

# Distributed Estimation of Graph 4-Profiles

Ethan R. Elenberg  
The University of Texas  
Austin, Texas, USA  
elenberg@utexas.edu

Michael Borokhovich  
The University of Texas  
Austin, Texas, USA  
michaelbor@utexas.edu

Karthikeyan Shanmugam  
The University of Texas  
Austin, Texas, USA  
karthiksh@utexas.edu

Alexandros G. Dimakis  
The University of Texas  
Austin, Texas, USA  
dimakis@austin.utexas.edu

## ABSTRACT

We present a novel distributed algorithm for counting all four-node induced subgraphs in a big graph. These counts, called the 4-profile, describe a graph's connectivity properties and have found several uses ranging from bioinformatics to spam detection. We also study the more complicated problem of estimating the local 4-profiles centered at each vertex of the graph. The local 4-profile embeds every vertex in an 11-dimensional space that characterizes the local geometry of its neighborhood: vertices that connect different clusters will have different local 4-profiles compared to those that are only part of one dense cluster.

Our algorithm is a local, distributed message-passing scheme on the graph and computes all the local 4-profiles in parallel. We rely on two novel theoretical contributions: we show that local 4-profiles can be calculated using compressed two-hop information and also establish novel concentration results that show that graphs can be substantially sparsified and still retain good approximation quality for the global 4-profile.

We empirically evaluate our algorithm using a distributed GraphLab implementation that we scaled up to 640 cores. We show that our algorithm can compute global and local 4-profiles of graphs with millions of edges in a few minutes, significantly improving upon the previous state of the art.

## 1. INTRODUCTION

Graph  $k$ -profiles are local statistics that count the number of small subgraphs in a big graph.  $k$ -profiles are a natural generalization of triangle counting and are increasingly popular for several problems in big graph analytics. Globally, they form a concise graph description that has found several applications for the web [4, 21] as well as social and biological networks [29, 22]. Furthermore, as we explain, the *local* profile of a vertex is an embedding in a low-dimensional feature space that reveals local structural information. Mathemat-

ically,  $k$ -profiles are of significant recent interest since they are connected to the emerging theory of graph homomorphisms, graph limits and graphons [5, 29, 19].

There are 4 possible graphs on 3 vertices, labeled  $H_0; \dots; H_3$  in Figure 1a. The (global) 3-profile of a graph  $G(V; E)$  is a vector having one coordinate for each distinct  $H_i$  that counts how many times that  $H_i$  appears as an induced subgraph of  $G$ . For example, the graph  $G = K_4$  (the complete graph on 4 vertices) has the 3-profile  $[0; 0; 0; 4]$  since it contains 4 triangles and no other (induced) subgraphs. The graph  $C_5$  (the cycle on 5 vertices, *i.e.* a pentagon) has the 3-profile  $[0; 5; 5; 0]$ . Note that the sum of the  $k$ -profile is always  $\binom{|V|}{k}$ , the total number of subgraphs. Estimating 3-profiles of big graphs is a topic that has received attention from several communities recently (e.g. see [29, 31, 13, 7] and references therein).

In this paper we are interested in the significantly more challenging problem of estimating 4-profiles. Figure 1b shows the 11 possible graphs on 4 vertices,<sup>1</sup> labeled as  $F_i$ ,  $i = 0; \dots; 10$ . Given a big graph  $G(V; E)$  we are interested in estimating the global 4-profile, *i.e.* count how many times each  $F_i$  appears as an induced subgraph of  $G$ . In addition to global graph statistics, we are interested in local 4-profiles: given a specific vertex  $v_0$ , the local 4-profile of  $v_0$  is an 11-dimensional vector, with each coordinate  $i$  counting how many induced  $F_i$ 's contain  $v_0$ . In Figure 2 we show an example of the local 4-profile of a vertex.

The local 4-profile of a vertex can be seen as an embedding in an 11-dimensional space that characterizes the local geometry of its neighborhood: vertices that connect different clusters will have different local 4-profiles compared to those that are only part of one dense cluster. A very naive estimation of 4-profiles requires examining  $\binom{n}{4}$  possible subgraphs. Furthermore, for estimating each local 4-profile independently, this computation has to be repeated  $n$  times, once for each vertex. Note that the local 4-profiles may be rescaled and added together to obtain the global 4-profile. Since some of the 4-profile subgraphs are disconnected (like  $F_0; F_1; F_5$ ), local 4-profiles contain information beyond the local neighborhood of a vertex. Therefore, in a distributed setting, it seems that global communication is required.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.  
WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.  
ACM 978-1-4503-4143-1/16/04.  
<http://dx.doi.org/10.1145/2872427.2883082>.

<sup>1</sup>Actually there are 17 local subgraphs when considering vertex automorphisms. This is discussed in Section 2 in detail. For the purpose of initial exposition, we will ignore vertex automorphisms.

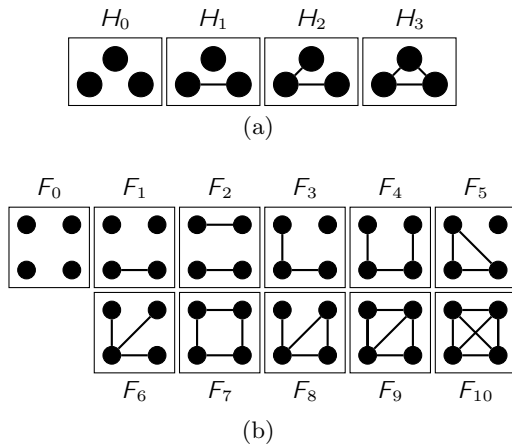


Figure 1: (a) The 4 possible non-isomorphic graphs on 3 vertices used to calculate the 3-profile of a graph  $G$ . The 3-profile counts how many times each  $H_i$  appears in  $G$ . (b) The 11 non-isomorphic graphs on 4 vertices used to calculate the 4-profile of a graph.

## 1.1 Our Contributions

Surprisingly, we show that very limited global information is sufficient to calculate all local 4-profiles and that it can be re-used to calculate all the local 4-profiles in parallel. Specifically, we introduce a distributed algorithm to estimate all the local 4-profiles and the global profile of a big graph. This restrictive setting does not allow communication between nonadjacent vertices, a key component of previous centralized, shared-memory approaches. Our algorithm relies on two novel theoretical results:

**Two-hop histograms are sufficient:** Our algorithm operates by having each vertex first perform local message-passing to its neighbors and then solve a novel system of equations for the local 4-profile. Focusing on a vertex  $v_0$ , the first easy step is to calculate its local 3-profile. It can be shown that the local 3-profile combined with the full two-hop connectivity information is sufficient to estimate the local 4-profile for each vertex  $v_0$ . This is not immediately obvious, since naively counting the 3-path (an automorphism of  $F_4$ ) would require 3-hop connectivity information.

However, we show that less information needs to be communicated. Specifically, we prove that the triangle list combined with what we call the *two-hop histogram* is sufficient: for each vertex  $v_i$  that is 2-hops from  $v_0$ , we only need the *number of distinct paths* connecting it to  $v_0$ , not the full two hop neighborhood. If the two-hop neighborhood is a tree, this amounts to no compression. However, for real graphs the two-hop histogram saves a factor of 3x to 5x in communication in our experiments. This enables (Section 4) an even more significant running time speedup of 5 – 10 times on several distributed experiments using 12 – 20 compute nodes.

**Profile Sparsification:** One idea that originated from triangle counting [27, 28] is to first perform random subsampling of edges to create a sparse graph called a *triangle sparsifier*. Then count the number of triangles in the sparse graph and rescale appropriately to estimate the number in the original graph. The main challenge is proving that the randomly sparsified graph has a number of triangles suffi-

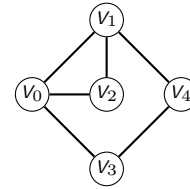


Figure 2: An example for local profiles. The global 3-profile is  $[0;3;6;1]$ . The global 4-profile is  $[0;0;0;0;2;0;0;1;2;0;0]$ . The local 4-profile of  $v_0$  is  $[0;0;0;0;1;0;0;1;2;0;0]$ . The first 1 in the profile corresponds to the subgraph  $F_4$ . Notice that  $v_0$  participates in only one  $F_4$ , jointly with vertices  $v_2; v_3; v_4$ .

ciently concentrated around its expectation. Recently this idea was generalized to 3-profile sparsifiers in [7], with concentration results for estimating the full 3-profile. These papers rely on Kim-Vu polynomial concentration techniques [16] that scale well in theory, but typically the estimated errors are orders of magnitude larger than the measured quantities for reasonable graph sizes. In this paper, we introduce novel concentration bounds for global  $k$ -profile sparsifiers that use a novel information theoretic technique called read- $k$  functions [11]. Our read- $k$  bounds allow usable concentration inequalities for sparsification factors of approximately 0.4 or higher (Section 4.1). Note that removing half the edges of the graph does not accelerate the running time by a factor of 2, but rather by a factor of nearly 8, as shown in our experiments.

