Optimal Quasi-clique: Hardness, Equivalence with Densest-k-Subgraph, and Quasi-partitioned Community Mining

Aritra Konar[1], **Nicholas D. Sidiropoulos**[2]
[1]KU Leuven, Leuven, Belgium
[2]University of Virginia, Charlottesville, USA
aritra.konar@kuleuven.be, nikos@virginia.edu.

Abstract

Dense subgraph discovery (DSD) is a key primitive in graph mining that typically deals with extracting cliques and near-cliques. In this paper, we revisit the optimal quasi-clique (OQC) formulation for DSD and establish that it is NP–hard. In addition, we reveal the hitherto unknown property that OQC can be used to explore the entire spectrum of densest subgraphs of all distinct sizes by appropriately varying a single hyperparameter, thereby forging an intimate link with the classic densest-k-subgraph problem (DkS). We corroborate these findings on real-world graphs by applying the simple greedy algorithm for OQC with improved hyperparameter tuning, to quickly generate high-quality approximations of the size-density frontier. Our findings indicate that OQC not only extracts high quality (near)-cliques, but also large and loosely-connected subgraphs that exhibit well defined local community structure. The latter discovery is particularly intriguing, since OQC is not explicitly geared towards community detection.

Introduction

Dense subgraph detection (DSD) is a key primitive in graph mining that aims to extract highly interconnected subsets of vertices from a graph. Applications of the problem range from discovering regulatory motifs in genomic DNA, mining trending topics in social media, finding functional modules in gene co-expression networks, and communities in social networks – see (Cadena, Chen, and Vullikanti 2018; Lanciano et al. 2023) and references therein. In recent years, DSD has also found application in spotting fraudulent behavior in user-product graphs (Hooi et al. 2016) and financial transaction networks (Zhang et al. 2017; Li et al. 2020; Chen and Tsourakakis 2022).

Directly maximizing subgraph density (defined as the fraction of the maximum number of possible edges in a subgraph) admits trivial solutions such as a single edge. This motivates using alternative surrogates for density maximization. The classic Densest Subgraph (DSG) problem (Goldberg 1984) aims to extract a dense vertex subset that maximizes the average induced degree. DSG can be solved exactly in polynomial-time via maximum-flow (Goldberg 1984). In practice, a simple vertex peeling-based greedy approximation algorithm (Charikar 2000) is used, as it enjoys linear-time complexity and provides a 0.5-approximation guarantee for DSG. Recently, "multi-pass" generalizations of the greedy algorithm have been developed which exhibit superior performance (Boob et al. 2020; Chekuri, Quanrud, and Torres 2022). Another well-known formulation is the core decomposition (Seidman 1983), which is tantamount to maximizing the *minimum* induced degree - the resulting vertex subset is known as the maxcore, which can be obtained via a slight modification of the greedy peeling algorithm for DSG. These approaches suffer from an inherent limitation - there is no means of explicitly controlling the size of the extracted subgraphs. Hence, one cannot rule out the possibility that these extracted subgraphs will have low density. Unfortunately, such cases can occur on real-world graphs. For example, the peeling algorithm for DSG can output the entire graph as the solution (Tsourakakis et al. 2013). Meanwhile, empirical studies have revealed that the maxcores typically do not form a dense quasi-clique (Shin, Eliassi-Rad, and Faloutsos 2016).

If the density of DSG/maxcore proves to be unsatisfactory, the Densest-k-Subgraph (DkS) problem (Feige, Peleg, and Kortsarz 2001) can be employed - given a pre-specified size parameter k, extract the densest size-k vertex subset (i.e., the one which harbors the maximum number of induced edges). By solving the problem for various k, we obtain a collection of the densest subgraphs of distinct sizes, from which a solution of desired density can be selected. We designate the entire spectrum of such subgraphs (i.e., the densest of each distinct size) the *optimal size-density frontier*. Unfortunately, this extra flexibility comes at a price - DkS is NP–hard and is notoriously difficult to approximate in the worst-case (Manurangsi 2017). Notwithstanding this fact, practical algorithms which work well for this problem on real graphs include (Papailiopoulos et al. 2014; Konar and Sidiropoulos 2021). However, a limitation of these approaches is that they entail solving an optimization problem for each k, which can prove computationally expensive when generating candidate solutions of various sizes. An alternative is the recent Generalized Mean Densest Subgraph (GMDSG) framework (Veldt, Benson, and Kleinberg 2021), which employs a single parameter p for computing generalized means of degree sequences of a subgraph. By vary-

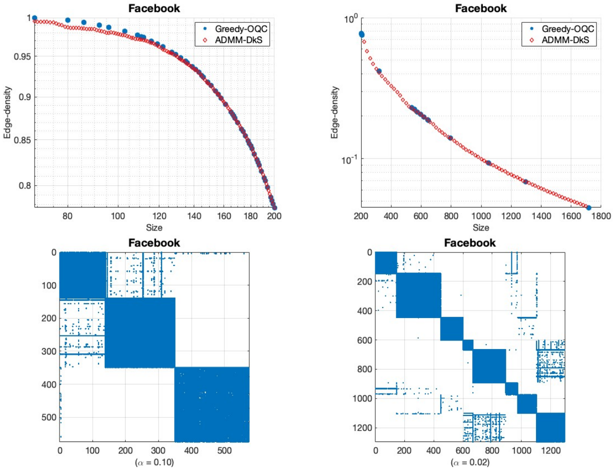

Figure 1: (Top panel): Size-density frontiers generated using greedyOQC (blue) and DkS (red) on the Facebook dataset. (Left): Subgraphs in the range spanned by $\alpha \in [0.33, 0.99]$ and (right): in the range $\alpha \in [0.01, 0.33)$. Denser subgraphs mined by greedyOQC correpsond to larger values of α. (Bottom panel): Visualizing local communities in loosely-knit subgraphs extracted by OQC via the block diagonal structure of their adjacency matrices. (Left): Size = 574, density = 0.21. (Right): Size = 1297, density = 0.07.

ing p, one can extract a family of dense subgraphs which obey different notions of density, with DSG and maxcore corresponding to the choices of $p = 1$ and $p = -\infty$, respectively. For $p \geq 1$, GMDSG can be solved optimally in polynomial-time via maximum-flow, and is also amenable to high-quality approximation via a generalized greedy peeling algorithm. While being a useful generalization of DSG, presently it is not known whether the solution of GMDSG (for a given p) corresponds to the densest subgraph of that size; i.e., whether it equals the solution of DkS (in terms of density) for a given k.

