

# RECEIPT: REfine CoarsE-grained IndePendent Tasks for Parallel Tip decomposition of Bipartite Graphs

Kartik Lakhotia

Ming Hsieh Department of Electrical Engineering  
University of Southern California  
klakhoti@usc.edu

Viktor Prasanna

Ming Hsieh Department of Electrical Engineering  
University of Southern California  
prasanna@usc.edu

Rajgopal Kannan

U.S Army Research Lab  
Los Angeles, CA 90094  
rajgopal.kannan.civ@mail.mil

Cesar A. F. De Rose

School of Technology  
Pontifical Catholic University of Rio Grande do Sul  
cesar.derose@puccrs.br

## ABSTRACT

Tip decomposition is a crucial kernel for mining dense subgraphs in bipartite networks, with applications in spam detection, analysis of affiliation networks etc. It creates a hierarchy of vertex-induced subgraphs with varying densities determined by the participation of vertices in butterflies (2, 2–biclques). To build the hierarchy, existing algorithms iteratively follow a *delete-update*(peeling) process: deleting vertices with the minimum number of butterflies and correspondingly updating the butterfly count of their 2-hop neighbors. The need to explore 2-hop neighborhood renders tip-decomposition computationally very expensive. Furthermore, the inherent sequentiality in peeling only minimum butterfly vertices makes derived parallel algorithms prone to heavy synchronization.

In this paper, we propose a novel parallel tip-decomposition algorithm – REfine CoarsE-grained Independent Tasks (RECEIPT) that relaxes the peeling order restrictions by partitioning the vertices into multiple independent subsets that can be concurrently peeled. This enables RECEIPT to simultaneously achieve a high degree of parallelism and dramatic reduction in synchronizations. Further, RECEIPT employs a hybrid peeling strategy along with other optimizations that drastically reduce the amount of wedge exploration and execution time.

We perform detailed experimental evaluation of RECEIPT on a shared-memory multicore server. It can process some of the largest publicly available bipartite datasets *orders of magnitude faster* than the state-of-the-art algorithms – achieving up to 1100× and 64× reduction in the number of thread synchronizations and traversed wedges, respectively. Using 36 threads, RECEIPT can provide up to 17.1× self-relative speedup.

## PVLDB Reference Format:

Kartik Lakhotia, Rajgopal Kannan, Viktor Prasanna, and Cesar A. F. De Rose. RECEIPT: REfine CoarsE-grained IndePendent Tasks for Parallel Tip decomposition of Bipartite Graphs. PVLDB, 14(3): 404–417, 2021.  
doi:10.14778/3430915.3430929

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing [info@vldb.org](mailto:info@vldb.org). Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 14, No. 3 ISSN 2150-8097.  
doi:10.14778/3430915.3430929

## PVLDB Artifact Availability:

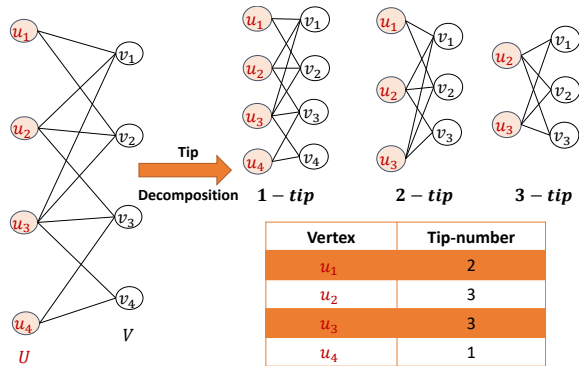
The source code, data, and/or other artifacts have been made available at <https://github.com/kartiklakhotia/RECEIPT>.

## 1 INTRODUCTION

Dense subgraph mining is a fundamental problem used for anomaly detection, spam filtering, social network analysis, trend summarizing and several other real-world applications [9, 14, 20, 44, 56, 68]. Many of these modern day applications use bipartite graphs to effectively represent two-way relationship structures, such as consumer-product purchase history [26], user-group memberships in social networks [38], author-paper networks [41], etc. Consequently, mining cohesive structures in bipartite graphs has gained tremendous interest in recent years [51, 54, 63, 64, 69].

Many techniques have been developed to uncover hierarchical dense structures in unipartite graphs, such as truss and core decomposition [7, 20, 27, 36, 45, 49, 60, 67]. Such off-the-shelf analytics can be conveniently utilized for discovering dense parts in projections of bipartite graphs as well [42]. However, this approach results in a loss of information and a blowup in the size of the projection graphs [51]. To prevent these issues, researchers have explored the role of butterflies (2, 2–biclques) to mine dense subgraphs directly in bipartite networks [51, 69]. A butterfly is the most basic unit of cohesion in bipartite graphs. Recently, Sariyuce et al. conceptualized  $k$ -tip as a vertex-induced subgraph with at least  $k$  butterflies incident on every vertex in one of the bipartite vertex sets (fig.1). They show that  $k$ -tips can unveil hierarchical dense regions in bipartite graphs more effectively than unipartite approaches applied on projection graphs. As a space-efficient representation for the  $k$ -tip hierarchy, Sariyuce et al. further define the notion of *tip number* of a vertex  $u$  as the largest  $k$  for which a  $k$ -tip contains  $u$ . In this paper, we study the problem of finding tip numbers of vertices in a bipartite graph, also known as *tip decomposition*.