**System implementation and evaluation:** We implemented our algorithm using GraphLab PowerGraph [12] and tested it in multicore and distributed systems scaling up to 640 cores. The benefits of two-hop histogram compression and sparsification allowed us to compute the global and local 4-profiles of very large graphs. For example, for a graph with 5 million vertices and 40 million edges we estimated the global 4-profile in less than 10 seconds. For computing all local 4-profiles on this graph, the previous state of the art [13] required 1200 seconds while our distributed algorithm required less than 100 seconds.

## 1.2 Related Work

The problem of counting triangles in a graph has been addressed in distributed [24] and streaming [4] settings, and this is a standard analytics task for graph engines [23]. The Doulion algorithm [27] estimates a graph’s triangle count via simple edge subsampling. Other recent work analyzes more complex sampling schemes [25, 6] and extends to approximately counting certain 4-subgraphs [1, 15]. Mapreduce algorithms for clique counting were introduced by Finocchi et al. [10]. Our approach is similar to that of [7], which calculates all 3-subgraphs and a subset of 4-subgraphs distributedly using *edge pivots*. In this work we introduce the 2-hop histogram to compute all 4-subgraphs.

Concentration inequalities for the number of triangles in a random graph have been studied extensively. The standard method of martingale bounded differences (McDiarmid’s inequality) is known to yield weak concentrations around the mean for this problem. The breakthrough work of Kim and Vu [16] provides superior asymptotic bounds by analyzing the concentration of multivariate polynomials. This

was later improved and generalized in [14], and applied to subsampled triangle counting in [28]. Our analysis uses a different technique called read- $k$  functions [11] that produces sharper concentration results for practical problem sizes.<sup>2</sup>

Previous systems of equations relating clique counts to other 4-subgraphs appear in [17], [31], [13], and [2]. However, these are applied in a centralized setting and depend on information collected from nonadjacent vertices. In this work, we use additional equations to solve the same system by sharing only local information over adjacent vertices. The connected 4-subgraphs, or graphlets [22], have found applications in fields such as bioinformatics [26] and computational neuroscience [9]. In [30], authors use *all* global 4-subgraphs to analyze neuronal networks. We evaluate our algorithm against Orca [13], a centralized 4-graphlet counting algorithm, as well as its GPU implementation [20]. Notice that while Orca calculates only connected 4-subgraphs, our algorithm calculates all the connected and the disconnected 4-subgraphs for each vertex.

Concurrent with the writing of this paper, a parallel algorithm for 4-subgraph counting was introduced in [2]. Our algorithm differs by working within GraphLab PowerGraph’s Gather-Apply-Scatter framework instead of the native, multithreaded C++ implementation of [2]. In terms of empirical performance, both our work and [2] show similar running time improvements of one order of magnitude over Orca. A more detailed comparison would depend on the hardware and datasets used. More importantly, our work focuses on a distributed (as opposed to multicore parallel) framework, and for our setting minimizing communication is critical.

Our theoretical results are significantly different from [2] and may be useful in improving that system also. Specifically, [2] explicitly counts the number of 4-cycles ( $F_7$  in Figure 1b) whereas our results show that it is possible to use only two-hop histograms instead. This results in less communication overhead, but this benefit is perhaps not as significant for shared-memory multicore platforms. Our second theoretical result, the novel sparsification concentration bounds, can be used for any subgraph estimation algorithm and quantify a provable tradeoff between speed and accuracy.

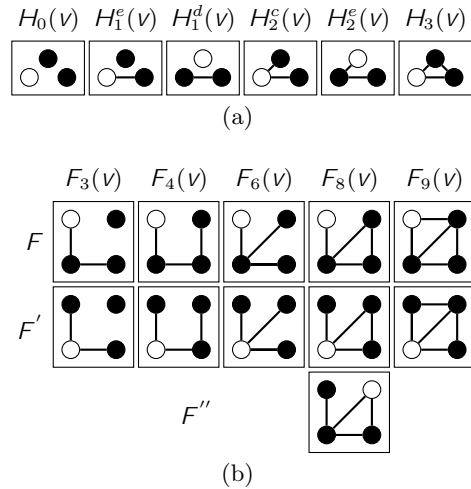
## 2. DISTRIBUTED ALGORITHM

In this section, we describe 4-Prof-Dist, our algorithm for computing the exact 4-profiles in a distributed manner. To the best of our knowledge, this is the first *distributed* algorithm for calculating 4-profiles. The key insight is to cast existing and novel equations into the GraphLab PowerGraph framework [12] to get implicit connectivity information about vertices outside the 1-hop neighborhood. Specifically, we construct the local 4-profile from local 3-profile, local 4-clique count, and additional histogram information which describes the number of paths to all 2-hop neighbors.

**Theorem 1.** *There is a distributed algorithm that computes the exact local 4-profile of a graph given each vertex has stored its local 3-profile, triangle list, and 2-hop histogram.*

Note that the local 4-profiles at each vertex can be added and appropriately rescaled (using the symmetries of each

<sup>2</sup>Even though concentrations using Kim-Vu become tighter asymptotically, this happens for graphs with well over  $10^{13}$  edges (see also Figure 5).



**Figure 3: Unique (a) 3-subgraphs and (b) 4-subgraphs from the perspective of the white vertex  $v$ .  $F_8$  is the only subgraph with a third vertex automorphism  $F_8''$  because no other subgraph contains vertices with 3 different degrees.**

subgraph, also called automorphism orbits [22]) to obtain the global 4-profile.

4-Prof-Dist is implemented in the Gather-Apply-Scatter (GAS) framework [12]. A distributed algorithm in this framework has 3 main phases: Gather, Apply and Scatter. Every vertex and edge has stored data which is acted upon. During the Gather phase, a vertex can access all its adjacent edges and neighbors and *gather* data they possess, *e.g.*, neighbor ID, using a custom reduce operation  $\oplus$  (*e.g.*, addition, concatenation). The accumulated information is available for a vertex at the next phase, Apply, in which it can change its own data. In the final Scatter phase, every edge sees the data of its (incident) vertices and operates on it to modify the edge data. All nodes start each phase simultaneously, and if needed, the whole GAS cycle is repeated until the algorithm’s completion.

4-Prof-Dist solves a slightly larger problem of keeping track of counts of 17 unique subgraphs up to vertex automorphism (see Figure 3). We will describe a full rank system of equations which is sufficient to calculate the local 4-profile at every  $v \in V$ . The following subsections each explain a component of 4-Prof-Dist. These separate routines are combined efficiently in Algorithm 1 to calculate the local 4-profile in a small number of GAS cycles.

### 2.1 Edge Pivot Equations

The majority of our equations relate the local 4-profile to neighboring local 3-profiles with *edge pivots* [7]. At each vertex  $v$ , each combinatorial equation relates a linear combination of the local 4-subgraph counts to the count of a pair of 3-subgraphs sharing an edge  $va$ . Some of these equations appear in a centralized setting in previous literature ([17], [31], [13], [2]). In our algorithm, the 3-subgraph pair count accumulates at  $v$  as all incident edges  $va$  *pivot* over it. The edges fixed by a specific 3-subgraph pair correspond to common edges among a subset of 4-subgraphs. Before that, in an initial GAS round, each vertex  $v$  must gather the ID of each vertex in its neighborhood, *i.e.*  $a \in \Gamma(v)$ , and the

following quantities must be stored at each edge  $va$  during the Scatter phase:

$$\begin{aligned} n_{1,va}^e &= |V| - (|\Gamma(v)| + |\Gamma(a)| - |\Gamma(v) \cap \Gamma(a)|); \\ n_{2,va}^e &= |\Gamma(v) \setminus \{\Gamma(a) \cup a\}| = |\Gamma(v)| - |\Gamma(v) \cap \Gamma(a)| - 1; \\ n_{2,va}^e &= n_{2,av}^c; \quad n_{3,va} = |\Gamma(v) \cap \Gamma(a)|; \end{aligned} \quad (1)$$

**Gather:** The above quantities are summed at each vertex  $v$  to calculate the local 3-profile at  $v$ . For example,  $n_{3,v} = \frac{1}{2} \sum_a n_{3,va}$ . In addition, we gather the sum of functions of pairs of these quantities forming 13 *edge pivot* equations:

