

Fast computation of distance-generalized cores using sampling

Nikolaj Tatti¹

¹HIIT, University of Helsinki, Finland.

Contributing authors: nikolaj.tatti@helsinki.fi;

Abstract

Core decomposition is a classic technique for discovering densely connected regions in a graph with large range of applications. Formally, a \mathbf{k} -core is a maximal subgraph where each vertex has at least \mathbf{k} neighbors. A natural extension of a \mathbf{k} -core is a (\mathbf{k}, \mathbf{h}) -core, where each node must have at least \mathbf{k} nodes that can be reached with a path of length \mathbf{h} . The downside in using (\mathbf{k}, \mathbf{h}) -core decomposition is the significant increase in the computational complexity: whereas the standard core decomposition can be done in $\mathcal{O}(\mathbf{m})$ time, the generalization can require $\mathcal{O}(\mathbf{n}^2 \mathbf{m})$ time, where \mathbf{n} and \mathbf{m} are the number of nodes and edges in the given graph. In this paper we propose a randomized algorithm that produces an ϵ -approximation of (\mathbf{k}, \mathbf{h}) core decomposition with a probability of $1 - \delta$ in $\mathcal{O}(\epsilon^{-2} \mathbf{h} \mathbf{m} (\log^2 \mathbf{n} - \log \delta))$ time. The approximation is based on sampling the neighborhoods of nodes, and we use Chernoff bound to prove the approximation guarantee. We also study distance-generalized dense subgraphs, show that the problem is **NP**-hard, provide an algorithm for discovering such graphs with approximate core decompositions, and provide theoretical guarantees for the quality of the discovered subgraphs. We demonstrate empirically that approximating the decomposition complements the exact computation: computing the approximation is significantly faster than computing the exact solution for the networks where computing the exact solution is slow.

Keywords: distance-generalized core decomposition, sampling, approximation algorithm, Chernoff bounds

1 Introduction

Core decomposition is a classic technique for discovering densely connected regions in a graph. The appeal of core decomposition is a simple and intuitive definition, and the fact that the core decomposition can be computed in linear time. Core decomposition has a large range of applications such as graph visualization [1], graph modeling [4], social network analysis [23], internet topology modeling [7], influence analysis [19, 27], bioinformatics [2, 16], and team formation [5].

More formally, a k -core is a maximal subgraph such that every vertex has at least k degree. We can show that k -core form a nested structure: $(k + 1)$ -core is a subset of k -core, and that the core decomposition can be discovered in linear time [3]. Core decomposition has been extended to directed [14], multi-layer [12], temporal [13], and weighted [24] networks.

A natural extension of core decomposition, proposed by Bonchi et al. [6], is a distance-generalized core decomposition or (k, h) -core decomposition, where the degree is replaced by the number of nodes that can be reached with a path of length h . Here, h is a user parameter and $h = 1$ reduces to a standard core decomposition. Using distance-generalized core decomposition may produce a more refined decomposition [6]. Moreover, it can be used when discovering h -clubs, distance-generalized dense subgraphs, and distance-generalized chromatic numbers [6].

Studying such structures may be useful in graphs where paths of length h reveal interesting information. For example, assume a authorship network, where an edge between a paper and a researcher indicate that the researcher was an author of the paper. Then paths of length 2 contain co-authorship information.

The major downside in using the distance-generalized core decomposition is the significant increase in the computational complexity: whereas the standard core decomposition can be done in $\mathcal{O}(m)$ time, the generalization can require $\mathcal{O}(n^2m)$ time, where n and m are the number of nodes and edges in the given graph.

To combat this problem we propose a randomized algorithm that produces an ϵ -approximation of (k, h) core decomposition with a probability of $1 - \delta$ in

$$\mathcal{O}\left(\frac{hm \log n/\delta}{\epsilon^2} \log \frac{n\epsilon^2}{\log n/\delta}\right) \subseteq \mathcal{O}(\epsilon^{-2}hm(\log^2 n - \log \delta))$$

time.

The intuition behind our approach is as follows. In order to compute the distance-generalized core decomposition we need to discover and maintain h -neighborhoods for each node. We can discover the h -neighborhood of a node v by taking the union of the $(h - 1)$ -neighborhood of the adjacent nodes, which leads to a simple dynamic program. The computational bottleneck comes from the fact that these neighborhoods may become too large. So, instead of computing the complete neighborhood, we have a carefully selected budget M .

The moment the neighborhood becomes too large, we delete (roughly) half of the nodes, and to compensate for the sampling we multiply our size estimate by 2. This procedure is repeated as often as needed. Since we are able to keep the neighbor samples small, we are able to compute the decomposition faster.

We use Chernoff bounds to determine an appropriate value for M , and provide algorithms for maintaining the h -neighborhoods. The maintenance require special attention since if the h -neighborhood becomes too small we need to bring back the deleted nodes.

Finally, we study distance-generalized subgraphs, a notion proposed by Bonchi et al. [6] that extends a notion of dense subgraphs. Here the density is the ratio of h -connected node pairs and nodes. We show that the problem is **NP**-hard and propose an algorithm based on approximate core maps, extending the results by Bonchi et al. [6].

The rest of the paper is organized as follows. In Section 2 we introduce preliminary notation and formalize the problem. In Section 3 we present a naive version of the algorithm that yields approximate results but is too slow. We prove the approximation guarantee in Section 4, and speed-up the algorithm in Section 5. In Section 6 we study distance-generalized dense subgraphs. We discuss the related work in Section 7. Finally, we compare our method empirically against the baselines in Section 8 and conclude the paper with discussion in Section 9.

This work extends the conference paper [26].

2 Preliminaries and problem definition

In this section we establish preliminary notation and define our problem.

Assume an undirected graph $G = (V, E)$ with n nodes and m edges. We will write $A(v)$ to be the set of nodes adjacent to v . Given an integer h , we define an h -path to be a sequence of *at most* $h + 1$ adjacent nodes. An h -neighborhood $N(v; h, X)$ is then the set of nodes that are reachable with an h -path in a set of nodes X . If $X = V$ or otherwise clear from context, we will drop it from the notation. Note that $N(v; 1) = A(v) \cup \{v\}$.

We will write $\deg_h(v; X) = |N(v; h, X)| - 1$, where X is a set of nodes and $v \in X$. We will often drop X from the notation if it is clear from the context.

A k -core is the maximal subgraph of G for which all nodes have at least a degree of k . Discovering the cores can be done in $\mathcal{O}(m)$ time by iteratively deleting the vertex with the smallest degree [23].

Bonchi et al. [6] proposed to extend the notion of k -cores to (k, h) -cores. Here, given an integer h , a (k, h) -core is the maximal graph H of G such that $|N(v; h)| - 1 \geq k$ for each $v \in V(H)$, that is, we can reach at least k nodes from v with a path of at most h nodes. The *core number* $c(v)$ of a vertex v is the largest k such that v is contained in (k, h) -core H . We will call H as the *core graph* of v and we will refer to c as the *core map*.

Note that discovering $(k, 1)$ -cores is equal to discovering standard k -cores. We follow the same strategy when computing (k, h) -cores as with standard

cores: we iteratively find and delete the vertex with the smallest degree [6]. We will refer to the exact algorithm as EXACTCORE. While EXACTCORE is guaranteed to produce the correct result the computational complexity deteriorates to $\mathcal{O}(n^2m)$. The main reason here is that the neighborhoods $N(v; h)$ can be significantly larger than just adjacent nodes $A(v)$.

In this paper we consider approximating cores.

Definition 2.1 (approximative (k, h) -core). *Given a graph G an integer h and approximation guarantee ϵ , an ϵ -approximative core map $c' : V \rightarrow \mathbb{N}$ maps a node to an integer such that $|c'(v) - c(v)| \leq \epsilon c(v)$ for each $v \in V$.*

We will introduce an algorithm that computes an ϵ -approximative core map with a probability of $1 - \delta$ in quasilinear time.

3 Naive, slow algorithm

In this section we introduce a basic idea of our approach. This version of the algorithm will be still too slow but will approximate the cores accurately. We will prove the accuracy in the next section, and then refine the subroutines to obtain the needed computational complexity.

The bottleneck for computing the cores is maintaining the h -neighborhood $N(v; h)$ for each node v as we delete the nodes. Instead of maintaining the complete h -neighborhood we will keep only certain nodes if the neighborhood becomes too large. We then compensate the sampling when estimating the size of the h -neighborhood.

Assume that we are given a graph G , an integer h , approximation guarantee ϵ , and a probability threshold δ . Let us define numbers $C = \log(2n/\delta)$ and

$$M = 1 + \frac{4(2 + \epsilon)}{\epsilon^2}(C + \log 8) \quad . \quad (1)$$

The quantity M will act as an upper bound for the sampled h -neighborhood, while C will be useful when analyzing the properties of the algorithm. We will see later that these specific values will yield the approximation guarantees.

We start the algorithm by sampling the rank of a node from a geometric distribution $r[v] = \text{geo}(1/2)$. Note that ties are allowed. During the algorithm we maintain two key variables $B[v, i]$ and $k[v, i]$ for each $v \in V$ and each index $i = 1, \dots, h$. Here,

$$B[v, i] = \{u \in N(v; i) \mid r[u] \geq k[v, i]\}$$

is a subset of i -neighborhood $N(v; i)$ consisting of nodes whose rank $r[u] \geq k[v, i]$. The threshold $k[v, i]$ is set to be as small as possible such that $|B[v, i]| \leq M$.

We can estimate $c(v)$ from $B[v, h]$ and $k[v, h]$ as follows: Consider the quantity $d = |B[v, h] \setminus \{v\}|2^{k[v, h]}$. Note that for an integer k the probability of a

vertex v having a rank $r[v] \geq k$ is 2^{-k} . This hints that d is a good estimate for $c(v)$. We show in the next section that this is indeed the case but d is lacking an important property that we need in order to prove the correctness of the algorithm. Namely, d can increase while we are deleting nodes. To fix this pathological case we estimate $c(v)$ with $\max(d, M2^{k[v,h]-1})$ if $k[v, h] > 0$, and with d if $k[v, h] = 0$. The pseudo-code for the estimate is given in Algorithm 1.