In this paper, our primary goal is to highlight an alternative means of mining subgraphs from the optimal size-density frontier as opposed to employing DkS. To this end, we revisit the optimal quasi-clique formulation (OQC) proposed in (Tsourakakis et al. 2013). Similar to GMDSG, the framework employs a single parameter α to quantify subgraph density; in particular how unexpected the density of a subgraph is w.r.t. to a random subgraph model. In (Tsourakakis et al. 2013), a greedy peeling algorithm was developed for OQC and tested using $\alpha = 1/3$ to demonstrate that it outperforms DSG on real-world graphs. However, the merits of such a parameter choice have not been formally investigated. In fact, the precise role played by α remains ill-understood. Loosely speaking, the OQC formulation (2) can be viewed as a "regularized" counterpart of DkS, with α serving as a trade-off parameter between subgraph size and density. Building on this intuition, we provide several important insights regarding the problem. Our contributions can be summarized as follows.

1. **Hardness:** We prove that OQC is NP–hard for undirected, unweighted graphs, thereby settling a longstanding conjecture regarding the complexity of the problem originally posed in (Tsourakakis et al. 2013).
2. **Equivalence with DkS:** We demonstrate that the densities of the maximizers of OQC obtained by continuous variation of the parameter α equal those of the maximizers of DkS obtained by variation of the discrete size parameter k in DkS. In other words, by varying their respective parameters, both formulations generate the optimal size-density frontier in a graph [1]. In order to establish our result, we prove the existence of sub-intervals of α where the maximizers of OQC are the densest subgraphs of a particular size-k. We remark that such an equivalence between non-convex, combinatorial problems is surprising, since unlike establishing equivalences between regularized and constrained variants of *continuous* problems, we cannot appeal to strong duality (Boyd and Vandenberghe 2004), or to penalty-based approaches (Bertsekas 2014).
3. **Quickly exploring the size-density frontier:** Since both DkS and OQC are difficult problems to solve exactly, in practice, there can be a difference in the quality of the subgraphs extracted by them. An implication of our results is that the greedy peeling algorithm for OQC (Tsourakakis et al. 2013) is a natural baseline for benchmarking the performance of DkS methods. In addition to its linear-time complexity, an attractive feature of this peeling method is that the peeling order does not depend on α. Hence, by running the method *once* to obtain the order, different values of α can be used in a post-processing step to select subgraphs of different densities and sizes. This is in stark contrast to methods for DkS, which have to be run for each distinct k. An illustrative example of the performance of greedy peeling and the convex relaxation algorithm (Konar and Sidiropoulos 2021) for DkS on the Facebook dataset (obtained from (**?**)) is provided in Figure 1. In the top panels, we display the size-density frontiers for OQC and DkS for two ranges of subgraph sizes. Notice how closely the curves match, with the peeling method exhibiting slightly better densities for subgraph sizes less than 110. Additionally, we noted that increasing α beyond $1/3$ generally improves the density performance; e.g., with $\alpha = 1/3$ we obtain a subgraph of size 200 with density 0.78 whereas for $\alpha = 0.99$ we obtain a clique of size 69. It can also be observed that the frontier generated by OQC is coarser compared to DkS - this is due to the nature of the peeling algorithm (see Experiments for a more detailed discussion).
4. **Large and sparse quasi-cliques can also be interesting:** DSD is mostly concerned with extracting cliques and near-cliques, which reside in the high density region of the optimal size-density frontier. Thus, the task of mining larger, less cohesive subgraphs is *apriori* not well motivated. Unexpectedly, it turns out that in real-world graphs, quasi-cliques with density as low as 7% can exhibit well-defined, non-trivial *local community structure*.

[1] Such a property is currently not known for GMDSG.

This is illustrated in the bottom panel of Figure 1 - as α is decreased, the peeling algorithm "zooms-out" to reveal sparsely connected subgraphs which harbor loosely interconnected communities of smaller dense subgraphs. This discovery is surprising, since the objective function of OQC *does not explicitly promote community structure*. Our results on other real-world graph reveals a similar pattern (see Experiments).

The Optimal Quasi-clique Problem

Consider an undirected, unweighted graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ on n vertices with m edges. Given a subset of vertices $\mathcal{S} \subseteq \mathcal{V}$, let $e(\mathcal{S})$ denote the number of edges in the subgraph $\mathcal{G}_{\mathcal{S}}$ induced by $\mathcal{S}$. The density of $\mathcal{S}$ is defined as $\rho(\mathcal{S}) := e(\mathcal{S})/\binom{|\mathcal{S}|}{2}$. The optimal quasi-clique (OQC) formulation proposed in (Tsourakakis et al. 2013) aims at finding the subgraph that maximizes the following objective function

$$f_\alpha(\mathcal{S}) := e(\mathcal{S}) - \alpha\binom{|\mathcal{S}|}{2}. \quad (1)$$

The first term encourages the subgraph induced by $\mathcal{S}$ to have a large number of edges while the second term penalizes large subgraph sizes. The regularization parameter $\alpha \in (0, 1)$ plays a balancing act in trading off subgraph density for size. The objective function admits the following interpretation - the second term can be viewed as the number of edges that appear in expectation in a random Erdos-Renyi graph defined on the vertex subset $\mathcal{S}$, where $\alpha \in (0, 1)$ denotes the probability of an edge connecting a pair of vertices. Thus, $f_\alpha(\mathcal{S})$ assigns a greater reward to subgraphs $\mathcal{G}_{\mathcal{S}}$ which exhibit a large surplus of edges with respect to the random subgraph model. Overall, OQC aims to solve the following optimization problem

$$\max_{\mathcal{S}\subseteq\mathcal{V}} f_\alpha(\mathcal{S}) \quad (2)$$

The choice of the parameter α affects the size and density of the extracted solution. Intuitively, selecting a small value of α allows large, non-dense subgraphs to exhibit a large edge surplus. As the value of α is increased, dense subgraphs of smaller size are favored. In (Tsourakakis et al. 2013), it was recommended to set $\alpha = 1/3$.