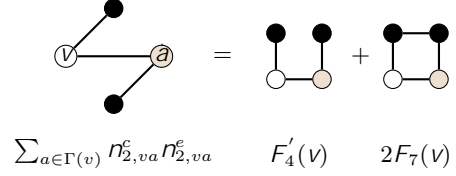
$$\begin{aligned} \sum_{a \in \Gamma(v)} \binom{n_{1,va}^e}{2} &= F_1(v) + F_2(v); \\ \sum_{a \in \Gamma(v)} \binom{n_{2,va}^c}{2} &= 3F_6'(v) + F_8'(v); \\ \sum_{a \in \Gamma(v)} \binom{n_{3,va}}{2} &= F_9'(v) + 3F_{10}(v); \\ \sum_{a \in \Gamma(v)} n_{1,va}^e n_{2,va}^c &= 2F_3'(v) + F_4'(v); \\ \sum_{a \in \Gamma(v)} n_{1,va}^e n_{3,va} &= 2F_5(v) + F_8''(v); \\ \sum_{a \in \Gamma(v)} n_{2,va}^c n_{2,va}^e &= F_4'(v) + 2F_7(v); \\ \sum_{a \in \Gamma(v)} n_{2,va}^c n_{3,va} &= 2F_8'(v) + 2F_9'(v); \\ n_{1,v}^d |\Gamma(v)| &= F_2(v) + F_4(v) + F_8(v); \end{aligned} \quad (2)$$

The primed notation differentiates between subgraphs of different automorphism orbits, as in Figure 3. By accumulating pairs of 3-profile structures as in (2), we receive aggregate connectivity information about vertices more than 1 hop away. Consider the sixth equation as an example. The product between  $n_{2,va}$  and  $n_{2,va}^e$  subgraphs along edge  $va$  forms 4-node graphs with the following structural constraints: three vertex pairs are connected, two vertex pairs are disjoint, and one pair may be either connected or disjoint.  $F_4'(v)$  and  $F_7(v)$  satisfy these constraints and differ on the unconstrained edge. Thus, as shown in Figure 4, they both contribute to the sum of  $n_{2,va}^c n_{2,va}^e$ .

The following edge pivot equations are linearly independent when solving for the local 4-profile only. Note the last 2 equations require calculating the local 3-profile:

$$\begin{aligned} \sum_{a \in \Gamma(v)} \binom{n_{2,va}^e}{2} &= F_6(v) + F_8(v); \\ \sum_{a \in \Gamma(v)} n_{1,va}^e n_{2,va}^e &= F_3(v) + F_4(v); \\ \sum_{a \in \Gamma(v)} n_{2,va}^e n_{3,va} &= F_8''(v) + 2F_9(v); \\ \sum_{a \in \Gamma(v)} n_{3,a} - n_{3,va} &= F_8(v) + 2F_9(v) + 3F_{10}(v); \\ \sum_{a \in \Gamma(v)} n_{2,a}^e - n_{2,va}^c &= F_4(v) + 2F_7(v) + F_8''(v) + 2F_9'(v); \end{aligned} \quad (3)$$

**Apply:** Store the left hand sides of all 13 equations at  $v$ .



**Figure 4: Edge pivot equation for vertex  $v$  counting triangles as edges  $va$  pivot about their common vertex  $v$ . The subgraphs  $F_4'(v)$  and  $F_7(v)$  differ by one edge.**

## 2.2 Clique Counting

The aim of this subtask is to count 4-cliques that contain the vertex  $v$ . For this, we accumulate a list of triangles at each vertex  $v$ . Then, at the Scatter stage for every  $va$ , it is possible to check if neighbors common to  $v$  and  $a$  have an edge between them. This implies a 4-clique.

**Scatter:** In addition to the intersection size  $|\Gamma(v) \cap \Gamma(a)|$  at each edge  $va$  as before, we now require the intersection list  $\{b : b \in \Gamma(v); b \in \Gamma(a)\}$  as a starting point.

**Gather, Apply:** The intersection list is gathered at each vertex  $v$ . This produces all pairs of neighbors in  $\Gamma(v)$  which are adjacent, *i.e.* all triangles containing  $v$ . It is stored as  $\Delta(v)$  during the Apply stage at  $v$ .

**Gather, Apply:** Each edge  $va$  computes the number of 4-cliques by counting how many pairs in  $\Delta(a)$  contain exactly two neighbors of  $v$ . We use a similar equation to calculate  $F_8(v)$  concurrently:

$$\begin{aligned} \sum_{a \in \Gamma(v)} |\{b:c \in \Delta(a) : b \in \Gamma(v); c \in \Gamma(v)\}| &= 3F_{10}(v); \\ \sum_{a \in \Gamma(v)} |\{b:c \in \Delta(a) : b \notin \Gamma(v); c \in \Gamma(v)\}| &= F_8(v); \end{aligned} \quad (4)$$

At the Apply stage, store the left hand sides as vertex data.

## 2.3 Histogram 2-hop Information

Instead of calculating the number of cycles  $F_7(v)$  directly, we can simply construct another linearly independent equation and add it to our system. Let each vertex  $a$  have a vector of (vertex ID, count) pairs  $(p; c_a[p])$  for each of its adjacent vertices  $p$ . Initially,  $c_a[p] = 1$  and this *histogram* contains the same information as  $\Gamma(a)$ . For any  $a \in \Gamma(v)$  and  $p \in \Gamma(v)$ ,  $c_a[p] = 1 \Leftrightarrow vap$  forms a 2-path. Thus,  $v$  can collect these vectors to determine the total number of 2-paths from  $v$  to  $p$ . This lets us calculate a linear combination involving cycle subgraph counts with an equation that is linearly independent from the others in our system.

**Gather:** At each  $v$ , take a union of histograms from each neighbor  $a$ , resolving duplicate entries with the reduce operation  $(p; c_{a_1}) \oplus (p; c_{a_2}) = (p; c_{a_1} + c_{a_2})$ .

**Apply:** Given the gathered histogram vector,  $\{\oplus_{a \in \Gamma(v)} c_a[p]\}_{p \in \Gamma(v)}$ , calculate the number of non-induced 4-cycles involving  $p$  and two neighbors:

$$\sum_{p \in \Gamma(v)} \binom{\oplus_{a \in \Gamma(v)} c_a[p]}{2} = F_7(v) + F_9(v); \quad (5)$$

Next, we upper bound savings from our 2-hop histogram by analyzing the improvement when the only information

---

**Algorithm 1** 4-Prof-Dist

---

- 1: **Input:** Graph  $G(V; E)$  with  $|V|$  vertices,  $|E|$  edges.
  - 2: **Gather:** For each vertex  $v$  union over edges of the ‘other’ vertex in the edge,  $\cup_{a \in \Gamma(v)} a = \Gamma(v)$ .
  - 3: **Apply:** Store the gather as vertex data  $\mathbf{v.nb}$ , size automatically stored.
  - 4: **Scatter:** For each edge  $e_{va}$ , compute and store scalars in (1).
  - 5: **Gather:** For each edge  $e_{va}$ , sum edge scalar data of neighbors in (2) - (3) and combine two-hop histograms.
  - 6: **Apply:** For each vertex  $v$ , sum over  $\rho \in \Gamma(v)$  in (5), store other data in array  $\mathbf{v.u}$ . No Scatter.
  - 7: **Gather:** For each vertex  $v$  collect pairs of connected neighbors in  $\Delta(v)$ .
  - 8: **Apply:** Store connected neighbor (triangle) list as vertex data  $\mathbf{v.conn}$ . No Scatter.
  - 9: **Gather:** For each vertex  $v$  sum (4).
  - 10: **Apply:** Append data to array  $\mathbf{v.u}$ . Multiply  $\mathbf{v.u}$  by a matrix to solve system of equations.
  - 11: **return**  $[\mathbf{v}: \mathbf{v.N0} \mathbf{v.N1} \mathbf{v.N2} \dots \mathbf{v.N10}]$
- 

transmitted across the network to a vertex  $v$  is each non-neighboring vertex and its final count  $\oplus_{a \in \Gamma(v)} c[\rho]$ . Let

$$h_v = |\Gamma(\Gamma(v)) \setminus \{\Gamma(v) \cup v\}|:$$

For each  $v$ , the difference between full and histogram information is at most  $\sum_{a \in \Gamma(v)} (|\Gamma(a)| - 1) - 2h_v$ . The exact benefit of (5) depends on the internal implementation of the reduce operation  $\oplus$  as pairs of neighbors are gathered.

Counting the number of distinct pairs of 2-paths to each 2-hop neighbor, *i.e.*  $\frac{1}{2}(c[\rho]^2 - c[\rho])$ , requires counting the second moment of  $c$  taken over  $h_v$  terms. Due to a result by Alon ([3], Proposition 3.7), the memory required to count this value exactly (moreover, to approximate it deterministically) is  $\Omega(h_v)$ . Thus, up to implementation details, our memory use is optimal.