Algorithm 1: ESTIMATE(v), estimates $|N(v, h)| - 1$ using $B[v, h]$ and $k[v, h]$.

```

1  $k \leftarrow k[v, h]$ ;
2  $d \leftarrow |B[v, h] \setminus \{v\}|2^k$ ;
3 if  $k > 0$  then  $d \leftarrow \max(d, M2^{k-1})$ ;
4 return  $d$ ;
```

To compute $B[v, i]$ we have the following observation.

Proposition 3.1. *For any $v \in V$ and any $i = 1, \dots, h$, we have*

$$B[v, i] = \{u \in T \mid r[u] \geq k[v, i]\}, \quad \text{where} \\ T = \{v\} \cup \{u \in B[w, i-1] \mid w \in A(v)\} \quad .$$

Moreover, $k[v, i] \geq k[w, i-1]$ for any $w \in A(v)$.

Proof Let $w \in A(v)$. Since $N(w, i-1) \subseteq N(v, i)$, we have $k[v, i] \geq k[w, i-1]$. Consequently, $B[v, i] \subseteq T \subseteq N(v, i)$, and by definition of $B[v, i]$, the claim follows. \square

The proposition leads to COMPUTE, an algorithm for computing $B[v, i]$ given in Algorithm 2. Here, we form a set T , a union of sets $B[w, i-1]$, where $w \in A(v)$. After T is formed we search for the threshold $k[v, i] \geq \max_{w \in A(v)} k[w, i-1]$ that yields at most M nodes in T , and store the resulting set in $B[v, i]$.

Algorithm 2: Naive version of COMPUTE(v, i). Recomputes $B[v, i]$ and $k[v, i]$ from scratch.

```

1  $T \leftarrow (t_j)_{j=1} \leftarrow \{v\} \cup \bigcup_{w \in A(v)} B[w, i-1]$  sorted by  $r[\cdot]$ ;
2  $k \leftarrow \max \{k[w, i-1] \mid w \in A(v)\}$ ;
3 if  $|T| > M$  then
4    $k \leftarrow \max(k, r[t_{\lfloor M \rfloor + 1}] + 1)$ ;
5  $B[v, i] \leftarrow \{u \in T \mid r[u] \geq k\}$ ;
6  $k[v, i] \leftarrow k$ ;
```

As the node, say u , is deleted we need to update the affected nodes. We do this update in Algorithm 3 by recomputing the neighbors $v \in A(u)$, and see if $B[v, i]$ and $k[v, i]$ have changed; if they have, then we recompute $B[w, i + 1]$ for all $w \in A(v)$, and so on.

Algorithm 3: Naive version of UPDATE(u). Deletes u and updates the affected $B[v, i]$ and $k[v, i]$.

```

1 delete  $u$  from  $G$ ;
2  $U \leftarrow \emptyset$ ;
3 foreach  $i = 1, \dots, h$  do
4   | add neighbors of  $u$  to  $U$ ;
5   |  $W \leftarrow \emptyset$ ;
6   | foreach  $v \in U$  do
7     | COMPUTE( $v, i$ );
8     | if  $B[v, i]$  or  $k[v, i]$  has changed then
9       |   | add neighbors of  $v$  in  $W$ ;
10    |   | if  $i = h$  then  $d[v] \leftarrow \text{ESTIMATE}(v)$  ;
11  |  $U \leftarrow W$ ;
```

The main algorithm CORE, given in Algorithm 4, initializes $B[v, i]$ using COMPUTE, deletes iteratively the nodes with smallest estimate $d[v]$ while updating the sets $B[v, i]$ with UPDATE.

Algorithm 4: CORE(G, ϵ, C) approximative core decomposition. Setting $C = \log(2n/\delta)$ yields an ϵ -approximation with $1 - \delta$ probability.

```

1 foreach  $v \in V$  do
2   |  $r[v] \leftarrow \text{geo}(1/2)$ ;
3   |  $B[v, 0] \leftarrow \{v\}$ ;
4   |  $k[v, 0] \leftarrow 0$ ;
5  $M \leftarrow 1 + \frac{4(2+\epsilon)}{\epsilon^2}(C + \log 8)$ ;
6 foreach  $i = 1, \dots, h$  do
7   | foreach  $v \in V$  do COMPUTE( $v, i$ ) ;
8  $c \leftarrow 0$ ;
9 while graph is not empty do
10  |  $u \leftarrow \arg \min_v d[v]$  (use  $k[v, h]$  as a tie breaker);
11  |  $c \leftarrow \max(c, d[u])$ ;
12  | output  $u$  with  $c$  as the core number;
13  | UPDATE( $u$ );
```

4 Approximation guarantee

In this section we will prove the approximation guarantee of our algorithm. The key step is to show that ESTIMATE produces an accurate estimate. For notational convenience, we need the following definition.

Definition 4.1. Assume d integers $X = (x_1, \dots, x_d)$ and an integer M . Define

$$S_i = |\{j \in [d] \mid x_j \geq i\}| \text{ and } T_i = |\{j \in [d] \mid x_j \geq i, j \geq 2\}|$$

to be the number of integers larger than or equal to i . Let $k \geq 0$ be the smallest integer for which $S_k \leq M$. Define

$$\Delta(X; M) = \begin{cases} \max(T_k 2^k, M 2^{k-1}), & \text{if } k > 0, \\ T_k 2^k, & \text{if } k = 0 \end{cases}.$$

Note that if $R = (r[w] \mid w \in N(v; h))$ with $r[v]$ being the first element in R , then $\Delta(R; M)$ is equal to the output of ESTIMATE(v).

Our first step is to show that $\Delta(X; M)$ is monotonic.

Proposition 4.1. Assume $M > 0$. Let x_1, \dots, x_d be a set of integers. Select $d' \leq d$. Then

$$\Delta(x_1, \dots, x_{d'}; M) \leq \Delta(x_1, \dots, x_d; M) \quad .$$

Note that this claim would not hold if we did not have the $M 2^{k-1}$ term in the definition of $\Delta(X; M)$.

Proof Let k , S_i , and T_i be as defined for $\Delta(x_1, \dots, x_d; M)$ in Definition 4.1. Also, let k' , S'_i , and T'_i be as defined for $\Delta(x_1, \dots, x_{d'}; M)$ in Definition 4.1.

Since $S'_i \leq S_i$, we have $k' \leq k$. If $k' = k$, the claim follows immediately since also $T'_i \leq T_i$. If $k' < k$, then

$$\Delta(x_1, \dots, x_d; M) \geq M 2^{k-1} \geq M 2^{k'} \geq T'_{k'} 2^{k'}$$

and

$$\Delta(x_1, \dots, x_{d'}; M) \geq M 2^{k-1} \geq M 2^{k'-1},$$

proving the claim. \square

Next we formalize the accuracy of $\Delta(X; M)$. We prove the claim in Appendix.

Proposition 4.2. Assume $0 < \epsilon \leq 1/2$. Let $\mathcal{R} = R_1, \dots, R_d$ be independent random variables sampled from geometric distribution, $\text{geo}(1/2)$. Assume $C > 0$ and define M as in Eq. 1. Then

$$|\Delta(\mathcal{R}; M) - (d-1)| \leq \epsilon(d-1) \tag{2}$$

with probability $1 - \exp(-C)$.

We are now ready to state the main claim.

Proposition 4.3. *Assume graph G with n nodes, $\epsilon > 0$, and $C > 0$. For each node $v \in V$, let $c(v)$ be the core number reported by EXACTCORE and let $c'(v)$ be the core number reported by CORE. Then with probability $1 - 2ne^{-C}$*

$$|c(v) - c'(v)| \leq \epsilon c(v),$$

for every node in V . Moreover, if $c(v) \leq M$, where M is given in Eq. 1, then $c(v) = c'(v)$.

We will prove the main claim of the proposition with two lemmas. In both proofs we will use the variable τ_v which we define to be the value of $d[v]$ when v is deleted by CORE.

The first lemma establishes a lower bound.

Lemma 4.1. *The lower bound $c'(v) \geq (1 - \epsilon)c(v)$ holds with probability $1 - ne^{-C}$.*

Proof For each node $v \in V$, let R_v be a rank, an independent random variable sampled from geometric distribution, $geo(1/2)$.

Let H_v be the core graph of v as solved by EXACTCORE. Define $S_v = N(v, h) \cap H_v$ to be the h -neighborhood of v in H_v . Note that $c(v) \leq |S_v| - 1$. Let \mathcal{R}_v be the list of ranks ($R_w; w \in S_v$) such that R_v is always the first element.

Proposition 4.2 combined with the union bound states that

$$|\Delta(\mathcal{R}_v; M) - (|S_v| - 1)| \leq \epsilon(|S_v| - 1) \quad . \quad (3)$$

holds with probability $1 - ne^{-C}$ for every node v . Assume that these events hold.

To prove the claim, select a node v and let w be the first node in H_v deleted by CORE. Let F be the graph right before deleting w by CORE. Then

$$\begin{aligned} c'(v) &\geq c'(w) && (\text{CORE picked } w \text{ before } v \text{ or } w = v) \\ &\geq \tau_w \\ &\geq \Delta(\mathcal{R}_w; M) && (H_w \subseteq H_v \subseteq F \text{ and Prop. 4.1}) \\ &\geq (1 - \epsilon)(|S_w| - 1) && (\text{Eq. 3}) \\ &\geq (1 - \epsilon)c(w) && (S_w = N(w, h) \cap H_w) \\ &\geq (1 - \epsilon)c(v), && (w \in H_v) \end{aligned}$$

proving the lemma. □

Next, we establish the upper bound.

