SAND2014-19934R

# Finding Hierarchical and Overlapping Dense Subgraphs using Nucleus Decompositions

Ahmet Erdem Sarıyüce[†] [*], C. Seshadhri[‡], Ali Pınar[‡], Ümit V. Çatalyürek[†]
sariyuce.1@osu.edu, scomand@sandia.gov, apinar@sandia.gov, umit@bmi.osu.edu
[†]*The Ohio State University, Columbus, OH, USA*
[‡]*Sandia National Labs, Livermore, CA, USA*

## ABSTRACT

Finding dense substructures in a graph is a fundamental graph mining operation, with applications in bioinformatics, social networks, and visualization to name a few. Yet most standard formulations of this problem (like clique, quasi-clique, k-densest subgraph) are NP-hard. Furthermore, the goal is rarely to find the "true optimum", but to identify many (if not all) dense substructures, understand their distribution in the graph, and ideally determine a hierarchical structure among them. Current dense subgraph finding algorithms usually optimize some objective, and only find a few such subgraphs without providing any hierarchy. It is also not clear how to account for overlaps in dense substructures.

We define the *nucleus decomposition* of a graph, which represents the graph as a *forest of nuclei*. Each nucleus is a subgraph where smaller cliques are present in many larger cliques. The forest of nuclei is a hierarchy by containment, where the edge density increases as we proceed towards leaf nuclei. Sibling nuclei can have limited intersections, which allows for discovery of overlapping dense subgraphs. With the right parameters, the nuclear decomposition generalizes the classic notions of $k$-cores and $k$-trusses.

We give provable efficient algorithms for nuclear decompositions, and empirically evaluate their behavior in a variety of real graphs. The tree of nuclei consistently gives a global, hierarchical snapshot of dense substructures, and outputs dense subgraphs of higher quality than other state-of-the-art solutions. Our algorithm can process graphs with tens of millions of edges in less than an hour.

## Categories and Subject Descriptors

F.2.2 [**Nonnumerical Algorithms and Problems**]: Computations on Discrete Structures; G.2.2 [**Graph Theory**]: Graph Algorithms

[*]Work done while the author was interning at Sandia National Labs, Livermore.



## General Terms

Algorithms

## Keywords

$k$-core, $k$-truss, decomposition, hierarchy, overlapping subgraphs, dense subgraph discovery

## 1. INTRODUCTION

Graphs are widely used to model relationships in a wide variety of domains such as sociology, bioinformatics, infrastructure, the WWW, to name a few. One of the key observations is that while real-world graphs are often globally sparse, they are locally dense. In other words, the average degree is often quite small (say at most 10 in a million vertex graph), but vertex neighborhoods are often dense. The classic notions of transitivity [47] and clustering coefficients [48] measure these densities, and are high for many real-world graphs [35, 40].

Finding dense subgraphs is a critical aspect of graph mining [30]. It has been used for finding communities and spam link farms in web graphs [29, 20, 13], graph visualization [2], real-time story identification [4], DNA motif detection in biological networks [18], finding correlated genes [49], epilepsy prediction [26], finding price value motifs in financial data [14], graph compression [8], distance query indexing [27], and increasing the throughput of social networking site servers [21]. This is closely related to the classic sociological notion of group cohesion [6, 17]. There are tangential connections to classic community detection, but the objectives are significantly different. Community definitions involve some relation of inner versus outer connections, while dense subgraphs purely focus on internal cohesion.

### 1.1 The challenges of dense subgraphs

Our input is a graph $G = (V, E)$. For vertex set $S$, we use $E(S)$ to denote the set of edges internal to $S$. The *edge density* of $S$ is $\rho(S) = |E(S)|/\binom{|S|}{2}$, the fraction of edges in $S$ with respect to the total possible. The aim is to find a set $S$ with high density subject to some size constraint. Typically, we are looking for large sets of high density.

In general, one can define numerous formulations that capture the main problem. The maximum clique problem is finding the largest $S$ where $\rho(S) = 1$. Finding the densest $S$ of size at least $k$ is the $k$-densest subgraph problem. Quasi-cliques, as defined recently by Tsourakakis et al. [43], are sets that are almost cliques, up to some fixed "defect." Unfortunately, most formulations of finding dense subgraphs

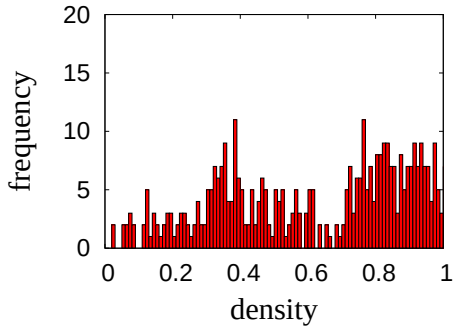



Figure 1: Density histogram of `facebook` $(3,4)$-nuclei. 145 nuclei have density of at least 0.8 and 359 nuclei are with the density of more than 0.25.

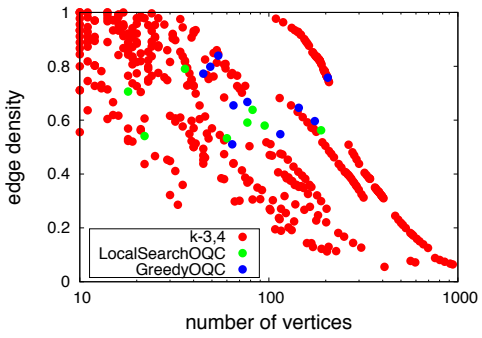



Figure 2: Size vs. density plot for `facebook` $(3,4)$-nuclei. 50 nuclei are larger than 30 vertices with the density of at least 0.8. There are also 138 nuclei larger than 100 vertices with density of at last 0.25.

are NP-hard, even to approximate [24, 16, 28].

For graph analysis, one rarely looks for just a single (or the optimal, for whatever notion) dense subgraph. We want to find many dense subgraphs and understand the relationships among them. Ideally, we would like to see if they nest within each other, if the dense subgraphs are concentrated in some region, and if they occur at various scales of size and density. Our paper is motivated by the following questions.

- How do we attain a global, hierarchical representation of many dense subgraphs in a real-world graph?
- Can we define an efficiently solvable objective that directly provides *many* dense subgraphs? We wish to avoid heuristics, as they can be difficult to predict formally.

## 1.2 Our results

**Nucleus decompositions:** Our primary theoretical contribution is the notion of of *nuclei* in a graph. Roughly speaking, an $(r,s)$-nucleus, for fixed (small) positive integers $r < s$, is a maximal subgraph where every $r$-clique is part of many $s$-cliques. (The real definition is more technical and involves some connectivity properties.) Moreover, nuclei that do not contain one another cannot share an $r$-clique. This

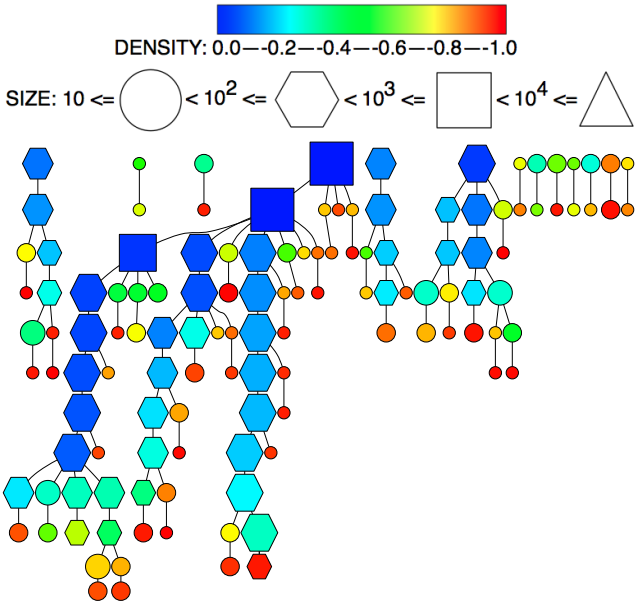



Figure 3: $(3,4)$-nuclei forest for `facebook`. Legends for densities and sizes are shown at the top. Long chain paths are contracted to single edges. In the uncontracted forest, there are 47 leaves and 403 nuclei. Branching depicts the different regions in the graph, 13 connected components exist in the top level. Sibling nuclei have limited overlaps up to 7 vertices.

is inspired by and is a generalization of the classic notion of $k$-cores, and also $k$-trusses (or triangle cores).

