

Better Fewer but Better: Community Search with Outliers

Francesco Bonchi
ISI Foundation, Turin, Italy
Eurecat, Barcelona, Spain
francesco.bonchi@isi.it

Lorenzo Severini
UniCredit Services
Rome, Italy
loseverini@gmail.com

Mauro Sozio
Telecom Paris, IP Paris
Paris, France
sozio@telecom-paris.fr

Abstract—Given a set of vertices in a network, that we believe are of interest for the application under analysis, *community search* is the problem of producing a subgraph potentially explaining the relationships existing among the vertices of interest. In practice this means that the solution should add some vertices to the query ones, so to create a connected subgraph that exhibits some “cohesiveness” property. This problem has received increasing attention in recent years: while several cohesiveness functions have been studied, the bulk of the literature looks for a solution subgraphs containing *all* the query vertices. However, in many exploratory analyses we might only have a reasonable belief about the vertices of interest: if only one of them is not really related to the others, forcing the solution to include all of them might hide the existence of much more cohesive and meaningful subgraphs, that we could have found by allowing the solution to detect and drop the outlier vertex. In this paper we study the problem of *community search with outliers*, where we are allowed to drop up to k query vertices, with k being an input parameter. We consider three of the most used measures of cohesiveness: the minimum degree, the diameter of the subgraph and the maximum distance with a query vertex. By optimizing one and using one of the others as a constraint we obtain three optimization problems: we study their hardness and we propose different exact and approximation algorithms.

I. INTRODUCTION

Community search is a fundamental graph mining problem which has recently received a great deal of attention: given a graph and a set of query vertices, we wish to find a subset of vertices containing the query ones such that the induced subgraph is connected and optimizes some cohesiveness measure. The extracted subgraph may provide useful insights on the relationships existing among the query vertices and other vertices in the graph. For instance, given a set of suspected terrorists in a social networks, which other individuals in the social network should we monitor? Given a set of proteins of interest in a protein-protein interaction network, which other proteins can participate in pathways with them? This is an exploratory data analysis task: in most applications we might be given some query vertices that *we believe* might have some interesting connection or participate in some relevant pattern or substructure. However, it might be the case that *not all of them* are relevant for the discovery task at hand. By forcing the solution to connect them all, we might produce much larger and less cohesive solutions, hiding the really interesting ones that we could instead find by just allowing some query vertices not to belong to the solution. While several cohesiveness

functions have been studied, the bulk of the literature (briefly surveyed next) enforce that *all* query vertices be present in the output solution. Before presenting our contributions, we provide some background definition and present the relevant related literature.

Background and related work. Given a graph $G = (V, E)$ and a set of query vertices $Q \subseteq V$, a wide family of problems requires to find a connected subgraph H of G , that contains all query vertices Q , while exhibiting some nice properties of cohesiveness, compactness or density. Several problems fit under this framework, such as *community search* [2], [4], [16], *seed set expansion* [13], and *connectivity subgraphs* [1], [5], [15], [17]. Kloumann and Kleinberg [13] provide a systematic evaluation of different methods for *seed set expansion* on graphs with known community structure, assuming that the seed set Q is made of vertices belonging to the same community. Faloutsos et al. [5] address the problem of finding a subgraph that connects two query vertices ($|Q| = 2$) and contains at most b other vertices, optimizing a measure of proximity based on *electrical-current flows*. Tong and Faloutsos [17] extend [5] by introducing the concept of *Center-piece Subgraph* dealing with query sets of any size. Sozio and Gionis [16] developed a framework for solving a wide range of community search optimization problems. The most basic problem studied in [16] consists of finding a connected subgraph containing Q and maximizing the minimum degree. They developed an efficient algorithm for solving such a problem optimally, however, it is well-known that such a variant of the problem suffers from the *free rider* effect. To alleviate such a problem other constraints on the output graph can be enforced (so-called *monotone* functions). In our work, we focus on the variant where a distance constraint between the nodes in the graph and the query nodes is enforced, additionally. We adapt such a variant to the case with outliers while maintaining the nice property that it can be solved optimally. More recently, Cui *et al.* [4] devise a local-search approach to improve the efficiency of the method defined in [16], but only for the special case of a single query vertex. The case of multiple query vertices has instead been addressed by Barbieri *et al.* [2], who exploit core decomposition as a preprocessing step to improve efficiency. Ruchansky et al. [15] introduce the parameter-free problem of extracting the

Minimum Wiener Connector, that is the connected subgraph containing Q which minimizes the pairwise sum of shortest-path distances among its vertices. Recent approaches also introduce the flexibility of having query vertices belonging to different communities [3], [19]. Finally, community search has been formalized for attributed graphs [6], [11], spatial graphs [7] and temporal graphs [8], [18]. All these approaches look for a connected subgraph of the input graph containing *all* the query vertices. Three recent approaches allow some forms of outliers in community search. Akoglu et al. [1] study the problem of finding *pathways*, i.e., connection subgraphs for a large query set Q , in terms of the Minimum Description Length (MDL) principle. According to MDL, a pathway is simple when only a few bits are needed to relay which edges should be followed to visit all of Q . Given a graph G and a query set Q , Gionis et al. [9] study the problem of finding a connected subgraph of G that has more vertices that belong to Q than vertices that do not. Ruchansky et al. [14] introduce the problem of finding the *minimum inefficiency subgraph*: they show that the problem is NP-hard and develop an efficient greedy algorithm. The minimum inefficiency subgraph is not required to be connected: as such one could consider the query vertices that ends up disconnected as outliers. None of these three approaches enforces an upper bound on the number of outliers that be dropped, while no theoretical guarantee is provided. Instead our work studies the general problem with the input parameter on the maximum number of allowed outliers, and presents algorithms with strong theoretical guarantees.