## 2.4 Normalization and Symmetry

Our final local equation comes from summing the local 4-profile across all 17 automorphisms:

$$\sum_i^{17} F_i(v) = \binom{|V| - 1}{3}; \quad (6)$$

To calculate the global 4-profile, we utilize global symmetry and scaling equations. Let  $F_i = \sum_{v \in V} F_i(v)$ . Globally, each subgraph count is in exact proportion with the same subgraph counted from a different vertex automorphism. The ratio depends on the subgraph’s degree distribution:

$$\begin{aligned} F_3 &= 2F'_3; & F_4 &= F'_4; & F_6 &= 3F'_6; \\ F_8 &= F'_8; & F''_8 &= 2F_8; & F_9 &= F'_9; \end{aligned} \quad (7)$$

Global symmetry makes the equation for  $F_8$  and the system (3) linearly dependent. We sum across vertices, inverting a single  $11 \times 11$  system to yield the final global 4-profile

$[N_0 :: N_{10}]$  by scaling appropriately:

$$\begin{aligned} N_0 &= \frac{F_0}{4}; & N_1 &= \frac{F_1}{2}; & N_2 &= \frac{F_2}{4}; & N_3 &= F'_3; \\ N_4 &= \frac{F'_4}{2}; & N_5 &= \frac{F_5}{3}; & N_6 &= F'_6; & N_7 &= \frac{F_7}{4}; \\ N_8 &= F_8; & N_9 &= \frac{F_9}{2}; & N_{10} &= \frac{F_{10}}{4}. \end{aligned} \quad (8)$$

## 3. SPARSIFIER AND CONCENTRATION

In this section, we describe the process for approximating the exact number of subgraphs in a graph  $G$ . Denote the exact counts by  $[N_0 :: N_{10}]^T$  and the estimates by  $[X_0 :: X_{10}]^T$ .

We are sparsifying the original graph  $G$  by keeping each edge independently with probability  $\rho$ . Denote the random subsampled graph by  $\tilde{G}$  and its global 4-profile by  $[Y_0 :: Y_{10}]^T$ . Clearly each triangle survives with probability  $\rho^3$  and each 4-clique survives with  $\rho^6$ . Therefore, in expectation,  $\mathbb{E}[Y_{10}] = \rho^6 N_{10}$  and  $X_{10} = \frac{1}{\rho^6} Y_{10}$  is unbiased.

This simple correspondence does not hold for other subgraphs: each triangle in  $\tilde{G}$  can only be a triangle in  $G$  that survived edge removals, but other subgraphs of  $\tilde{G}$  could be originating from multiple subgraphs of  $G$  depending on the random sparsification process. We can, however, relate the original 4-profile vector to the expected subsampled 4-profile vector by a matrix multiplication. Let  $F(abcd)$  and  $\tilde{F}(abcd)$  represent the induced 4-subgraph on the vertices  $abcd$  before and after subsampling, respectively. Then define  $\mathbf{H}$  by  $H_{ij} = \mathbb{P}(\tilde{F}(abcd) = F_i \mid F(abcd) = F_j)$ . Thus, we form an unbiased estimator, *i.e.*  $\mathbb{E}[X_i] = N_i; i = 1 :: 10$ , by inverting the edge sampling matrix.

For 3-profiles, this process is described by the following system of equations:

$$\begin{bmatrix} \mathbb{E}[Y_0] \\ \mathbb{E}[Y_1] \\ \mathbb{E}[Y_2] \\ \mathbb{E}[Y_3] \end{bmatrix} = \begin{bmatrix} 1 & 1 - \rho & (1 - \rho)^2 & (1 - \rho)^3 \\ 0 & \rho & 2\rho(1 - \rho) & 3\rho(1 - \rho)^2 \\ 0 & 0 & \rho^2 & 3\rho^2(1 - \rho) \\ 0 & 0 & 0 & \rho^3 \end{bmatrix} \begin{bmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \end{bmatrix}; \quad (9)$$

For 4-profiles, the vectors are 11 dimensional and a similar linear system can be explicitly computed – we include the equations in the Appendix. This matrix turns out to be invertible and we can therefore calculate the 4-profile exactly if we have access to the expected values of the sparsified 4-profile. Of course, we can only obtain one sample random graph and calculate that 4-profile, which will be an accurate estimate if the 4-profile quantities are sufficiently concentrated around their expectation.

### 3.1 Graph k-profile Concentration

Previous work used this idea of graph sparsification for triangle counting [28] and 3-profiles [7]. The main concentration tool used was the Kim and Vu polynomial concentration [16, 28] which unfortunately gives very loose bounds for practical graph sizes. Figure 5 compares the accuracy bound derived in this section to the bound predicted by [16]. Clearly the Kim-Vu concentration does not provide meaningful bounds for the experiments in Section 4.1. However, our results match observed sparsifier accuracy much more closely.

Our novel concentration results exploit the fact that partial derivatives of the desired quantities are sparse in the number of edge variables. This allows us to use a novel information theoretic concentration technique called read- $k$

functions [11]. For simplicity, we only explain the concentration of 4-cliques ( $F_{10}$  subgraphs) here. We establish the general result for all 11 4-profile variables in the Appendix. Additional details can be found in the extended version of this paper [8]. Our main concentration result is as follows:

**Theorem 2.** *Let  $G$  be a graph with  $N_{10}$  4-cliques, and let  $k_{10}$  be the maximum number of 4-cliques sharing a common edge. Let  $X_{10}$  be the 4-clique estimate obtained from subsampling each edge with probability  $0 < \rho \leq 1$ , choose  $0 < \gamma < 1$ , and choose  $\rho_{RK} > 0$ . If*

$$\rho \geq \left( \frac{\log(2\rho_{RK})k_{10}}{2\rho_{RK}N_{10}} \right)^{1/12};$$

then  $|N_{10} - X_{10}| \leq \rho_{RK}N_{10}$  with probability at least  $1 - \frac{1}{m^\gamma}$ .

**Proof.** Our proof relies on read- $k$  function families [11], a recent characterization of dependencies among functions of random variables. Rather than Lipschitz bounding the value of each partial derivative, as in [16, 28, 7], this technical tool benefits from the fact that each first partial derivative is sparse in the number of edge variables.

**Definition 1 (read- $k$  families).** *Let  $X_1, \dots, X_m$  be independent random variables. For  $j \in [r]$ , let  $P_j \subseteq [m]$  and let  $f_j$  be a Boolean function of  $\{X_i\}_{i \in P_j}$ . Assume that  $|\{j \mid i \in P_j\}| \leq k$  for every  $i \in [m]$ . Then the random variables  $Z_j = f_j(\{X_i\}_{i \in P_j})$  are called a read- $k$  family.*

Each variable only affects  $k$  of the  $r$  Boolean functions. Let  $G$  be a graph with  $N_{10}$  4-cliques and a maximum of  $k_{10}$  4-cliques sharing a common edge. The corresponding 4-clique estimator  $X_{10}$  follows this exact structure. Each edge sampling variable appears in at most  $k_{10}$  of the  $N_{10}$  terms. We now state the main result required for our analysis. Note that when applied to estimating the number of 4-cliques, the bound is independent of the number of edges. Therefore, it is much stronger than arguments involving Lipschitz bounded functions such as McDiarmid's inequality.

**Proposition 1 (Read- $k$  Concentration [11]).** *Let  $Z_1, \dots, Z_r$  be a family of read- $k$  indicator variables with  $\mathbb{P}(Z_i = 1) = p_i$ , and let  $\rho$  be the average of  $p_1, \dots, p_r$ . Then for any  $\gamma > 0$ ,*

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^r Z_i \geq (\rho + \gamma)r\right) &\leq \exp\left(-D(\rho + \gamma \parallel \rho) \frac{r}{k}\right) \\ \mathbb{P}\left(\sum_{i=1}^r Z_i \leq (\rho - \gamma)r\right) &\leq \exp\left(-D(\rho - \gamma \parallel \rho) \frac{r}{k}\right); \end{aligned}$$

where  $D(x \parallel y) = x \log\left(\frac{x}{y}\right) + (1-x) \log\left(\frac{1-x}{1-y}\right)$  is the Kullback-Leibler divergence of  $x$  and  $y$ . Both bounds are less than  $\exp(-2\gamma^2 r/k)$ .

Let  $Y_{10} = \sum_{\square(a,b,c,d) \in \mathcal{H}_{10}} t_{ab} t_{bc} t_{cd} t_{da} t_{ac} t_{bd}$ . Then

$$\begin{aligned} \mathbb{P}(|Y_{10} - \rho^6 N_{10}| \geq \rho_{RK} N_{10}) &\leq 2 \exp\left(-\frac{2\rho_{RK} N_{10}}{k_{10}}\right) \\ \Rightarrow \mathbb{P}(|X_{10} - N_{10}| \geq \rho_{RK} N_{10}) &\leq 2 \exp\left(-\frac{2\rho^{12} \rho_{RK} N_{10}}{k_{10}}\right); \end{aligned}$$

The claim follows by setting the right hand side less than and solving for  $\rho$ .  $\square$

Next, we state conditions under which our method outperforms the Kim and Vu concentration results [16].

**Corollary 1.** *Let  $G$  be a graph with  $m$  edges. If  $\rho = \Omega(1/\log m)$  and  $\gamma = \Omega(1/m)$ , then read- $k$  provides better triangle sparsifier accuracy than Kim-Vu. If additionally  $k_{10} \leq N_{10}^{5/6}$ , then read- $k$  provides better 4-clique sparsifier accuracy than Kim-Vu.*

**Proof.** We prove this result for the case of 4-cliques only because the case for triangles is similar. First we must derive a similar 4-clique concentration bound using the techniques in [16, 28].