We show that the $(r,s)$-nuclei (for any $r < s$) form a hierarchical decomposition of a graph. The nuclei are progressively denser as we go towards the leaves in the decomposition. We provide an exact, efficient algorithm that finds all the nuclei and builds the hierarchical decomposition. In practice, we observe that $(3,4)$-nuclei provide the most interesting decomposition. We find the $(3,4)$-nuclei for a large variety of more than 20 graphs. Our algorithm is feasible in practice, and we are able to process a 39 million edge graph is less than an hour (using commodity hardware).

**Dense subgraphs from $(3,4)$-nuclei:** The $(3,4)$-nuclei provide a large set of dense subgraphs for range of densities and sizes. For example, there are 403 $(3,4)$-nuclei (of size at least 10 vertices) in a `facebook` network of 88K edges. We show the density histogram of these nuclei in Fig. 1, plotting the number of nuclei with a given density. Observe that we get numerous dense subgraphs, and many with density fairly close to 1. In Fig. 2, we present a scatter plot of vertex size vs density of the $(3,4)$-nuclei. Observe that we obtain dense subgraphs over a wide range of sizes. For comparison, we also plot the output of recent dense subgraph algorithms from Tsourakakis et al [43]. (These are arguably the state-of-the-art. More details in next section.) Observe that $(3,4)$-nuclei give dense subgraphs of comparable quality. In some cases, the output of [43] is very close to a $(3,4)$-nucleus.

**Representing a graph as forest of $(3,4)$-nuclei:** We build the forest of $(3,4)$-nuclei for all graphs experimented on. An example output is that of Fig. 3, the forest of $(3,4)$-nuclei for the `facebook` network. Each node of the forest is a $(3,4)$-nucleus, and tree edges indicate containment. More

generally, an ancestor nucleus contains all descendant nuclei. By the properties of $(3,4)$-nuclei, any two incomparable nodes do not share a triangle. So the branching in the forest represents different regions of the graph. (All nuclei of less than 10 vertices is omitted. For presentation, we contract long chain paths in the tree to single edges, so the forest has less than 403 nodes.)

In the nuclei figures, densities are color-coded, with hotter colors indicating higher density. The log of sizes are coded by shape (circles comprise between 10 and 100 vertices, hexagons between 100 and 1000 vertices, etc.) For a fixed shape, relative size corresponds to relative size in vertices. We immediately see the hierarachy of dense structures. Observe the colors becoming hotter as we go towards to leaves, which are mostly red (density $> 0.8$). We see numerous hexagons and large circles of color between light blue to green. These indicate the larger parent subgraphs of moderate density (actually density of say 0.25 is fairly high for a subgraph having many hundreds of vertices).

The branching is also significant, and we can group together the dense subgraphs according to the hierarchy. We observe such branching in all our experiments, and show more such results later in the paper. The $(3,4)$-nuclei provide a simple, hierarchical visualization of dense substructures. They are well-defined and their exact computation is algorithmically feasible and practical.

We also want to emphasize the overlap between sibling nuclei. While sibling nuclei cannot share triangles, but they can share edges. We observe roughly 20 pairs of $(3,4)$-nuclei having intersections of 4-6 vertices. For larger graphs, we observe many more pairs of intersecting nuclei (with larger intersections).

The rest of the paper is organized as follows: §2 summarizes the related work, §3 introduces the main definitions and the lemma about the nucleus decomposition, §4 gives the algorithm to generate a nucleus decomposition and provides a complexity analysis, §5 contains the results of extensive experiments we have, and §6 concludes the paper by discussing the future directions.

## 2. PREVIOUS WORK

**Dense subgraph algorithms:** As discussed earlier, most formulations of the densest subgraph problem are NP-hard. Some variants such as maximum average degree[22, 19] and the recently defined triangle-densest subgraph [44] are polynomial time solvable. Linear time approximation algorithms have been provided by Asashiro et al. [5], Charikar [9], and Tsourakakis [44]. There are numerous recent practical algorithm for various such objectives: Andersen and Chellapilla's use of cores for dense subgraphs [3], Rossi et al.'s surprising heuristic for clique [34], Tsourakakis et al.'s notion of quasi-cliques [43]. These algorithms are extremely efficient and produce excellent output. For comparison's sake, we consider Tsourakakis et al. [43] as the state-of-the-art, which was compared with previous core-based heuristics and is much superior to prior art. Indeed, their algorithms are elegant, extremely efficient, and provide high quality output (and much faster than ours. More discussion in §5.4). These methods are tailored to finding one (or a few) dense subgraphs, and do not give a global/hierarchical view of the structure of dense subgraphs. We believe it would be worthwhile to relate their methods with our notion of nuclei, to design even better algorithms.

**$k$-cores and $k$-trusses:** The concepts of $k$-cores and $k$-trusses form the inspiration for our work. A $k$-core is a maximal subgraph where each vertex has minimum degree $k$, while a $k$-truss is a subgraph where each edge participates in at least $k$ triangles. The first definition of $k$-cores was given by Erdős and Hajnal [15]. It has been rediscovered numerous times in the context of graph orientations and is alternately called the coloring number and degeneracy [31, 38]. The first linear time algorithm for computing $k$-cores was given by Matula and Beck [32]. The earliest applications of cores to social networks was given by Seidman [38], and it is now a standard tool in the analysis of massive networks. The notions of $k$-truss or triangle-cores were independently proposed by by Cohen [11], Zhang and Parthasarathy [50], and Zhao and Tung [51] for finding clusters and for network visualization. They all provide efficient algorithms for these decompositions, and Cohen [11] and Wang and Cheng [45] explicitly focus on massive scale. In [46], Wang et al. proposed DN-graph, a similar concept to $k$-truss, where each edge should be involved in $k$ triangles, and adding or removing a vertex from DN-graph breaks this constraint. Apart from the $k$-core and $k$-truss definitions, $k$-plex and $k$-club subgraph definitions have drawn a lot of interest as well. In a $k$-plex, each vertex is missing no more than $k-1$ edges to its neighbors [39]. It can be seen as a complementary $k$-core definition. In a $k$-club, the shortest path from any vertex to other vertex is not more than $k$ [33]. All these methods find subgraphs of moderate density, but give a global decomposition to visualize a graph.

## 3. NUCLEUS DECOMPOSITION

Our main theoretical contribution is the notion of nucleus decompositions. We have an undirected, simple graph $G$. We use $K_r$ to denote an $r$-clique and start with some technical definitions.

DEFINITION 1. *Let $r < s$ be positive integers and $\mathcal{S}$ be a set of $K_s$s in $G$.*

- *$K_r(\mathcal{S})$ the set of $K_r$s contained in some $S \in \mathcal{S}$.*
- *The number of $S \in \mathcal{S}$ containing $R \in K_r(\mathcal{S})$ is the* $\mathcal{S}$-degree *of that $K_r$.*
- *Two $K_r$s $R, R'$ are* $\mathcal{S}$-connected *if there exists a sequence $R = R_1, R_2, \ldots, R_k = R'$ in $K_r(\mathcal{S})$ such that for each $i$, some $S \in \mathcal{S}$ contains $R_i \cup R_{i+1}$.*

These definitions are simple generalizations of the standard notion of the vertex degree and connectedness. Indeed, setting $r = 1$ and $s = 2$ ($\mathcal{S}$ is a set of edges) yields exactly that. Our main definition follows.

DEFINITION 2. *Let $k$ and $r < s$ be positive integers. A* $k$-$(r,s)$-nucleus *is a maximal union $\mathcal{S}$ of $K_s$s such that:*

- *The $\mathcal{S}$-degree of any $R \in K_r(\mathcal{S})$ is at least $k$.*
- *Any $R, R' \in K_r(\mathcal{S})$ are $\mathcal{S}$-connected.*

We simply refer to $(r,s)$-nuclei when $k$ is unspecified. Note that we treat nuclei as a union of cliques, though eventually, we look at this as a subgraph. Our theoretical treatment is more convenient in the former setting, and hence we stick with this definition. In our applications, we simply look at nuclei as subgraphs.