Tip decomposition can be employed in several real-world applications that utilize dense subgraphs. It can find (hierarchical) groups of researchers with common affiliations from author-paper networks [51]. It can be used to detect communities of spam reviewers from user-rating graphs in databases of e-commerce companies; such reviewers collaboratively rate selected products, appearing in close-knit structures [17, 39] that tip decomposition can unveil. It can be used for document clustering, graph summarization and link



**Figure 1: Tip decomposition of vertex set  $U$  in a bipartite graph.**  $u_4$  and  $u_1$  participate in 1 and 2 butterflies, respectively. Although  $u_3$  participates in 5 butterflies in original graph, only 3 of them are with  $u_2$  with which it creates a 3-tip.

prediction as dense  $k$ -tips unveil groups of nodes with connections to common and similar sets of neighbors [12, 14, 32, 40].

Existing sequential and tip decomposition algorithms [51, 54] employ a bottom-up peeling approach. Vertices with the minimum butterfly count are peeled (deleted), the count of their 2-hop neighbors with shared butterflies is decremented, and the process is then iterated. However, exploring 2-hop neighborhood for every vertex requires traversing a large number of wedges, rendering tip decomposition computationally intensive. For example, the  $TrU$  dataset has 140 million edges but peeling it requires traversing 211 trillion wedges, which is intractable for a sequential algorithm. For such high complexity analytics, parallel computing is often used to scale to large datasets [33, 43, 55]. In case of tip decomposition, the bottom-up peeling approach used by the existing parallel algorithms severely restricts their scalability. It mandates constructing levels of the  $k$ -tip hierarchy in non-decreasing order of tip numbers, thus imposing sequential restrictions on the order of vertex peeling. Note that the parallel threads need to synchronize at the end of every peeling iteration. Further note that the range of tip numbers can be quite large and the number of iterations required to peel all vertices with a given tip number is variable. Taken together, the conventional approach of parallelizing the workload within each iteration requires major synchronization, rendering it ineffective. For example,  $PARBUTTERFLY$  experiences  $\approx 1.5$  million synchronization rounds for peeling  $TrU$  [54]. These observations motivate the need for an algorithm that exploits the parallelism available across multiple levels of a  $k$ -tip hierarchy to reduce the amount of synchronizations.

In this paper, we propose the REfine CoarsE-grained IndePendent Tasks (RECEIPT) algorithm, that adopts a novel two-step approach to drastically reduce the amount of parallel peeling iterations and in turn, the amount of synchronization. The key insight that drives the development of RECEIPT is that the tip-number  $\theta_u$  of a vertex  $u$  only depends on the butterflies shared between  $u$  and other vertices with tip numbers *greater than or equal to*  $\theta_u$ . Thus, vertices with smaller tip numbers can be peeled in any order without affecting the correctness of  $\theta_u$  computation. To this purpose, RECEIPT divides tip decomposition into a *two-step computation* where each step exposes parallelism across a different dimension.

In the first step, it creates few *non-overlapping* ranges that represent lower and upper bounds on the vertices' tip numbers. To find vertices with a lower bound  $\theta$ , all vertices with upper bound smaller than  $\theta$  can be peeled simultaneously, providing *sufficient vertex-workload per parallel peeling iteration*. The small number of ranges ensures *little synchronization* in this step. In the second step, RECEIPT *concurrently processes vertex subsets* corresponding to different ranges to compute the exact tip numbers<sup>1</sup>. The absence of overlap between tip number ranges enables each of these subsets to be peeled *independently* of other vertices in the graph.

RECEIPT's two-step approach further enables development of novel optimizations that radically decrease the amount of wedge exploration. Thus, RECEIPT combines both *computational efficiency* and *parallel performance* for fast decomposition of large bipartite graphs. Overall, our contributions can be summarized as follows:

- (1) We propose a novel RefinE CoarsE-grained Independent Tasks (RECEIPT) algorithm for tip decomposition in bipartite graphs. RECEIPT is the *first* algorithm that parallelizes workload across the levels of subgraph hierarchy created by tip decomposition.
- (2) We show that RECEIPT is theoretically work efficient and dramatically reduces thread synchronization. As an example, it incurs only 1335 synchronization rounds while processing  $TrU$ , which is  $1105\times$  less than the existing parallel algorithms.
- (3) We develop novel optimizations enabled by the two-step approach of RECEIPT. These optimizations drastically reduce the amount of wedge exploration and improve computational efficiency of RECEIPT. For instance, we traverse 3297B wedges to tip decompose  $TrU$ , which is  $64\times$  less than the state-of-the-art.

We conduct detailed experiments using some of the largest public real-world bipartite graphs. RECEIPT extends the limits of current practice by feasibly computing tip decomposition for these datasets. For example, it can process the  $TrU$  graph in 46 minutes whereas the state-of-the-art does not finish in 10 days. Using 36 threads, we achieve up to  $17.1\times$  parallel speedup.

## 2 BACKGROUND

In this section, we will discuss state-of-the-art algorithms for counting per-vertex butterflies (sec.2.1). Counting is used to initialize the *support* (running count of incident butterflies) of each vertex during tip-decomposition, and also constitutes a crucial optimization in RECEIPT. Hence, it is imperative to analyze counting algorithms. We will also discuss the bottom-up peeling approach for tip-decomposition used by the existing algorithms (sec.2.2).

Table 1 lists the notations used in this paper. Note that we decompose either  $U$  or  $V$  vertex set at a time. For clarity of description, we assume that  $U$  is the primary set to process and use the word "wedge" to imply a wedge with endpoints in  $U$ . However, for empirical analysis, we will individually decompose both sets in each graph.

### 2.1 Per-vertex butterfly counting

A butterfly (2,2-bicliques/quadrangle) is a combination of two wedges with common endpoints. A simple way to count butterflies is to explore all wedges and combine the ones with common end points. However, counting per-vertex butterflies using this procedure is

<sup>1</sup>A vertex subset for a given range is peeled by a single thread in the second step.