Problems studied and results. In this paper, we study the *community search with outliers* problem where we are allowed to drop up to k query vertices from the input graph, with k being provided in input. We focus on the most widely-used cohesiveness functions considered in the literature, such as the *diameter* of the output graph [12], and its *minimum degree* [16], while studying their generalization to the case with outliers. By optimizing one measure and using the other as a constraint we obtain two optimization problems. Our work is the first one to propose algorithms with strong theoretical guarantees for the problem of community search with outliers. More in details:

- For the problem of minimizing the diameter under a minimum degree constraint when at most k query vertex can be dropped, we develop a 2-approximation algorithm. When k is a constant and there is no constraint on the minimum degree our algorithm boasts almost linear time in the size of the input graph (modulo a logarithmic factor). We then show that there is no algorithm with an approximation guarantee better than 2, unless $P=NP$, making our result tight.
- For the problem of maximizing the minimum degree under a diameter constraint when at most k query vertex can be dropped, we show that it NP-hard and we develop a bicriteria algorithm with approximation guarantees.
- Finally, we also study a variant where, instead of a constraint on the diameter (i.e., maximum distance among

any pair of vertices in the solution, regardless they are query vertices or not), we have a constraint on the maximum distance between a query vertex and any other vertex in the solution subgraph H . We will show that a solution for such a problem can be found efficiently in polynomial time.

Our theoretical results are complemented with an extensive experimental evaluation on real-world graphs, confirming the fact that by allowing outliers to be removed, much more cohesive solutions can be found. We also assess the running time of our algorithms on large real-world graphs while we show their effectiveness in identifying outlier query vertices.

II. PROBLEM DEFINITIONS AND COMPLEXITY

We are given an undirected graph $G = (V_G, E_G)$, a set of query vertices $Q \subseteq V_G$, and a positive integer $k < |Q|$. The shortest-path distance $d_G(u, v)$ between two vertices u, v in G is defined as the length of the shortest path connecting u, v in G . In the case when u, v are not in a same connected component we let $d(u, v) := +\infty$. The diameter of a graph G , denoted with $diam(G)$, is defined as the maximum shortest-path distance between any two vertices in G . In this section, we shall define the main problems studied in the rest of the paper. We call our problems CSO-MAXMINDEG-DIAM, CSO-MAXMINDEG-DIST, CSO-MINDIAM-MINDEG, where CSO indicates the general problem (community search with outliers), the first part after '-' indicates the objective function (e.g. MINDIAM), while the last part indicates the main constraint (e.g. MINDEG). We first consider the variant where we wish to minimize the diameter while satisfying a constraint on the minimum degree. The problem of minimizing the diameter has been studied in [12] (without outliers). In our work, we use a similar proof strategy, while following the framework developed in [16].

Problem 1 (CSO-MINDIAM-MINDEG): Given a graph $G = (V_G, E_G)$, a set of query nodes $Q \subseteq V_G$, an integer $k \geq 0$ such that $k \leq |Q| - 1$, an integer $\delta_{min} \geq 0$, find an induced subgraph $H = (V_H, E_H)$ of G such that:

- 1) H is connected;
- 2) $|V_H \cap Q| \geq |Q| - k$;
- 3) the minimum degree in H is at least δ_{min} ;
- 4) $diam(H)$ is minimized among all subgraphs satisfying 1-3;

Theorem 1: CSO-MINDIAM-MINDEG (Problem 1) is NP-hard even when $\delta_{min} = 1$.

Proof: We reduce the decision version of Maximum Clique problem (MDC) to CSO-MINDIAM-MINDEG. Notice that a graph is a clique if and only if its diameter is one. Given a graph $G = (V_G, E_G)$ and an integer h , we wish to determine whether G contains a clique of size at least h . We construct an instance of MDC problem with $Q = V_G$ and $h = |Q| - k$. Let $H = (V_H, E_H)$ be the solution of MDC in this instance. By definition of MDC, $V_H \subseteq V_G$ and $|V_H| \geq |Q| - k$. If $diam(H) = 1$, since H is an induced subgraph of G , G contains a clique of size $h = |Q| - k$. On the other hand,

consider the case when $\text{diam}(H) > 1$. Since, by definition, H is the subgraph with size $|Q| - k$ with minimum diameter, this means that G does not contain any clique of size at least $h = |Q| - k$. ■

From the fact that the diameter is an integer, it follows that any approximation algorithm with an approximation guarantee strictly better than 2 would compute an optimum solution to the MDC. The following corollary follows.

Corollary 1: CSO-MINDIAM-MINDEG cannot be approximated in polynomial time within a factor of $(2 - \epsilon)$ for any $\epsilon > 0$, unless $\mathbf{P}=\mathbf{NP}$.

We next consider the community search problem where the objective is to find a subgraph with maximum minimum degree while satisfying a constraint on its diameter and being allowed to remove up to k query vertices. We provide the following formal definition.

Problem 2 (CSO-MAXMINDEG-DIAM): Given a graph $G = (V_G, E_G)$, a set of query vertices $Q \subseteq V_G$, an integer $k \geq 0$ such that $k \leq |Q| - 1$, an integer $\text{diam}_{\max} \geq 1$, find an induced subgraph $H = (V_H, E_H)$ of G such that:

- 1) H is connected;
- 2) $|V_H \cap Q| \geq |Q| - k$;
- 3) $\text{diam}(H) \leq \text{diam}_{\max}$
- 4) the minimum degree is maximized among all subgraphs satisfying 1-3;

Theorem 2: CSO-MAXMINDEG-DIAM (Problem 2) is NP-hard even when $\text{diam}_{\max} = 1$.

We omit the proof due to the space limit: we will present it in the extended version. For Problem 2 we provide a bicriteria algorithm (Algorithm 3, Theorem 6) in Section III. Finally, we also consider a variant where there is a constraint on the maximum distance between a query vertex and any other vertex in the solution subgraph H . We will show that a solution for such a problem can be found efficiently in polynomial time.

Problem 3 (CSO-MAXMINDEG-DIST): Given a graph $G = (V_G, E_G)$, a set of query vertices $Q \subseteq V_G$, an integer $k \geq 0$ such that $k \leq |Q| - 1$, an integer $d_{\max} \geq 1$, find an induced subgraph $H = (V_H, E_H)$ of G such that:

- 1) H is connected;
- 2) $|V_H \cap Q| \geq |Q| - k$;
- 3) for any $v \in V_H$, $d(V_H \cap Q, v) := \min_{q \in V_H \cap Q} d(q, v) \leq d_{\max}$
- 4) the minimum degree is maximized among all subgraphs satisfying 1-3;

For Problem 3 we provide a polynomial time exact algorithm (Algorithm 2, Theorem 5) in Section III.