Intuitively, a nucleus is a tightly connected cluster of cliques. For large $k$, we expect the cliques in $\mathcal{S}$ to intersect heavily,

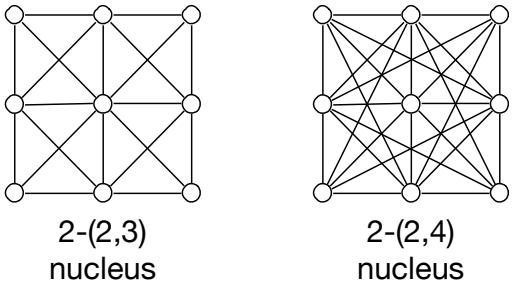



Figure 4: Having same number of vertices, 2-$(2,4)$ nucleus is denser than 2-$(2,3)$.

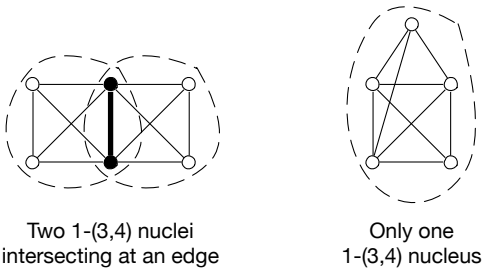



Figure 5: The left figure shows two $(3,4)$-nuclei overlapping at an edge. The right figure has only one $(3,4)$-nucleus

creating a dense subgraph. For a fixed $k, r$ and same number of vertices, the density of the nuclei increases, as we increase $s$. Consider the example of Fig. 4, where there is a 2-$(2,3)$-nucleus and a 2-$(2,4)$-nucleus on the same number of vertices. Since in the latter case, we need every edge to participate in at least 2 $K_4$s, the resulting density is much higher.

As stated earlier, our definitions are directly inspired by $k$-cores and $k$-trusses. Set $r = 1, s = 2$. A $k$-$(1,2)$-nucleus is a maximal (induced) connected subgraph with minimum vertex degree $k$. *This is exactly a $k$-core.* Setting $r = 2, s = 3$ gives maximal subgraphs where every edge participates in at least $k$ triangles, and edges are triangle-connected. This is essentially the definition of $k$-trusses or triangle-cores.

So far we only discussed the degree constraint of nuclei. Note that a nucleus is not just connected in the usual (edge) sense, but requires the stronger property of being $\mathcal{S}$-connected. The standard definitions of trusses or triangle-cores omit the triangle-connectedness. For us, this is critical. Two cliques of distinct $(r,s)$-nuclei *can* intersect. For example, when $r > 2$, nuclei can have edge overlaps. This allows for finding even denser subgraphs, as Fig. 5 shows. In the left, cores, trusses, etc. pick up the entire graph. But there are actually 2 different 1-$(3,4)$-nuclei (each $K_4$) intersecting at an edge. The $(3,4)$-nuclei are denser than the graph itself. Note that any edge disjoint decomposition would not find two dense subgraphs.

Critically, the set of $(r,s)$-nuclei form a laminar family. A laminar family is a set system where all pairwise intersections are trivial (either empty or contains one of the sets).

LEMMA 1. *The family of $(r,s)$-nuclei form a laminar family.*

PROOF. Consider $k$-$(r,s)$-nucleus $\mathcal{S}$ and $k'$-$(r,s)$-nucleus $\mathcal{S}'$, where $k \leq k'$. Suppose they had a non-empty intersection, so some $K_s$ $S$ is contained in both $\mathcal{S}$ and $\mathcal{S}'$. Observe that $K_r$s in $K_r(\mathcal{S})$ are connected to $K_r$s in $K_r(\mathcal{S}')$. Furthermore, the $(\mathcal{S} \cup \mathcal{S}')$-degree of member of $K_r(\mathcal{S} \cup \mathcal{S}')$ is at least $k$. Hence $\mathcal{S} \cup \mathcal{S}'$ satisfies the two conditions of being a nucleus, except maximality. By $\mathcal{S}$ is a $k$-$(r,s)$-nucleus, so $\mathcal{S} \cup \mathcal{S}' = \mathcal{S}$. So any non-empty intersection is trivial. □

Consider two nuclei that not ancestor-descendant. By the above lemma, these two two nuclei (considered as subgraphs of $G$) cannot share a $K_s$. Actually, the argument above proves that they cannot even share a $K_r$. This is the key disjointness property of nuclei.

Every laminar family is basically a hierarchical set system. Alternately, every laminar family can be represented by a forest of containment. For every nucleus $\mathcal{S}$, any other nucleus intersecting $\mathcal{S}$ is either contained in $\mathcal{S}$ or contains $\mathcal{S}$. Furthermore, all these sets are nested in each other. It makes sense to talk of the smallest sized nucleus containing $\mathcal{S}$. This leads to the main construct we use to represent nuclei.

DEFINITION 3. *Fix $r < s$. Define the* forest of $(r,s)$-nuclei *as follows. There is a node for each $(r,s)$ nucleus. The parent of every nucleus is the smallest (by cardinality) other nucleus containing it.*

In our figures, we will only show the internal nodes of out degree at least 2, and contract any path of out degree 1 vertices into a single path. This preserves all the branching of the forest.

# 4. GENERATING NUCLEUS DECOMPOSITIONS

Our primary algorithmic goal is to construct the tree of nuclei. The algorithm is a direct adaptation of the classic Matula-Beck result of getting $k$-cores in linear time [32]. There are numerous technicalities involved in generalizing the proof. Intuitively, we do the following. Construct a graph $\mathcal{H}$ where the nodes are all $K_r$s of $G$ and there is an edge connecting two $K_r$s if they are contained in a single $K_s$ of $G$. We then perform a core decomposition on $\mathcal{H}$. Actually, this does not work. Edges of $G$ (obviously) contain exactly 2 vertices of $G$, and the procedure above exactly produces nuclei for $r = 1, s = 2$. In general, a $K_s$ contains $\binom{s}{r}$ $K_r$s, and the graph analogy above is incorrect. At some level, we are performing a hypergraph version of Matula-Beck. The proofs therefore need to be adapted to this setting.

Analogous to $k$-cores, the main procedure `set-k` assigns a number, denoted by $\kappa(\cdot)$, to each $K_r$ in $G$.

It is convenient to denote the set of $K_r$s in $G$ by $R_1, R_2, \ldots$, where $R_i$ is the $i$th processed $K_r$ in `set-k`. We will refer to this index as *time*. When we say "at time $t$", we mean at the beginning of the iteration where $R_t$ is processed.

CLAIM 1. *The sequence $\{\kappa(R_i)\}$ is monotonically non-decreasing.*

PROOF. This holds because the loop goes $R$ in non-decreasing order of $d(R)$ and Step 11 ensures that no new value of $\delta(\cdot)$ decreases below the current $k(R)$. □

- Because of Claim 1, we can define *transition time* $t_i$ to be the first time when the $k$-value becomes $i$. Formally, $t_i$ is the unique index such that $k(R_{t_i}) = i$ and $k(R_{t_i - 1}) < i$.

**Algorithm 1:** `set-k`$(G, r, s)$

```
1  Enumerate all K_r s and K_s s in G(V, E);
2  For every K_r R, initialize δ(R) to be the number of K_s
   containing R;
3  Mark every K_r as unprocessed;
4  for each unprocessed K_r R with minimum δ(R) do
5      κ(R) = δ(R);
6      Find set S of K_s s containing R;
7      for each S ∈ S do
8          if any K_r R' ⊂ S is processed then
9              Continue;
10         for each K_r R' ⊂ S, R' ≠ R do
11             if δ(R') > δ(R) then
12                 δ(R') = δ(R') − 1 ;
13     Mark R as processed;
14 return array k(·) ;
```

- We say $K_s$ $S$ is *unprocessed at time $t$* if all $R \in K_r(S)$ are unprocessed at time $t$. The set of $K_s$s is denoted by $\mathcal{S}_t$.
- The *supergraph* $\mathcal{G}_t$ has node set $K_r(\mathcal{S}_t)$, and $R, R' \in K_r(\mathcal{S}_t)$ are connected by a link if $R \cup R'$ is contained in some $K_s$ of $\mathcal{S}_t$. Links are associated with elements of $\mathcal{S}_t$ (and there may be multiple links between $R$ and $R'$).

We prove an auxiliary claim relating the $\delta(\cdot)$ values to $\mathcal{S}_t$.

Claim 2. *At time $t$, for any unprocessed $K_r$ $R$, $\delta(R)$ is at least the $\mathcal{S}_t$-degree of $R$. If $t = t_k$ (for some $k$), then $\delta(R)$ is exactly the $\mathcal{S}_t$-degree of $R$.*

Proof. Pick unprocessed $R'$. The value of $\delta(R)$ is initially the number of $K_s$s containing $R'$. It is decremented only in Step 12, which happens only when a processed $K_s$ containing $R'$ is found. (Sometimes, the decrement will still not happen, because of Step 11.) Hence, the value of $\delta(R')$ at time $t$ is at least the number of unprocessed $K_s$s containing $R'$.