**Table 1: Frequently used notations**

$G(W = (U, V), E)$	bipartite graph $G$ with vertices $W$ and edges $E$ $U$ and $V$ are two disjoint vertex sets in $G$
$n, m$	no. of vertices and edges in $G$ ; $n =  W , m =  E $
$\alpha$	arboricity of $G$ [10]
$N_u$	neighboring vertices of $u$
$d_u$	degree of vertex $u$ ; $d_u =  N_u $
$\triangleright_u / \triangleright_U$	support (# butterflies) of $u$ / vertices in set $U$
$\triangleright_{u_1, u_2} = \triangleright_{u_2, u_1}$	# butterflies shared between $u_1$ and $u_2$
$\wedge_U$	# wedges with endpoints in set $U$
$\theta_u$	tip number of vertex $u$
$\theta^{max}$	maximum tip number for a vertex
$P$	number of vertex subsets created by RECEIPT
$T$	number of threads

extremely expensive with  $O\left(\sum_{u \in U} \sum_{v \in N_u} d_v\right)$  complexity (if we use vertices in  $U$  as end points).

Chiba and Nishizeki [10] proposed a vertex-priority quadrangle counting algorithm which traverses  $O\left(\sum_{(u,v) \in E} \min(d_u, d_v)\right) = O(\alpha \cdot m)$  wedges with  $O(1)$  work per wedge. This is the most efficient butterfly counting approach in terms of computational complexity. Wang et al.[63] further propose a cache efficient version of the vertex-priority algorithm that reorders the vertices in decreasing order of degree. Their algorithm only traverses wedges where degree of the last vertex is greater than the degree of the start and middle vertices. It can be easily parallelized by concurrently processing multiple start vertices [54, 63], as shown in alg.1. In RECEIPT, we adopt this parallel variant for per-vertex counting by adding the contributions from traversed wedges to start, mid and end-points. Each thread is provided a  $\theta(|W|)$  array for wedge aggregation (line 5, alg.1). This is similar to the best performing batch aggregation mode in the PARBUTTERFLY framework [54].

**Algorithm 1** Counting per-vertex butterflies (pvBcnt)

---

**Input:** Bipartite Graph  $G(W = (U, V), E)$   
**Output:** Butterfly counts  $\triangleright_u \forall u \in W$

- 1: Relabel vertices in  $G$  in descending order of degree
- 2: **for each**  $u \in U \cup V$  **do in parallel**
- 3:     Sort  $N_u$  in ascending order of new labels
- 4:      $\triangleright_w \leftarrow 0$
- 5: **for each**  $sp \in U \cup V$  **do in parallel**
- 6:     Initialize  $wdg\_arr$  array to all zeros
- 7:      $nze \leftarrow \{\phi\}, nzw \leftarrow \{\phi\}$
- 8:     **for each**  $mp \in N_{sp}$
- 9:         **for each**  $ep \in N_{mp}$
- 10:             **if**  $(ep \geq mp)$  or  $(ep \geq sp)$  **then** break
- 11:              $wdg\_arr[ep] \leftarrow wdg\_arr[ep] + 1$
- 12:              $nze \leftarrow nze \cup \{ep\}, nzw \leftarrow nzw \cup \{(mp, ep)\}$
- 13:     **for each**  $u \in nze$   $\triangleright$  same side contribution
- 14:          $bcnt \leftarrow \binom{wdg\_arr[u]}{2}$
- 15:         Atomically add  $bcnt$  to  $\triangleright_u$  and  $\triangleright_{sp}$
- 16:     **for each**  $(u, v) \in nzw$   $\triangleright$  opp. side contribution
- 17:          $bcnt \leftarrow wdg\_arr[v] - 1$
- 18:         Atomically add  $bcnt$  to  $\triangleright_u$

---

## 2.2 Tip Decomposition

Sariyuce et al.[51] introduced  $k$ -tips as vertex-induced subgraphs with large number of butterflies. They formally define it as follows:

**DEFINITION 1.** A bipartite subgraph  $H = (U', V, E) \subseteq G$ , induced on  $U$ , is a  **$k$ -tip** iff

- each vertex  $u \in U'$  participates in at least  $k$  butterflies,
- each vertex-pair  $(u, u') \in U'$  is connected by a series of butterflies,
- $H$  is maximal i.e. no other  $k$ -tip subsumes  $H$ .

$k$ -tips are hierarchical – a  $k$ -tip overlaps with  $k'$ -tips for all  $k' < k$ . Therefore, storing all  $k$ -tips is inefficient and often, infeasible. **Tip number**  $\theta_u$  is defined as the maximum  $k$  for which  $u$  is present in a  $k$ -tip. Tip numbers provide a space-efficient representation of  $k$ -tip hierarchy with quick retrieval. In this paper, we study the problem of finding tip numbers, known as **tip decomposition**.

Algorithms in current practice use Bottom-Up Peeling (BUP) for tip-decomposition, as shown in alg.2. It initializes vertex support using per-vertex butterfly counting, and then iteratively peels the vertices with minimum support until no vertex remains. When a vertex  $u \in U$  is peeled, its support at that instant is recorded as its tip number  $\theta_u$ . Further, for every vertex  $u'$  with  $\triangleright_{u, u'} > 0$  shared butterflies, the support of  $u'$  is decreased by  $\triangleright_{u, u'}$  (capped at  $\theta_u$ ). Thus, tip numbers are assigned in a non-decreasing order. The complexity of bottom-up peeling (alg.2), dominated by wedge traversal (lines 8-10), is  $O\left(\sum_{u \in U} \sum_{v \in N_u} d_v\right)$ .