Hardness

Regarding the computational complexity of problem (2), little is known - it has long been suspected to be NP–hard (Tsourakakis et al. 2013), but a formal proof has remained elusive thus far. We point out that a generalization of the OQC problem studied in (Cadena, Vullikanti, and Aggarwal 2016), where the edges of $\mathcal{G}$ are allowed to have arbitrary weights, has been shown to be NP–hard. An analogous result for undirected graphs, where each edge has unit weight, however, is not presently known. Our first major contribution settles the matter by furnishing a proof of NP–hardness via a reduction from the decision version of the MAXCLIQUE problem, which is known to be NP–complete (Karp 1972). Given $\mathcal{G}$ and a positive integer $k \geq 3$, the decision variant of the MAXCLIQUE problem asks whether the maximum clique size is at least k. We demonstrate that for every choice of k, there exists a sub-interval of $\alpha \in (0, 1)$ for which there is a one-to-one correspondence between the solutions of problem (2) and MAXCLIQUE. Hence, OQC is at least as hard as solving an arbitrary decision instance of MAXCLIQUE. Our reduction utilizes the following key result in extremal graph theory.

Fact 1 (**Turán's theorem** (Turan 1941)). *Every graph on n vertices that does not contain a k-clique, can have at most the following number of edges.*

$$\tau(n, k) := \left(1 - \frac{1}{k-1}\right)\frac{n^2}{2} \quad (3)$$

In other words, if the number of edges in a n-vertex graph exceeds $\tau(n, k)$, then it must contain a k-clique. We adopt the following approach: if a graph contains a k-clique, then it is a subset of a subgraph of size at least k with "sufficiently" large edge density. That is, if we are able to locate an induced subgraph $\mathcal{G}_{\mathcal{S}}$ such that the number of edges induced $e(\mathcal{S})$ exceeds the threshold $\tau(|\mathcal{S}|, k)$, then that subgraph must harbor a k-clique. Since such a k-clique in $\mathcal{G}_{\mathcal{S}}$ is also a k-clique in $\mathcal{G}$, it then follows that an affirmative answer to the instance of MAXCLIQUE has been determined. The task of detecting such a subgraph and tying it to the solution of the OQC problem (2) is formalized in the following result.

Theorem 1. *The optimal quasi-clique problem on undirected graphs* (2) *is NP-hard.*

Proof. We briefly sketch the outline. Given a decision instance of MAXCLIQUE, we utilize Turán's theorem to determine the smallest constant $\alpha \in (0, 1)$ for which a subgraph $\mathcal{G}_{\mathcal{S}}$ induced by $\mathcal{S} \subseteq \mathcal{V}$ obeys the inequality

$$\alpha\binom{|\mathcal{S}|}{2} \geq \tau(|\mathcal{S}|, k) = \left(1 - \frac{1}{k-1}\right)\frac{|\mathcal{S}|^2}{2}. \quad (4)$$

For such a choice of α, if it additionally holds that the edge-surplus $f_\alpha(\mathcal{S}) > 0$, then we have the implication $e(\mathcal{S}) > \tau(|\mathcal{S}|, k)$. This in turn implies that $\mathcal{G}_{\mathcal{S}}$ harbors a k-clique. It turns out that a sufficient choice of α is

$$\alpha_k := 1 - \frac{1}{(k-1)^2}. \quad (5)$$

We can show that solving (2) with $\alpha = \alpha_k$ and examining whether the size of the solution is greater than or equal to, or smaller than k, corresponds to solving any decision instance of MAXCLIQUE. □

Remark 1. When testing for the presence of a k-clique, the above result remains unchanged if the threshold $\alpha = \alpha_k$ is replaced by any value of α in the sub-interval $[\alpha_k, 1)$. This is because for a fixed value of k, α_k is the smallest constant that satisfies the inequality (4). Clearly, any value of α which exceeds this threshold is also a valid choice.

Unveiling the Role of α

As mentioned previously, the choice of α plays a key role in determining the quality of the subgraph extracted by OQC (in terms of size and density). However, the question of which subgraphs of $\mathcal{G}$ correspond to maximizers of OQC for a given value of $\alpha \in (0,1)$ has not been formally investigated. This is the main object of our study.

Given a parameter $k \in [K] := \{2, \cdots, n\}$, we denote the optimal (i.e., maximum) density across all size-k subgraphs as $\rho_k^* := \max_{|\mathcal{S}|=k} \rho(\mathcal{S})$. The collection of pairs $\{(k, \rho_k^*)\}_{k \in [K]}$ then corresponds to the *optimal size-density* frontier of $\mathcal{G}$; i.e., each frontier point denotes the maximum subgraph density of a given size. Regarding the relationship among the optimal density values $\{\rho_k^*\}_{k\in[K]}$, the following result is known (Kawase and Miyauchi 2018, Lemma 1).

Lemma 1. *For any graph $\mathcal{G}$, the optimal size-k subgraph density ρ_k^* is a monotonically non-increasing function of the size; i.e., it always holds that*

$$\rho_k^* \geq \rho_{k+1}^*, \forall\, k \in [K]. \tag{6}$$

Let ω denote the size of the maximum clique in $\mathcal{G}$, and $\mathcal{C}_\omega$ be a subset of vertices that constitute a maximum clique. The result implies that for every size $k \leq \omega$, the maximum density $\rho_k^* = 1$. This is because a fixed size subgraph attains a density of 1 (the maximum possible value) if and only if it is a clique, and the maximum clique contains all cliques of smaller size. For sizes $k > \omega$, the optimal density ρ_k^* is bounded away from 1, i.e., the densest subgraphs in this range of sizes are quasi-cliques. Furthermore, by virtue of Lemma 1, the density ρ_k^* of these optimal quasi-cliques is a monotone non-increasing function of size k.