III. ALGORITHMS AND ANALYSIS

In this section we describe the algorithms for the three problems described in the previous section.

Minimizing the Diameter with Degree Constraint. We develop an algorithm for computing an approximate solution for CSO-MINDIAM-MINDEG, where the objective is to minimize the diameter of the output graph while enforcing

a lower bound on the minimum degree of the vertices. Our algorithm follows the framework developed in [16], however, the variant with outliers has not been studied before, to the best of our knowledge. The main procedure of our algorithm (Algorithm 1) receives a vertex $q \in Q$ in input and computes a feasible subgraph (if any) containing q with “small” diameter. In particular, we shall show that if q belongs to an optimum solution O for CSO-MINDIAM-MINDEG then our algorithm computes a 2-approximation for such a problem. It proceeds as follows. Let $G_0 := G$. At each step $t \geq 1$ the graph G_t is obtained from G_{t-1} as follows. If there is a vertex violating the constraint on the degree, then such a vertex and all its edges are removed from G_{t-1} . Otherwise, a vertex at maximum distance from q (and all its edges) is removed from G_{t-1} . Ties are broken arbitrarily. The algorithm terminates as soon as the set of edges is empty or more than k query vertices have been removed, at which step it computes the graph H with minimum diameter among all G_t ’s that are feasible. If none of the G_t ’s is feasible, it returns “unfeasible”. A pseudocode of our algorithm is shown in Algorithm 1.

Algorithm 1 ALGO-MINDIAM-MINDEG (q)

Input: $G = (V_G, E_G)$, $Q \subseteq V$, $\delta_{\min} \geq 0$, $q \in Q$

Output: a subgraph $H \subseteq G$

- 1: $G_0 \leftarrow G$, $t \leftarrow 0$
 - 2: **while** $E(V_t) \neq \emptyset$ and $|V_t \cap Q| \geq |Q| - k$ **do**
 - 3: $t \leftarrow t + 1$
 - 4: **if** there is v in G_{t-1} with $\delta(v) < \delta_{\min}$ **then**
 - 5: let v be such a node
 - 6: **if** $v = q$ **break**
 - 7: **else**
 - 8: let v be a vertex with maximum $d_{G_{t-1}}(v, q)$
 - 9: $G_t \leftarrow (V_{t-1} \setminus \{v\}, E(V_{t-1} \setminus \{v\}))$
 - 10: **if** there is no feasible graph among all G_t ’s return *unfeasible*
 - 11: **else** return a graph with min max distance from q among all feasible G_t ’s.
-

The following lemma proves the approximation guarantees of our algorithm.

Lemma 3: If there is no feasible solution for CSO-MINDIAM-MINDEG, Algorithm 1 returns *unfeasible*. Otherwise, let $O = (V_O, E_O)$ be an optimum solution for CSO-MINDIAM-MINDEG. If $q \in V_O$, then Algorithm 1 computes a 2-approximation solution for CSO-MINDIAM-MINDEG.

We omit the proof due to the space limit. The following theorem states our main result.

Theorem 4: Running Algorithm 1 for each of $k + 1$ query vertices chosen arbitrarily, while outputting a feasible solution with minimum diameter computes a 2-approximation algorithm for the CSO-MINDIAM-MINDEG, while requiring $O(k \cdot \max(|V_G| + |E_G|, |V_G| \log |V_G|))$ operations when $\delta_{\min} = 1$ and $O(|V_G| \cdot (|V_G| + |E_G|))$ operations in the general case.

Proof: The 2-approximation guarantee follows from Lemma 3 and from the fact that at least one of the $k+1$ query vertices considered by ALGO-MINDIAM-MINDEG(q) must belong to an optimum solution (if any). To bound the number of operations required by the algorithm, observe that we need exactly one BFS for each of the $k+1$ query vertices in the case when there is no constraint on the minimum degree. This follows from the fact that vertices are deleted in non-increasing order of distance from q , which implies that the distance between any remaining vertex and q is not affected. Therefore, the vertices can be sorted upfront according to their distance from q , requiring $O(k \cdot \max(|V_G| + |E_G|, |V_G| \log |V_G|))$ operations in this special case and $O(|V_G| \cdot (|V_G| + |E_G|))$ operations in the general case. ■

Observe that the running time of our algorithm is almost linear (modulo a logarithmic factor) when there is no constraint on the minimum degree and k is a constant. Otherwise, $O(|V_G| \cdot (|V_G| + |E_G|))$ operations are needed in the worst case, however, the actual number of operations required in real-world graphs are significantly smaller since we can remove in batch all the vertices v with $\delta(v) < \delta_{\min}$.

Maximizing Minimum Degree with Distance Constraint.

We develop an algorithm for computing an optimum solution to CSO-MAXMINDEG-DIST. Let $G_0 := G$. At each step $t \geq 1$ the graph G_t is obtained from G_{t-1} as follows. For any vertex v in $G_{t-1} := (V_{t-1}, E_{t-1})$, let Q_v be the set of query vertices in the same connected component of v in G_{t-1} . If there is a vertex v such that $|Q_v| < |Q| - k$ or $d_H(v, Q_v) := \min_{q \in Q_v} d_H(v, q) > d_{\max}$, then remove v and all its edges. Otherwise, remove a vertex v with minimum degree in G_{t-1} , breaking ties arbitrarily. The algorithm terminates when G_t becomes empty, at which step it computes the graph H with maximum minimum degree among all feasible connected components of all G_t 's. If there are several graphs H with the same maximum minimum degree then it considers the graph H with the largest number of query vertices. It then produces in output a connected component in H . If none of the G_t 's is feasible, it returns “unfeasible”. A pseudocode of our algorithm is shown in Algorithm 2. The following theorem proves the optimality of our algorithm.

Theorem 5: If there is no feasible solution for CSO-MAXMINDEG-DIST, Algorithm 2 returns unfeasible. Otherwise it computes an optimum solution for CSO-MAXMINDEG-DIST.