**Algorithm 2** Tip decomposition using bottom-up peeling (BUP)

---

**Input:** Bipartite graph  $G(W = (U, V), E)$   
**Output:** Tip numbers  $\theta_u \forall u \in U$

- 1:  $\{\triangleright_U, \triangleright_V\} \leftarrow \text{pvBcnt}(G)$   $\triangleright$  Initial count
- 2: **while**  $U \neq \{\phi\}$  **do**  $\triangleright$  Peel
- 3:     let  $u \in U$  be the vertex with minimum support  $\triangleright_u$
- 4:      $\theta_u \leftarrow \triangleright_u, U \leftarrow U \setminus \{u\}$
- 5:     update( $u, \theta_u, \triangleright_U, G$ )
- 6: **function** update( $u, \theta_u, \triangleright_U, G$ )
- 7:     Initialize hashmap  $wdg\_arr$  to all zeros
- 8:     **for each**  $v \in N_u$
- 9:         **for each**  $u' \in N_v \setminus \{u\}$
- 10:              $wdg\_arr[u'] \leftarrow wdg\_arr[u'] + 1$
- 11:     **for each**  $u' \in wdg\_arr$   $\triangleright$  Update Support
- 12:          $\triangleright_{u, u'} \leftarrow \binom{wdg\_arr[u']}{2}$   $\triangleright$  shared butterflies
- 13:          $\triangleright_{u'} \leftarrow \max\{\theta_u, \triangleright_{u'} - \triangleright_{u, u'}\}$

---

**2.2.1 Challenges.** Tip decomposition is computationally very expensive and parallel computing is widely used to accelerate such workloads. However, the state-of-the-art parallel tip decomposition framework PARBUTTERFLY [13, 54] only utilizes parallelism within each peeling iteration by concurrently peeling all vertices with minimum support value. This restrictive approach is adopted due to the following sequential dependency between iterations – *support updates computed in an iteration guide the choice of vertices to peel in the subsequent iterations*. As shown in [54], PARBUTTERFLY is work-efficient with a complexity of  $O\left(\sum_{u \in U} \sum_{v \in N_u} d_v + \rho_v \log^2 m\right)$ , where  $\rho_v$  is the number of peeling iterations. However, its scalability is limited in practice due to the following drawbacks:

(1) Alg. 2 incurs large number of iterations and low workload per individual iteration. The resulting synchronization and load imbalance render simple parallelization ineffective for acceleration.

**Objective 1** – Design an efficient parallel algorithm for tip decomposition that reduces thread synchronizations.

(2) Alg. 2 explores all wedges with end-points in  $U$ . This is computationally expensive and can make it infeasible even for a scalable parallel algorithm, to execute in reasonable time.

**Objective 2** – Reduce the amount of wedge traversal.

### 3 REFINE COARSE-GRAINED INDEPENDENT TASKS ALGORITHM

In this section, we present a novel shared-memory parallel algorithm for tip decomposition – Refine CoarsE-grained IndePendent Tasks (RECEIPT), that drastically reduces the number of parallel peeling iterations (objective 1, sec.2.2.1). The fundamental insight underlying RECEIPT is that  $\theta_u$  only depends on the number of butterflies that  $u$  shares with vertices having tip numbers *no less than*  $\theta_u$ . Therefore, if we initialize the support  $\bowtie_u$  to total butterflies of  $u$  in  $G$ , only the following are required to compute  $\theta_u$ :

- The *aggregate* effect of peeling all vertices  $v$  with  $\theta_v < \theta_u$  on  $\bowtie_u$ :  $\bowtie_u$  would be  $\geq \theta_u$  after such deletion.
- The effect of deleting vertices  $v$  with  $\theta_v = \theta_u$  on  $\bowtie_u$ :  $\bowtie_u$  would be  $\leq \theta_u$  after such deletion.

This insight allows us to eliminate the major bottleneck for parallel tip decomposition i.e. the constraint of deleting only minimum support vertices in any peeling iteration. In order to find vertices with tip number greater than or equal to  $\theta$ , all vertices with tip numbers less than  $\theta$  can be peeled simultaneously, providing sufficient parallelism. However, for every  $\theta \in \{0, 1 \dots \theta^{max}\}$ , peeling all the vertices  $v$  with  $\theta_v < \theta$  and computing corresponding support updates will make the algorithm extremely inefficient. To avoid this inefficiency, RECEIPT follows a 2-step approach as shown in fig.2.

In the first step, it divides the spectrum<sup>2</sup> of tip numbers  $[0, \theta^{max}]$  into  $P$  smaller ranges  $R_1, R_2 \dots R_P$ , where  $P$  is a user-defined parameter. A range  $R_i$  is defined by the set of integers in  $[\theta(i), \theta(i+1))$  with boundary conditions  $\theta(1) = 0$  and  $\theta(P+1) > \theta^{max}$ . Note that the ranges have no overlap i.e. for any pair  $(i, j)$ ,  $i \neq j$  implies  $R_i \cap R_j = \{\phi\}$ . Corresponding to each range  $R_i$ , RECEIPT CD also finds the subset of vertices  $U_i$  whose tip-numbers belong to that range i.e.  $U_i = \cup_{u \in U} \{u \mid \theta_u \in R_i\}$ . In other words, instead of finding the exact tip number  $\theta_u$  of a vertex  $u$ , the first step in RECEIPT computes *bounds* on  $\theta_u$ , by finding its range affiliation using peeling. Therefore, this step is named **Coarse-grained Decomposition (RECEIPT CD)**. The absence of overlap between the ranges allows each subset to be peeled independently of others for exact tip number computation in a later step.

RECEIPT CD has a stark difference from conventional bottom-up approach: instead of peeling vertices with minimum support, every iteration concurrently peels *all vertices* with support value in a *broad range*. Setting  $P \ll \theta^{max}$  ensures a large amount of vertices peeled per iteration (*sufficient parallel workload*) and significantly less number of iterations (*dramatically less synchronization*) compared to parallel variants of bottom-up peeling (alg.2).