**Lemma 1.** *Let  $G$  be a graph with  $m$  edges and  $N_{10}$  cliques, and  $k_{10}$  be the maximum number of 4-cliques sharing a common edge. Let  $a_6 = 8^6 \sqrt{6!}$ ,  $0 < \rho \leq 1$ , and  $\gamma_{KV} > 0$ . Let  $X_{10}$  be the 4-clique estimate obtained from subsampling each edge with probability  $\rho$ . If*

$$\frac{\rho}{\max\left\{\sqrt[6]{1=N_{10}}, \sqrt[3]{k_{10}=N_{10}}\right\}} \geq \frac{a_6^2 \log^{12}(m^{5+\gamma})}{2^{\gamma_{KV}}};$$

then  $|N_{10} - X_{10}| \leq \gamma_{KV} N_{10}$  with probability at least  $1 - \frac{1}{m^\gamma}$ .

The proof of this lemma is a straightforward application of the main result in [16]. It can be found in the extended version of this paper [8].

Now we are ready to prove the corollary by comparing Theorem 2 and Lemma 1. Fix  $\rho; \gamma > 0$  and  $\gamma > 1$  such that  $\rho = \Omega(1/\log m)$  and  $\gamma = m^{-\gamma} = \Omega(1/m)$ . Now we analyze the bounds  $\gamma_{KV}$  and  $\rho_{RK}$ . For any graph and  $a_6$  defined in Lemma 1,

$$\begin{aligned} \frac{1}{a_6^2} \leq 1; \quad \frac{1}{(5+\gamma)^{12}} \leq 1; \\ \log(2^{1/\gamma} m) \leq 2 \log m; \quad \left(\frac{k_{10}}{N_{10}}\right)^{2/3} \leq 1; \end{aligned} \tag{10}$$

We further require  $k_{10} \leq N_{10}^{5/6}$ . Then the condition on  $\rho$  with (10) implies

$$\begin{aligned} \rho^{11} &\geq 1/\log^{11}(m) \\ &\geq \frac{\log(2^{1/\gamma} m)}{2a_6^2(5+\gamma)^{12} \log^{12}(m)} \min\left\{k_{10}=N_{10}^{5/6}; (k_{10}=N_{10})^{2/3}\right\}; \end{aligned}$$

Rearranging terms,

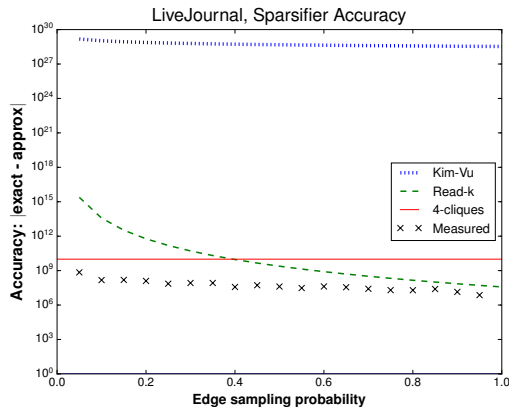
$$\begin{aligned} \frac{2}{\gamma_{KV}} &= \frac{a_6^2 \log^{12}(m^{5+\gamma}) \max\left\{\sqrt[6]{1=N_{10}}, \sqrt[3]{k_{10}=N_{10}}\right\}}{\rho} \\ &\geq \frac{2}{\rho_{RK}} = \frac{\log(2m^\gamma)k_{10}=N_{10}}{2\rho^{12}}; \end{aligned}$$

$\square$

We note that the asymptotic condition on  $\rho$  in Corollary 1 includes a constant term much less than 1. This is due to the looseness of inequalities in (10) and implies that Theorem 2 is superior to Lemma 1 over all  $\rho$  values of practical interest. While these bounds contain the quantities we wish to estimate, they provide guidelines for the performance of sampling heuristics. We also investigate this in Section 4.1 for some realistic graphs.

## 4. EXPERIMENTS

Let us now describe the implementation and experimental results of our algorithm. We implement 4-Prof-Dist on GraphLab v2.2 (PowerGraph) [12] and measure its running time and accuracy on large input graphs.<sup>3</sup> First, we show that edge sampling yields very good approximation results for global 4-profile counts and achieves substantial execution speedups and network traffic savings when multiple machines are in use. Due to its distributed nature, we can show 4-Prof-Dist runs substantially faster when using multiple CPU cores and/or machines. Notice that multi-core and multiple machines can not speed up some centralized algorithms, e.g., Orca [13], which we use as a baseline for our results. Note also that Orca produces only a partial 4-subgraph count, *i.e.* it calculates only connected 4-subgraphs, while 4-Prof-Dist calculates all 17 per vertex.



**Figure 5: Comparison of 4-clique sparsifier concentration bounds with accuracy measured in edge sampling experiments on the LiveJournal graph.**

**The systems:** We perform the experiments on two systems. The first system is a single powerful server, further referred to as Asterix. The server is equipped with 256 GB of RAM and two Intel Xeon E5-2699 v3 CPUs, 18 cores each. Since each core has two hardware threads, up to 72 logical cores are available to the GraphLab engine. The second system is an EC2 cluster on AWS.<sup>4</sup> The cluster is comprised of 20 c3.8xlarge machines, each having 60 GB RAM and 32 virtual CPUs.

**The data:** In our experiments we use two real graphs representing different datasets: social networks (LiveJournal: 4,846,609 vertices, 42,851,237 edges) and a WWW graph of Notre Dame (WEB-NOTRE: 325,729 vertices, 1,090,108 edges) [18]. Notice that the above graphs are originally directed, but since our work deals with undirected graphs, all duplicate edges (*i.e.*, bi-directional) were removed and directionality is ignored.

### 4.1 Results

**Accuracy:** The first result is that our edge sampling approach greatly improves running time while maintaining a very good approximation of the *global* 4-profile. In Figure 6a we can see that the running time decreases drastically

<sup>3</sup>Available at <http://github.com/eelenberg/4-profiles>

<sup>4</sup>Amazon Web Services - <http://aws.amazon.com>

when the sampling probability decreases. At the same time, Figure 6b shows that the mean ratio of true to estimated global 4-profiles is within  $\pm 2.5\%$ . Similar to [15], which uses a more complex sampling scheme to count connected 4-subgraphs, this ratio is usually much less than 1%. We show here only profiles  $F_7 - F_{10}$  since their counts are the smallest and were observed to have the lowest accuracy. In Figure 5 we compare theoretical concentration bounds on a logarithmic scale and show the benefit of Theorem 2. While the guarantees provided by Kim-Vu [16] bounds are very loose (the additive error is bounded by numbers which are orders of magnitude larger than the true value), the read- $k$  approach is much closer to the measured values. We can see that for large sampling probabilities ( $p \geq 0.5$ ), the measured error is at most 2 orders of magnitude smaller than the value predicted by Theorem 2.

**2-hop histogram:** Now we compare two methods of calculating the left hand side of (5) from Section 2.3. We show that a simple implementation in which a vertex gathers its full 2-hop neighborhood (*i.e.*, IDs of its neighbors’ neighbors) is much less efficient than the *two-hop histogram* approach used in 4-Prof-Dist (see Section 2.3). In Figures 7 and 9 we can see that the histogram approach is an order of magnitude faster for various numbers of machines, and that its network requirements are up to 5x less than that of the simple implementation. Moreover, our algorithm could handle much larger graphs while the simple implementation ran out of memory.

**Running time:** Finally, we show that 4-Prof-Dist can run much faster than the current state of the art graphlet counting implementations. The algorithm and the GraphLab platform on which it runs are both distributed in nature. The latter allows 4-Prof-Dist to exploit multiple cores on a single machine as well as a cluster of machines. Figure 6c shows running time as a function of CPU cores. We compare this result to the running time of a single core, C++ implementation of Orca [13]. Our 4-Prof-Dist algorithm becomes faster after only 25 cores and is 2x faster using 60 cores. Moreover, 4-Prof-Dist allows scaling to a large number of machines. In Figure 8 we can see how the running time for the LiveJournal graph decreases when the number of machines increases. Since Orca cannot benefit from multiple machines, we see that 4-Prof-Dist runs up to 12x faster than Orca. This gap widens as the cluster grows larger. In [20], the authors implemented a GPU version of Orca using CUDA. However, the reported speedup is about 2x which is much less than we show here on the AWS cluster (see Figure 8 for  $p = 1$ ). We also note a substantial running time benefit of the sampling approach for *global* 4-profiles. In Figures 8 and 10, we see that with  $p = 0.1$  we can achieve order of magnitude improvements in both speed and network traffic. This sampling probability maintains very good accuracy, as shown in Figure 6b.