Proof: If there is no feasible solution then the algorithm outputs unfeasible, in that, the algorithm makes sure that only feasible solutions are produced in output. If there is a feasible solution, then there must be a smallest step t such that there is $v \in O$ with $v \in V_t \setminus V_{t+1}$, as the algorithm eventually removes all vertices in the graph. Let Q_v be the set of query vertices in the same connected component of v in G_t . It holds that a) $|Q_v| \geq |Q| - k$ and b) $d_{G_t}(v, Q_v) \leq d_{\max}$. This follows from the facts that O is an induced subgraph of G_t , $d_O(u, v) \geq d_{G_t}(u, v)$ for all u, v in O , and any connected component in O is also connected in G_t . As a result, every vertex in G_t satisfies

Algorithm 2 ALGO-MAXMINDEG-DIST

Input: $G = (V_G, E_G)$, $Q \subseteq V$, $d_{\max} > 0$

Output: a subgraph H of G

- 1: $G_0 \leftarrow G$, $t \leftarrow 0$
 - 2: **while** $E(V_t) \neq \emptyset$ **do**
 - 3: $t \leftarrow t + 1$
 - 4: Let Q_v be the set of query vertices in the same connected component of v in G_{t-1} .
 - 5: **if** there is a vertex v such that $|Q_v| < |Q| - k$ or $d_{G_{t-1}}(v, Q_v) > d_{\max}$ **then**
 - 6: let v be such a node
 - 7: **else**
 - 8: let v be a vertex with minimum degree in G_t
 - 9: $G_t \leftarrow (V_{t-1} \setminus \{v\}, E(V_{t-1} \setminus \{v\}))$
 - 10: **If** there is no feasible graph among all G_t 's return *unfeasible*
 - 11: **else** return a subgraph with maximum minimum degree among all feasible connected components of all G_t 's
-

a) and b), for otherwise the algorithm would have removed a vertex violating at least one such a constraint. Moreover, v has minimum degree in G_t . Let C be the connected component in G_t containing O . The following chain of inequalities proves the optimality of C :

$$\delta^{\min}(O) \leq \delta_O(v) \leq \delta_C(v) = \delta^{\min}(C),$$

where the first inequality follows from $v \in O$, the second inequality follows from the fact that O is contained in C , while the equality follows from the way v is chosen. ■

Implementation Details. Our algorithm requires to compute a BFS at each step, which might be cumbersome in the case when the input graph is large. Such a problem can be alleviated by a preprocessing step which proceeds as follows. First, we run a BFS starting from each of the query vertices and compute the minimum distance between every vertex in the input graph and the query vertices. Then, all vertices violating the distance constraints are removed all together. It is easy to see that none of those vertices can be part of any optimum solution. We shall refer to such a preprocessing step as *pruning*. After that, Algorithm 2 is executed in the resulting graph. In most cases of interest, the pruning phase allows to filter out a large fraction of vertices in the input graph, allowing Algorithm 2 to be executed efficiently. Let $H = (V_H, E_H)$ be the graph obtained after the pruning phase. The running time of the pruning phase is $O(|Q| \cdot (|V_G| + |E_G|))$, while the running time of our algorithm after pruning is $O(|Q| \cdot (|V_H| + |E_H|) \cdot |V_H|)$. Our experiments on real-world data show that our algorithm is much more efficient in practice than what the worst-case analysis suggests.

Maximizing Minimum Degree with Diameter Constraint.

We develop an algorithm for CSO-MAXMINDEG-DIAM where the objective is to maximize the minimum degree subject to a constraint on the diameter of the graph. The pseudocode of the main procedure is shown in Algorithm 3.

Similarly to the algorithm for CSO-MAXMINDEG-DIAM, we run Algorithm 3 for each of $k+1$ query vertices (chosen

Algorithm 3 ALGO-MAXMINDEG-DIAM(q)**Input:** $G = (V_G, E_G), Q \subseteq V, diam_{max} \geq 1, q \in Q$ **Output:** a subgraph H of G

- 1: $G_0 \leftarrow G, t \leftarrow 0$
- 2: **while** $E(V_t) \neq \emptyset$ and $|V_t \cap Q| \geq |Q| - k$ **do**
- 3: $t \leftarrow t + 1$
- 4: **if** there is v in G_{t-1} with $d(q, v) > diam_{max}$ **then**
- 5: let v be such a node
- 6: **else**
- 7: let v be a vertex with minimum degree in G_{t-1}
- 8: **if** $v = q$ **break**
- 9: $G_t \leftarrow (V_{t-1} \setminus \{v\}, E(V_{t-1} \setminus \{v\}))$
- 10: **If** there is no feasible graph among all G_t 's return *unfeasible*
- 11: **else** return a graph with max min degree among all feasible G_t 's.

arbitrarily) and output the solution with max min degree. The following theorem states the approximation guarantees of our algorithm.

Theorem 6: If there is no feasible solution for CSO-MAXMINDEG-DIAM, Algorithm 3 returns *unfeasible*. Otherwise, let $O = (V_O, E_O)$ be an optimum solution for CSO-MAXMINDEG-DIAM and let H be the solution computed by our algorithm for CSO-MAXMINDEG-DIAM. It holds that 1) $\delta_{min}(H) \geq \delta_{min}(O)$, and 2) $diam(H) \leq 2 \cdot diam(O)$.

The proof is similar to the proof of Theorem 4 and it is omitted for space constraints. Our algorithm can be seen as a bicriteria algorithm (see for example [10]). That is, an algorithm for a problem with multiple objective functions, which in our case are the minimum degree and the diameter of the output graph. The worst case complexity is $O(k \cdot |V_G| \cdot (|V_G| + |E_G|))$ since, for each deleted vertex we should recompute all the distances from q , for each query vertex q . However, if $\max\{d(q, v)\} > diam_{max}$, we can speed up the computation removing in batch all the vertices v such that $d(q, v) > diam_{max}$.

IV. EXPERIMENTS

In this section we assess the efficiency and the effectiveness of the three community search with outliers methods. In particular we focus our analysis on the following main aspects:

- 1) the capability of CSO-MAXMINDEG-DIAM, CSO-MAXMINDEG-DIST and CSO-MINDIAM-MINDEG to detect query vertices that are outliers;
- 2) comparison with the Minimum Inefficiency Subgraph proposed in [14] in the outlier classification task;
- 3) characteristics of the solution subgraphs produced, varying the parameter k , i.e., the number of query vertices that we are allowed to drop, and varying the constraints;
- 4) characteristics of the solution subgraphs produced varying the query set Q w.r.t. its size, and the distance among its vertices;
- 5) runtime.