<sup>2</sup>RECEIPT does not assume prior knowledge of maximum tip number value and computes an upper bound during the execution.

The next step finds the exact tip numbers of vertices and is termed **Fine-grained Decomposition (RECEIPT FD)**. The key idea behind RECEIPT FD is as follows – for each vertex  $u$ , if we know the number of butterflies it shares with vertices in its own subset  $U_i$  and in subsets with higher tip number ranges ( $\cup_{j>i} \{U_j\}$ ), then every subset can be peeled independently of others. RECEIPT FD exploits this independence to concurrently process multiple vertex subsets by simultaneously employing sequential bottom up peeling on the subgraphs induced by these subsets. Setting  $P \gg T$  ensures that RECEIPT FD can be efficiently parallelized. Thus, RECEIPT avoids strict sequential dependencies across peeling iterations and systematically parallelizes tip decomposition.

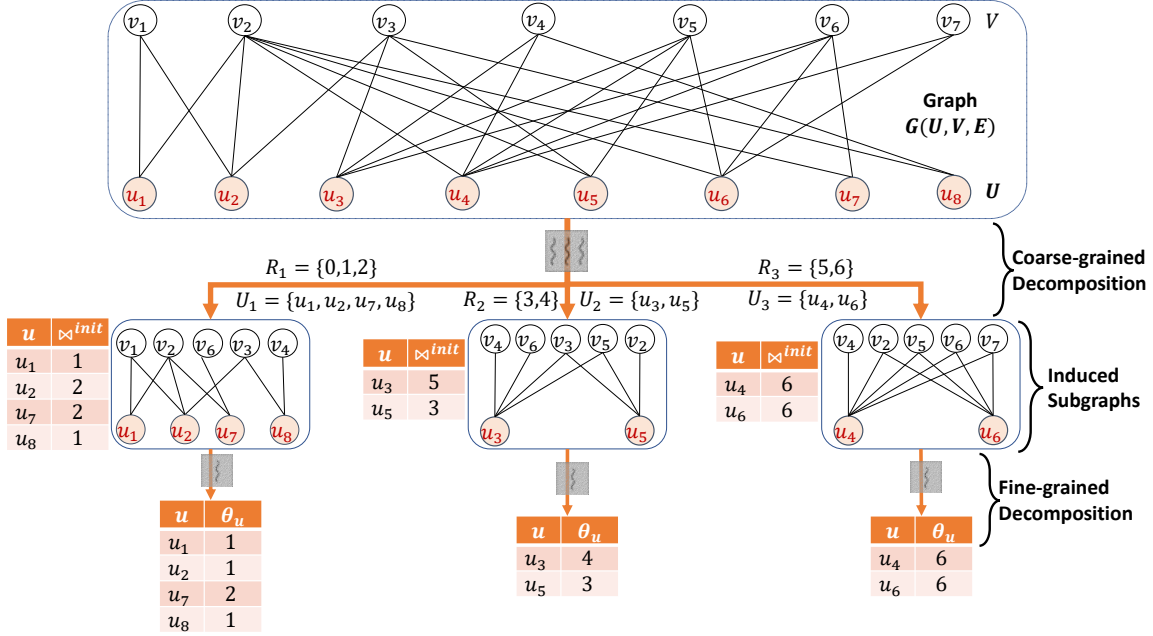
Since each butterfly has two vertices in  $U$ , butterflies with vertices in different subsets are not preserved in the induced subgraphs. Hence, butterfly counting on a subgraph induced on  $U_i$  will not account for butterflies shared between  $u \in U_i$  and vertices in subsets with higher tip number ranges. However, we note that when  $U_{i-1}$  is completely peeled in RECEIPT CD, the support of a vertex  $u \in U$  reflects the number of butterflies it shares with remaining vertices i.e.  $\cup_{j \geq i} \{U_j\}$ . This is precisely the initial butterfly count for  $u$  as required in RECEIPT FD. Hence, we store these values during RECEIPT CD (in  $\bowtie^{init}$  vector as shown in fig.2) and use them for support initialization in RECEIPT FD.

RECEIPT’s two-step approach can potentially double the workload as each wedge may get traversed in both steps. However, we note that since each subset is processed independently of others, support updates are *not communicated between the subsets*. Hence, when peeling a subset  $U_i$ , we only need to traverse wedges with both endpoints in  $U_i$ . Therefore, RECEIPT FD first creates an *induced subgraph* on  $U_i$  and only explores the wedges in that subgraph. This dramatically reduces the amount of work done in RECEIPT FD. For example, in fig.2, the original graph  $G$  has 38 wedges with endpoints in  $U$ , whereas the three subgraphs collectively have only 11 such wedges which will be traversed during RECEIPT FD.

#### 3.1 Coarse-grained Decomposition

Alg.3 depicts the pseudocode for Coarse-grained Decomposition (RECEIPT CD). It takes a bipartite graph  $G(U, V, E)$  as input and partitions the vertex set  $U$  into  $P$  subsets with different tip number ranges. Prior to creating vertex subsets, RECEIPT CD uses per-vertex counting (alg.1) to initialize the support of vertices in  $U$ .

