B| \geq 1} \frac{w(C \setminus B, B)}{|C \setminus B|} \tag{8a}$$

and, for $\beta = 1$,

$$\max_{A \subseteq V : A \supsetneq C} \frac{w(A, A) - w(C, C)}{|A \setminus C|} \overset{\text{(iii)}}{\leq} \alpha \overset{\text{(iv)}}{<} \frac{w(C, C)}{|C| - 1}. \tag{8b}$$

□

PROOF (ii) and (iii) follow from (7a) and (7b) with $\beta = 0$ and 1, respectively. (i) and (iv) follow from (7b) and (7a) by choosing $A = V$ with $\beta = 0$ and $B \subsetneq C : |B| = 1$ with $\beta = 1$, respectively. ■

V. COMPUTATION

In this section, we will derive a polynomial-time algorithm that returns the proposed community hierarchy $\mathcal{C}$ for any given $\beta \in [0, 1]$. The implementation is available at [1].

A. Divide-and-conquer

We first rewrite (2) as a two-step minimization

$$\hat{f}(\alpha) = \min_{t \in V} \hat{f}_t(\alpha), \text{ where} \tag{9a}$$

$$\hat{f}_t(\alpha) := \min_{C \subseteq V : t \in C} f_\alpha(C) \tag{9b}$$

and f_α is defined in (2b). The minimization in (9b) has a unique inclusion-wise minimum solution due to a well-known result in submodular function minimization (see [10]), and the fact that f (and thereby f_α) is submodular, i.e., $\forall C_1, C_2 \subseteq V$,

$$f(C_1) + f(C_2) \geq f(C_1 \cup C_2) + f(C_1 \cap C_2). \tag{10}$$

This can be seen by rewriting f, defined in (2c), using the identity $w(V \setminus C, C) + w(C, C) = \sum_{i \in C} d_i$ as

$$f(C) = w(V \setminus C, C) - \beta \sum_{i \in C} d_i, \tag{11}$$

which is submodular in C since the graph cut $w(V \setminus C, C)$ is submodular (and the sum $\sum_{i \in C} d_i$ is modular) [10].

Definition 3 For $\alpha \geq 0$, $\beta \in [0, 1]$, and $t \in V$, a community candidate, or candidate for short, $C_{\alpha,t}$ is defined as the inclusion-wise minimum solution to the minimization (9b). □

(Similar to communities, we make the dependency on β implicit for notational simplicity.) For a given $\beta \in [0, 1]$, let T^*_α be the set of optimal solutions t to the minimization (9a), then we have the following proposition.

Proposition 2 *The set of inclusion-wise minimal solutions to (2) is equal to the set of minimal candidates $C_{\alpha,t}$ s.t. $t \in T^*_\alpha$. In other words, the set of communities $\mathcal{S}_\alpha$ can be written as*

$$\text{minimal}\{C_{\alpha,t} \mid t \in T^*_\alpha\} \setminus \{\{i\} \mid i \in V\}, \tag{12}$$

PROOF Follows directly from Definitions 2 and 3. ■

Example 4 Consider the undirected graph in Fig. 1d with $V := \{0, 1, 2, 3\}$. It is straightforward to enumerate all the web communities (1) as

$$\{1, 2\}, \{0, 1, 2\}, \{1, 2, 3\}, \{0, 1, 2, 3\}. \tag{13}$$

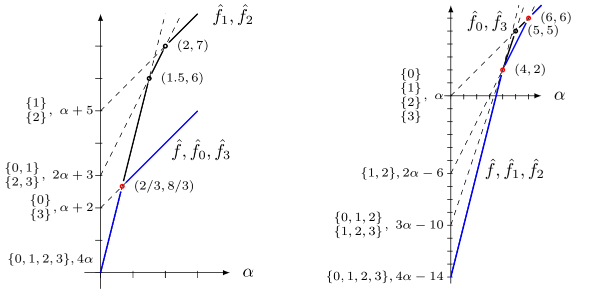
For $\beta = 0$, Fig. 2a shows the plots (against α) of the functions $\hat{f}(\alpha)$ (blue), $\hat{f}_t(\alpha)$ (solid), and $f_\alpha(C)$ (dashed black).[1] For $\beta = 1$, the same functions are shown in Fig. 2b. The candidates are determined by the solid lines, e.g., for $\beta = 0$, we have $C_{\alpha,t}$ as

$$\begin{array}{ccccl} t=0 & t=1 & t=2 & t=3 & \\ \{0\} & \{1\} & \{2\} & \{3\}, & \alpha \geq 2 \\ \{0\} & \{0,1\} & \{2,3\} & \{3\}, & \alpha \in [1.5, 2) \\ \{0\} & V & V & \{3\}, & \alpha \in [\frac{2}{3}, 1.5) \\ V & V & V & V, & \alpha < 2/3\ . \end{array} \tag{14}$$