Lemma 4.2. *The upper bound $c'(v) \leq (1 + \epsilon)c(v)$ holds with probability $1 - ne^{-C}$.*

Proof For each node $v \in V$, let R_v be an independent random variable sampled from geometric distribution, $\text{geo}(1/2)$.

Consider the exact algorithm EXACTCORE for solving the (k, h) core problem. Let H_v be the graph induced by the existing nodes right before v is removed by EXACTCORE. Define $S_v = N(v, h) \cap H_v$ to be the h -neighborhood of v in H_v . Note that $c(v) \geq |S_v| - 1$. Let \mathcal{R}_v be the list of ranks $(R_w; w \in S_v)$ such that R_v is the first element.

Proposition 4.2 combined with the union bound states that

$$|\Delta(\mathcal{R}_v; M) - (|S_v| - 1)| \leq \epsilon(|S_v| - 1) \quad . \quad (4)$$

holds with probability $1 - ne^{-C}$ for every node v . Assume that these events hold.

Select a node v . Let W be the set containing v and the nodes selected *before* v by CORE. Select $w \in W$. Let F be the graph right before deleting w by CORE. Let u be the node in F that is deleted first by EXACTCORE. Let β be the value of $d[u]$ when w is deleted by CORE. Then

$$\begin{aligned} \tau_w &\leq \beta && (\text{CORE picked } w \text{ over } u \text{ or } w = u) \\ &\leq \Delta(\mathcal{R}_u; M) && (F \subseteq H_u \text{ and Proposition 4.1}) \\ &\leq (1 + \epsilon)(|S_u| - 1) && (\text{Eq. 4}) \\ &\leq (1 + \epsilon)c(u) && (S_u = N(u, h) \cap H_u) \\ &\leq (1 + \epsilon)c(v) \quad . && (v \in F \subseteq H_u) \end{aligned}$$

Since this bound holds for *any* $w \in W$, we have

$$c'(v) = \max_{w \in W} \tau_w \leq (1 + \epsilon)c(v),$$

proving the lemma. \square

We are now ready to prove the proposition.

Proof of Proposition 4.3 The probability that one of the two above lemmas does not hold is bounded by the union bound with $2ne^{-C}$, proving the main claim.

To prove the second claim note that when $d[v] \leq M$ then $d[v]$ matches accurately the number of the remaining nodes that can be reached by an h -path from a node v . On the other hand, if there is a node w that reaches more than M nodes, we are guaranteed that $d[w] \geq M$ and $k[w, h] > 0$, implying that CORE will always prefer deleting v before w . Consequently, at the beginning CORE will select the nodes in the same order as EXACTCORE and reports the same core number as long as there are nodes with $d[v] \leq M$ or, equally, as long as $c(v) \leq M$. \square

5 Updating data structures faster

Now that we have proven the accuracy of CORE, our next step is to address the computational complexity. The key problem is that COMPUTE is called too often and the implementation of UPDATE is too slow.

As CORE progresses, the set $B[v, i]$ is modified in two ways. The first case is when some nodes become too far away, and we need to delete these nodes from $B[v, i]$. The second case is when we have deleted enough nodes so that we can lower $k[v, i]$ and introduce new nodes. Our naive version of UPDATE calls COMPUTE for both cases. We will modify the algorithms so that COMPUTE is

called only to handle the second case, and the first case is handled separately. Note that these modifications do not change the output of the algorithm.

First, we change the information stored in $B[v, i]$. Instead of storing just a node u , we will store (u, z) , where z is the number of neighbors $w \in A(v)$, such that u is in $B[w, i - 1]$. We will store $B[v, i]$ as a linked list *sorted* by the rank. In addition, each node $u \in B[w, i - 1]$ is augmented with an array $Q = (q_v \mid v \in A(w))$. An entry q_v points to the location of u in $B[v, i]$ if u is present in $B[v, i]$. Otherwise, q_v is set to null.

We will need two helper functions to maintain $B[v, i]$. The first function is a standard merge sort, $\text{MERGESORT}(X, Y)$, that merges two sorted lists in $\mathcal{O}(|X| + |Y|)$ time, maintaining the counters and the pointers.

The other function is $\text{DELETE}(X, Y)$ that removes nodes in Y from X , which we will use to remove nodes from $B[v, i]$. The deletion is done in by reducing the counters of the corresponding nodes in X by 1, and removing them when the counter reaches 0. It is vital that we can process a single node $y \in Y$ in constant time. This will be possible because we will be able to use the pointer array described above.

Let us now consider calling COMPUTE . We would like to minimize the number of calls of COMPUTE . In order to do that, we need additional bookkeeping. The first additional information is $m[v, i]$ which is the number of neighboring nodes $w \in A(v)$ for which $k[w, i - 1] = k[v, i]$. Proposition 3.1 states that $k[v, i] \geq k[w, i - 1]$, for all $w \in A(v)$. Thus, if $m[v, i] > 0$, then there is a node $u \in A(v)$ with $k[v, i] = k[u, i - 1]$ and so recomputing $B[v, i]$ will not change $k[v, i]$ and will not add new nodes in $B[v, i]$.

Unfortunately, maintaining just $m[v, i]$ is not enough. We may have $k[v, i] > k[w, i - 1]$ for any $w \in A(v)$ *immediately* after COMPUTE . In such case, we compute sets of nodes

$$X_w = \{u \in B[w, i - 1] \mid r[u] = k[v, i] - 1\},$$

and combine them in $D[v, i]$, a union of X_w along with the counter information similar to $B[v, i]$, that is,

$$D[v, i] = \{(u, z) \mid z = |\{w \in A(v) \mid u \in X_w\}| > 0\} \quad .$$

The key observation is that as long as $|B[v, i]| + |D[v, i]| > M$, the level $k[v, i]$ does not need to be updated.

There is one complication, namely, we need to compute $D[v, i]$ in $\mathcal{O}(M \deg v)$ time. Note that, unlike $B[v, i]$, the set $D[v, i]$ can have more than M elements. Hence, using MERGESORT will not work. Moreover, a stock k -ary merge sort requires $\mathcal{O}(M \deg(v) \log \deg(v))$ time. The key observation to avoid the additional log factor is that $D[v, i]$ does not need to be sorted. More specifically, we first compute an array

$$a[u] = |\{w \in A(v) \mid u \in X_w\}|,$$

and then extract the non-zero entries to form $D[v, i]$. We only need to compute the non-zero entries so we can compute these entries in $\mathcal{O}(\sum |X_w|) \subseteq \mathcal{O}(M \deg v)$ time. Moreover, since we do not need to keep them in order we can extract the non-zero entries in the same time. We will refer this procedure as UNION, taking the sets X_w as input and forming $D[v, i]$.

We need to maintain $D[v, i]$ efficiently. In order to do that we augment each node $u \in B[w, i - 1]$ with an array $(q_v \mid v \in A(w))$, where q_v points to the location of u in $D[v, i]$ if $u \in D[v, i]$.

The pseudo-code for the updated COMPUTE is given in Algorithm 5. Here we compute $B[v, i]$ and $k[v, i]$ first by using MERGESORT iteratively and trimming the resulting set if it has more than M elements. We proceed to compute $m[v, i]$ and $D[v, i]$. If $m[v, i] = 0$, we compute $D[v, i]$ with UNION. Note that if $m[v, i] > 0$, we leave $D[v, i]$ empty. The above discussion leads immediately to the computational complexity of COMPUTE.

Proposition 5.1. *COMPUTE(v, i) runs in $\mathcal{O}(M \deg v)$ time.*

The pseudo-code for UPDATE is given in Algorithm 6. Here, we maintain a stack U of tuples (v, Y) , where v is the node that requires an update, and Y are the nodes that have been deleted from $B[v, i]$ during the previous round. First, if $|B[v, i]| + |D[v, i]| \leq M$ and $m[v, i] = 0$, we run COMPUTE(v, i). Next, we proceed by reducing the counters of Z in $B[w, i + 1]$ and $D[w, i + 1]$ for each $w \in A(v)$. We also update $m[w, i + 1]$. Finally, we add (w, Z) to the next stack, where Z are the deleted nodes in $B[w, i + 1]$.

Proposition 5.2. *UPDATE maintains $B[v, i]$ correctly.*

Proof As CORE deletes nodes from the graph, Proposition 3.1 guarantees that $B[v, i]$ can be modified only in two ways: either node u is deleted from $B[v, i]$ when u is no longer present in any $B[w, i - 1]$ where $w \in A(v)$, or $k[v, i]$ changes and new nodes are added.

The first case is handled properly as UPDATE uses DELETE whenever a node is deleted from $B[w, i - 1]$.

The second case follows since if $|B[v, i]| + |D[v, i]| > M$ or $m[v, i] > 0$, then we know that COMPUTE will not change $k[v, i]$ and will not introduce new nodes in $B[v, i]$. \square

Proposition 5.3. *Assume a graph G with n nodes and m edges. Assume $0 < \epsilon \leq 1/2$, constant C , and the maximum path length h . The running time of CORE is bounded by*

$$\mathcal{O}\left(hmM \log \frac{n}{M}\right) = \mathcal{O}\left(hmC\epsilon^{-2} \log \frac{n\epsilon^2}{C}\right)$$

with a probability of $1 - n \exp(-C)$, where M is defined in Eq. 1.