The first step in computing a subset  $U_i$  is to find the tip number range  $R_i = [\theta(i), \theta(i+1))$ . Ideally, the ranges should be computed such that the number of wedges in the induced subgraphs are uniform for all  $P$  subsets (to ensure load balance during RECEIPT FD). However, induced subgraphs are not known prior to actual partitioning. Secondly, exact tip numbers are not known either and hence, vertices in  $U_i$  cannot be determined prior to partitioning, for different values of  $\theta(i+1)$ . Considering these challenges, RECEIPT CD uses two proxies for range determination (lines 16-21): (1) the number of wedges in the original graph as a proxy for the wedges in induced subgraphs, and (2) current support of vertices as a proxy for tip numbers. Now, to balance the wedge counts across subsets, RECEIPT CD aggregates the *wedge count (in G)* of vertices in bins corresponding to their *current support* and computes a prefix sum over the bins (Ref: proof of theorem 3). For any unique support



**Figure 2:** Graphical illustration of tip decomposition using RECEIPT. Coarse-grained Decomposition partitions  $U$  into three vertex-subsets  $U_1, U_2$  and  $U_3$  whose tip numbers belonging to ranges  $R_1, R_2$  and  $R_3$ , respectively. It processes each peeling iteration in parallel. Fine-grained decomposition creates subgraphs induced on  $U_1, U_2$  and  $U_3$ , initializes vertex support using  $\triangleright^{init}$  and peels each of them sequentially. It processes multiple subgraphs concurrently. Note that  $G$  has 38 wedges whereas the induced subgraphs collectively have only 11 wedges.

value  $\theta$ , the prefix output now represents the total wedge count of vertices with support  $\leq \theta$  (line 19). The upper bound  $\theta(i+1)$  is then chosen such that the prefix output for  $\theta(i+1)$  is close to (but no less than) the average wedges per subset (line 20).

After finding the range, RECEIPT CD iteratively peels the vertices and adds them to subset  $U_i$  (lines 9-14). In each iteration, it peels *activeSet* – the set of all vertices with support in the entire range  $R_i$ . This is unlike bottom-up peeling where vertices with a single (minimum) support value are peeled in an iteration. Thus, RECEIPT CD enjoys higher workload per iteration which enables efficient parallelization. For the first peeling iteration of every subset  $U_i$ , RECEIPT CD scans all vertices in  $U$  to initialize *activeSet* (line 9). In subsequent iterations, *activeSet* is constructed by tracking only those vertices whose support has been updated. For correctness of parallel algorithm, the update routine used in RECEIPT CD (line 13) uses atomic operations to decrease vertex supports.

Apart from tip number ranges and partitions of  $U$ , RECEIPT CD also outputs an array  $\triangleright_U^{init}$ , which is used to initialize support in RECEIPT FD (sec.3). Before peeling for a subset begins, it copies the current support of remaining vertices into  $\triangleright_U^{init}$  (line 7). Thus, for a vertex  $u \in U_i$ ,  $\triangleright_u^{init}$  indicates the support of  $u$  (a) after all vertices in  $U_{i-1}$  are peeled, and (b) before any vertex in  $U_i$  is peeled.

**3.1.1 Adaptive Range Determination.** Since range determination uses current support of vertices as a proxy for tip numbers, *tgt* – the target wedge count for a subset  $U_i$ , is covered by the vertices added to  $U_i$  in the very first peeling iteration. Hence,  $\sum_{u \in U_i} w[u] \geq \text{tgt}$ , where  $w[u]$  is the wedge count of  $u$  in  $G$ . Note that as the support of other vertices decreases after this iteration, more vertices may get added to  $U_i$  and total wedge count of vertices in the final subset

---

### Algorithm 3 Coarse-grained Decomposition (RECEIPT CD)

---

**Input:** Bipartite graph  $G(U, V, E)$ , # partitions  $P$   
**Output:** Ranges  $\{\theta(1), \theta(2) \dots \theta(P+1)\}$ , Vertex Subsets  $\{U_1, U_2 \dots U_P\}$ , Support initialization vector  $\triangleright_U^{init}$   
 $w[u] \leftarrow$  number of wedges in  $G$  with endpoint  $u$

- 1:  $U_i \leftarrow \{\phi\} \forall i \in \{1, 2 \dots P\}$
- 2:  $\{\triangleright_U, \triangleright_V\} \leftarrow \text{pvBcnt}(G)$
- 3:  $\theta(1) \leftarrow 0, i \leftarrow 1$
- 4:  $\text{tgt} \leftarrow \frac{\sum_{u \in U} w[u]}{P}$   $\triangleright$  average wedges per subset
- 5: **while**  $U \neq \{\phi\}$  **and**  $i \leq P$  **do**
- 6:   **for each**  $u \in U$  **do in parallel**  $\triangleright$  Support Init
- 7:      $\triangleright_u^{init} \leftarrow \triangleright_u$
- 8:    $\theta(i+1) \leftarrow \text{findHi}(U, \triangleright_U, w, \text{tgt})$   $\triangleright$  Upper Bound
- 9:    $\text{activeSet} \leftarrow$  vertices with support in  $[\theta(i), \theta(i+1))$
- 10:   **while**  $\text{activeSet} \neq \{\phi\}$  **do**  $\triangleright$  Peel Range
- 11:      $U_i \leftarrow U_i \cup \text{activeSet}, U \leftarrow U \setminus \text{activeSet}$
- 12:     **for each**  $u \in \text{activeSet}$  **do in parallel**
- 13:        $\text{update}(u, \theta(i), \triangleright_U, G)$   $\triangleright$  Ref: alg.2
- 14:      $\text{activeSet} \leftarrow$  vertices with support in  $[\theta(i), \theta(i+1))$
- 15:    $i \leftarrow i + 1$
- 16: **function**  $\text{findHi}(U, \triangleright_U, w, \text{tgt})$
- 17:   Initialize hashmap  $\text{work}$  to all zeros
- 18:   **for each**  $\triangleright \in \triangleright_U$
- 19:      $\text{work}[\triangleright] \leftarrow \sum_{u \in U} (w[u] \cdot \mathbb{1}(\triangleright_u \leq \triangleright))$
- 20:    $\theta \leftarrow \arg \min(\triangleright)$  such that  $\text{work}[\triangleright] \geq \text{tgt}$
- 21:   return  $\theta + 1$