Note that each entry in (14) is the inclusion-wise minimum solution to (10). Propositions 2 asserts that, for a given β, the set of communities is specified by the blue line as

$$\overset{\beta=0}{\begin{cases} \emptyset, & \alpha \geq 2/3 \\ \{0,1,2,3\}, & \alpha < 2/3 \end{cases}}, \quad \overset{\beta=1}{\begin{cases} \emptyset, & \alpha \geq 6 \\ \{1,2\}, & \alpha \in [4,6) \\ \{0,1,2,3\}, & \alpha < 4 \end{cases}}. \tag{15}$$

All the communities at $\beta = 0$ (trivial in this example) are web communities as dictated by Corollary 1. Note also that the web community $\{1, 2\}$ is captured at $\beta = 1$ but not $\beta = 0$, i.e., when the cost function (2c) rewards the internal influence. This situation may persist in general, and shows the benefit of choosing $\beta > 0$. Namely, if there is a dense subgraph that is moderately connected with the rest of the graph, then it is reasonable to favor such a subgraph over a sparse one, even if the sparse subgraph exhibits weaker connection to the rest of the graph.



(a) $\beta = 0$

(b) $\beta = 1$

Fig. 2: The plots of $\hat{f}_t(\alpha)$ (solid) against α for the undirected graph in Fig. 1d. The optimal $\hat{f}_t(\alpha)$ define $\hat{f}(\alpha)$ (solid blue).

In the example above, the non-trivial community at $\beta = 1$ is a web community, which is desirable. However, for the same reason mentioned in the example, a community at $\beta > 0$ may still be desirable even if it is not a web community. For instance, consider the graph in Fig. 1e. The set $\{0, 1\}$ is not a web community since node 1 is more connected with the rest of the graph than it is with node 0. Despite this, it is still desirable to consider this set as a community since it is the densest 2-subgraph. This set is indeed a community according to our formulation and can be returned at $\beta = 1$ and $\alpha \in [2, 6)$. □

B. Using maxflow algorithm

Since $f(C)$ is a submodular function, the minimization problem (2) can be solved using any submodular function minimization (SFM) algorithm. However, a generic SFM algorithm is computationally expensive. Below we reduce the problem of finding the candidate of $t \in V$ at any α, β to the mincut problem of the following augmented graph.

Definition 4 Let α, β be as in Definition 2, $t \in V$, and let $s \notin V$ be some additional node. The (α, β, t)-augmented digraph is the digraph on $V \cup \{s\}$ whose edge weight $w_{\alpha,\beta,t} : (V \cup \{s\})^2 \to \mathbb{R}_{\geq 0}$ is defined as

$$w_{\alpha\beta t}(i,j) := \begin{cases} w(i,j), & i \in V, j \in V \backslash \{t\} \\ w(i,j) + \beta d_i, & i \in V, j = t \\ \alpha, & i = s, j \in V \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Theorem 3 *The candidate $C_{\alpha,t}$ is the unique inclusion-wise minimum set C such that $(\{s\} \cup V \backslash C, C)$ is an s–t mincut of the (α, β, t)-augmented digraph.* □

[1] The figure shows $f_\alpha(C)$ only for the relevant subsets $C \subseteq V$ achieving (9b), i.e., the candidates. In other words, all suppressed lines lie above the curve of $\hat{f}_t$ for all $t \in V$.

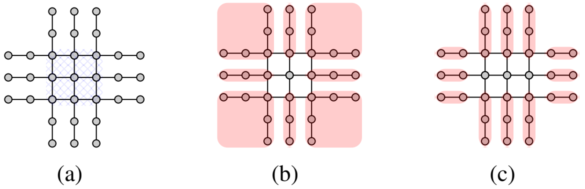

(a) (b) (c)

Fig. 3: An example graph that exhibits a grid-like center that connects to threads of nodes along the grid's perimeter. Communities and cut-clusters are highlighted using a crosshatch (blue) and no-pattern (red) marks, respectively. (a) Shows the returned community for $\alpha \in [\frac{31}{20}, \frac{27}{16})$ and $\beta = 0.7$, and (b) the cut-clusters for $\alpha \in [\frac{1}{8}, \frac{1}{2})$ and (c) the cut-clusters for $\alpha \in [\frac{1}{2}, 1)$. There are no other non-trivial (i.e., the entire set) solutions.