Our second major contribution establishes that solving OQC with varying α is equivalent to mining subgraphs which correspond to different points on the optimal size-density frontier. To be precise, we show that for every unique density value in the set $\{\rho_k^*\}_{k\in[K]}$, there exists a sub-interval of $\alpha \in (0,1)$ for which the solution of OQC corresponds to the largest subgraph of that density value. For example, our result implies that there is a range of α for which the maximizers of problem (2) are the maximum cliques, which correspond to the largest subgraphs in $\mathcal{G}$ with density 1. As expected, our results show that large values of α enable OQC to mine maximum cliques and optimal near-cliques lying on the optimal size-density frontier, with smaller values extracting larger subgraphs of lower density on this frontier.

Extracting the Maximum Clique

We provide sufficient conditions on α such that the optimal solution of problem (2) coincides with the set of maximum cliques in $\mathcal{G}$. First, we establish the following warm-up result. Consider a vertex subset $\mathcal{S}$ of size at most ω with density $\rho(\mathcal{S}) \in [0,1]$. Then, for any choice of $\alpha \in (0,1)$ in the edge-surplus function (1), the following statement is true.

Lemma 2. *For any subgraph of size $|\mathcal{S}| \leq \omega$, it always holds that*

$$f_\alpha(\mathcal{S}) \leq f_\alpha(\mathcal{C}_\omega). \tag{7}$$

Since the maximum clique size ω is unique, the inequality (7) is satisfied with equality if and only if $\mathcal{S}$ constitutes a maximum clique in $\mathcal{G}$. We conclude from the above result that all subgraphs of $\mathcal{G}$ which lie in the "shadow" of the maximum clique, i.e., which are dominated in size and density by $\mathcal{C}_\omega$, are always sub-optimal for (2), irrespective of the choice of $\alpha \in (0,1)$. Consequently, if $\mathcal{S}_\alpha^*$ denotes the optimal solution of (2), for every value of $\alpha \in (0,1)$, it must hold that $|\mathcal{S}_\alpha^*| \geq \omega$ and

$$f_\alpha(\mathcal{S}_\alpha^*) \geq f_\alpha(\mathcal{C}_\omega). \tag{8}$$

Going forward, we are interested in determining for what range of values of α are the above pair of inequalities satisfied with equality, which implies that the optimal solution of (2) coincides with the maximum clique. We expect the required value of α to be large in order for the edge-surplus attained by the maximum cliques in $\mathcal{G}$ to dominate that of all other subgraphs. Let $\rho_{\omega+1}^*$ denote the density of the densest quasi-clique larger than ω. Define the threshold

$$\hat{\alpha} := \rho_{\omega+1}^* - (1 - \rho_{\omega+1}^*) \cdot c_0, \tag{9}$$

where $c_0 := \omega(\omega-1)/(n-\omega)(n+\omega+1)$. Then, we have the following result.

Theorem 2. *For all $\alpha \in (\hat{\alpha}, 1)$, the maximizers of OQC are the maximum cliques in $\mathcal{G}$.*

Remark 2. We point out that extracting a maximum clique $\mathcal{C}_\omega$ corresponds to extracting all points on the optimal size-density frontier $\{(k,1)\}_{k\leq\omega}$, since $\mathcal{C}_\omega$ contains all cliques of smaller sizes.

Extracting Optimal Quasi-Cliques Larger than the Maximum Clique

Define the set $[L] := \{1, \cdots, n-\omega\}$. For a fixed parameter $\ell \in [L]$, let $\mathbb{Q}_\ell$ denote the set of all quasi-cliques in $\mathcal{G}$ of size $\omega + \ell$. Let $\mathcal{Q}_\ell^* \in \mathbb{Q}_\ell$ denote an *optimal* quasi-clique of size $\omega + \ell$ that attains the maximum density $\rho_{\omega+\ell}^*$; i.e., we have $\mathcal{Q}_\ell^* \in \arg\max_{|\mathcal{S}|=\omega+\ell} \rho(\mathcal{S})$. Next, we show that α can be selected such that the maximizers of OQC correspond to optimal quasi-cliques $\{\mathcal{Q}_\ell^*\}_{\ell\in[L]}$. Note that such optimal quasi-cliques correspond to the points $(\omega+\ell, \rho_{\omega+\ell}^*)$ on the optimal size-density frontier. Our analysis requires the following assumption.

Assumption 1: *Every optimal density value in the range $\{\rho_{\omega+\ell}^*\}_{\ell\in[L]}$ is unique.*

In other words, the optimal density values are not repeated for subgraph sizes larger than ω. While reasonable, this condition does not hold without loss of generality (e.g., in a 4-cycle, $\rho_3^* = \rho_4^*$). Nevertheless, its primary utility is to keep derivations simple; it can be relaxed at the expense of more cumbersome technical arguments.

Warm-up: We first consider the case of $\ell = 1$, which corresponds to extracting $\mathcal{Q}_1^*$. The extension to the general case will be described afterwards. In order for $\mathcal{Q}_1^*$ to be the unique maximizer of (2), α should satisfy each of the following conditions.

1. $f_\alpha(\mathcal{Q}_1^*) > f_\alpha(\mathcal{Q}_1), \forall\ \mathcal{Q}_1 \in \mathbb{Q}_1 \setminus \mathcal{Q}_1^*$. This reflects the fact that $\mathcal{Q}_1^*$ is required to have the maximum edge-surplus amongst all quasi-cliques $\mathbb{Q}_1$ of size $\omega + 1$.