---

$U_i$  can be significantly higher than  $tgt$ . This could result in some subsets having very high workload. Moreover, it is possible that  $U$  gets completely deleted in  $\ll P$  subsets, thus restricting the parallelism available during RECEIPT FD. To avoid this scenario, we implement two-way adaptive range determination:

- (1) Instead of statically computing an average target  $tgt$ , we dynamically update  $tgt$  for every subset based on the wedge count of remaining vertices in  $U$  and the remaining number of subsets to create. *If some subsets cover a large number of wedges, the target for future subsets is automatically reduced, thereby preventing a situation where all vertices get peeled in much less than  $P$  subsets.*
- (2) A subset  $U_i$  can cover significantly larger number of wedges than the target  $tgt$ . RECEIPT CD assumes predictive local behavior i.e. subset  $U_{i+1}$  will exhibit similar behavior to  $U_i$ . Therefore, to balance the wedge counts of subsets, RECEIPT dynamically scales the target wedge count for  $U_{i+1}$  with a scaling factor  $s_i = \frac{tgt}{\sum_{u \in U_i} w[u]} \leq 1$ . Note that  $s_i$  quantifies the overshooting of target wedges in  $U_i$ .

After  $P$  partitions, if some vertices still remain in  $U$ , RECEIPT CD puts all of them in a single subset  $U_{P+1}$  and increments  $P$ .

### 3.2 Fine-grained Decomposition

Alg.4 presents the pseudocode for Fine-grained Decomposition (RECEIPT FD), which takes as input the vertex subsets and the tip number ranges created by RECEIPT CD, and computes the exact tip numbers. It creates a task queue of subset IDs from which threads are exclusively allocated vertex subsets to peel (line 4). Before peeling  $U_i$ , a thread initializes the support of vertices in  $U_i$  from the  $\succ_u^{init}$  vector and induces a subgraph  $G_i$  on  $W_i = (U_i, V)$  (lines 5-6). Thereafter, sequential bottom-up peeling is applied on  $G_i$  for tip decomposition of  $U_i$  (lines 7-10).

---

#### Algorithm 4 Fine-grained Decomposition (RECEIPT FD)

---

**Input:** Bipartite graph  $G(U, V, E)$ , # partitions  $P$ , Vertex Subsets  $\{U_1, U_2 \dots U_P\}$ , Support initialization vector  $\succ_u^{init}$ , # threads  $T$

**Output:** Tip number  $\theta_u$  for each  $u \in U$

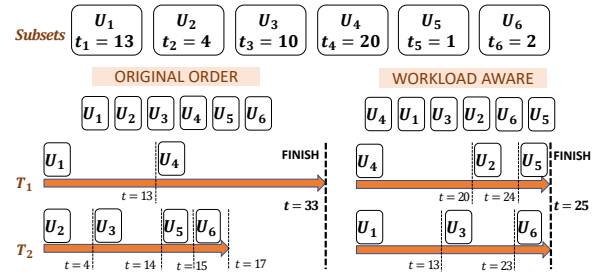
- 1: Insert the integers  $i \in \{1, 2 \dots P\}$  in a queue  $Q$
- 2: **for**  $thread\_id = 1, 2 \dots T$  **do in parallel**
- 3:     **while**  $Q$  is not empty **do**
- 4:         Atomically pop  $i$  from  $Q$
- 5:          $G_i \leftarrow$  subgraph induced by  $W_i = (U_i, V)$
- 6:          $\succ_{U_i} \leftarrow \succ_{U_i}^{init}$   $\triangleright$  Initialize Support
- 7:         **while**  $U_i \neq \{\phi\}$  **do**  $\triangleright$  Peel
- 8:              $u \leftarrow \arg \min_{u \in U_i} \{\succ_u\}$
- 9:              $\theta_u \leftarrow \succ_u$ ,  $U_i \leftarrow U_i \setminus \{u\}$
- 10:            update( $u, \theta_u, \succ_{U_i}, G_i$ )

---

**3.2.1 Parallelization Strategy.** While adaptive range determination(sec.3.1.1) tries to create subsets with uniform wedges in  $G$ , the actual work per subset in FD depends on the wedges in induced subgraphs  $G_i(U_i, V, E_i)$  that can be non-uniform. Therefore, to improve load balance across threads, we use parallelization strategies inspired from Longest Processing Time scheduling rule which is a known  $\frac{4}{3}$ -approximation algorithm [22]. However, exact processing

time for peeling an induced subgraph  $G_i$  is unknown. Instead, we use the number of wedges with endpoints in  $U_i$  as a proxy along with runtime task scheduling as given below: :

- *Dynamic task allocation*  $\rightarrow$  Threads atomically pop unique subset IDs from the task queue when they become idle during runtime (line 4). Thus, all threads are busy until every subset is scheduled.
- *Workload-aware Scheduling*  $\rightarrow$  We sort the subset IDs in task queue in decreasing order of their wedge counts. Thus, the subsets with highest workload (wedges) get scheduled first and the threads processing them naturally receive fewer tasks in the future. Fig.3 shows how workload-aware scheduling can tremendously improve the efficiency of dynamic allocation.