## 5. CONCLUSIONS

We introduced a novel distributed algorithm for estimating 4-profiles of large graphs. We relied on two theoretical results that can be of independent interest: that 4-profiles can be estimated with limited 2-hop information and that randomly erasing edges gives sharper approximation compared to previous analysis. We showed that our scheme outperforms the previous state of the art and can exploit cloud infrastructure to scale.

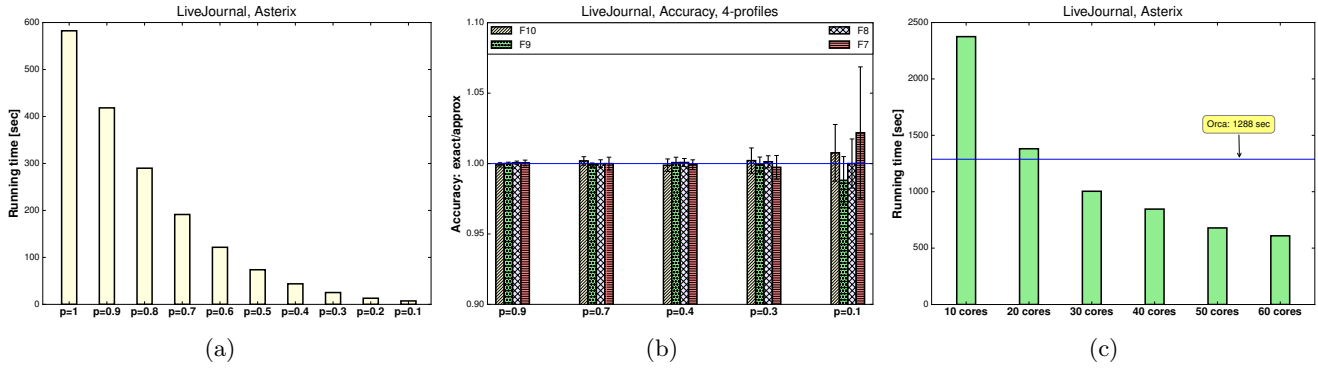


Figure 6: LiveJournal graph, Asterix system. All the results are averaged over 10 iterations. (a) – Running time as a function of sampling probability. (b) – Accuracy of the  $F_7 - F_{10}$  global counts, measured as ratio of the exact count to the estimated probability. (c) – Comparison of running times of Orca and our exact 4-Prof-Dist algorithm. Clearly, 4-Prof-Dist benefits from the use of multiple cores.

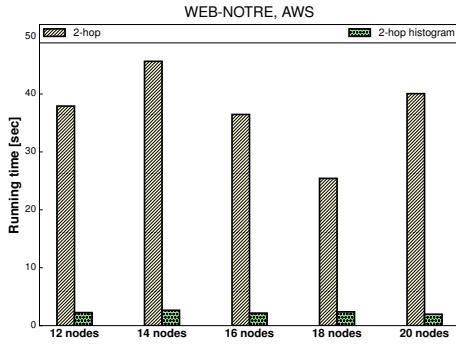


Figure 7: AWS cluster of up to 20 machines (nodes), results averaged over 10 iterations. Running time comparing naive 2-hop implementation and 2-hop histogram approach on the Notre Dame web graph.

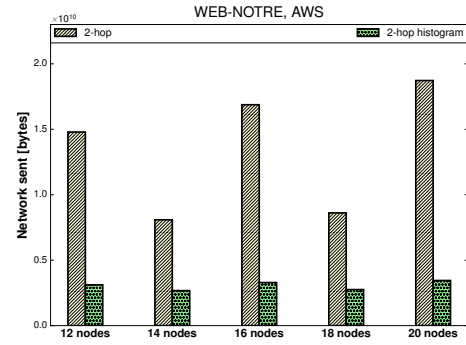


Figure 9: Network usage comparing naive 2-hop implementation and 2-hop histogram approach on the Notre Dame web graph.

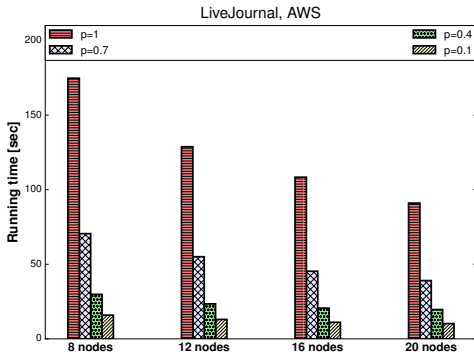


Figure 8: Running time of 4-Prof-Dist for various number of compute nodes and sampling probability  $\rho$ , on the LiveJournal graph.

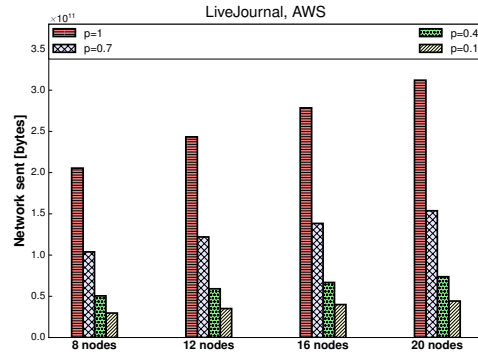


Figure 10: Network usage of 4-Prof-Dist for various number of compute nodes and sampling probability  $\rho$ , on the LiveJournal graph.

**Acknowledgements:** We would like to thank the anonymous reviewers for their useful comments. This research has been supported by NSF Grants CCF 1344179, 1344364, 1407278, 1422549 and ARO YIP W911NF-14-1-0258.

## 6. REFERENCES

[1] N. K. Ahmed, N. Duffield, J. Neville, and R. Kompella. Graph Sample and Hold: A Framework for Big-Graph Analytics. In *KDD*, 2014.



- [2] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield. Efficient Graphlet Counting for Large Networks. In *IEEE International Conference on Data Mining*, 2015.
- [3] N. Alon, Y. Matias, and M. Szegedy. The Space Complexity of Approximating the Frequency Moments. In *STOC*, pages 20–29, 1996.
- [4] L. Becchetti, P. Boldi, C. Castillo, and A. Gionis. Efficient Semi-Streaming Algorithms for Local Triangle Counting in Massive Graphs. In *KDD*, 2008.
- [5] C. Borgs, J. Chayes, and K. Vesztegombi. Counting Graph Homomorphisms. *Topics in Discrete Mathematics*, pages 315–371, 2006.
- [6] T. Eden, A. Levi, D. Ron, and C. Seshadhri. Approximately Counting Triangles in Sublinear Time. In *FOCS*, pages 614–633, 2015.
- [7] E. R. Elenberg, K. Shanmugam, M. Borokhovich, and A. G. Dimakis. Beyond Triangles: A Distributed Framework for Estimating 3-profiles of Large Graphs. In *KDD*, pages 229–238, 2015.
- [8] E. R. Elenberg, K. Shanmugam, M. Borokhovich, and A. G. Dimakis. Distributed Estimation of Graph 4-profiles. <http://arxiv.org/abs/1510.02215>, 2015.
- [9] F. Fei, B. Jie, and D. Zhang. Frequent and Discriminative Subnetwork Mining for Mild Cognitive Impairment Classification. *Brain Connectivity*, 4(5):347–360, June 2014.
- [10] I. Finocchi, M. Finocchi, and E. G. Fusco. Clique Counting in MapReduce: Algorithms and Experiments. *ACM Journal of Experimental Algorithmics*, 20(1), 2015.
- [11] D. Gavinsky, S. Lovett, M. Saks, and S. Srinivasan. A Tail Bound for Read-k Families of Functions. *Random Structures & Algorithms*, 47(1):99–108, 2015.
- [12] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin. PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs. In *10th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 17–30, 2012.
- [13] T. Hočevár and J. Demšar. A Combinatorial Approach to Graphlet Counting. *Bioinformatics*, 30(4):559–65, Feb. 2014.
- [14] S. Janson, K. Oleszkiewicz, and A. Ruciński. Upper Tails for Subgraph Counts in Random Graphs. *Israel Journal of Mathematics*, 142(1):61–92, 2004.
- [15] M. Jha, C. Seshadhri, and A. Pinar. Path Sampling: A Fast and Provable Method for Estimating 4-Vertex Subgraph Counts. In *WWW*, pages 495–505, 2015.
- [16] J. H. Kim and V. H. Vu. Concentration of Multivariate Polynomials and Its Applications. *Combinatorica*, 20(3):417–434, 2000.
- [17] M. Kowaluk, A. Lingas, and E.-M. Lundell. Counting and Detecting Small Subgraphs via Equations. *SIAM Journal of Discrete Mathematics*, 27(2):892–909, 2013.
- [18] J. Leskovec and A. Krevl. SNAP Datasets: Stanford Large Network Dataset Collection. <http://snap.stanford.edu/data>, June 2014.
- [19] L. Lovász. *Large Networks and Graph Limits*, volume 60. American Mathematical Soc., 2012.
- [20] A. Milinković, S. Milinković, and L. Lazić. A Contribution to Acceleration of Graphlet Counting. In *Infotech Jahorina Symposium*, volume 14, pages 741–745, 2015.
- [21] D. O’Callaghan, M. Harrigan, J. Carthy, and P. Cunningham. Identifying Discriminating Network Motifs in YouTube Spam. Feb. 2012.
- [22] N. Przulj. Biological Network Comparison Using Graphlet Degree Distribution. *Bioinformatics*, 23(2):177–183, 2007.
- [23] N. Satish, N. Sundaram, M. A. Patwary, J. Seo, J. Park, M. A. Hassaan, S. Sengupta, Z. Yin, and P. Dubey. Navigating the Maze of Graph Analytics Frameworks using Massive Graph Datasets. In *SIGMOD*, pages 979–990, 2014.
- [24] T. Schank. *Algorithmic Aspects of Triangle-Based Network Analysis*. PhD thesis, 2007.
- [25] C. Seshadhri, A. Pinar, and T. G. Kolda. Triadic Measures on Graphs: The Power of Wedge Sampling. In *Proceedings of the SIAM Conference on Data Mining*, pages 10–18, 2013.
- [26] N. Shervashidze, K. Mehlhorn, and T. H. Petri. Efficient Graphlet Kernels for Large Graph Comparison. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 488–495, 2009.
- [27] C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos. DOULION: Counting Triangles in Massive Graphs with a Coin. In *SIGKDD*, 2009.
- [28] C. E. Tsourakakis, M. Kolountzakis, and G. L. Miller. Triangle Sparsifiers. *Journal of Graph Theory and Applications*, 15(6):703–726, 2011.
- [29] J. Ugander, L. Backstrom, M. Park, and J. Kleinberg. Subgraph Frequencies: Mapping the Empirical and Extremal Geography of Large Graph Collections. In *WWW*, pages 1307–1318, 2013.
- [30] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. Structural Properties of the Caenorhabditis elegans Neuronal Network. 7(2), 2011.
- [31] V. V. Williams, J. Wang, R. Williams, and H. Yu. Finding Four-Node Subgraphs in Triangle Time. *SODA*, pages 1671–1680, 2014.