Suppose $t = t_k$. For any preceding $\hat{t} < t$, the current $\kappa(\cdot)$ value is always at most $k$. For unprocessed (at time $t$) $R$, $\delta(R) > k$. Hence the decrement of Step 12 will always happen, and $\delta(R)$ is exactly the $\mathcal{S}_t$-degree of $R$. □

Claim 3. *Every $k$-$(r,s)$-nucleus is contained in $\mathcal{S}_{t_k}$.*

Proof. Consider $k$-$(r,s)$-nucleus $\mathcal{S}$. Take the first $R \in K_r(\mathcal{S})$ that is processed. At this time (say $t$), no $S \in \mathcal{S}$ can be processed. Hence, $\mathcal{S} \subseteq \mathcal{S}_t$. By Claim 2, $d(R)$ is at least the $\mathcal{S}_t$-degree of $R$, which is at least the $\mathcal{S}$-degree of $R$. The latter is at least $k$, since $\mathcal{S}$ is a $k$-$(r,s)$-nucleus. By definition of $t_k$, $t \geq t_k$ and hence $\mathcal{S}_t \subseteq \mathcal{S}_{t_k}$. Thus, $\mathcal{S} \subseteq \mathcal{S}_{t_k}$. □

The main lemma shows that the output of `set-k` essentially tells us the nuclei.

Lemma 2. *The $k$-$(r,s)$-nuclei are exactly the links (which are $K_s$s) of connected components of $\mathcal{G}_{t_k}$.*

Proof. Consider $k$-$(r,s)$-nucleus $\mathcal{S}$. By Claim 3, it is contained in $\mathcal{S}_{t_k}$. By the nucleus definition, $\mathcal{S}$ is connected (as links) in $\mathcal{G}_{t_k}$. Let $\mathcal{S}'$ be the (set of links) connected component of $\mathcal{G}_{t_k}$ containing $\mathcal{S}$. By Claim 2, at time $t_k$, for any $R \in K_r(\mathcal{S}')$, $\delta(R)$ is exactly the $\mathcal{S}_{t_k}$-degree of $R$. Since $\mathcal{S}'$ is a connected component of $\mathcal{G}_{t_k}$, the $\mathcal{S}_{t_k}$-degree is the $\mathcal{S}'$-degree, which in turn is at least $k$. In other words, $\mathcal{S}'$ satisfies both conditions of being a $k$-$(r,s)$-nucleus, except maximality. By maximality of $\mathcal{S}$, $\mathcal{S} = \mathcal{S}'$. □

**Building the forest of nuclei:** From Lem. 2, it is fairly straightforward to get all the nuclei. First run `set-k` to get the processing times and the $k(\cdot)$ values. We can then get all $t_k$ times as well. Suppose for any $K_r$ in $G$, we can access all the $K_s$s containing it. Then, it is routine to traverse $\mathcal{G}_{t_k}$ to get the links of connected components. To avoid traversing the same component repeatedly, we produce nuclei in reverse order of $k$. In other words, suppose all connected components of $\mathcal{G}_{t_{k+1}}$ have been determined. For $\mathcal{G}_{t_k}$, it suffices to determine the connected components involving nodes processed in time $[t_k, t_{k+1})$. Any time a traversal encounters a node in $\mathcal{G}_{t_{k+1}}$, we need not traverse further. This is because all other connected nodes of $\mathcal{G}_{t_{k+1}}$ are already known from previous traversals. We do not get into the data structure details here, but it suffices to visit all nodes and links of $\mathcal{G}_0$ exactly once.

## 4.1 Bounding the complexity

There are two options of implementing this algorithm. The first is faster, but has forbiddingly large space. The latter is slower, but uses less space. In practice, we implement the latter algorithm. We use $ct_r(v)$ for the number of $K_r$s containing $v$ and $ct_r(G)$ for the total number of $K_r$s in $G$. We denote by $RT_r(G)$ the running time of an arbitrary procedure that enumerates all $K_r$s in $G$.

Theorem 1. *It is possible to build the forest of nuclei in $O(RT_r(G) + RT_s(G))$ time with $O(ct_r(G) + ct_s(G))$ space.*

Proof. This is the obvious implementation. The very first step of `set-k` requires the clique enumeration. Suppose we store the global supergraph $\mathcal{G} = \mathcal{G}_0$. This has a node for every $K_r$ in $G$ and a link for every $K_s$ in $G$. The storage is $O(ct_r(S) + ct_s(G))$. From this point onwards, all remaining operations are linear in the storage. This is by the analysis of the standard core decomposition algorithm of Matula and Beck [32]. Every time we process a $K_r$, we can delete it and all incident links from $\mathcal{G}$. Every link is touched at most a constant number of times during the entire running on `set-k`. As explained earlier, we can get all the nuclei by a single traversal of $\mathcal{G}$. □

Theorem 2. *It is possible to build the forest of nuclei in $O(RT_r(G) + \sum_v ct_r(v) d(v)^{s-r})$ time with $O(ct_r(G))$ space.*

Proof. Instead of explicitly building $\mathcal{G}$, we only build adjacency lists when required. The storage is now only $O(ct_r(G))$. In other words, given a $K_r$ $R$, we find all $K_s$s containing $R$ only when $R$ is processed/traversed. Each $R$ is processed or traversed at most once in `set-k` and the forest building. Suppose $R$ has vertices $v_1, v_2, \ldots, v_r$. We can find all $K_s$s containing $R$ by looking at all $(s-r)$-tuples in each of the neighborhoods of $v_i$. (Indeed, it suffices to look at just one such neighborhood.) This takes time at most $\sum_R \sum_{v \in R} d(v)^{s-r} = \sum_v \sum_{R \ni v} d(v)^{s-r} = \sum_v ct_r(v) d(v)^{s-r}$. □

Let us understand these running times. When $r < s \leq 3$, it clearly benefits to go with Thm. 1. Triangle enumeration is a well-studied problem and there exist numerous optimized, parallel solutions for the problem. In general, the classic triangle enumeration of Chiba and Nishizeki takes $O(m^{3/2})$ [10] and is much better in practice [12, 37, 42]. This completely bounds the time and space complexities.

For our best results, we build the $(3,4)$-nuclei, and the number of $K_4$s is too large to store. We go with Thm. 2.

| | \|V\| | \|E\| | Description | $\sum_v c_3(v)d(v)$ | $(3,4)$ time | [43] Density (size) | $(3,4)$-nucleus Density (size) |
|---|---|---|---|---|---|---|---|
| dolphins | 62 | 159 | Biological | $2.2K$ | $< 1$ | 0.68(8) | 0.71(8) |
| polbooks | 105 | 441 | US Politics Books | $23.8K$ | $< 1$ | 0.67(13) | 0.62(13) |
| adjnoun | 112 | 425 | Adj. and Nouns | $17.6K$ | $< 1$ | 0.60(15) | 0.22(32) |
| football | 115 | 613 | World Soccer 98 | $26.3K$ | $< 1$ | 0.89(10) | 0.89(10) |
| jazz | 198 | $2.74K$ | Musicians | $2.3M$ | $< 1$ | 1.00(30) | 1.00(30) |
| uelegans n. | 297 | $2.34K$ | Biological | $418K$ | $< 1$ | 0.61(21) | 0.91(10) |
| celegans m. | 453 | $2.04K$ | Biological | $565K$ | $< 1$ | 0.67(17) | 0.64(18) |
| email | $1.13K$ | $5.45K$ | Email | $1.2M$ | $< 1$ | 1.00(12) | 1.00(12) |
| facebook | $4.03K$ | $88.23K$ | Friendship | $712M$ | 93 | 0.83(54) | 0.98(109) |
| protein_inter. | $9.67K$ | $37.08K$ | Protein Inter. | $35M$ | $< 1$ | 1.00(11) | 1.00(11) |
| as-22july06 | $22.96K$ | $48.43K$ | Autonomous Sys. | $199M$ | $< 1$ | 0.58(12) | 1.00(18) |
| twitter | $81.30K$ | $2.68M$ | Follower-Followee | $1.8B$ | 396 | 0.85(83) | 1.00(26) |
| soc-sign-epinions | $131.82K$ | $841.37K$ | Who-trust-whom | $1.4B$ | 242 | 0.71(79) | 1.00(112) |
| coAuthorsCiteseer | $227.32K$ | $814.13K$ | CoAuthorship | $2.1B$ | 50.1 | 1.00(87) | 1.00(87) |
| citationCiteseer | $268.49K$ | $1.15M$ | Citation | $297M$ | 3.4 | 0.71(10) | 1.00(13) |
| web-NotreDame | $325.72K$ | $1.49M$ | Web | $33.9B$ | 671 | 1.00(151) | 1.00(155) |
| amazon0601 | $403.39K$ | $3.38M$ | CoPurchase | $802M$ | 23 | 1.00(11) | 1.00(11) |
| web-Google | $875.71K$ | $5.10M$ | Web | $11.4B$ | 163 | 1.00(46) | 1.00(33) |
| com-youtube | $1.13M$ | $2.98M$ | Social | $451M$ | 43 | 0.49(119) | 0.92(24) |
| as-skitter | $1.69M$ | $11.09M$ | Autonomous Sys. | $1.6B$ | $1,036$ | 0.53(319) | 0.94(91) |
| wikipedia-2005 | $1.63M$ | $19.75M$ | Wikipedia Link | $741B$ | $1,312$ | 0.53(33) | 0.82(14) |
| wiki-Talk | $2.39M$ | $5.02M$ | Wikipedia User | $136B$ | 605 | 0.48(321) | 0.59(95) |
| wikipedia-200609 | $2.98M$ | $37.26M$ | Wikipedia Link | $2,015B$ | $2,830$ | 0.49(376) | 0.62(103) |
| wikipedia-200611 | $3.14M$ | $39.38M$ | Wikipedia Link | $2,197B$ | $3,039$ | 1.00(55) | 1.00(32) |