For our purposes we use four real-world networks described in Table I.

TABLE I
CHARACTERISTICS OF DATASETS USED.

	$ V $	$ E $	density	avg deg	diam.
amazon	334,863	925,872	1.6e-5	5.5	44
dblp	317,080	1,049,866	2.1e-5	6.62	23
youtube	1,138,499	2,990,443	4.6e-6	5.27	21
ljournal	3,997,962	34,681,189	4.3e-6	17.3	16

TABLE II

OUTLIERS IDENTIFICATION CAPABILITIES OF THE THREE ALGORITHMS ON THREE DATASETS, WITH $n = 10$ AND $m = k \in [1, 4]$. THE TABLE REPORTS THE AVERAGE NUMBER OF OUTLIERS DETECTED (AND THE PERCENTAGE W.R.T. THE TOTAL NUMBER OF OUTLIERS PRESENT). A DASH REPORTS AN EXPERIMENT WHICH DOES NOT HAVE A FEASIBLE SOLUTION IN AT LEAST 50% OF THE RUNS.

Dataset	Problem [Constraint]	$k = 1$	$k = 2$	$k = 3$	$k = 4$
dblp		1 (100%)	1.9 (95%)	2.8 (93.3%)	3.8 (99%)
amazon	CSO-MAXMINDEG-DIAM	0.9 (90%)	1.9 (95%)	2.9 (96.7%)	3.7 (97.9%)
youtube	[$diam_{max} = 4$]	0.6 (69%)	1.3 (65%)	1.6 (53.3%)	3.73 (93%)
dblp		0.5 (50%)	1.3 (65%)	1.8 (60%)	2.7 (67.5%)
amazon	CSO-MAXMINDEG-DIST	0.9(90%)	1.9 (95%)	2.9 (96.7%)	3.9 (97.5%)
youtube	[$d_{max} = 2$]	0.4 (40%)	0.6 (30%)	0.9 (30%)	1.4 (35%)
dblp		0.3 (30%)	0.8 (40%)	1.2 (40%)	1.8(45%)
amazon	CSO-MINDIAM-MINDEG	-	0.7(35%)	0.9 (30%)	1.4 (35%)
youtube	[$\delta_{min} = 3$]	0.5(90%)	1.0 (50%)	1.5 (50%)	2.0 (50%)

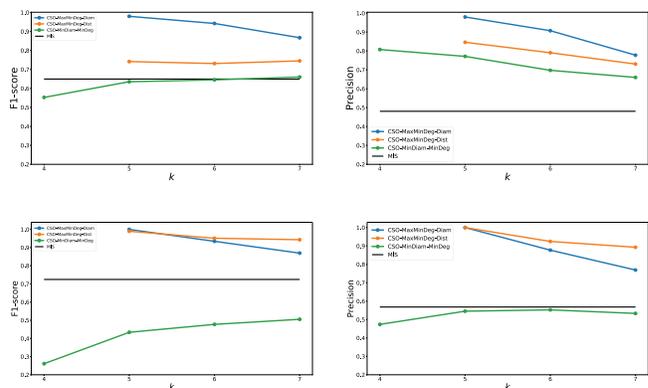


Fig. 1. Average F1-Score and average Precision of the solutions on dblp (top) and amazon (bottom) for CSO-MINDIAM-MINDEG ($diam_{max} = 3$), CSO-MAXMINDEG-DIST ($d_{max} = 2$), CSO-MAXMINDEG-DIAM ($\delta_{min} = 3$) and MIS.

All datasets are publicly-available, undirected and unweighted: youtube¹ and ljournal¹ are social networks, dblp¹ is a collaboration network and amazon¹ is a co-purchasing network. All our datasets, come with auxiliary *ground-truth communities* information. This information is needed to create query workloads with known number of outliers (i.e., query vertices coming from a different community w.r.t. the majority of the other query vertices), as it will be described in details later. The largest dataset, ljournal, is used only to report runtime. All algorithms are implemented in C++, building on the open-source NetworKit framework². The experiments are conducted on a server equipped with 32 GB RAM and with a processor Intel 1.2 GHz CPU.

¹<http://snap.stanford.edu>

²<http://networkit.it.kit.edu/>

Capability to detect outliers. We first study the capability of CSO-MAXMINDEG-DIAM, CSO-MAXMINDEG-DIST and CSO-MINDIAM-MINDEG to detect outliers. We randomly select ten vertices (i.e. $n = 10$) inside the same community (randomly chosen). Then, select m vertices at random from different communities w.r.t. the one from which the first n query vertices have been selected. As these communities are ground-truth communities, we can consider the vertices coming from different communities as unrelated, and if there is a dominating group of vertices (the n vertices chosen from a unique community) then the others can be considered ground-truth outliers that we can control. In this experiment vary m in the range $[1, 4]$ and we set $m = k$: in this way, the number of effective outliers (i.e. m) is equal to the number of the query vertices that we are allowed to drop (i.e. k). We generate 20 query sets following the process just described and we report the average value. Since some query sets may not have any solution for some problem and for some constraint, we report the value only the first 10 sets return a solution. Table II reports the average number of outliers detected (and the percentage w.r.t. the total number of outliers present) for the three algorithms under three datasets. We can observe how in general, all the algorithms perform quite well in detecting the outliers injected in the query sets, with values ranging in between 30% and 100%.