Algorithm 5: Refined version of COMPUTE(v, i). Recomputes $B[v, i]$ and $k[v, i]$ from scratch.

```

1  $T \leftarrow (v, 1);$ 
2  $k \leftarrow \max(k[w, i - 1] \mid w \in A(v));$ 
3 foreach  $w \in A(v)$  do
4    $T \leftarrow \text{MERGESORT}(T, \{(u, c) \in B[w, i - 1] \mid r[u] \geq k\});$ 
5   if  $|T| > M$  then
6      $k \leftarrow \max(k, r[t_{\lfloor M \rfloor + 1}] + 1);$  //  $(t_j) = T$ 
7      $T \leftarrow \{u \in T \mid r[u] \geq k\};$ 
8  $B[v, i] \leftarrow T;$ 
9  $k[v, i] \leftarrow k;$ 
10  $m[v, i] \leftarrow |\{w \in A(v) \mid k[w, i - 1] = k[v, i]\}|;$ 
11  $D[v, i] \leftarrow \emptyset;$ 
12 if  $m[v, i] = 0$  then
13    $X_w \leftarrow \{u \in B[w, i - 1] \mid r[u] = k - 1\}$  for  $w \in A(v);$ 
14    $D[v, i] \leftarrow \text{UNION}(\{X_w \mid w \in A(v)\});$ 

```

Proof We will prove the proposition by bounding $R_1 + R_2$, where R_1 is the total time needed by COMPUTE and the R_2 is the total time needed by the inner loop in UPDATE.

We will bound R_1 first. Note that a single call of COMPUTE(v, i) requires $\mathcal{O}(M \deg v)$ time.

To bound the number of COMPUTE calls, let us first bound $k[v, i]$. Proposition 4.2 and union bound implies that

$$M2^{k[v, i]-1} \leq (1 + \epsilon)c(v) \leq 2n$$

holds for all nodes $v \in V$ with probability $1 - n \exp(-C)$. Solving for $k[v, i]$ leads to

$$k[v, i] \leq 2 + \log_2 \frac{n}{M} \in \mathcal{O}\left(\log \frac{n}{M}\right). \quad (5)$$

We claim that COMPUTE(v, i) is called at most twice per each value of $k[v, i]$. To see this, consider that COMPUTE(v, i) sets $m[v, i] = 0$. Then we also set $D[v, i]$ and we are guaranteed by the first condition on Line 9 of UPDATE that the next call of COMPUTE(v, i) will lower $k[v, i]$. Assume now that COMPUTE(v, i) sets $m[v, i] > 0$. Then the second condition on Line 9 of UPDATE guarantees that the next call of COMPUTE(v, i) either keeps $m[v, i]$ at 0 (and computes $D[v, i]$) or lowers $k[v, i]$.

In summary, the time needed by COMPUTE is bounded by

$$R_1 \in \mathcal{O}\left(\sum_{i,v} M \deg(v) \log \frac{n}{M}\right) = \mathcal{O}\left(hmM \log \frac{n}{M}\right).$$

Let us now consider R_2 . For each deleted node in $B[v, i]$ or for each lowered $k[v, i]$ the inner for-loop requires $\mathcal{O}(\deg v)$ time. Equation 5 implies that the total number of deletions from $B[v, i]$ is in $\mathcal{O}(M \log \frac{n}{M})$, and that we can lower $k[v, i]$ at most $\mathcal{O}(\log \frac{n}{M})$ times. Consequently,

$$R_2 \in \mathcal{O}\left(h \sum_v (M + 1) \log \frac{n}{M} \deg v\right) = \mathcal{O}\left(hmM \log \frac{n}{M}\right).$$

Algorithm 6: Refined version of UPDATE(u). Deletes u and updates the affected $B[v, i]$ and $k[v, i]$.

```

1  $U \leftarrow \emptyset$ ;
2 foreach  $i = 1, \dots, h$  do
3   add  $(u, B[u, i - 1])$  to  $U$ ;
4    $B[u, i - 1] \leftarrow \emptyset$ ;
5    $W \leftarrow \emptyset$ ;
6   while  $U$  is not empty do
7      $(v, Y) \leftarrow$  pop entry from  $U$ ;
8      $k \leftarrow k[v, i]$ ;
9     if  $|B[v, i]| + |D[v, i]| \leq M$  and  $m[v, i] = 0$  then
10      COMPUTE( $v, i$ );
11     if  $i = h$  then  $d[v] \leftarrow$  ESTIMATE( $v$ ) ;
12     if  $i < h$  and  $(Y \neq \emptyset$  or  $k[v, i] \neq k)$  then
13       foreach  $w \in A(v)$ ,  $w \neq u$  do
14          $X_1 \leftarrow$  pointers of  $Y$  in  $B[w, i + 1]$ ;
15          $X_2 \leftarrow$  pointers of  $Y$  in  $D[w, i + 1]$ ;
16         DELETE( $B[w, i + 1], X_1$ );
17         DELETE( $D[w, i + 1], X_2$ );
18          $Z \leftarrow$  nodes removed from  $B[w, i + 1]$ ;
19         if  $k[v, i] \neq k$  and  $k = k[w, i + 1]$  then
20            $m[w, i + 1] \leftarrow m[w, i + 1] - 1$ ;
21         add  $(w, Z)$  to  $W$ ;
22    $U \leftarrow W$ 
23 delete  $u$  from  $G$ ;
```

We have bounded both R_1 and R_2 proving the main claim. \square

Corollary 5.1. Assume real values $\epsilon > 0$, $\delta > 0$, a graph G with n nodes and m edges. Let $C = \log(2n/\delta)$. Then CORE yields ϵ approximation in

$$\mathcal{O}\left(\frac{hm \log n / \delta}{\epsilon^2} \log \frac{n\epsilon^2}{\log n / \delta}\right)$$

time with $1 - \delta$ probability.

Proposition 5.4. CORE requires $\mathcal{O}(hmM)$ memory.

Proof An entry in $B[v, i]$ requires $\mathcal{O}(\deg v)$ memory for the pointer information. An entry in $D[v, i]$ only requires $\mathcal{O}(1)$ memory. Since $|B[v, i]| \leq M$ and $|D[v, i]| \leq M \deg v$, the claim follows. \square

In order to speed-up the algorithm further we employ two additional heuristics. First, we can safely delay the initialization of $B[v, i]$ until every $B[w, i-1]$, where $w \in A(v)$, yields a core estimate that is below the current core number. Delaying the initialization allows us to ignore $B[v, i]$ during UPDATE. Second, if the current core number exceeds the number of remaining nodes, then we can stop and use the current core number for the remaining nodes. While these heuristics do not provide any additional theoretical advantage, they speed-up the algorithm in practice.

6 Distance-generalized dense subgraphs

In this section we will study distance-generalized dense subgraphs, a notion introduced by Bonchi et al. [6].

In order to define the problem, let us first define $E_h(X)$ to be the node pairs in X that are connected with an h -path in X . We exclude the node pairs of form (u, u) . Note that $E(X) = E_1(X)$.

We define the h -density of X to be the ratio of $E_h(X)$ and $|X|$,

$$dns(X; h) = \frac{|E_h(X)|}{|X|}.$$

We will sometimes drop h from the notation if it is clear from the context.

Problem 6.1 (DENSE). *Given a graph G and h find the subgraph X maximizing $dns(X; h)$.*

DENSE can be solved for $h = 1$ in polynomial time using fractional programming combined with minimum cut problems [15]. However, the distance-generalized version of the problem is **NP**-hard.

Proposition 6.1. *DENSE is **NP**-hard even for $h = 2$.*

To prove the result we will use extensively the following lemma.

Lemma 6.1. *Let X be the densest subgraph. Let $Y \subseteq X$ and $Z \cap X = \emptyset$. Then*

$$\frac{|E_h(X)| - |E_h(X \setminus Y)|}{|Y|} \geq dns(X) \geq \frac{|E_h(X \cup Z)| - |E_h(X)|}{|Z|}$$

Proof Due to optimality $dns(X) \geq dns(X \setminus Y)$. Then

$$\frac{|E_h(X)| - |E_h(X \setminus Y)|}{|Y|} \geq \frac{|E_h(X)| - dns(X)(|X| - |Y|)}{|X| - (|X| - |Y|)} = dns(X).$$

Similarly, $dns(X) \geq dns(X \cup Z)$ implies

$$\frac{|E_h(X \cup Z)| - |E_h(X)|}{|Z|} \leq \frac{dns(X)(|X| + |Z|) - |E_h(X)|}{(|X| + |Z|) - |X|} = dns(X),$$

proving the claim. \square

Proof of Proposition 6.1 To prove the claim we will reduce 3DMATCH to our problem. In an instance of 3DMATCH we are given a universe $U = u_1, \dots, u_n$ of size n and m sets \mathcal{C} of size 3 and ask whether there is an exact cover of U in \mathcal{C} .

We can safely assume that C_1 does not intersect with any other set. Otherwise, we can add a new set and 3 new items without changing the outcome of the instance.

In order to define the graph, let us first define $k = 12m$ and $\ell = 3k(3k - 1)/2 + 6k - 1$. Note that $k \geq 12$.

For each $u_i \in U$, we add k nodes a_{ij} , where $j = 1, \dots, k$. For each a_{ij} , we add 2ℓ unique nodes that are all connected to a_{ij} . We will denote the resulting star with S_{ij} . We will select a non-center node from S_{ij} and denote it by b_{ij} . Finally, we write $S'_{ij} = S_{ij} \setminus \{a_{ij}, b_{ij}\}$.

For each set $C_t \in \mathcal{C}$, we add a node, say p_t , and connect it to b_{ij} for every $u_i \in C$ and $j = 1, \dots, k$. We will denote the collection of these nodes with P . We connect every node in P to p_1 .

Let X be the nodes of the densest subgraph for $h = 2$. Let $Q = P \cap X$ and let \mathcal{R} be the corresponding sets in \mathcal{C} .

To simplify the notation we will need the following counts of node pairs. First, let us define α to be the number of node pairs in a single S_{ij} connected with a 2-path,

$$\alpha = E_2(S_{ij}) = \binom{2\ell + 1}{2}.$$

Second, let us define the number of node pairs connected with a 2-path using a single node $p_t \in P$. Since p_t connects $3k$ nodes b_{ij} and reaches $3k$ nodes b_{ij} and $3k$ nodes a_{ij} , we have

$$\beta = \binom{3k}{2} + 6k.$$

Finally, consider W consisting of a single p_t and the corresponding $3k$ stars. Let us write $\gamma = 3k\alpha + \beta$ to be the number of node pairs connected by a 2-path in W .