Table 1: Important statistics for the real-world graphs of different types and sizes. Largest graph in the dataset has more than $39M$ edges. Times are in seconds. Density of subgraph $S$ is $|E(S)|/\binom{|S|}{2}$ where $E(S)$ is the set of edges internal to $S$. Sizes are in number of vertices.

The storage is now at most the number of triangles, which is manageable. The running time is basically bounded by $O(\sum_v ct_r(v)d(v))$. The number of triangles incident to $v$, $ct_3(v)$ is $cc(v)d(v)^2$, where $cc(v)$ is the clustering coefficient of $v$. We therefore get a running time of $O(\sum_v cc(v)d(v)^3)$. This is significantly superlinear, but clustering coefficients generally decay with degree [35, 40]. Overall, the implementation can be made to scale to tens of millions of edges with little difficulty.

# 5. EXPERIMENTAL RESULTS

We applied our algorithm to large variety of graphs, obtained from from SNAP [41] and UF Sparse Matrix Collection[1]. The vital statistics of these graphs are given in Tab. 1. All the algorithms in our framework are implemented in C++ and compiled with `gcc 4.8.1` at -O2 optimization level. All experiments are performed on a Linux operating system running on a machine with two Intel Xeon E5520 2.27 GHz CPUs, with 48GB of RAM.

We computed the $(r,s)$-nuclei for all choices of $r < s \leq 4$, but do not present all results for space considerations. We mostly observe that the forest of $(3,4)$-nuclei provides the highest quality output, both in terms of hierarchy and density.

As mentioned earlier, we will now treat the nuclei as just induced subgraphs of $G$. A nucleus can be thought of a set of vertices, and we take all edges among these vertices (induced subgraph) to attain the subgraph. The *size* of a nucleus always refers to the number of vertices, unless otherwise specified. For any set $S$ of vertices, the density of the induced subgraph is $|E(S)|/\binom{|S|}{2}$, where $E(S)$ is the set of edges internal to $S$. *We ignore any nucleus with less than 10 vertices. Such nuclei are not considered in any of our results.*

For the sake of demonstration, we present detailed results on only 4 graphs (given in Tab. 1): `facebook`, `soc-sign-epinions`, `web-NotreDame`, and `wikipedia-200611`. This covers a variety of graphs, and other results are analogous.

## 5.1 The forest of nuclei

We were able to construct the forest of $(3,4)$-nuclei for all graphs in Tab. 1, but only give the forests for `facebook` (Fig. 3), `soc-sign-epinions` (Fig. 6), and `web-NotreDame` (Fig. 7). For the `web-NotreDame` figure, we could not present the entire forest, so we show some trees in the forest that had nice branching. The density is color coded, from blue (density 0) to red (density 1). The nuclei sizes, in terms of vertices, are coded by shape: circles correspond to at most $10^2$ vertices, hexagons in the range $[10^2, 10^3]$, squares in the range $[10^3, 10^4]$, and triangles are anything larger. The relative size of the shape, is the relative size (in that range) of the set.

Overall, we see that the $(3,4)$-nuclei provide a hierarchical representation of the dense subgraphs. The leaves are mostly red, and their densities are almost always $> 0.8$. But we obtain numerous nuclei of intermediate sizes and densities. In the `facebook` forest and to some extent in the `web-NotreDame` forest, we see hexagons of light blue to green (nuclei subgraphs of $> 100$ vertices of densities of at least 0.2). The branching is quite prominent, and the smaller dense nuclei tend to nest into larger, less dense nuclei. This held in every single $(3,4)$-nucleus forest we computed. This appears to validate the intuition that real-world networks have a hierarchical structure.

The (3.4)-nuclei figures provide a useful visualization of the dense subgraph structure. The `web-NotreDame` has a million edges, and it is not possible to see the graph as a whole. But the forest of nuclei breaks it down into meaningful parts, which can be visually inspected. The overall forest is large (about 900 nuclei), but the nesting structure makes it easy to absorb. We have not presented the results here, but even the `wikipedia-200611` graph of 38 million edges has about a forest of only 4000 nuclei (which we were

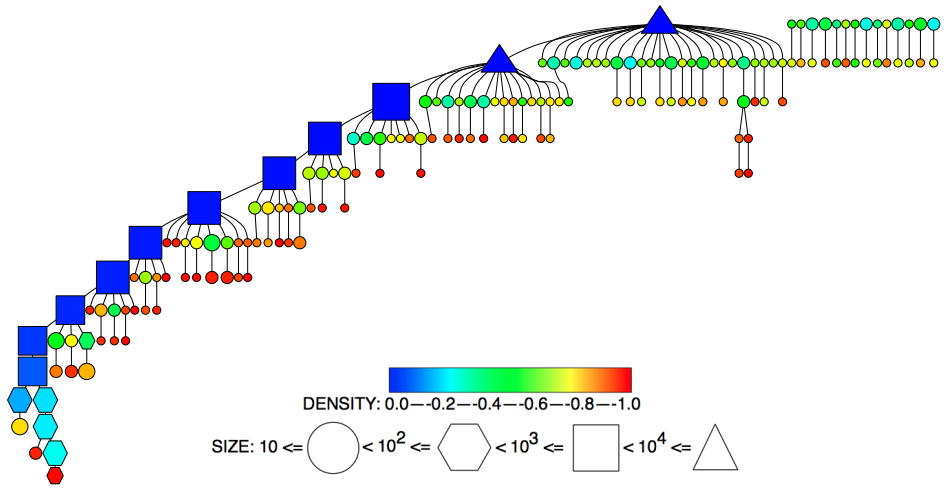



Figure 6: $(3,4)$-nuclei forest for `soc-sign-epinions`. There are 465 total nodes and 75 leaves in the forest. There is a clear hierarchical structure of dense subgraphs. Leaves are mostly red ($> 0.8$ density). There are also some light blue hexagons, representing subgraphs of size $\geq 100$ vertices with density of at least 0.2.