## APPENDIX

### A. IMPLEMENTATION DETAILS

To improve the practical performance of 4-Prof-Dist (see Algorithm 1 for pseudocode), we handle low and high degree vertices differently. As in GraphLab PowerGraph's standard triangle counting, cuckoo hash tables are used if the vertex degree is above a threshold. Now, we also threshold vertices to determine whether the 2-hop histogram in Section 2.3 will be either a vector or an unordered map. This is because sorting and merging operations on a vector scale poorly with increasing degree size, while an unordered map has constant lookup time. We found that this approach successfully trades off processing time and memory consumption.

### B. EXTENSION TO GLOBAL 4-PROFILE SPARSIFIER

Another advantage to read- $k$  function families is that they are simpler to extend to more complex subgraphs. We now state concentration results for the full 4-profile sparsifier evaluated experimentally in Section 4. Using the notation in Section 3, the edge sampling matrix  $\mathbf{H}$  is defined by the relations

$$\begin{bmatrix} \mathbb{E}[Y_0] \\ \vdots \\ \mathbb{E}[Y_{10}] \end{bmatrix} = \mathbf{H} \begin{bmatrix} N_0 \\ \vdots \\ N_{10} \end{bmatrix} \Rightarrow \begin{bmatrix} X_0 \\ \vdots \\ X_{10} \end{bmatrix} = \mathbf{H}^{-1} \begin{bmatrix} Y_0 \\ \vdots \\ Y_{10} \end{bmatrix} ;$$

Let  $t = \frac{p-1}{p}$ . Then the inverse sampling matrix is given by

$$\mathbf{H}^{-1} = \begin{bmatrix} S_{11} & S_{12} \\ \mathbf{0}_{4 \times 7} & S_{22} \end{bmatrix} ; \quad \text{where}$$

$$S_{11} = \begin{bmatrix} 1 & t & t^2 & t^2 & t^3 & t^3 & t^3 \\ 0 & \frac{1}{p} & \frac{2t}{p} & \frac{2t}{p} & \frac{3t^2}{p} & \frac{3t^2}{p} & \frac{3t^2}{p} \\ 0 & 0 & \frac{1}{p^2} & 0 & \frac{2t}{p^2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{p^2} & \frac{2t}{p^2} & \frac{3t}{p^2} & \frac{3t}{p^2} \\ 0 & 0 & 0 & 0 & \frac{1}{p^3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{p^3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{p^3} \end{bmatrix} ;$$

$$S_{12} = \begin{bmatrix} t^4 & t^4 & t^5 & t^6 \\ \frac{4t^3}{4t^3} & \frac{4t^3}{4t^3} & \frac{5t^4}{5t^4} & \frac{6t^5}{6t^5} \\ \frac{p}{2t^2} & \frac{p}{t^2} & \frac{p}{2t^3} & \frac{p}{3t^4} \\ \frac{4t^2}{p^2} & \frac{5t^2}{p^2} & \frac{8t^3}{p^2} & \frac{12t^4}{p^2} \\ \frac{4t}{p^3} & \frac{2t}{p^3} & \frac{6t^2}{p^3} & \frac{12t^3}{p^3} \\ 0 & \frac{t}{p^3} & \frac{2t^2}{p^3} & \frac{4t^3}{p^3} \\ 0 & \frac{t}{p^3} & \frac{2t^2}{p^3} & \frac{4t^3}{p^3} \end{bmatrix} ;$$

$$S_{22} = \begin{bmatrix} \frac{1}{p^4} & 0 & \frac{t}{p^4} & \frac{3t^2}{p^4} \\ 0 & \frac{1}{p^4} & \frac{4t}{p^4} & \frac{12t^2}{p^4} \\ 0 & 0 & \frac{1}{p^5} & \frac{6t}{p^5} \\ 0 & 0 & 0 & \frac{1}{p^6} \end{bmatrix} ;$$

and  $\mathbf{0}_{4 \times 7}$  is a  $4 \times 7$  matrix of zeros.

The binomial coefficients in these matrices influence our concentration bounds. A more detailed proof of the following result may be found in the extended version of this paper [8].

**Theorem 3** (4-profile sparsifier). *Consider the sampling process described above and in Section 3. Let  $X_i$ ;  $0 \leq i \leq 10$  (and  $\mathbf{X}$  be a vector of these estimates), be the actual estimates of 4-profiles. Let  $k_i$  be the maximum number of subgraphs  $F_i$  sharing a common edge. Let  $Y_i$ ;  $0 \leq i \leq 10$ , be the 4 profile counts of the sparsified graph. Then let  $N_i$ ;  $0 \leq i \leq 10$ , be the actual counts. Choose  $0 < \epsilon < 1$  and  $\delta > 0$ . Let  $C = (192)^2 = 2$  and*

$$k_\alpha = k_2 + k_3; \quad k_\beta = k_4 + k_5 + k_6; \quad k_\gamma = k_7 + k_8; \\ N_\alpha = N_2 + N_3; \quad N_\beta = N_4 + N_5 + N_6; \quad N_\gamma = N_7 + N_8;$$

If

$$\rho \geq \left( \frac{C \log(2=)(k_{10})}{2N_{10}} \right)^{1/12} ; \quad \rho \geq \left( \frac{C \log(2=)(k_9 + 6k_{10})}{2(N_9 + 6N_{10})} \right)^{1/10}$$

$$\rho \geq \left( \frac{C \log(2=)(k_8 + 4k_9 + 12k_{10})}{2(N_8 + 4N_9 + 12N_{10})} \right)^{1/8}$$

$$\rho \geq \left( \frac{C \log(2=)(k_7 + k_9 + 3k_{10})}{2(N_7 + N_9 + 3N_{10})} \right)^{1/8}$$

$$\rho \geq \left( \frac{C \log(2=)(k_6 + k_8 + 2k_9 + 4k_{10})}{2(N_6 + N_8 + 2N_9 + 4N_{10})} \right)^{1/6}$$

$$\rho \geq \left( \frac{C \log(2=)(k_5 + k_8 + 2k_9 + 4k_{10})}{2(N_5 + N_8 + 2N_9 + 4N_{10})} \right)^{1/6}$$

$$\rho \geq \left( \frac{C \log(2=)(k_4 + 4k_7 + 2k_8 + 6k_9 + 12k_{10})}{2(N_4 + 4N_7 + 2N_8 + 6N_9 + 12N_{10})} \right)^{1/6}$$

$$\rho \geq \left( \frac{C \log(2=)}{2} \right)^{1/4} \times$$

$$\left( \frac{k_3 + 2k_4 + 3k_5 + 3k_6 + 4k_7 + 5k_8 + 8k_9 + 12k_{10}}{N_3 + 2N_4 + 3N_5 + 3N_6 + 4N_7 + 5N_8 + 8N_9 + 12N_{10}} \right)^{1/4}$$

$$\rho \geq \left( \frac{C \log(2=)(k_2 + k_4 + 2k_7 + k_8 + 2k_9 + 3k_{10})}{2(N_2 + N_4 + 2N_7 + N_8 + 2N_9 + 3N_{10})} \right)^{1/4}$$

$$\rho \geq \left( \frac{C \log(2=)(k_1 + 2k_\alpha + 3k_\beta + 4k_\gamma + 5k_9 + 6k_{10})}{2(N_1 + 2N_\alpha + 3N_\beta + 4N_\gamma + 5N_9 + 6N_{10})} \right)^{1/2}$$

$$n_0 \leq |V|^2 \left( |V|^2 - \frac{C \log(2=)}{2} \right);$$

then  $\|\mathbf{X}\|_\infty \leq \binom{|V|}{4}$  with probability at least  $1 - \delta$ .