We will prove the proposition with a series of claims.

Claim 1: $\text{dns}(X) > \ell$. The density of W as defined above is

$$\text{dns}(W) = \frac{3k\alpha + \beta}{3k(2\ell + 1) + 1} > \frac{3k\alpha + \ell}{3k(2\ell + 1) + 1} = \ell.$$

Thus, $\text{dns}(X) \geq \text{dns}(W) > \ell$.

Claim 2: \mathcal{R} is disjoint. To prove the claim, assume otherwise and let C_t , with $t > 1$, be a set overlapping with some other set in \mathcal{R} .

Let us bound the number of node pairs that are *solely* connected with p_t . The node p_t connects $3k + 1$ nodes in V . Out of these nodes at least $k + 1$ nodes are connected by another node in X . In addition, p_t reaches to a_{ij} and b_{ij} , where $u_i \in C_t$ and $j = 1, \dots, k$, that is, $6k$ nodes in total. Finally, p_t may connect to every other node in P , at most $m - 1$ nodes, and every a_{ij} connected to p_1 , at most $3k$ nodes. In summary, we have

$$\begin{aligned} |E_2(X)| - |E_2(X \setminus \{p_t\})| &\leq \binom{3k + 1}{2} - \binom{k + 1}{2} + 6k + m - 1 + 3k \\ &= \ell - k^2/2 + 5k/2 + m + 3k < \ell \leq \text{dns}(X). \end{aligned}$$

Lemma 6.1 with $Y = \{p_t\}$ now contradicts the optimality of X . Thus, \mathcal{R} is disjoint.

Claim 3: Either $S_{ij} \subseteq X$ or $S_{ij} \cap X = \emptyset$. To prove the claim assume that $S_{ij} \cap X \neq \emptyset$.

Assume that $b_{ij} \notin X$. Then $S_{ij} \cap X$ is a disconnected component with density less than ℓ , contradicting Lemma 6.1. Assume that $b_{ij} \in X$ and $a_{ij} \notin X$. Then deleting b_{ij} will reduce at most $3k + m - 1 < \ell$ connected node pairs, contradicting Lemma 6.1.

Assume that $b_{ij}, a_{ij} \in X$. If $S'_{ij} \cap X = \emptyset$, then deleting a_{ij} will reduce at most 2 connected node pairs, contradicting Lemma 6.1. Assume now there are $u \in S'_{ij} \cap X$ and $w \in S'_{ij} \setminus X$. Then $|E_2(X \cup \{w\})| - |E_2(X)| > |E_2(X)| - |E_2(X \setminus \{u\})|$, contradicting Lemma 6.1. Consequently, $S_{ij} \subseteq X$.

Claim 4: If $p_t \in X$, then X contains every corresponding S_{ij} . To prove the claim assume otherwise.

Assume first that there are no corresponding S_{ij} in X for p_t . If $t > 1$, then p_t reaches to at most $m - 1 + 3k$ nodes. If $t = 1$, then p_1 connects at most $m - 1$ nodes and reaches to at most $(m - 1)(3k + 1)$ nodes.

Both cases lead to

$$|E_2(X)| - |E_2(X \setminus \{p_t\})| \leq \binom{m-1}{2} + (m-1)(3k+1) < \ell < \text{dns}(X),$$

contradicting Lemma 6.1.

Assume there is at least one corresponding S_{ij} in X but not all, say $S_{i'j'}$ is missing. Then

$$|E_2(X)| - |E_2(X \setminus S_{ij})| < |E_2(X \cup S_{i'j'})| - |E_2(X)|,$$

contradicting Lemma 6.1.

Claim 5: No S_{ij} without corresponding p_t is included in X . To prove the claim note that such S_{ij} is disconnected and has density of ℓ , contradicting Lemma 6.1.

The previous claims together show that the density of X is equal to

$$\text{dns}(X) = \frac{|Q|\gamma + (|Q| - 1)(6k) + \binom{|Q|}{2}}{|Q|(3k(2\ell + 1) + 1)},$$

which is an increasing function of $|Q|$. Since \mathcal{R} is disjoint and maximal, the 3DMATCH instance has a solution if and only if \mathcal{R} is a solution. \square

One of the appealing aspects of $\text{dns}(X; h)$ for $h = 1$ is that we can 2-approximate in linear time [8]. This is done by ordering the nodes with EXACTCORE, say v_1, \dots, v_n and then selecting the densest subgraph of the form v_1, \dots, v_i .

The approximation guarantee for $h > 1$ is weaker even if use EXACTCORE. Bonchi et al. [6] showed that $2\text{dns}(Y) \geq \sqrt{2\text{dns}(X) + 1/4} - 1/2$ when we use EXACTCORE.

Using CORE instead of EXACTCORE poses additional challenges. In order to select a subgraph among n candidates, we need to estimate the density of its subgraph. We cannot use $d[v]$ used by CORE as these are the values that CORE uses to determine the order.

Assume that CORE produced order of vertices v_1, \dots, v_n , first vertices deleted first. To find the densest graph among the candidate, we essentially repeat CORE except now we delete the nodes using the order v_1, \dots, v_n . We then estimate the number of edges with the identity

$$2|E_h(X)| = \sum_{v \in X} \deg_h(v; X) \quad .$$

We will refer to this algorithm as ESTDENSE. The pseudo-code for ESTDENSE is given in Algorithm 7.

Algorithm 7: ESTDENSE($G, v_1, \dots, v_n, \epsilon, C$) approximative dense subgraph. Setting $C = \log(n^2/\delta)$ yields an approximation with $1 - \delta$ probability.

```

1 foreach  $v \in V$  do
2    $r[v] \leftarrow \text{geo}(1/2)$ ;
3    $B[v, 0] \leftarrow \{v\}$ ;
4    $k[v, 0] \leftarrow 0$ ;
5  $M \leftarrow 1 + \frac{4(2+\epsilon)}{\epsilon^2}(C + \log 8)$ ;
6 foreach  $i = 1, \dots, h$  do
7   foreach  $v \in V$  do COMPUTE( $v, i$ ) ;
8  $R \leftarrow \sum_v d[v]$ ;
9 foreach  $i = 1, \dots, n$  do
10   estimate  $\text{dns}(v_i, \dots, v_n)$  with  $R/(n - i + 1)$ ;
11   UPDATE( $v_i$ );
12   keep  $R$  updated when recomputing  $d[v]$  with ESTIMATE;
13 return densest tested subgraph;
```

The algorithm yields to the following guarantee.

Proposition 6.2. Assume $\epsilon > 0, C > 0$ and h . Define $\gamma = \frac{1-\epsilon}{1+\epsilon}$. For any given k , let C_k be the (k, h) -core. Define

$$\beta = \min_k \frac{|C_k|}{|C_{k\gamma}|}$$

to be the smallest size ratio between C_k and $C_{k\gamma}$.

Let X be the h -densest subgraph.

Let c' be an ϵ -approximative core map and let v_1, \dots, v_n be the corresponding vertex order. Let $Y = \text{ESTDENSE}(G, v_1, \dots, v_n, \epsilon, C)$ Then

$$2\text{dns}(Y) \geq \gamma\beta \left(\sqrt{2\text{dns}(X) + 1/4} - 1/2 \right)$$

with probability $1 - n^2 \exp(-C)$.

To prove the result we need the following lemma.

Lemma 6.2. For any given k , define $C'_k = \{v \mid c'(v) \geq k\}$. Then

$$\text{dns}(C'_{k(1-\epsilon)}) \geq \beta \text{dns}(C_k) \quad .$$

Proof Write $F = C'_{k(1-\epsilon)}$. Let $v \in C_k$. Then $c'(v) \geq (1-\epsilon)c(v) \geq (1-\epsilon)k$ and so $v \in F$. Thus $C_k \subseteq F$. Conversely, let $v \in F$. Then $(1+\epsilon)c(v) \geq c'(v) \geq k(1-\epsilon)$ and so $v \in C_{\gamma k}$. Thus $F \subseteq C_{\gamma k}$. The definition of β now implies

$$dns(F) = \frac{|E_h(F)|}{|F|} \geq \beta \frac{|E_h(F)|}{|C_k|} \geq \beta \frac{|E_h(C_k)|}{|C_k|}$$

proving the claim. \square

Proof of Proposition 6.2 Let c be the core map produced by EXACTCORE. For any given k , define $C'_k = \{v \mid c'(v) \geq k\}$.

Let $u \in X$ be the first vertex deleted by EXACTCORE. Let $b = \deg_h(u; X)$ be its h -degree. Write $X' = X \setminus \{u\}$. Since X is optimal,

$$\frac{|E_h(X)|}{|X|} \geq \frac{|E_h(X')|}{|X'|}.$$

Deleting u from X will delete b node pairs from $E_h(X)$ containing u . In addition, every node in the h -neighborhood of u may be disconnected from each other, potentially reducing the node pairs by $\binom{b}{2}$. In summary,

$$|E_h(X)| - |E_h(X')| \leq b + \binom{b}{2} = \binom{b+1}{2}.$$

Combining the two inequalities leads to

$$\binom{b+1}{2} \geq |E_h(X)| - \frac{|E_h(X)|(|X| - 1)}{|X|} = \frac{|E_h(X)|}{|X|} = dns(X).$$

Solving for b results in

$$b \geq \sqrt{2dns(X) + 1/4} - 1/2. \quad (6)$$

Let Z be the nodes right before u is deleted by EXACTCORE. Note that $c(u) \geq \deg_h(u; Z) \geq \deg_h(u; X) = b$.

Let C_k be the smallest core containing u , that is, $c(u) = k$. By definition, $\deg_h(v; C_k) \geq k \geq b$, for all $v \in C_k$.