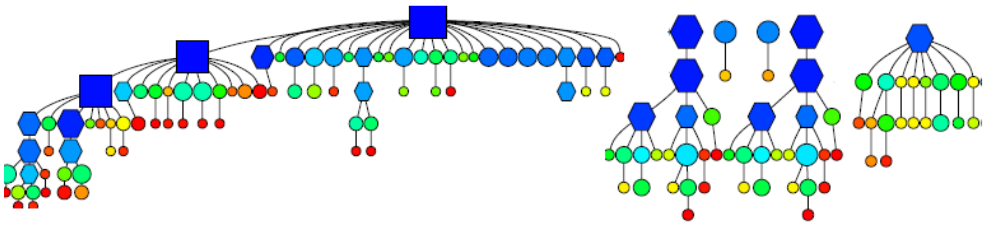
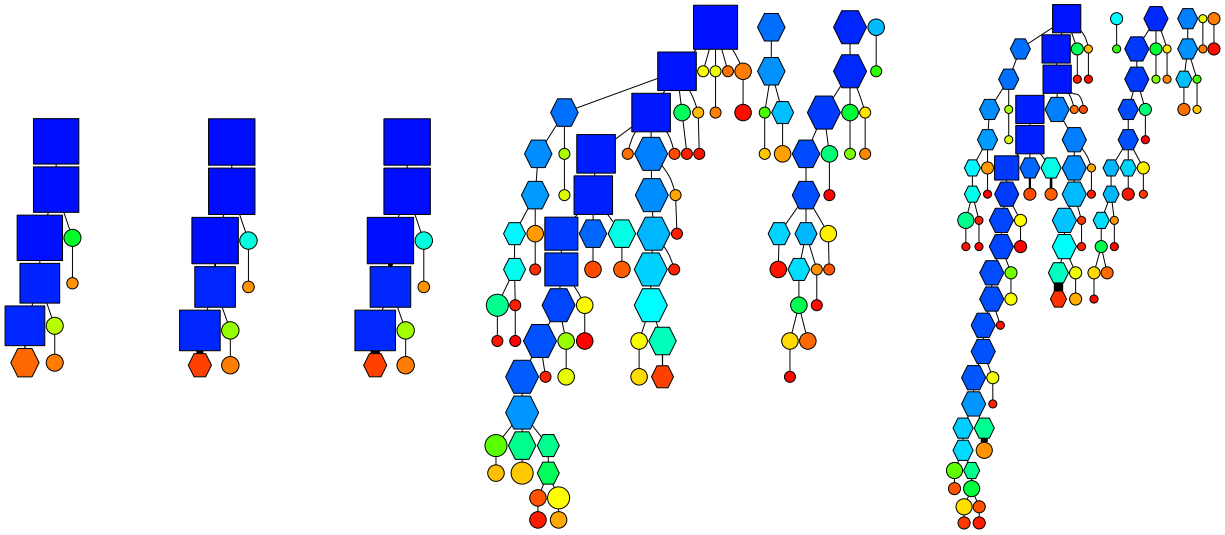
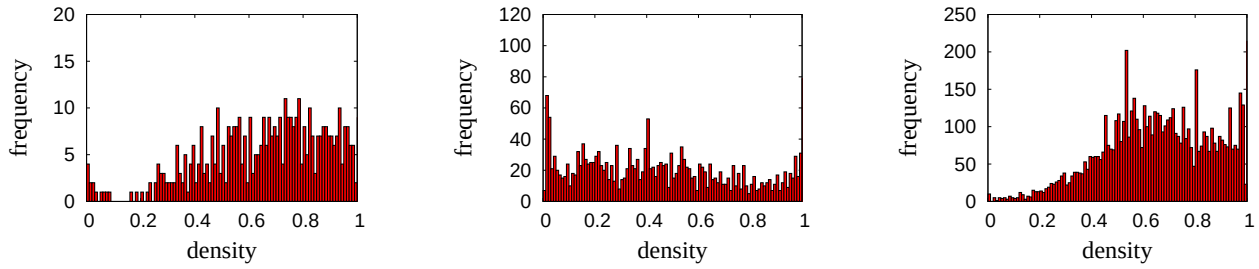


Figure 7: Part of the $(3,4)$-nuclei forest for `web-NotreDame`. In the entire forest, there are 2059 nodes and 812 leaves. 79 of the leaves are clique, up to size of 155. There is a nice branching structure leading to a decent hierarchy.



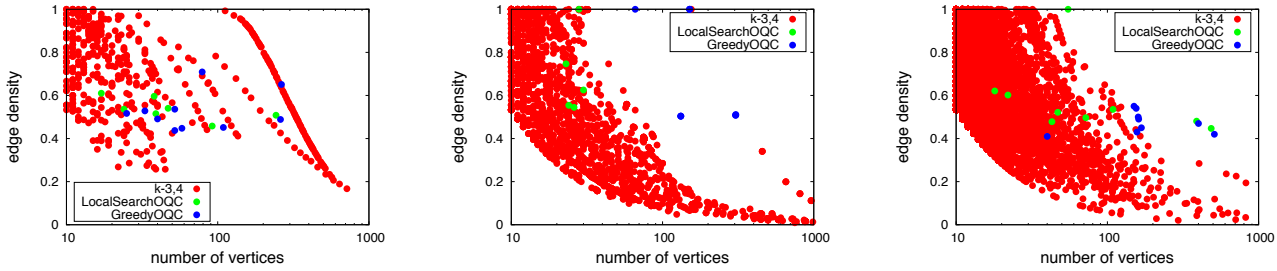
a: $(1,2)$-nuclei b: $(1,3)$-nuclei c: $(1,4)$-nuclei d: $(2,3)$-nuclei e: $(2,4)$-nuclei

Figure 8: $(r,s)$-nuclei forests for `facebook` when $r < s \leq 4$ (Except $(3,4)$, which is given in Fig. 3). For $r = 1$, trees are more like chains. Increasing $s$ results in larger number of internal nodes, which are contracted in the illustrations. There is some hierarchy observed for $r = 2$, but it is not as powerful as $(3,4)$-nuclei, i.e., branching structure is more obvious in $(3,4)$-nuclei.




(a) `soc-sign-epinions` (b) `web-NotreDame` (c) `wikipedia-200611`

Figure 9: Density histograms for nuclei of three graphs. $x$-axis (binned) is the density and $y$-axis is the number of nuclei (at least 10 vertices) with that density. Number of nuclei with the density above 0.8 is significant: 139 for `soc-sign-epinions`, 355 for `web-NotreDame`, and 1874 for `wikipedia-200611`. Also notice that, the mass of the histogram is shifted to right in `soc-sign-epinions` and `wikipedia-200611` graphs.




(a) `soc-sign-epinions` (b) `web-NotreDame` (c) `wikipedia-200611`

Figure 10: Density vs. size plots for nuclei of three graphs. State-of-the-art algorithms are depicted with `*OQC` variants, and they report one subgraph at each run. We ran them 10 times to get a general picture of the quality. Overall, $(3,4)$-nuclei is very competitive with the state-of-the-art and produces many number of subgraphs with high quality and non-trivial sizes.

able to easily visualized by a drawing tool).

Other choices of $r, s$ for the nuclei do not lead to much branching. We present all nucleus trees for $r < s \leq 4$ for the `facebook` graph in Fig. 8 (except $(3,4)$ which is given in Fig. 3). Clearly, when $r = 1$, the nucleus decomposition is boring. For $r = 2$, some structure arises, but not as dramatic of Fig. 3. Results vary over graphs, but for $r = 1$, there is pretty much just a chain of nuclei. For $r = 2$, some graphs show more branching, but we consistently see that for $(3,4)$-nuclei, the forest of nuclei is always branched.

## 5.2 Dense subgraph discovery

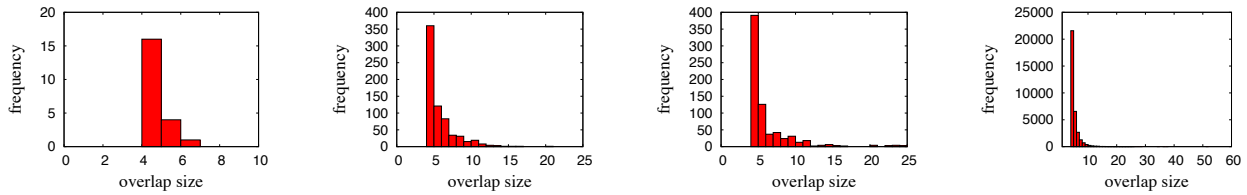
We plot the density histograms of the $(3,4)$-nuclei for various graphs in Fig. 9. The $x$-axis is (binned) density and the $y$-axis is the number of nuclei (all at least 10 vertices) with that density. It can be immediately seen that we find many non-trivial dense subgraphs. It is surprising to note how many near cliques (density $> 0.9$) we find. We tend to find more subgraphs of high density, and other than the `web-NotreDame` graph, the mass of the histogram is shifted to the right. The number of subgraphs of density at least 0.5 is in the order of hundreds (and more than a thousand for `wikipedia-200611`).