Outliers classification task evaluation. We next compare against the Minimum Inefficiency Subgraph (MIS) studied in [14]. It is important to note that the problem in [14] is totally parameterless, so that it decides automatically the number of outliers. In particular, the formulation of [14] always requires all the query vertices to be part of the solution, but on the other hand, it allows the solution subgraph to be *disconnected*. This way, query vertices which are not in the largest connected component (LCC) are considered outliers. We select $n = 10$ query vertices inside the same community (with maximum distance among them equal to 4) and $m = 5$ outliers. We run the Minimum Inefficiency Subgraph and CSO-MINDIAM-MINDEG ($d_{max} = 3$), CSO-MAXMINDEG-DIST ($d_{max} = 2$), CSO-MAXMINDEG-DIAM ($\delta_{min} = 3$) on *dblp* and *amazon* varying the parameter $k \in [4, 7]$. We repeat the experiment 10 times. From every MIS solution, we extract the largest connected component (LCC) to identify the query vertices which *are not* in the LCC and mark them as outliers. We compute the average F1-score and the average precision varying the parameter k and we report the results in Figure 1. Since MIS does not require any parameter in input, its score is constant for every k . We notice that, for $k \geq 5$, CSO-MINDIAM-MINDEG and CSO-MAXMINDEG-DIST achieve a better F1-score and a better precision on both *dblp* and *amazon*. On the other hand, CSO-MAXMINDEG-DIAM has worst performance since it returns larger solutions than the other methods due to the fact that *amazon* and *dblp* networks have a low average degree. For $k = 4$ CSO-MINDIAM-MINDEG and CSO-MAXMINDEG-DIST does not return any solution for at least one query set of vertices:

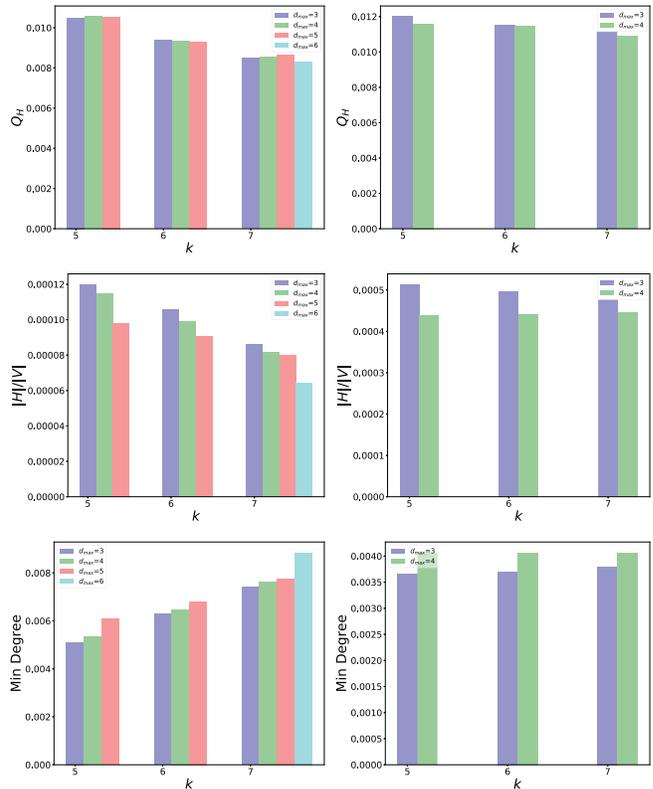


Fig. 2. Characteristics of the solution on *youtube* (left) and *amazon* (right) for CSO-MINDIAM-MINDEG with $k \in \{4, 6, 8\}$ and the minimum degree threshold δ_{min} in the range $[3, 6]$.

however in this case it is easy to see that the average F1-score is greater or equal to 0.5 since every vertex is classified as outlier.

Characterization varying constraints and k . We next study how the solutions change varying the number of allowed outlier vertices k . In this set of experiments we no longer set k to the known number of outliers in the query set. The query workload generation is as follows. We randomly pick a community and we select randomly $n = 10$ vertices from the community, with distance at most 4 from each other. Then we select $m = 4$ outliers randomly in the whole network. As before, we generate 20 query sets and we report the average value of the first 10 query sets. In particular we report three measures:

- 1) the number Q_H of query vertices contained in the solution subgraph H ;
- 2) the relative size of H (i.e the ratio between the number of vertices in H ($|H| = |V_H|$) divided the number of vertices in G ($|G| = |V_G|$);
- 3) the minimum degree within H .

The results are reported in Figure 2 for CSO-MINDIAM-MINDEG, Figure 3 for CSO-MAXMINDEG-DIAM, and Figure 4 for CSO-MAXMINDEG-DIST, over two datasets: *youtube* and *dblp*. For CSO-MINDIAM-MINDEG we vary the minimum degree threshold δ_{min} in the range $[3, 6]$;

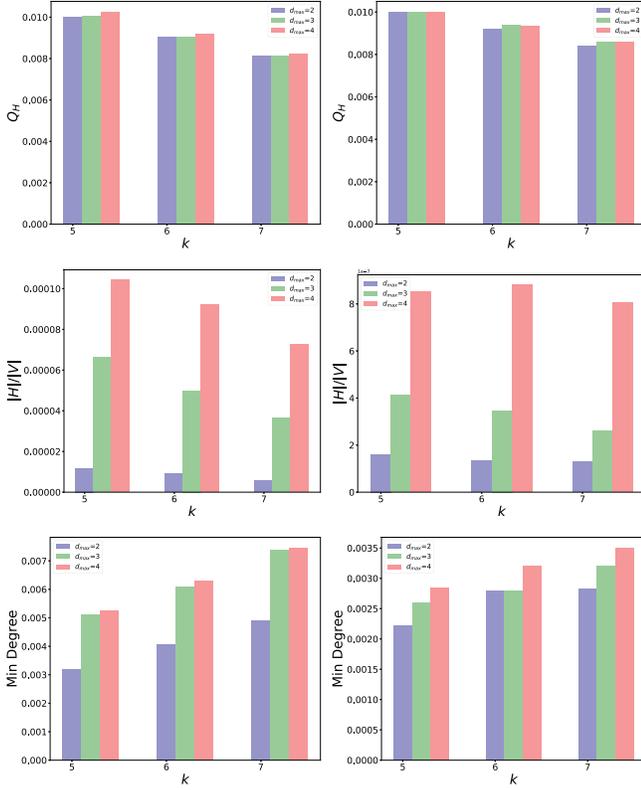


Fig. 3. Characteristics of the solution on youtube (left) and amazon (right) for CSO-MAXMINDEG-DIAM with $k \in \{4, 6, 8\}$ and the maximum diameter threshold $diam_{max}$ in the range $[2, 4]$.