2. $f_\alpha(\mathcal{Q}_1^*) > f_\alpha(\mathcal{Q}_\ell), \forall\ \mathcal{Q}_\ell \in \mathbb{Q}_+ \setminus \mathbb{Q}_1$. This ensures that the edge-surplus attained by $\mathcal{Q}_1^*$ dominates that of the quasi-cliques of size larger than $\omega + 1$.

3. $f_\alpha(\mathcal{Q}_1^*) > f_\alpha(\mathcal{C}_\omega)$. That is, the edge surplus of $\mathcal{Q}_1^*$ must exceed that of the maximum clique $\mathcal{C}_\omega$. Recall the assertion of Lemma 2, which states that for any choice of $\alpha \in (0, 1)$, we have $f_\alpha(\mathcal{C}_\omega) \geq f_\alpha(\mathcal{S})$, for all subgraphs of size $|\mathcal{S}| \leq \omega$. Hence, satisfying the condition $f_\alpha(\mathcal{Q}_1^*) > f_\alpha(\mathcal{C}_\omega)$ also guarantees that $f_\alpha(\mathcal{Q}_1^*) > f_\alpha(\mathcal{S})$ is satisfied, for all subgraphs $\mathcal{S}$ smaller than ω.

Define the thresholds

$$\mathsf{LB}(1) := \rho^*_{\omega+2} - (\rho^*_{\omega+1} - \rho^*_{\omega+2}) \cdot c_1, \tag{10a}$$

$$\mathsf{UB}(1) := \rho^*_{\omega+1} - (1 - \rho^*_{\omega+1})(\omega - 1)/2 \tag{10b}$$

where c_1 is a constant dependent on n, ω. We can show that the above thresholds define an open sub-interval where α satisfies the above three conditions. This leads to the following result.

Theorem 3. *For all $\alpha \in (\mathsf{LB}(1, \mathsf{UB}(1))$, the maximizers of OQC correspond to optimal quasi-cliques $\mathcal{Q}_1^*$.*

The general case: Next, we consider the extraction of optimal quasi-cliques of sizes $\ell \in \{2, \cdots, n - \omega\}$. Again, the following three conditions must be met for $\mathcal{Q}_\ell^*$ to be the unique maximizer of (2).

4. $f_\alpha(\mathcal{Q}_\ell^*) > f_\alpha(\mathcal{Q}_\ell), \forall\ \mathcal{Q}_\ell \in \mathbb{Q}_\ell \setminus \mathcal{Q}_\ell^*$. This condition is the same as **(1)**.

5. $f_\alpha(\mathcal{Q}_\ell^*) > f_\alpha(\mathcal{Q}_k), \forall\ \mathcal{Q}_k \in \mathbb{Q}_k, k \in [K]_\ell := \{\ell + 1, \cdots, n - \omega\}$. This is a generalization of condition **(2)** to ensure that the edge surplus of $f_\alpha(\mathcal{Q}_\ell^*)$ dominates that of all subgraphs of size larger than ℓ.

6. $f_\alpha(\mathcal{Q}_\ell^*) > f_\alpha(\mathcal{Q}_j), \forall\ \mathcal{Q}_j \in \mathbb{Q}_j, j \in \{1, \cdots, \ell - 1\}$. This condition generalizes **(3)** and ensures that the edge surplus of the optimal quasi-clique $\mathcal{Q}_\ell^*$ of size $\omega + \ell$ exceeds that of all quasi-cliques of smaller sizes.

Define the thresholds

$$\mathsf{LB}(\ell) := \rho^*_{\omega+(\ell+1)} - (\rho^*_{\omega+\ell} - \rho^*_{\omega+(\ell+1)})c_\ell, \tag{11a}$$

$$\mathsf{UB}(\ell) := \rho^*_{\omega+\ell} - (\rho^*_{\omega+(\ell-1)} - \rho^*_{\omega+\ell}) \cdot \left\lceil \frac{\omega + (\ell - 2)}{2} \right\rceil \tag{11b}$$

where c_ℓ is a constant dependent on n, ℓ, ω. By an appropriate generalization of the arguments underpinning Theorem 3, we can utilize the above thresholds to obtain the following result.

Theorem 4. *For all $\alpha \in (\mathsf{LB}(\ell), \mathsf{UB}(\ell))$, the maximizers of OQC correspond to optimal quasi-cliques $\mathcal{Q}_\ell^*$.*

Overall, our results demonstrate that for any graph $\mathcal{G}$, there exists a choice of α such that the maximizers of OQC correspond to the largest quasi-clique that attains a unique density on the optimal size-density frontier.

Relationship with Densest-k-Subgraph

In the previous section we demonstrated that there exists a choice of α in OQC which enables extraction of subgraphs comprising the optimal size-density frontier; i.e., subgraphs of $\mathcal{G}$ corresponding to the pairs $\{(k, \rho_k^*)\}_{k \in [K]}$. An alternate means of traversing this frontier is to employ the DENSEST-k-SUBGRAPH (DkS) formulation. Given a size parameter $k \in [K]$, DkS aims to find maximizers of the optimization problem $\max\limits_{|\mathcal{S}|=k} \rho(\mathcal{S})$. Clearly, any size-$k$ maximizer of DkS corresponds to the point (k, ρ_k^*) on the optimal size-density frontier. Thus, by varying k, DkS can be used to sweep the frontier comprising the pairs $\{(k, \rho_k^*)\}_{k \in [K]}$. This implies that in terms of the optimal density *value* attainable for each specific subgraph size, OQC and DkS are equivalent.

A natural follow-up question to consider then is the relationship between the maximizers of the twin formulations. Can the problems be viewed as being equivalent in this respect as well? To this end, we define the following notation. For a fixed value of $\alpha \in (0, 1)$, let $\mathbb{S}_\alpha^*$ denote the collection of maximizers of OQC; i.e., a subgraph $\mathcal{S}_\alpha^* \in \mathbb{S}_\alpha^*$ is a maximizer of OQC. Similarly, for a fixed size $k \in [K]$, let $\mathbb{S}_k^*$ denote the collection of maximizers of DkS.