**Proof.** We apply Proposition 1 a total of 11 times to the sampling-estimator system defined above by  $\mathbf{H}$  and  $\mathbf{H}^{-1}$ . In our context, each sampled subgraph count  $Y_i$  is a sum of functions in a read- $k_{Y_i}$  family, where  $k_{Y_i} \leq \min\{|V| - 2; N_i\}$ . Let  $k_{i,e}$  be the maximum number of subgraphs  $F_i$  sharing a common edge  $e$ , and let  $k_i = \max_e k_{i,e}$ , for  $i = 0; \dots; 10$ .

The  $Y_i$ 's have the following parameters:

$$\begin{aligned}
r_{Y_0} &= \binom{|V|}{4}; \quad k_{Y_0} = |V| \\
r_{Y_1} &= N_1 + 2N_2 + 2N_3 + 3N_4 + 3N_5 + 3N_6 + 4N_7 + \\
&\quad 4N_8 + 5N_9 + 6N_{10} \\
k_{Y_1} &= k_1 + 2k_2 + 2k_3 + 3k_4 + 3k_5 + 3k_6 + 4k_7 + 4k_8 + \\
&\quad 5k_9 + 6k_{10} \\
r_{Y_2} &= N_2 + N_4 + 2N_7 + N_8 + 2N_9 + 3N_{10} \\
k_{Y_2} &= k_2 + k_4 + 2k_7 + k_8 + 2k_9 + 3k_{10} \\
r_{Y_3} &= N_3 + 2N_4 + 3N_5 + 3N_6 + 4N_7 + 5N_8 + 8N_9 + 12N_{10} \\
k_{Y_3} &= k_3 + 2k_4 + 3k_5 + 3k_6 + 4k_7 + 5k_8 + 8k_9 + 12k_{10} \\
r_{Y_4} &= N_4 + 4N_7 + 2N_8 + 6N_9 + 12N_{10} \\
k_{Y_4} &= k_4 + 4k_7 + 2k_8 + 6k_9 + 12k_{10} \\
r_{Y_5} &= N_5 + N_8 + 2N_9 + 4N_{10}; \quad k_{Y_5} = k_5 + k_8 + 2k_9 + 4k_{10} \\
r_{Y_6} &= N_6 + N_8 + 2N_9 + 4N_{10}; \quad k_{Y_6} = k_6 + k_8 + 2k_9 + 4k_{10} \\
r_{Y_7} &= N_7 + N_9 + 3N_{10}; \quad k_{Y_7} = k_7 + k_9 + 3k_{10} \\
r_{Y_8} &= N_8 + 4N_9 + 12N_{10}; \quad k_{Y_8} = k_8 + 4k_9 + 12k_{10} \\
r_{Y_9} &= N_9 + 6N_{10}; \quad k_{Y_9} = k_9 + 6k_{10} \\
r_{Y_{10}} &= N_{10}; \quad k_{Y_{10}} = k_{10}
\end{aligned}$$

We apply Proposition 1 to each estimator. This is shown in the proof of Theorem 2 for  $Y_{10}$  and in the extended paper [8] for the other estimators. Rearranging to solve for  $\rho$ ,

$$\begin{aligned}
\rho &\geq \left( \frac{\log(2=)k_{10}}{2^2 N_{10}} \right)^{1/12}; \quad \rho \geq \left( \frac{\log(2=)(k_9 + 6k_{10})}{2^2 (N_9 + 6N_{10})} \right)^{1/10} \\
\rho &\geq \left( \frac{\log(2=)(k_8 + 4k_9 + 12k_{10})}{2^2 (N_8 + 4N_9 + 12N_{10})} \right)^{1/8} \\
\rho &\geq \left( \frac{\log(2=)(k_7 + k_9 + 3k_{10})}{2^2 (N_7 + N_9 + 3N_{10})} \right)^{1/8} \\
\rho &\geq \left( \frac{\log(2=)(k_6 + k_7 + 2k_9 + 4k_{10})}{2^2 (N_6 + N_8 + 2N_9 + 4N_{10})} \right)^{1/6} \\
\rho &\geq \left( \frac{\log(2=)(k_5 + k_7 + 2k_9 + 4k_{10})}{2^2 (N_5 + N_8 + 2N_9 + 4N_{10})} \right)^{1/6} \\
\rho &\geq \left( \frac{\log(2=)(k_4 + 4k_7 + 2k_8 + 6k_9 + 12k_{10})}{2^2 (N_4 + 4N_7 + 2N_8 + 6N_9 + 12N_{10})} \right)^{1/6} \\
\rho &\geq \left( \frac{\log(2=)}{2^2} \right)^{1/4} \times \\
&\quad \left( \frac{k_3 + 2k_4 + 3k_5 + 3k_6 + 4k_7 + 5k_8 + 8k_9 + 12k_{10}}{N_3 + 2N_4 + 3N_5 + 3N_6 + 4N_7 + 5N_8 + 8N_9 + 12N_{10}} \right)^{1/4} \\
\rho &\geq \left( \frac{\log(2=)(k_2 + k_4 + 2k_7 + k_8 + 2k_9 + 3k_{10})}{2^2 (N_2 + N_4 + 2N_7 + N_8 + 2N_9 + 3N_{10})} \right)^{1/4} \\
\rho &\geq \left( \frac{\log(2=)(k_1 + 2k_\alpha + 3k_\beta + 4k_\gamma + 5k_9 + 6k_{10})}{2^2 (N_1 + 2N_\alpha + 3N_\beta + 4N_\gamma + 5N_9 + 6N_{10})} \right)^{1/2};
\end{aligned}$$

where

$$\begin{aligned}
k_\alpha &= k_2 + k_3; \quad k_\beta = k_4 + k_5 + k_6; \quad k_\gamma = k_7 + k_8; \\
N_\alpha &= N_2 + N_3; \quad N_\beta = N_4 + N_5 + N_6; \quad N_\gamma = N_7 + N_8.
\end{aligned}$$

The final condition comes from the result for  $Y_0$ :

$$n_0 \leq \binom{|V|}{4} - \frac{\log(2=)|V|^2}{2^2} \leq |V|^2 \left( |V|^2 - \frac{\log(2=)}{2^2} \right):$$

Plugging into our estimators (given by  $\mathbf{H}^{-1}$ ), we get the following error bounds:

$$\begin{aligned}
X_0 &\leq (n_1 + n_2 + n_3) + (n_1 + 2n_2 + 3n_3 + n_2 + 3n_3 + n_3) \\
&\leq (2n_1 + 4n_2 + 8n_3) \leq 8 \binom{|V|}{3} \\
X_1 &\leq (N_1 + \dots + 192N_{10}) \leq 192 \binom{|V|}{4} \\
X_2 &\leq (N_2 + \dots + 48N_{10}) \leq 48 \binom{|V|}{4} \\
X_3 &\leq (N_3 + 4N_4 + 6N_5 + \dots + 192N_{10}) \leq 192 \binom{|V|}{4} \\
X_4 &\leq (N_4 + \dots + 96N_{10}) \leq 96 \binom{|V|}{4} \\
X_5 &\leq (N_5 + \dots + 32N_{10}) \leq 32 \binom{|V|}{4} \\
X_6 &\leq (N_6 + \dots + 32N_{10}) \leq 32 \binom{|V|}{4} \\
X_7 &\leq (N_7 + 2N_9 + 12N_{10}) \leq 12 \binom{|V|}{4} \\
X_8 &\leq (N_8 + 4N_9 + 12N_{10}) + 4(N_9 + 6N_{10}) + 12(N_{10}) \\
&\leq (N_8 + 8N_9 + 48N_{10}) \leq 48 \binom{|V|}{4} \\
X_9 &\leq (N_9 + 6N_{10}) + 6(N_{10}) \\
&\leq (N_9 + 12N_{10}) \leq 12 \binom{|V|}{4} \\
X_{10} &\leq N_{10}:
\end{aligned}$$

Thus the maximum deviation in any estimator is less than  $192 \binom{|V|}{4}$ . Substituting  $=^2 = =^2=(192)^2 = =^2=2C$  completes the proof.  $\square$