Let $F = C'_{k(1-\epsilon)}$. Lemma 6.2 now states that

$$2dns(F) \geq 2\beta dns(C_k) = \beta \frac{1}{|C_k|} \sum_{v \in C_k} \deg_h(v; C_k) \geq \beta k \geq \beta b. \quad (7)$$

Let $d'(Z)$ be the estimated density for a subgraph Z .

Proposition 4.2 and the union bound state that

$$dns(Y) \geq \frac{1}{1+\epsilon} d'(Y) \geq \frac{1}{1+\epsilon} d'(F) \geq \gamma dns(F) \quad (8)$$

with probability $1 - n^2 e^{-C}$. Eqs. 6–8 prove the inequality in the claim. \square

ESTDENSE is essentially CORE so we can apply Proposition 5.3.

Corollary 6.1. *Assume real values $\epsilon > 0$, $\delta > 0$, a graph G with n nodes and m edges. Let $C = \log(n^2/\delta)$. Then ESTDENSE runs in*

$$\mathcal{O}\left(\frac{hm \log n / \delta}{\epsilon^2} \log \frac{n\epsilon^2}{\log n / \delta}\right)$$

time and Proposition 6.2 holds with $1 - \delta$ probability.

Finally, let us describe a potentially faster variant of the algorithm that we will use in our experiments. The above proof will work even if replace C_k with the most inner (exact) core. Since $F = C'_{k(1-\epsilon)}$ we can prune all the vertices for which $c'(v) < k(1-\epsilon)$. The problem is that we do not know k but we can lower bound it with $k \geq k'/(1+\epsilon)$, where $k' = \max_v c'(v)$. In summary, before running ESTIMATE we remove all the vertices for which $c'(v) < \gamma k'$.

7 Related work

The notion of distance-generalized core decomposition was proposed by Bonchi et al. [6]. The authors provide several heuristics to significantly speed-up the baseline algorithm (a variant of an algorithm proposed by Batagelj and Zaveršnik [3]). Despite being significantly faster than the baseline approach, these heuristics still have the computational complexity in $\mathcal{O}(nn'(n' + m'))$, where n' and m' are the numbers of nodes and edges in the largest h -neighborhood. For dense graphs and large values of h , the sizes n' and m' can be close n and m , leading to the computational time of $\mathcal{O}(n^2m)$. We will use these heuristics as baselines in Section 8.

All these algorithms, as well as ours, rely on the same idea of iteratively deleting the vertex with the smallest $\deg_h(v)$ and updating these counters upon the deletion. The difference is that the previous works maintain these counters exactly—and use some heuristics to avoid updating unnecessary nodes—whereas we approximate $\deg_h(v)$ by sampling, thus reducing the computational time at the cost of accuracy.

A popular variant of decomposition is a k -truss, where each edge is required to be in at least k triangles [9, 17, 28–30]. Saryüce and Pinar [21], Saryüce et al. [22] proposed (r, s) nucleus decomposition, an extension of k -cores where the notion nodes and edges are replaced with r -cliques and s -cliques, respectively. Saryüce and Pinar [21] points out that there are several variants of k -trusses, depending on the connectivity requirements: Huang et al. [17] requires the trusses to be triangle-connected, Cohen [9] requires them to be connected, and Zhang and Parthasarathy [29] allows the trusses to be disconnected.

A k -core is the largest subgraph whose smallest degree is at least k . A similar concept is the densest subgraph, a subgraph whose average degree is the largest [15]. Such graphs are convenient variants for discovering dense communities as they can be discovered in polynomial time [15], as opposed to, e.g., cliques that are inapproximable [31].

Interestingly, the same peeling algorithm that is used for core decomposition can be use to 2-approximate the densest subgraph [8]. Tatti [25] proposed a variant of core decomposition so that the densest subgraph is equal to the inner core. This composition is solvable in polynomial time and can be approximated using the same peeling strategy.

A distance-generalized clique is known as h -club, which is a subgraph where every node is reachable by an h -path from every node [20]. Here the path

must be inside the subgraph. Since cliques are 1-clubs, discovering maximum h -clubs is immediately an inapproximable problem. Bonchi et al. [6] argued that (k, h) decomposition can be used to aid discovering maximum h -clubs.

Using sampling for parallelizing (normal) core computation was proposed by Esfandiari et al. [10]. Here, the authors sparsify the graph multiple times by sampling edges. The sampling probability depends on the core numbers: larger core numbers allow for more aggressive sparsification. The authors then use Chernoff bounds to prove the approximation guarantees. The authors were able to sample edges since the degree in the sparsified graph is an estimate of the degree in the original graph (multiplied by the sampling probability). This does not hold for (k, h) core decomposition because a node $w \in N(v; h)$ can reach v with several paths.

Approximating h -neighborhoods can be seen as an instance of a cardinality estimation problem. A classic approach for solving such problems is HyperLogLog [11]. Adopting HyperLogLog or an alternative approach, such as [18], is a promising direction for a future work, potentially speeding up the algorithm further. The challenge here is to maintain the estimates as the nodes are removed by CORE.

8 Experimental evaluation

Our two main goals in experimental evaluation is to study the accuracy and the computational time of CORE.

8.1 Datasets and setup

We used 8 publicly available benchmark datasets. *CaAstro* and *CaHep* are collaboration networks between researchers.¹ *RoadPa* and *RoadTX* are road networks in Pennsylvania and Texas.¹ *Amazon* contains product pairs that are often co-purchased in a popular online retailer.¹ *Youtube* contains user-to-user links in a popular video streaming service.² *Hyves* and *Douban* contain friendship links in a Dutch and Chinese social networks, respectively.³ The sizes of the graphs are given in Table 1.

We implemented CORE in C++⁴ and conducted the experiments using a single core (2.4GHz). For CORE we used 8GB RAM and for ESTDENSE we used 50GB RAM. In all experiments, we set $\delta = 0.05$.

8.2 Accuracy

In our first experiment we compared the accuracy of our estimate $c'(v)$ against the correct core numbers $c(v)$. As a measure we used the maximum relative

¹<http://snap.stanford.edu>

²<http://networkrepository.com/>

³<http://konect.cc/>

⁴<http://version.helsinki.fi/dacs>

Table 1: Sizes and computational times for the benchmark datasets. Here, n is the number of nodes m is the number of edges, M is the internal parameter of CORE given in Eq. 1. The running times for the baselines LUB and LB are taken from [6]. Dashes indicate that the experiments did not finish in 20 hours. For *Youtube* and *Hyves*, LUB was run with 52 CPU cores. The remaining experiments are done with a single CPU core.

| Dataset | n | m | M | $h = 2$ | | |
|----------------|-----------|-----------|-----|---------|-------|--------|
| | | | | CORE | LB | LUB |
| <i>CaHep</i> | 12 008 | 118 489 | 607 | 3.53s | 0.95s | 1.19s |
| <i>CaAstro</i> | 18 772 | 198 050 | 625 | 5.2s | 5.52s | 5.17s |
| <i>RoadPA</i> | 1 088 092 | 1 541 898 | 787 | 7.72s | 3.18s | 36.14s |
| <i>RoadTX</i> | 1 393 383 | 1 921 660 | 797 | 10.39s | 4.21s | 56.89s |
| <i>Amazon</i> | 334 863 | 925 872 | 740 | 4.96s | 2.51s | 12.98s |
| <i>Douban</i> | 154 908 | 327 162 | 709 | 6.39s | 4.3s | 6.76s |
| <i>Hyves</i> | 1 402 673 | 2 777 419 | 797 | 1m22s | 1m53s | 7m21s |
| <i>Youtube</i> | 495 957 | 1 936 748 | 756 | 1m11s | 1m43s | 3m12s |

| Dataset | $h = 3$ | | | $h = 4$ | | |
|----------------|---------|---------|--------|---------|---------|---------|
| | CORE | LB | LUB | CORE | LB | LUB |
| <i>CaHep</i> | 10.39s | 2m8s | 1m33s | 22.34s | 15m41s | 2m3s |
| <i>CaAstro</i> | 23.65s | 9m20s | 1m31s | 20.81s | 80m35s | 6m13s |
| <i>RoadPA</i> | 13.85s | 6.75s | 1m59s | 23.64s | 11.47s | 2m20s |
| <i>RoadTX</i> | 18.28s | 8.44s | 3m4s | 30.9s | 13.9s | 3m28s |
| <i>Amazon</i> | 17.26s | 29.27s | 51.92s | 1m15s | 4m56s | 3m11s |
| <i>Douban</i> | 57.78s | 31m4s | 3m41s | 1m34s | 912m42s | 59m17s |
| <i>Hyves</i> | 7m6s | 702m44s | 62m5s | 12m3s | — | 800m39s |
| <i>Youtube</i> | 5m12s | — | 53m12s | 4m26s | — | 155m11s |

error

$$\max_{v \in V} \frac{|c'(v) - c(v)|}{c(v)}.$$

Note that Proposition 4.3 states that the error should be less than ϵ with high probability.

The error as a function of ϵ for *CaHep* and *CaAstro* datasets is shown in Figure 1 for $h = 3, 4$. We see from the results that the error tends to increase as a function of ϵ . As ϵ decreases, the internal value M increases, reaching the point where the maximum core number is smaller than M . For such values, Proposition 4.3 guarantees that CORE produces correct results. We see, for example, that this value is reached with $\epsilon = 0.20$ for *CaHep*, and $\epsilon = 0.15$ for *CaAstro* when $h = 3$, and $\epsilon = 0.35$ for *Amazon* when $h = 4$.