for CSO-MAXMINDEG-DIAM we vary the maximum diameter threshold $diam_{max}$ in the range $[2, 4]$; for CSO-MAXMINDEG-DIST we vary the threshold on the maximum distance to the query vertices in $[1, 4]$. The first high-level observation is that, for all the three problems studied and under all settings, *increasing the number of allowed outliers produces “better” (i.e., more cohesive) solutions*: smaller in size, denser, with higher minimum degree. This was the starting motivation at the basis of this work. Similar observations can be done w.r.t. the selectivity of constraints. For instance, for CSO-MINDIAM-MINDEG (Figure 2), the higher is the minimum degree threshold δ_{min} the more cohesive the solutions found: i.e., denser and smaller. For CSO-MAXMINDEG-DIAM (Figure 3), the smaller is the maximum diameter threshold $diam_{max}$, the more cohesive the solutions found. Finally for CSO-MAXMINDEG-DIST (Figure 4) imposing a smaller maximum distance to the query vertices, produces more cohesive solutions.

Characterization with different query sets. We next study how the solutions change varying the characteristics of the query set Q : in particular the number of query vertices and the distance among query vertices. In Figure 5 we select $n = 10$ vertices in the same community and $m = 5$ outliers and we varying the distance among the query vertices. We set the input parameter $k = 5$. We notice that the increase the distance has

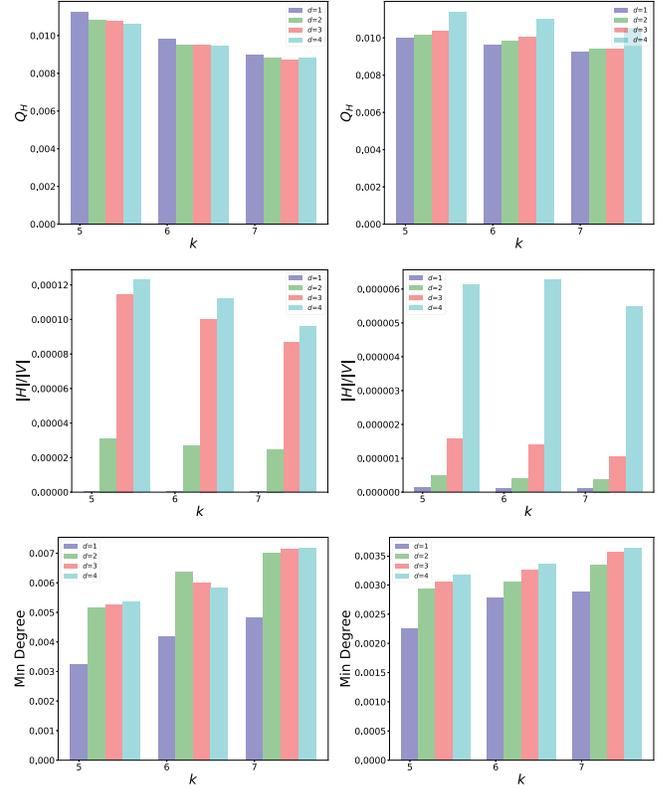


Fig. 4. Characteristics of the solution on youtube (left) and amazon (right) for CSO-MAXMINDEG-DIST with $k \in \{4, 6, 8\}$ and the threshold on the maximum distance to the query vertices in $[1, 4]$.

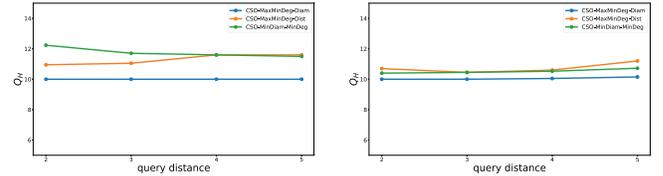


Fig. 5. Characteristics of the solution on amazon (left) and dblp (right) graphs with $k = 5$ varying the maximum distance among the query vertices inside the same community for CSO-MINDIAM-MINDEG with $\delta_{min} = 4$, CSO-MAXMINDEG-DIAM with $diam_{max} = 4$, and CSO-MAXMINDEG-DIST with $d_{max} = 4$.

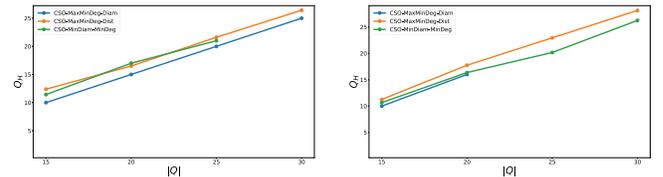


Fig. 6. Characteristics of the solution on amazon (left) and dblp (right) graphs with $k = 5$ varying the number of query vertices inside the same community for CSO-MINDIAM-MINDEG with $\delta_{min} = 4$, CSO-MAXMINDEG-DIAM with $diam_{max} = 4$, and CSO-MAXMINDEG-DIST with $d_{max} = 4$.

a low effect on the value of Q_H . In Figure 6 we set $m = 5$ and we vary the parameter n selecting different sets of query

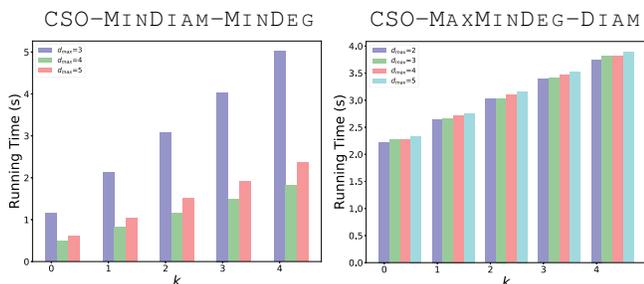


Fig. 7. Runtime on amazon graph with $n = 5$ and $m = k \in [0, 4]$ for different values of the constraints.