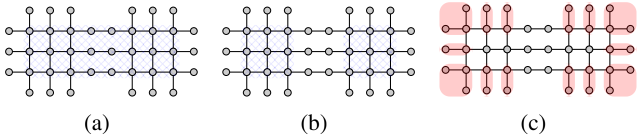

(a) (b) (c)

Fig. 4: Similar to Fig. 3. (a) Shows the returned community for $\alpha \in [\frac{37}{20}, \frac{58}{25})$ and $\beta = 0.85$, (b) the communities for $\alpha \in [\frac{58}{25}, \frac{75}{32})$ and $\beta = 0.85$, and (c) the cut-clusters for $\alpha \in [\frac{4}{41}, 1)$. There are no other, non-trivial, solutions.

PROOF (SKETCH) The theorem follows by showing that for any $C \subseteq V : t \in C$, the incut function of the augmented graph evaluated at C is equal to $f(C) + \alpha \cdot |C|$ plus some constant independent of C, i.e., (2) is equivalent to the s–t mincut problem on the augmented graph. ■

For a given α and β, the set $\mathcal{C}$ of communities can be computed as follows:

1) Run the maxflow algorithm n times, namely, for each $t \in V$ run the maxflow algorithm on the (α, β, t)-augmented graph to obtain $\hat{f}_t(\alpha)$ and its unique minimum solution $C_{\alpha,t}$.
2) Compute $\hat{f}(\alpha)$ as the minimum $\hat{f}_t(\alpha)$ over $t \in V$, and retain only the associated non-singleton $C_{\alpha,t}$ for $\mathcal{S}_\alpha$.

The second step can run in linear time and so the first step gives the overall complexity.

To compute the communities for all $\alpha \geq 0$, we can use the parametric maxflow algorithm of [11] to compute $\hat{f}_t(\alpha)$ for all $\alpha \geq 0$. The running time of the parametric algorithm is indeed the same as that of the push-relabel maxflow algorithm. Hence, we only need n maxflow computations on the augmented graphs to obtain $\hat{f}_t(\alpha)$ and $C_{\alpha,t}$ for all $\alpha \geq 0$ and $t \in V$. We can then compute $\hat{f}(\alpha)$ as the minimum $\hat{f}_t(\alpha)$ over $t \in V$, and retain the associated non-singleton $C_{\alpha,t}$ for $\mathcal{S}_\alpha$. This step can be computed in $O(n^2)$ because there are at most n line segments for each $\hat{f}_t(\alpha)$ and $\hat{f}(\alpha)$.

VI. Experimental results

To illustrate the benefits of the proposed communities over cut-clustering for larger networks than those in Fig. 1, we implemented and tested both algorithms on the graphs in Figs. 3 and 4. The graph in Fig. 3 contains a grid-like center, with peripheral-like chains of nodes attached to the grid's perimeter. The graph in Fig. 4 is similarly constructed, but with two grid-like centers. In both figures, it is desirable to differentiate the (denser) grid-like centers from the (sparser) peripherals.

The desired center is identified as a community in Fig. 3a, while cut-clustering only returns undesirable solutions, Figs. 3b and 3c. (See the figure's caption for details.) Note that the desired center is not a web community since each of its four corner nodes does not satisfy (1), which demonstrates the benefit of choosing $\beta > 0$. This also explains the undesirable behaviour of cut-clustering since a cut-cluster allows at most one of its members to violate (1) [6, Lemma 3.1]. (See [1, Appendix A] for more details.) Similar observations can be made using Fig. 4, where it is not hard to see that the same issues extend to larger, and other types of, graphs. More experimental results on synthetic and real-world data sets can be found in [1, Appendix B].

VII. Conclusion

We revisited the web communities detection problem. Our treatment applies to graphs and digraphs alike, and extends in a straightforward manner to submodular functions in general. Our theoretical derivations showed that the new formulation is efficiently computable, has a hierarchical structure, provides quality guarantees on the returned communities, and addresses some limitations of cut-clustering.

References

[1] C. Chan, A. Al-Bashabsheh, H. D. S. Tam, and C. Zhao, "Finding Better Web Communities in Digraphs via Max-Flow Min-Cut," GitHub repository: https://github.com/ccha23/Better-Web-Communities, Jan. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2545047

[2] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.

[3] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of web communities," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000, pp. 150–160.

[4] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[5] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066133, 2004.

[6] G. W. Flake, R. E. Tarjan, and K. Tsioutsiouliklis, "Graph clustering and minimum cut trees," *Internet Mathematics*, vol. 1, no. 4, pp. 385–408, 2004.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[9] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[10] S. Fujishige, *Submodular functions and optimization*, 2nd ed. Elsevier, 2005.

[11] G. Gallo, M. D. Grigoriadis, and R. E. Tarjan, "A fast parametric maximum flow algorithm and applications," *SIAM Journal on Computing*, vol. 18, no. 1, pp. 30–55, 1989.