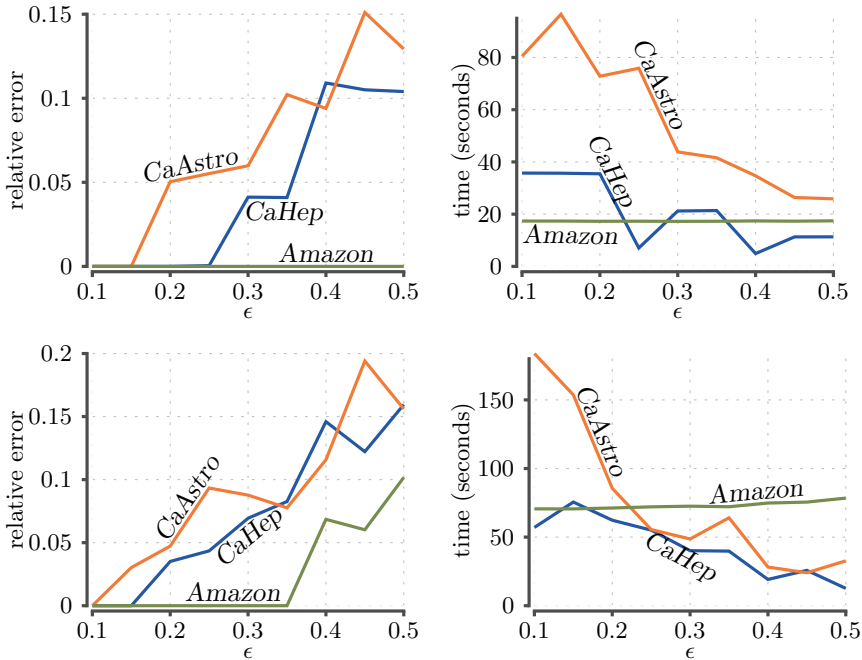


Figure 1: Relative error and computational time as a function of ϵ for *CaHep*, *CaAstro*, and *Amazon* datasets and $h = 3$ (top row) and $h = 4$ (bottom row).

8.3 Computational time

Our next experiment is to study the computational time as a function of ϵ ; the results are shown in Figure 1. From the results we see that generally computational time increases as ϵ decreases. The computational time flattens when we reach the point when $c(v) \leq M$ for every M . In such case, the lists $B[v, i]$ match exactly to the neighborhoods $N(v, i)$ and do not change if M is increased further. Consequently, decreasing ϵ further will not change the running time. Interestingly, the running time increases slightly for *Amazon* and $h = 4$ as ϵ increases. This is most likely due to the increased number of COMPUTE calls for smaller values of M .

Next, we compare the computational time of our method against the baselines LB and LUB proposed by Bonchi et al. [6]. As our hardware setup is similar, we used the running times for the baselines reported by Bonchi et al. [6]. Here, we fixed $\epsilon = 0.5$. The results are shown in Table 1.

We see from the results that for $h = 2$ the results are dominated by LB. This is due to the fact that most, if not all, nodes will have $c(v) \leq M$. In such case, CORE does not use any sampling and does not provide any speed up. This is especially the case for the road networks, where the core number stays low even for larger values of h . On the other hand, CORE outperforms the baselines in cases where $c(v)$ is large, whether due to a larger h or due

to denser networks. As an extreme example, LUB required over 13 hours with 52 CPU cores to compute core for *Hyves* while CORE provided an estimate in about 12 minutes using only 1 CPU core.

Interestingly enough, CORE solves *CaAstro* faster when $h = 4$ than when $h = 3$. This is due to the fact that we stop when the current core value plus one is equal to the number of remaining nodes.

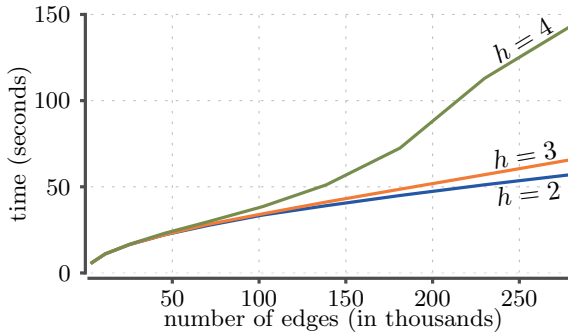


Figure 2: Computational time as a function of number of edges applied to synthetic data.

To further demonstrate the effect of the network size on the computation time we generate a series of synthetic datasets. Each dataset is stochastic blockmodel with 10 blocks of equal size, C_1, \dots, C_{10} . To add a hierarchical structure we set the probability of an edge between nodes in C_i and C_j with $i < j$ to be $10^{-6}i^2$. We vary the number of nodes from 10 000 to 100 000. The computational times for our method, with $h = 2, 3, 4$ and $\epsilon = 0.5$, are shown in Figure 2. As expected, the running times increase as the number of edges increase. Moreover, larger h require more processing time. We should stress that while Corollary 5.1 bounds the running time as quasi-linear, in practice the trend depends on the underlying model.

8.4 Dense subgraphs

Finally, we used ESTDENSE to estimate the densest subgraph for $h = 2, 3, 4$. We set $\epsilon = 0.5$ and $\delta = 0.05$. The results, shown in Table 2, are as expected. Both the density and the size of the h -densest subgraphs increase as the function of h . The dense subgraphs are generally smaller and less dense for the sparse graphs, such as, road networks.

In our experiments, the running times for ESTDENSE were generally smaller but comparable to the running times of CORE. The speed-up is largely due to the pruning of nodes with smaller core numbers. The exception was *Youtube* with $h = 3$, where ESTDENSE required over 23 minutes. The slowdown is due to CORE using lazy initialization of $B[v, i]$ whereas ESTDENSE needs $B[v, h]$

Table 2: Densities and sizes of discovered dense subgraphs for the benchmark datasets.

| <i>Dataset</i> | <i>h</i> = 2 | | <i>h</i> = 3 | | <i>h</i> = 4 | |
|----------------|-------------------------|----------|-------------------------|----------|-------------------------|----------|
| | <i>dns</i> (<i>X</i>) | <i>X</i> | <i>dns</i> (<i>X</i>) | <i>X</i> | <i>dns</i> (<i>X</i>) | <i>X</i> |
| <i>CaHep</i> | 494.77 | 1 383 | 1 372.14 | 3 998 | 3 121.32 | 7 069 |
| <i>CaAstro</i> | 570.55 | 3 525 | 2 955.16 | 10 321 | 6 280.64 | 15 100 |
| <i>RoadPA</i> | 5.18 | 19 496 | 10.74 | 4 407 | 18.94 | 15 556 |
| <i>RoadTX</i> | 6 | 65 | 10.6 | 908 | 18.28 | 19 530 |
| <i>Amazon</i> | 274.5 | 550 | 407.25 | 2 192 | 851.96 | 22 476 |
| <i>Douban</i> | 384.69 | 4 133 | 3 435.75 | 13 853 | 17 142.94 | 73 840 |
| <i>Hyves</i> | 15 135.53 | 31 884 | 31 832.31 | 224 136 | 142 173.06 | 448 330 |
| <i>Youtube</i> | 13 572.56 | 25 413 | 44 338.23 | 162 379 | 132 376.1 | 315 211 |

to be computed in order to obtain $d[v]$. This is also the reason why ESTDENSE requires more memory in practice.

9 Concluding remarks

In this paper we introduced a randomized algorithm for approximating distance-generalized core decomposition. The major advantage over the exact approximation is that the approximation can be done in $\mathcal{O}(\epsilon^{-2}hm(\log^2 n - \log \delta))$ time, whereas the exact computation may require $\mathcal{O}(n^2m)$ time. We also studied distance-generalized dense subgraphs by proving that the problem is **NP**-hard and extended the guarantee results of [6] to approximate core decompositions.

The algorithm is based on sampling the h -neighborhoods of the nodes. We prove the approximation guarantee with Chernoff bounds. Maintaining the sampled h -neighborhood requires carefully designed bookkeeping in order to obtain the needed computational complexity. This is especially the case since the sampling probability may change as the graph gets smaller during the decomposition.

In practice, the sampling complements the exact algorithm. For the setups where the exact algorithm struggles, our algorithm outperforms the exact approach by a large margin. Such setups include well-connected networks and values h larger than 3.

An interesting direction for future work is to study whether the heuristics introduced by Bonchi et al. [6] can be incorporated with the sampling approach in order to obtain even faster decomposition method.

Acknowledgments.

This research is supported by the Academy of Finland project MALSOME (343045).

A Proof of Proposition 4.2

We start with stating several Chernoff bounds.

Lemma A.1 (Chernoff bounds). *Let X_1, \dots, X_d be d independent Bernoulli random variables with $P(X_i = 1) = \mu$. Let $S = \sum_{i=1}^d X_i$. Then*

$$P(S \geq (1 + \epsilon)d\mu) < \exp\left(-\frac{\epsilon^2}{2 + \epsilon}d\mu\right), \quad (9)$$

$$P(S \leq (1 - \epsilon)d\mu) < \exp\left(-\frac{\epsilon^2}{2}d\mu\right), \quad \text{and} \quad (10)$$

$$P(|S - d\mu| \geq \epsilon d\mu) < 2 \exp\left(-\frac{\epsilon^2}{2 + \epsilon}d\mu\right). \quad (11)$$

Proof Eqs. 9–10 are standard multiplicative Chernoff bounds. Eq. 11 is obtained with a union bound of Eqs. 9–10, completing the claim. \square

To prove Proposition 4.2 we first need the following technical lemma.

Lemma A.2. *Assume $0 < \epsilon \leq 1/2$. Let R_1, \dots, R_d be independent random variables sampled from geometric distribution, $\text{geo}(1/2)$. Define*

$$S_i = |\{j \in [d] \mid R_j \geq i\}| \text{ and } T_i = |\{j \in [d] \mid R_j \geq i, j \geq 2\}|$$

to be the number of variables $\{R_j\}$ larger than or equal to i . Assume $C > 0$ and define M as in Eq. 1. Assume that $M \leq d$. Let $\ell \geq 1$ be an integer such that

$$M2^{\ell-1} \leq d < M2^\ell. \quad (12)$$

Then with probability $1 - \exp(-C)$ we have

$$\ell = 1 \quad \text{or} \quad S_{\ell-2} > M, \quad \text{and} \quad S_{\ell+1} \leq M, \quad (13)$$

and

$$|T_k - \mu_k(d-1)| \leq \epsilon \mu_k(d-1), \quad (14)$$

where $k = \ell - 1, \ell, \ell + 1$ and $\mu_k = 2^{-k}$.