Theorem 5. *For every $\alpha \in (0, 1)$, there exists a value of $k \in [K]$ such that $\mathbb{S}_\alpha^* \subseteq \mathbb{S}_k^*$. However, there exist maximizers of DkS which are not maximizers of OQC.*

The second case corresponds to scenarios where the optimal density value is repeated across successive subgraph sizes. Hence, the two formulations are not entirely equivalent w.r.t their maximizers. However, we can show that when such an event occurs, the maximizers of OQC correspond to the largest quasi-clique among all optimal quasi-cliques that attain the same density value (across successive sizes). Additionally, we can also show that the largest such quasi-clique contains all the quasi-cliques of smaller sizes (with the same density).

Lemma 3. *Let $\mathcal{Q}_k^*$ and $\mathcal{Q}_{k+1}^*$ be optimal quasi-cliques with densities $\rho_k^* = \rho_{k+1}^*$. Then, $\mathcal{Q}_{k+1}^*$ harbors a size-k optimal quasi-clique with density ρ_k^*.*

Note that the above result generalizes the fact that the maximum clique contains cliques of all sizes lesser than ω.

Experiments

In principle, both OQC and DkS can be employed to mine dense subgraphs of differing sizes from the optimal size-density frontier $\{(k, \rho_k^*)\}_{k \in [K]}$ of $\mathcal{G}$. However, these problems are NP–hard in the worst-case. In light of this fact, we resort to employing approximation algorithms for each formulation, which are not guaranteed to find optimal solutions in general. Thus, in practice, depending on the effectiveness of the selected algorithm, the quality of the subgraphs extracted (in terms of size and density) using the two formulations can be different. In this section, we conduct an empirical comparison of the subgraphs extracted by approximation methods for DkS and OQC on real-world graphs and provide guidelines regarding which formulation to use.

Lovász Relaxation for DkS: We employ the recent convex relaxation approach of (Konar and Sidiropoulos 2021)

Dataset	n	m	Network Type
FACEBOOK	4K	88K	Social
SOC-GOOGLE	211K	1.14M	Social
WEB-STANFORD	281K	2.31M	Web graph
MATHSCINET	332K	820K	Co-authorship
CA-DBLP	540K	15M	Co-authorship
WEB-GOOGLE	875K	5.10M	Web graph
PATENTS	3.7M	16.7M	Citation graph

Table 1: Summary of graph statistics: the number of vertices (n), the number of edges (m), and network type.

wherein the Lovász extension of the supermodular objective function of DkS is maximized over the convex hull of the sum-to-k constraints. The resulting problem is solved using the Alternating Direction Method of Multipliers (ADMM) (Condat 2013). As the solution is not guaranteed to be integral, a rounding post-processing step is used to obtain the candidate subgraph of the desired size k.

Greedy peeling for OQC: We employ the greedy vertex-peeling algorithm originally proposed in (Tsourakakis et al. 2013). Starting from the entire graph $\mathcal{G}$, the algorithm repeatedly peels off the lowest degree vertex until no vertices are left to remove. In the process, a sequence of nested subgraphs is generated, and the one which attains the largest edge surplus is returned as the solution. The algorithm can be implemented efficiently in $O(n + m)$ time. In (Tsourakakis et al. 2013), the choice of $\alpha = 1/3$ was recommended to select the subgraph with the largest edge-surplus. As our theoretical analysis reveals that increasing the value of α is more suitable for detecting dense quasi-cliques, in our experiments we employ larger values of α. In practice, given a graph $\mathcal{G}$, it is difficult to determine the exact sub-interval of α required to a extract quasi-clique of a desired size since we do not know *apriori* all the parameters required for constructing the requisite sub-interval of α, including for what range of subgraph sizes are the optimal density values repeated. Consequently, we resort to using empirically chosen values of α. Note that fine-tuning the selection of α can be accomplished in a post-processing step independently of the algorithm; i.e., the algorithm has to be executed *once* in order to obtain a ranking of the vertices based on the iteration index where they were eliminated (this procedure does not depend on the value of α). Thereafter, different values of α can be tested to extract the best solution relative to the corresponding edge-surplus function. In this manner, the algorithm can be employed to quickly generate an approximation of the optimal size-density curve of $\mathcal{G}$. This is an advantage enjoyed by the algorithm over its DkS counterpart, which needs to be run for each desired value of subgraph size k.

Datasets, pre-processing and implementation: We used a collection of datasets (summarized in Table **??**) obtained from standard repositories (Leskovec and Krevl 2014) to test the performance of all methods. Each dataset is pre-processed by symmetrizing any directed arcs, removing self-loops, and extracting the largest connected component. All our experiments were performed in Matlab on a Macbook equipped with 16GB RAM and an M2 processor. The code for the ADMM algorithm for solving DkS was the same employed in (Konar and Sidiropoulos 2021).

Performance on Real-world graphs: After running the GREEDYOQC algorithm on a dataset, we perform a grid search on α in the range $[0.01, 0.99]$, in increments of 0.01. Each value of α defines a different edge-surplus function, using which the subgraph with the largest edge surplus amongst the family of nested subgraphs generated by GREEDYOQC is selected. The smallest and largest size subgraphs obtained by this procedure are then set to be the lower and upper limits on k in the ADMM algorithm for DkS respectively. The size-density frontiers generated by these two different methods on the aforementioned datasets is depicted in Figure 2. We make the following general observations.

1. There exist "gaps" in the size-density frontier generated by GREEDYOQC. This is because for each choice of α, the solution is always restricted to be chosen from the same family of n nested subgraphs generated by the peeling process. We empirically confirmed that this can result in the same subgraph in the family attaining the largest edge-surplus for successive values of α. Owing to these "resolution limits", the subgraphs extracted can correspond to a coarse approximation of the optimal size-density frontier in terms of the range of subgraph sizes spanned. The results also showcase that larger values of α do indeed retrieve denser subgraphs; in particular the originally recommended (Tsourakakis et al. 2013) choice of $\alpha = 1/3$ can be sub-optimal in this regard.
2. Since the ADMM-based relaxation of DkS is designed to output a subgraph of a distinct size, it does not exhibit gaps in its generated size-density frontier. Thus, it offers a more fine-grained approximation of the optimal size-density frontier in terms of subgraph sizes compared to GREEDYOQC. However, this comes at the cost of extra computational time as the algorithm has to be run for each distinct value of k.
3. For subgraph sizes corresponding to the intersection of the twin size-density frontiers, the two algorithms are closely matched in general. However, for smaller subgraph sizes (≤ 100), the ADMM algorithm can perform worse than GREEDYOQC, which attains high-quality solutions in the range.