An alternate presentation of the dense subgraphs is a scatter plot of all $(3,4)$-nuclei with size in vertices versus density. This is given in Fig. 2 and Fig. 10, where the red dots correspond to the nuclei. We see that dense subgraphs are obtained in all scales of size, which is an extremely important feature. Nuclei capture more than just the densest (or high density) subgraphs, but find large sets of lower density (say around 0.2). Note that 0.2 is a significant density for sets of hundreds of vertices.

### 5.2.1 *Comparisons with previous art*

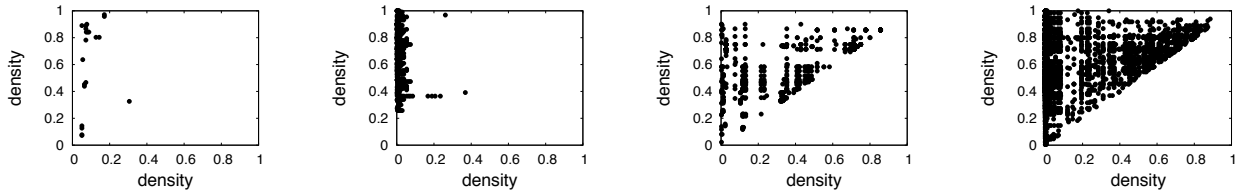
How does the quality of dense subgraphs found compare to the state-of-the-art? In the scatter plots of Fig. 2 and Fig. 10, we also show the output of two algorithms of [43] in green and blue. The idea of [43] is to approximate *quasi-cliques*, and their result provides two every elegant algorithms for this process. (We collectively refer to them as OQC.) OQC algorithms only give a single output, so we performed multiple runs to get many dense subgraphs. This is consistent with what was done in [43]. OQC algorithms clearly beat previous heuristics and it is fair to say that [43] is the state-of-the-art.

The $(3,4)$-nucleus decomposition does take significantly




(a) `facebook` (b) `soc-sign-epinions` (c) `web-NotreDame` (d) `wikipedia-200611`

Figure 11: Histograms over non-trivial overlaps for $(3,4)$-nuclei. Child-ancestor intersections are omitted. Overlap size is in terms of the number of vertices. Most overlaps are small in size. We also observe that $(2,s)$-nuclei, where $2 < s$, give almost no overlaps.




(a) `facebook` (b) `soc-sign-epinions` (c) `web-NotreDame` (d) `wikipedia-200611`

Figure 12: Overlap scatter plots for $(3,4)$-nuclei. Each axis shows the edge density of a participating nucleus in the pair-wise overlap. Larger density is shown on the $y$-axis. $(3,4)$-nuclei is able to get overlaps between very dense subgraphs, especially in `web-NotreDame` and `wikipedia-200611`. In `wikipedia-200611` graph, there are 1424 instances of pair-wise overlap between two nuclei, where each nucleus has the density of at least 0.8.

longer that the algorithms of [43]. But we always get much denser subgraphs in all runs. Moreover, the sizes are comparable if not larger than the output of [43]. Surprisingly, in `facebook` and `soc-sign-epinions`, some of the best outputs of OQC are very close to $(3,4)$-nuclei. Arguably, the $(3,4)$-nuclei perform worst on `wikipedia-200611`, where OQC find some larger and denser instances than $(3,4)$-nuclei. Nonetheless, the smaller $(3,4)$-nuclei are significantly denser. We almost always can find fairly large cliques.

In Tab. 1, we consider the OQC output vs $(3,4)$-nuclei for all graphs. Barring 4 instances, there is a $(3,4)$-nucleus that is larger and denser than the OQC output. In all cases but one (adjnoun), there is a $(3,4)$-nucleus of density (of non-trivial size) higher than the the OQC output. The nuclei have the advantage of being the output of a fixed, deterministic procedure, and not a heuristic that may give different outputs on different runs. We mention that OQC algorithms have a significant running time advantage over finding $(3,4)$-nuclei, for a single subgraph finding.

## 5.3 Overlapping nuclei

A critical aspect of nuclei is that they can overlap. Grappling with overlap is a major challenge when dealing with graph decompositions. We believe one of the benefits of nuclei is that they naturally allow for (restricted) overlap. As mentioned earlier, no two $(r,s)$-nuclei can contain the same $K_r$. This is a significant benefit of setting $r = 3, s = 4$ over other choices.

In Fig. 11, we plot the histogram over non-trivial overlaps for $(3,4)$-nuclei. (We naturally do not consider a child nucleus intersecting with an ancestor.) For a given overlap size in vertices, the frequency is the number of pairs of $(3,4)$-nuclei with that overlap. This is shown for 4 different graphs. The total number of pair-wise overlaps (the sum of frequencies) is typically around half the total number of $(3,4)$-nuclei. We observed that the Jaccard similarities are less than 0.1 (usually smaller). This suggest that we have large nuclei with some overlap.

There are bioinformatics applications for finding vertices that are present in numerous dense subgraphs [25]. The $(3,4)$-nuclei provide many such vertices. In Fig. 12, we give a scatter plot of all intersecting nuclei, where nuclei are indexed by density. For two intersecting nuclei of density $\alpha > \beta$, we put a point $(\alpha, \beta)$. We only plot pairs where the overlap is at least 5 vertices. Especially for `web-NotreDame` and `wikipedia-200611`, we get significant overlaps between dense clusters.

In contrast, for all other settings of $r, s$, we get almost no overlap. When $r < 3$, nuclei can only overlap at vertices, and this is too stringent to allow for interesting overlap.

## 5.4 Runtime results

Tab. 1 presents the runtimes in seconds for the entire construction. To provide some context, we describe runtimes for varying choices of $r, s$. For $r = 1, s = 2$ ($k$-cores), the decomposition is linear and extremely fast. For the largest graph (`wikipedia-200611`) we have, with $39M$ edges, it takes only 4.26 seconds. For $r = 2, s = 3$ (trusses), the time can be two orders of magnitude higher. And for $(3,4)$-nuclei, it is an additional order of magnitude higher. Nonetheless, our most

expensive run took less than an hour on the `wikipedia-200611` graph, and the final decomposition is quite insightful. It provides about 6000 nuclei with more than 10 vertices, most of them of have density of at least 0.4. The algorithms of [43] take roughly a minute for `wikipedia-200611` to produce only one dense subgraph.

The theoretical running time analysis of Thm. 2 gives a running time bound of $\sum_v c_3(v)d(v)$. In Tab. 1, we show this value for the various graphs. In general, we note that this value roughly correlates with the running time. For graphs where the running time is in many minutes, this quantity is always in the billions. For the large wiki graphs where the $(3,4)$-nucleus decomposition is most expensive, this is in the trillions.

# 6. FURTHER DIRECTIONS

The most important direction is in the applications of nucleus decompositions. We are currently investigating bioinformatics applications, specifically protein-protein and protein-gene interaction networks. Biologists often want a global view of the dense substructures, and we believe the $(3,4)$-nuclei could be extremely useful here. In our preliminary analyses, we wish to see if the nuclei pick out specific functional units. If so, that would provide strong validation of dense subgraph analyses for bioinformatics.

It is natural to try even larger values of $r, s$. Preliminary experimentation suggested that this gave little benefit in either the forest or the density of nuclei. Also, the cost of clique enumeration becomes forbiddingly large. It would be nice to argue that $r = 3, s = 4$ is a sort of sweet spot for nucleus decompositions. Previous theoretical work suggests that any graph with a sufficient triangle count undergo special "community-like" decompositions [23]. That might provide evidence to why triangle based nuclei are enough.

A faster algorithm for the $(3,4)$-nuclei is desirable. Clique enumeration is a well-studied problem [7], and we hope techniques from these results may provide ideas here. Of course, as we said earlier, any method based on storing $K_4$s is infeasible (space-wise). We hope to devise a clever algorithm or data structure that quickly determines the $K_4$s a triangle participates in.

Last but not least, we seek for incremental algorithms to maintain the $(r,s)$-nuclei for a stream of edges. There are existing techniques for streaming $k$-core algorithms [36] and we believe that similar methods can be adapted for $(r,s)$-nuclei maintenance.

# 7. ACKNOWLEDGEMENTS

We are grateful to Charalampos Tsourakakis for sharing his code base for [43]. This work was funded by the DARPA GRAPHS program. Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

# 8. REFERENCES

[1] University of florida sparse matrix collection. http://www.cise.ufl.edu/research/sparse/matrices/.