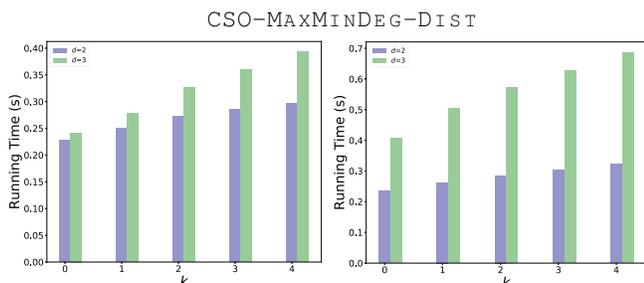


Fig. 8. Runtime on amazon (left) and dblp (right) graph with $n = 5$ and $m = k \in [0, 4]$ for different values of the constraints.

vertices (without any distance constraint): in particular we use $n \in \{10, 15, 20, 25\}$. On the x -axis we report the size of the query set $|Q| = n + m$. In this case the number of query vertices in the solution grows linearly with respect to the total number of query in input.

Runtime and parallelization Finally, we report runtime performance of the three algorithms in Figure 7 and in Figure 8. We can observe that the runtime generally grows with the value of k . Moreover, for all the three problems as the constraints are more selective (and thus the solutions more compact), the runtime decreases. We also tested runtime on the largest dataset, i.e., `1journal`, containing approximately 4M vertices and 34M edges. For all problems we produce the same workload of 20 query sets using $n = 5$ vertices belonging to the same community, $m = 4$ vertices belonging to different communities, and setting the input parameter $k = 4$. For CSO-MINDIAM-MINDEG ($\delta_{min} = 20$) the runtime was 8333.5 seconds (slightly more than 2 hours), while for CSO-MAXMINDEG-DIAM ($diam_{max} = 2$) the runtime was 56.8 seconds, and for CSO-MAXMINDEG-DIST ($d_{max} = 2$) was 21.63 seconds. We can improve the scalability of the algorithms for CSO-MINDIAM-MINDEG (respectively, CSO-MAXMINDEG-DIAM) by straightforward parallelization of the subgraph extraction in the following way: we assign to each processor a query vertex, run ALGO-MAXMINDEG-DIAM (respectively ALGO-MINDIAM-MINDEG) in parallel on each processor and compute the best solution among all the processors.

V. CONCLUSIONS

We study several variants of the community search problem, where we are allowed to drop up to k query vertices. We adopt three measures of cohesiveness: the minimum degree, the diameter, and the maximum distance with a query vertex. This leads to the formulation of three natural optimization problems, for each of which we either develop an efficient exact algorithm or prove its hardness and develop an approximation algorithm. Our work is the first one to propose algorithms with strong theoretical guarantees for the problem of community search with outliers. Our experimental evaluation shows the effectiveness of our approach in identifying outlier query vertices, while showing that much more cohesive solutions can be found when we are allowed to remove outliers.

Acknowledgments. Francesco Bonchi acknowledges support from Intesa Sanpaolo Innovation Center. Lorenzo Severini is with UniCredit Group. Mauro Sozio has been working in the frame of a cooperation between Huawei Technologies France SASU and Telecom Paris (Grant no. YBN2018125164). The content of this paper and the views expressed here, are solely responsibility of the authors. The funders had no role in study design, preparation of the manuscript or decision to publish.

REFERENCES

- [1] L. Akoglu, D. H. Chau, C. Faloutsos, N. Tatti, H. Tong, J. Vreeken, and L. Tong. Mining connection pathways for marked nodes in large graphs. In *SDM*, 2013.
- [2] N. Barbieri, F. Bonchi, E. Galimberti, and F. Gullo. Efficient and effective community search. *DAMI*, 29(5), 2015.
- [3] Y. Bian, Y. Yan, W. Cheng, W. Wang, D. Luo, and X. Zhang. On multi-query local community detection. pages 9–18, 2018.
- [4] W. Cui, Y. Xiao, H. Wang, and W. Wang. Local search of communities in large graphs. In *SIGMOD*, pages 991–1002, 2014.
- [5] C. Faloutsos, K. S. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *KDD*, 2004.
- [6] Y. Fang, R. Cheng, Y. Chen, S. Luo, and J. Hu. Effective and efficient attributed community search. 26(6):803–828, 2017.
- [7] Y. Fang, R. Cheng, X. Li, S. Luo, and J. Hu. Effective community search over large spatial graphs. 10(6):709–720, 2017.
- [8] E. Galimberti, M. Ciaperoni, A. Barrat, F. Bonchi, C. Cattuto, and F. Gullo. Span-core decomposition for temporal networks: Algorithms and applications. *TKDD*, 2020.
- [9] A. Gionis, M. Mathioudakis, and A. Ukkonen. Bump hunting in the dark: Local discrepancy maximization on graphs. In *ICDE*, 2015.
- [10] F. Grandoni, J. Könemann, A. Panconesi, and M. Sozio. A primal-dual bicriteria distributed algorithm for capacitated vertex cover. *SIAM J. Comput.*, 38(3):825–840, 2008.
- [11] X. Huang and L. V. Lakshmanan. Attribute-driven community search. 10(9):949–960, 2017.
- [12] X. Huang, L. V. S. Lakshmanan, J. X. Yu, and H. Cheng. Approximate closest community search in networks. *PVLDB*, 9(4):276–287, 2015.
- [13] I. M. Kloumann and J. M. Kleinberg. Community membership identification from small seed sets. In *KDD*, 2014.
- [14] N. Ruchansky, F. Bonchi, D. García-Soriano, F. Gullo, and N. Kourtellis. To be connected, or not to be connected: That is the minimum inefficiency subgraph problem. In *CIKM 2017*.
- [15] N. Ruchansky, F. Bonchi, D. García-Soriano, F. Gullo, and N. Kourtellis. The minimum wiener connector problem. In *SIGMOD*, 2015.
- [16] M. Sozio and A. Gionis. The community-search problem and how to plan a successful cocktail party. In *KDD*, pages 939–948, 2010.
- [17] H. Tong and C. Faloutsos. Center-piece subgraphs: problem definition and fast solutions. In *KDD*, pages 404–413, 2006.
- [18] I. Tsalouchidou, F. Bonchi, and R. Baeza-Yates. Adaptive community search in dynamic networks. In *IEEE BigData 2020*.
- [19] Y. Yan, Y. Bian, D. Luo, D. Lee, and X. Zhang. Constrained local graph clustering by colored random walk. pages 2137–2146, 2019.