We conclude that GREEDYOQC can be use to quickly obtain a high-quality approximation of the optimal size-density frontier. However, since it is limited in its resolution (in terms of the range of subgraph sizes spanned), ADMM-DkS can be employed over a smaller range of interest in order to obtain a finer-grained approximation of the frontier.

Large and loose quasi-cliques matter too: In dense subgraph discovery, one is typically interested in exploring the "high" (density) end of the optimal size-density frontier of $\mathcal{G}$, which is comprised of cliques and near-cliques. Given that, scant attention has been paid to exploring the opposing "low" end of the frontier, consisting of large subgraphs with low density. At first, it may seem that there is no apparent reason for doing so, since the subgraphs comprising this regime are not dense to begin with. However, since

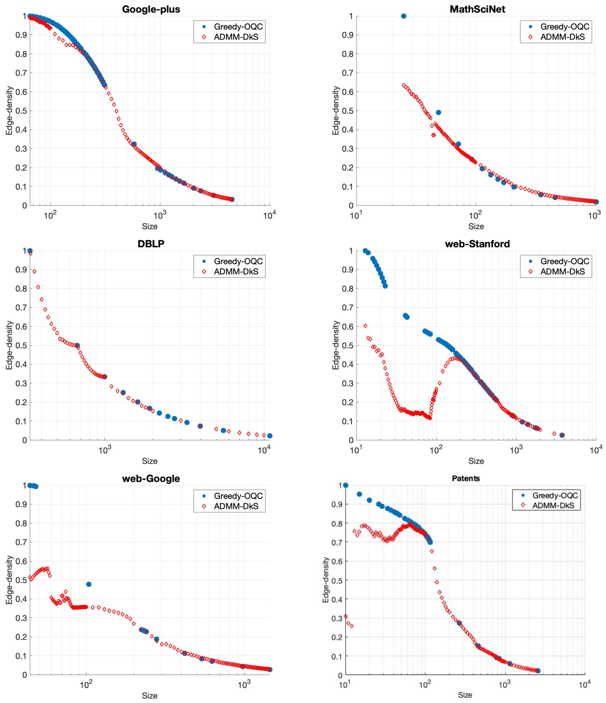

Figure 2: Size-density frontiers generated by GREEDYOQC and ADMM-DkS. For OQC, denser subgraphs correspond to smaller values of α.

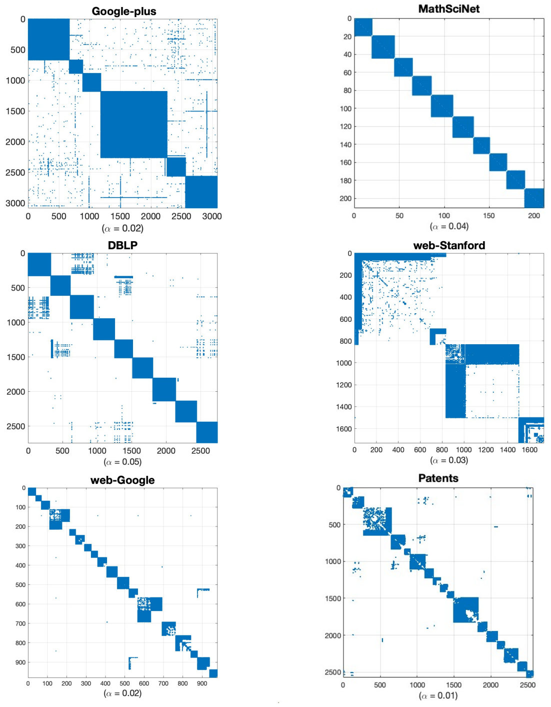

Figure 3: Presence of local communities in low-density subgraphs identified using OQC, as visualized by the block-diagonal structure of their respective adjacency matrices.

GREEDYOQC can be utilized to quickly explore any region of the frontier (by appropriate selection of α in the post-processing step), we analyzed the characteristics of the subgraphs comprising the low end. Our results indicate that subgraphs with density as low as $2 - 10\%$ can be interesting in their own right. Figure 3 depicts the sparsity pattern of the adjacency matrices of these extracted subgraphs across various datasets. Although these subgraphs are too large and sparse to be labelled dense (having only $5 - 10\%$ density), the block diagonal structure of their adjacency matrices reveals a striking property - *the presence of local community structure*. Evidently, these subgraphs are composed of multiple components of non-trivial size which exhibit sparse external connectivity and high internal cohesion. In order to reveal this community structure, we applied spectral clustering (Von Luxburg 2007) on the extracted subgraph.

It is well known that real-world graphs lack global community structure (Leskovec et al. 2009), and global partitioning methods such as normalized-cut (Shi and Malik 2000)[2] typically fail to find well connected clusters. Hence, a body of research has blossomed around local community detection (Spielman and Teng 2004; Andersen, Chung, and Lang 2006; Kloster and Gleich 2014; Orecchia and Zhu 2014; Veldt, Gleich, and Mahoney 2016; Wang et al. 2017) which use specialized techniques and algorithms tailored for detecting local communities. In that context, our results are surprising since (a): the edge-surplus function is a surrogate for maximizing the internal connectivity of a subgraph and not explicitly geared towards promoting community structure, and (b): it is not obvious *apriori* that running the peeling process simply based on removing the lowest degree vertex will "chip" away at the global structure in the right places to reveal local communities. It is striking that this indeed happens consistently across various real-world graphs.