[2] J. I. Alvarez-Hamelin, A. Barrat, and A. Vespignani. Large scale networks fingerprinting and visualization using the k-core decomposition. In *Advances in Neural Information Processing Systems 18*, pages 41–50. MIT Press, 2006.

[3] R. Andersen and K. Chellapilla. Finding dense subgraphs with size bounds. In *Workshop on Algorithms and Models for the Web-Graph (WAW)*, pages 25–37, 2009.

[4] A. Angel, N. Sarkas, N. Koudas, and D. Srivastava. Dense subgraph maintenance under streaming edge weight updates for real-time story identification. *Proc. VLDB Endow.*, 5(6):574–585, Feb. 2012.

[5] Y. Asahiro, K. Iwama, H. Tamaki, and T. Tokuyama. Greedily finding a dense subgraph. *J. Algorithms*, 34(2):203–221, Feb. 2000.

[6] D. J. Beal, R. Cohen, M. J. Burke, and C. L. McLendon. Cohesion and performance in groups: A meta-analytic clarification of construct relation. *Journal of Applied Psychology*, 88:989–1004, 2003.

[7] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, Sept. 1973.

[8] G. Buehrer and K. Chellapilla. A scalable pattern mining approach to web graph compression with communities. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 95–106, New York, NY, USA, 2008. ACM.

[9] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *Proceedings of the Third International Workshop on Approximation Algorithms for Combinatorial Optimization*, APPROX '00, pages 84–95, London, UK, UK, 2000. Springer-Verlag.

[10] N. Chiba and T. Nishizeki. Arboricity and subgraph listing algorithms. *SIAM J. Comput.*, 14:210–223, February 1985.

[11] J. Cohen. Trusses: Cohesive subgraphs for social network analysis. Natioanal Security Agency Technical Report, 2008.

[12] J. Cohen. Graph twiddling in a MapReduce world. *Computing in Science & Engineering*, 11:29–41, 2009.

[13] Y. Dourisboure, F. Geraci, and M. Pellegrini. Extraction and classification of dense communities in the web. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 461–470, New York, NY, USA, 2007. ACM.

[14] X. Du, R. Jin, L. Ding, V. E. Lee, and J. H. T. Jr. Migration motif: a spatial - temporal pattern mining approach for financial markets. In J. F. E. IV, F. Fogelman-SouliŐ, P. A. Flach, and M. Zaki, editors, *KDD*, pages 1135–1144. ACM, 2009.

[15] P. Erdős and A. Hajnal. On chromatic number of graphs and set-systems. *Acta Mathematica Hungarica*, 17:61–99, 1966.

[16] U. Feige. Relations between average case complexity and approximation complexity. In *Proceedings of Symposium on Theory of Computing*, pages 534–543, 2002.

[17] D. R. Forsyth. *Group Dynamics*. Cengage Learning, 2010.

[18] E. Fratkin, B. T. Naughton, D. L. Brutlag, and S. Batzoglou. Motifcut: regulatory motifs finding with

maximum density subgraphs. In *ISMB (Supplement of Bioinformatics)*, pages 156–157, 2006.

[19] G. Gallo, M. D. Grigoriadis, and R. E. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM J. Comput.*, 18(1):30–55, Feb. 1989.

[20] D. Gibson, R. Kumar, and A. Tomkins. Discovering large dense subgraphs in massive graphs. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 721–732. VLDB Endowment, 2005.

[21] A. Gionis, F. Junqueira, V. Leroy, M. Serafini, and I. Weber. Piggybacking on social networks. *Proc. VLDB Endow.*, 6(6):409–420, Apr. 2013.

[22] A. V. Goldberg. Finding a maximum density subgraph. Technical report, Berkeley, CA, USA, 1984.

[23] R. Gupta, T. Roughgarden, and C. Seshadhri. Decompositions of triangle-dense graphs. In *Innovations in Theoretical Computer Science (ITCS)*, pages 471–482, 2014.

[24] J. Håstad. Clique is hard to approximate within $n^{(1-\epsilon)}$. In *Acta Mathematica*, pages 627–636, 1996.

[25] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21(1):213–221, Jan. 2005.

[26] L. Iasemidis, D.-S. Shiau, W. Chaovalitwongse, J. Sackellares, P. Pardalos, J. Principe, P. Carney, A. Prasad, B. Veeramani, and K. Tsakalis. Adaptive epileptic seizure prediction system. *Biomedical Engineering, IEEE Transactions on*, 50(5):616–627, May 2003.

[27] R. Jin, Y. Xiang, N. Ruan, and D. Fuhry. 3-hop: a high-compression indexing scheme for reachability query. In U. Ćetintemel, S. B. Zdonik, D. Kossmann, and N. Tatbul, editors, *SIGMOD Conference*, pages 813–826. ACM, 2009.

[28] S. Khot. Ruling out ptas for graph min-bisection, dense k-subgraph, and bipartite clique. *SIAM Journal on Computing*, 36(4):1025–1071, 2006.

[29] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of the Eighth International Conference on World Wide Web*, WWW '99, pages 1481–1493, New York, NY, USA, 1999. Elsevier North-Holland, Inc.

[30] V. E. Lee, N. Ruan, R. Jin, and C. C. Aggarwal. A survey of algorithms for dense subgraph discovery. In C. C. Aggarwal and H. Wang, editors, *Managing and Mining Graph Data*, volume 40 of *Advances in Database Systems*, pages 303–336. Springer, 2010.

[31] D. Lick and A. White. k-degenerate graphs. *Canadian Journal of Mathematics*, 22:1082–1096, 1970.

[32] D. Matula and L. Beck. Smallest-last ordering and clustering and graph coloring algorithms. *Journal of ACM*, 30(3):417–427, 1983.

[33] R. Mokken. Cliques, clubs and clans. *Quality and Quantity*, 13(2):161–173, 1979.

[34] R. A. Rossi, D. F. Gleich, A. H. Gebremedhin, and M. M. A. Patwary. A fast parallel maximum clique algorithm for large sparse graphs and temporal strong components. *CoRR*, abs/1302.6256, 2013.

[35] A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, and B. Y. Zhao. Measurement-calibrated graph models for social network experiments. In *WWW '10*, pages 861–870. ACM, 2010.

[36] A. E. Sarıyüce, B. Gedik, G. Jacques-Silva, K.-L. Wu, and Ü. V. Çatalyürek. Streaming algorithms for k-core decomposition. In *39th International Conference on Very Large Data Bases (VLDB)*, Aug 2013.

[37] T. Schank and D. Wagner. Finding, counting and listing all triangles in large graphs, an experimental study. In *Experimental and Efficient Algorithms*, pages 606–609. Springer Berlin / Heidelberg, 2005.

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

[39] S. B. Seidman and B. Foster. A graph-theoretic generalization of the clique concept. *Journal of Mathematical Sociology*, 1978.

[40] C. Seshadhri, A. Pinar, and T. G. Kolda. Triadic measures on graphs: The power of wedge sampling. *Statistical Analysis and Data Mining*, 7(4):294–307, 2014.

[41] SNAP. Stanford network analysis package. http://snap.stanford.edu/snap, retrieved March, 2014.

[42] S. Suri and S. Vassilvitskii. Counting triangles and the curse of the last reducer. In *WWW'11*, pages 607–614, 2011.

[43] C. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. Tsiarli. Denser than the densest subgraph: Extracting optimal quasi-cliques with quality guarantees. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 104–112, New York, NY, USA, 2013. ACM.

[44] C. E. Tsourakakis. A novel approach to finding near-cliques: The triangle-densest subgraph problem. *CoRR*, abs/1405.1477, 2014.

[45] J. Wang and J. Cheng. Truss decomposition in massive networks. *Proceedings of the VLDB Endowment*, 5(9):812–823, 2012.

[46] N. Wang, J. Zhang, K.-L. Tan, and A. K. H. Tung. On triangulation-based dense neighborhood graph discovery. *Proc. VLDB Endow.*, 4(2):58–68, Nov. 2010.

[47] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[48] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[49] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 17+, 2005.

[50] Y. Zhang and S. Parthasarathy. Extracting analyzing and visualizing triangle k-core motifs within networks. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, ICDE '12, pages 1049–1060, Washington, DC, USA, 2012. IEEE Computer Society.

[51] F. Zhao and A. K. H. Tung. Large scale cohesive subgraphs discovery for social network visual analysis. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 85–96.

VLDB Endowment, 2013.