Proof First, note that Eq. 12 implies

$$2\mu_{\ell+1}d = \mu_\ell d < M \leq 4d\mu_{\ell+1} = 2^{-1}d\mu_{\ell-2}. \quad (15)$$

To prove the lemma, let us define the events

$$A_k = |T_k - \mu_k(d-1)| > \epsilon \mu_k(d-1),$$

and

$$B_1 = S_{\ell-2} \leq M \quad \text{and} \quad B_2 = S_{\ell+1} > M.$$

We will prove the result with union bound by showing that

$$\begin{aligned} & P(A_{\ell-1} \text{ or } A_\ell \text{ or } A_{\ell+1} \text{ or } B_1 \text{ or } B_2) \\ & \leq P(A_{\ell-1}) + P(A_\ell) + P(A_{\ell+1}) + P(B_1) + P(B_2) \\ & \leq 2/8e^{-C} + 2/8e^{-C} + 2/8e^{-C} + 1/8e^{-C} + 1/8e^{-C} \quad . \end{aligned}$$

To bound $P(A_k)$, observe that $P(R_j \geq k) = \mu_k$. The Chernoff bound now states that for $k \leq \ell + 1$ we have

$$\begin{aligned} P(A_k) &= P(|T_k - \mu_k(d-1)| > \epsilon \mu_k(d-1)) \\ &< 2 \exp \left(-\frac{\epsilon^2}{2+\epsilon} \mu_k(d-1) \right) & (\text{Eq. 11}) \\ &< 2 \exp \left(-\frac{\epsilon^2}{2+\epsilon} \mu_{\ell+1}(d-1) \right) & (k \leq \ell + 1) \\ &\leq 2 \exp \left(-\frac{\epsilon^2}{4(2+\epsilon)} (M-1) \right) & (\text{Eq. 15, } \mu_{\ell+1} \leq 1/4) \\ &= 2/8 \exp(-C) \quad . & (\text{Eq. 1}) \end{aligned}$$

Next, we bound B_1 , assuming $\ell > 1$ as otherwise we can ignore the term, with

$$\begin{aligned} P(S_{\ell-2} \leq M) &\leq P(S_{\ell-2} \leq 2^{-1} \mu_{\ell-2} d) & (\text{Eq. 15}) \\ &\leq P(S_{\ell-2} \leq (1-\epsilon) \mu_{\ell-2} d) & (\epsilon \leq 1/2) \\ &< \exp \left(-\frac{\epsilon^2}{2} \mu_{\ell-2} d \right) & (\text{Eq. 10}) \\ &\leq \exp(-\epsilon^2 M) & (\text{Eq. 15}) \\ &< \exp \left(-\frac{\epsilon^2}{4(2+\epsilon)} (M-1) \right) \\ &= 1/8 \exp(-C) & (\text{Eq. 1}) \end{aligned}$$

and B_2 with

$$\begin{aligned} P(S_{\ell+1} > M) &\leq P(S_{\ell+1} > 2\mu_{\ell+1} d) & (\text{Eq. 15}) \\ &\leq P(S_{\ell+1} > (1+2\epsilon)\mu_{\ell+1} d) & (\epsilon \leq 1/2) \\ &< \exp \left(-4\frac{\epsilon^2}{2+2\epsilon} \mu_{\ell+1} d \right) & (\text{Eq. 9}) \\ &\leq \exp \left(-\frac{\epsilon^2}{2+2\epsilon} M \right) & (\text{Eq. 15}) \\ &< \exp \left(-\frac{\epsilon^2}{4(2+\epsilon)} (M-1) \right) \\ &= 1/8 \exp(-C) \quad . & (\text{Eq. 1}) \end{aligned}$$

The bounds for $P(B_1)$, $P(B_2)$, and $P(A_k)$ complete the proof. \square

Proof of Proposition 4.2 Let S_i , T_i , and k be as defined in Definition 4.1 for $\Delta(\mathcal{R}; M)$. Let ℓ be as defined in Eq 12. We can safely assume that $M \leq d$.

Assume that the events in Lemma A.2 hold. Then Eq. 13 guarantees that $k = \ell - 1, \ell, \ell + 1$.

Write $Y_i = T_i 2^i$ and $Z_i = M 2^{i-1}$. Eq. 14 guarantees that

$$|Y_i - (d - 1)| \leq \epsilon(d - 1) \quad (16)$$

for $i = \ell - 1, \ell, \ell + 1$. If $k = 0$ or $Y_k \geq Z_k$, then $\Delta(\mathcal{R}; M) = Y_k$ and we are done.

Assume $k > 0$ and $Y_k < Z_k$. Then immediately

$$Z_k > Y_k \geq (1 - \epsilon)(d - 1) \quad .$$

To prove the other direction, first assume that $k > \ell - 1$. By definition of k , we have $S_{k-1} > M$ and consequently $T_{k-1} \geq M$. Thus,

$$Z_k = M 2^{k-1} \leq T_{k-1} 2^{k-1} = Y_{k-1} \leq (1 + \epsilon)(d - 1),$$

where the last inequality is given by Eq. 16. On the other hand, if $k = \ell - 1$, then

$$Z_k = M 2^{k-1} = M 2^{\ell-2} \leq d/2 \leq d - 1 \leq (1 + \epsilon)(d - 1),$$

where the second inequality holds since $k > 0$ implies that $d \geq 2$. In summary, Eq. 2 holds.

Since the events in Lemma A.2 hold with probability of $1 - \exp(-C)$, the claim follows. \square

References

- [1] J Ignacio Alvarez-Hamelin, Luca Dall’Asta, Alain Barrat, and Alessandro Vespignani. Large scale networks fingerprinting and visualization using the k-core decomposition. In *Advances in neural information processing systems*, pages 41–50, 2006.
- [2] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):1–27, 2003.
- [3] Vladimir Batagelj and Matjaž Zaveršnik. Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification*, 5(2):129–145, 2011.
- [4] Béla Bollobás. The evolution of random graphs. *Transactions of the American Mathematical Society*, 286(1):257–274, 1984.
- [5] Francesco Bonchi, Francesco Gullo, Andreas Kaltenbrunner, and Yana Volkovich. Core decomposition of uncertain graphs. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1316–1325, 2014.
- [6] Francesco Bonchi, Arijit Khan, and Lorenzo Severini. Distance-generalized core decomposition. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1006–1023, 2019.

- [7] Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, 2007.
- [8] Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. *APPROX*, 2000.
- [9] Jonathan Cohen. Trusses: Cohesive subgraphs for social network analysis. *National security agency technical report*, 16(3.1), 2008.
- [10] Hossein Esfandiari, Silvio Lattanzi, and Vahab Mirrokni. Parallel and streaming algorithms for k-core decomposition. In *International Conference on Machine Learning*, pages 1397–1406, 2018.
- [11] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *Discrete Mathematics and Theoretical Computer Science*, pages 137–156. Discrete Mathematics and Theoretical Computer Science, 2007.
- [12] Edoardo Galimberti, Francesco Bonchi, and Francesco Gullo. Core decomposition and densest subgraph in multilayer networks. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1807–1816, 2017.
- [13] Edoardo Galimberti, Alain Barrat, Francesco Bonchi, Ciro Cattuto, and Francesco Gullo. Mining (maximal) span-cores from temporal networks. In *Proceedings of the 27th ACM international Conference on Information and Knowledge Management*, pages 107–116, 2018.
- [14] Christos Giatsidis, Dimitrios M Thilikos, and Michalis Vazirgiannis. D-cores: measuring collaboration of directed graphs based on degeneracy. *Knowledge and information systems*, 35(2):311–343, 2013.
- [15] Andrew V Goldberg. Finding a maximum density subgraph. *University of California Berkeley Technical report*, 1984.
- [16] Patric Hagmann, Leila Cammoun, Xavier Gigandet, Reto Meuli, Christopher J. Honey, Van J. Wedeen, and Olaf Sporns. Mapping the structural core of human cerebral cortex. *PLoS, Biology*, 6(7):888–893, 2008.
- [17] Xin Huang, Hong Cheng, Lu Qin, Wentao Tian, and Jeffrey Xu Yu. Querying k-truss community in large and dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1311–1322, 2014.

- [18] Daniel M Kane, Jelani Nelson, and David P Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 41–52, 2010.
- [19] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010.
- [20] Robert J Mokken et al. Cliques, clubs and clans. *Quality & Quantity*, 13(2):161–173, 1979.
- [21] Ahmet Erdem Sariyüce and Ali Pinar. Fast hierarchy construction for dense subgraphs. *Proceedings of the VLDB Endowment*, 10(3), 2016.
- [22] Ahmet Erdem Sariyuce, C Seshadhri, Ali Pinar, and Umit V Catalyurek. Finding the hierarchy of dense subgraphs using nucleus decompositions. In *Proceedings of the 24th International Conference on World Wide Web*, pages 927–937, 2015.
- [23] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.
- [24] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106(16):6483–6488, 2009.
- [25] Nikolaj Tatti. Density-friendly graph decomposition. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(5):1–29, 2019.
- [26] Nikolaj Tatti. Fast computation of distance-generalized cores using sampling. In *ICDM*, 2021.
- [27] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012.
- [28] Jia Wang and James Cheng. Truss decomposition in massive networks. *Proceedings of the VLDB Endowment*, 5(9), 2012.
- [29] Yang Zhang and Srinivasan Parthasarathy. Extracting analyzing and visualizing triangle k-core motifs within networks. In *2012 IEEE 28th international conference on data engineering*, pages 1049–1060. IEEE, 2012.

- [30] Feng Zhao and Anthony KH Tung. Large scale cohesive subgraphs discovery for social network visual analysis. *Proceedings of the VLDB Endowment*, 6(2):85–96, 2012.
- [31] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 681–690, 2006.