Conclusions

We revisited the OQC problem and revealed that the densities of its solutions obtained by continuous variation of α is equivalent to that of the classic Densest-k-subgraph problem. This opened the door to utilizing the GREEDYOQC algorithm for mining dense subgraphs comprising the optimal size-density frontier. On real-world graphs, we demonstrated that the algorithm quickly generates a high-quality approximation of the frontier, to that generated by more computationally intensive baselines of DkS; albeit with possibly limited resolution. On turning the spotlight towards large, loosely connected quasi-cliques, we made the surprising discovery that they harbor well defined local communities, even though the OQC formulation does not explicitly promote community structure.

Acknowledgments

Supported by the KU Leuven Special Research Fund (BOF/STG-22-040) and the National Science Foundation, USA (IIS-1908070).

[2] (of which spectral clustering can be viewed as a relaxation)

References

Andersen, R.; Chung, F.; and Lang, K. 2006. Local graph partitioning using pagerank vectors. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, 475–486. IEEE.

Bertsekas, D. P. 2014. *Constrained optimization and Lagrange multiplier methods*. Academic press.

Boob, D.; Gao, Y.; Peng, R.; Sawlani, S.; Tsourakakis, C.; Wang, D.; and Wang, J. 2020. Flowless: Extracting densest subgraphs without flow computations. In *Proceedings of The Web Conference 2020*, 573–583.

Boyd, S. P.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Cadena, J.; Chen, F.; and Vullikanti, A. 2018. Graph anomaly detection based on Steiner connectivity and density. *Proc. of the IEEE*, 106(5): 829–845.

Cadena, J.; Vullikanti, A. K.; and Aggarwal, C. C. 2016. On dense subgraphs in signed network streams. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 51–60. IEEE.

Charikar, M. 2000. Greedy approximation algorithms for finding dense components in a graph. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, 84–95. Springer.

Chekuri, C.; Quanrud, K.; and Torres, M. R. 2022. Densest subgraph: Supermodularity, iterative peeling, and flow. In *Proc. of SODA*, 1531–1555. SIAM.

Chen, T.; and Tsourakakis, C. 2022. Antibenford subgraphs: Unsupervised anomaly detection in financial networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2762–2770.

Condat, L. 2013. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2): 460–479.

Feige, U.; Peleg, D.; and Kortsarz, G. 2001. The dense k-subgraph problem. *Algorithmica*, 29(3): 410–421.

Goldberg, A. V. 1984. *Finding a maximum density subgraph*. Technical report, University of California Berkeley, CA.

Hooi, B.; Song, H. A.; Beutel, A.; Shah, N.; Shin, K.; and Faloutsos, C. 2016. Fraudar: Bounding graph fraud in the face of camouflage. In *Proc. of SIGKDD*, 895–904. ACM.

Karp, R. M. 1972. Reducibility among combinatorial problems. In *Complexity of computer computations*, 85–103. Springer.

Kawase, Y.; and Miyauchi, A. 2018. The densest subgraph problem with a convex/concave size function. *Algorithmica*, 80: 3461–3480.

Kloster, K.; and Gleich, D. F. 2014. Heat kernel based community detection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1386–1395.

Konar, A.; and Sidiropoulos, N. D. 2021. Exploring the Subgraph Density-Size Trade-off via the Lovaśz Extension. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 743–751.

Lanciano, T.; Miyauchi, A.; Fazzone, A.; and Bonchi, F. 2023. A survey on the densest subgraph problem and its variants. arXiv:2303.14467.

Leskovec, J.; and Krevl, A. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. https://snap.stanford.edu/data. Accessed: 2023-08-15.

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.

Li, X.; Liu, S.; Li, Z.; Han, X.; Shi, C.; Hooi, B.; Huang, H.; and Cheng, X. 2020. Flowscope: Spotting money laundering based on graphs. In *Proc. of AAAI*, volume 34, 4731–4738.

Manurangsi, P. 2017. Almost-polynomial ratio ETH-hardness of approximating densest k-subgraph. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 954–961.

Orecchia, L.; and Zhu, Z. A. 2014. Flow-based algorithms for local graph clustering. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, 1267–1286. SIAM.

Papailiopoulos, D.; Mitliagkas, I.; Dimakis, A.; and Caramanis, C. 2014. Finding dense subgraphs via low-rank bilinear optimization. In *ICML*, 1890–1898.

Seidman, S. B. 1983. Network structure and minimum degree. *Social networks*, 5(3): 269–287.

Shi, J.; and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8): 888–905.

Shin, K.; Eliassi-Rad, T.; and Faloutsos, C. 2016. Corescope: Graph mining using k-core analysis—patterns, anomalies and algorithms. In *2016 IEEE 16th international conference on data mining (ICDM)*, 469–478. IEEE.

Spielman, D. A.; and Teng, S.-H. 2004. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, 81–90.

Tsourakakis, C.; Bonchi, F.; Gionis, A.; Gullo, F.; and Tsiarli, M. 2013. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 104–112.

Turan, P. 1941. On an extremal problem in graph theory. *Mat. Fiz. Lapok*, 48: 436–452.

Veldt, N.; Benson, A. R.; and Kleinberg, J. 2021. The generalized mean densest subgraph problem. In *Proc. of SIGKDD*, 1604–1614.

Veldt, N.; Gleich, D.; and Mahoney, M. 2016. A simple and strongly-local flow-based method for cut improvement. In *International Conference on Machine Learning*, 1938–1947. PMLR.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17: 395–416.

Wang, D.; Fountoulakis, K.; Henzinger, M.; Mahoney, M. W.; and Rao, S. 2017. Capacity releasing diffusion for speed and locality. In *International Conference on Machine Learning*, 3598–3607. PMLR.

Zhang, S.; Zhou, D.; Yildirim, M. Y.; Alcorn, S.; He, J.; Davulcu, H.; and Tong, H. 2017. Hidden: hierarchical dense subgraph detection with application to financial fraud detection. In *Proceedings of the 2017 SIAM international conference on data mining*, 570–578. SIAM.