**Figure 3: Benefits of Workload-aware Scheduling (WaS) in a 2-thread ( $T_1$  and  $T_2$ ) system. Top row shows vertex subsets with time required to peel them. Dynamic allocation without WaS finishes in 33 units of time compared to 25 units with WaS.**

### 3.3 Analysis

In this section, we will prove the correctness of tip numbers computed by RECEIPT. We will also analyze its computational complexity and show that RECEIPT is work-efficient. We will exploit comparisons with sequential BUP (alg.2) and hence, first establish the following lemma:

**LEMMA 1.** *In BUP, the support  $\succ_u$  of a vertex  $u$  at any time  $t$  before the first vertex with tip number  $\theta_u$  is peeled, depends on the cumulative effect of all vertices peeled till  $t$  and is independent of the order in which they are peeled.*

**PROOF.** Let  $\succ_u^G$  be the initial support of  $u$  (butterflies of  $u$  in  $G$ ),  $S$  be the set of vertices peeled till  $t$  and  $u' \in S$  be the most recently peeled vertex. Since BUP assigns tip numbers in a non-decreasing order, no vertex with tip number  $\geq \theta_u$  would be peeled till  $t$ . Hence,  $\theta_{u'} < \theta_u \leq \succ_u$ . Since  $\theta_{u'}$  was the minimum support in the latest peeling iteration,  $\succ_u = \succ_u^G + \sum_{v \in S} (-\succ_{v,u})$ . By commutativity of addition, this term is independent of the order in which  $-\succ_{v,u}$  are added i.e. the order in which vertices in  $S$  are peeled.  $\square$

Lemma 1 highlights that whether BUP peels a set of vertices  $S \subseteq U$  in its original order or in the same order as RECEIPT CD, the support of vertices with tip numbers higher than that of  $S$  would be the same. Next we show that parallel processing does not affect the correctness of support updates.

**LEMMA 2.** *Given a set  $S$  of vertices to be peeled in an iteration, the parallel peeling in RECEIPT CD (line 12-13, alg.3) correctly updates the support of vertices.*

PROOF. Let  $S$  be peeled in  $j^{\text{th}}$  iteration which is a part of peeling iterations for subset  $U_i$ . Parallel updates are correct if for any vertex  $u$  not yet assigned to any subset, support of  $u$  decreases by exactly  $\sum_{u' \in S} \varkappa_{u',u}$  in the  $j^{\text{th}}$  iteration. Since at most two vertices of a butterfly can be in  $S$  (two vertices of a butterfly are in  $V$ ), for any vertex pair  $(u_1, u_2)$ , either of the following is true in  $j^{\text{th}}$  iteration:

- *No vertex in the pair is peeled* – no updates are propagated from  $u_1$  to  $u_2$  and vice-versa.
- *Exactly one vertex in the pair is peeled* – without loss of generality, let  $u_1 \in S$ . Peeling  $u_1$  deletes exactly  $\varkappa_{u_1, u_2}$  unique butterflies incident on  $u_2$  because they share  $\varkappa_{u_1, u_2}$  butterflies and no vertex in  $U \setminus \{u_1, u_2\}$  (and hence in  $S$ ) participates in these butterflies. The update routine called for  $u_1$  (line 13, alg.3) also decreases support of  $u_2$  by exactly  $\varkappa_{u_1, u_2}$ , until  $\varkappa_{u_2} > \theta(i)$ . Since atomics are used to apply support updates, concurrent updates to same vertex do not conflict (sec.3.1). Thus, update routine calls for all vertices in  $S$  cumulatively decrease  $\varkappa_{u_2}$  by exactly  $\sum_{u' \in S} \varkappa_{u', u_2}$ , as long as  $\varkappa_{u_2} > \theta(i)$ . If  $\varkappa_{u_2} \leq \theta(i) < \theta(i+1)$ , vertex subset for  $u_2$  is already determined and any updates to  $\varkappa_{u_2}$  have no impact.
- *Both  $u_1$  and  $u_2$  are peeled* – vertex subset for  $u_1$  and  $u_2$  is determined. Any update to  $\varkappa_{u_1}$  or  $\varkappa_{u_2}$  has no effect.  $\square$

Now, we will show that RECEIPT CD correctly computes the tip number range for every vertex. Finally, we will show that RECEIPT accurately computes the exact tip numbers for all vertices. For clarity of explanation, we use  $\varkappa_u(j)$  to denote the support of a vertex  $u$  after  $j^{\text{th}}$  peeling iteration in RECEIPT CD.

LEMMA 3. *There cannot exist a vertex  $u$  such that  $u \in U_i$  and  $\theta_u \geq \theta(i+1)$ .*

PROOF. Let  $j$  be the first iteration in RECEIPT CD that wrongly peels a set of vertices  $S_w$ , and assigns them to subset  $U_i$  even though  $\theta_u \geq \theta(i+1) \forall u \in S_w$ . Let  $S_{hi} \supseteq S_w$  be the set of all vertices with tip numbers  $\geq \theta(i+1)$  and  $S_r$  be the set of vertices peeled up to iteration  $j-1$ . Since all vertices till  $j-1$  iterations have been correctly peeled,  $\theta_u < \theta(i+1) \forall u \in S_r$ . Hence,  $S_{hi} \subseteq U \setminus S_r$ .

Consider a vertex  $u \in S_w$ . Since  $u$  is peeled in iteration  $j$ ,  $\varkappa_u(j-1) < \theta(i+1)$ . From lemma 2,  $\varkappa_u(j-1)$  correctly represents the number of butterflies shared between  $u$  and  $U \setminus S_r$  (vertices remaining after  $j-1$  iterations). Since  $S_{hi} \subseteq U \setminus S_r$ ,  $u$  participates in at most  $\varkappa_u(j-1)$  butterflies with vertices in  $S_{hi}$ . By definition of tip-number (sec.2.2),  $\theta_u \leq \varkappa_u(j-1) < \theta(i+1)$ , which is a contradiction. Thus, no such  $u$  exists,  $S_w = \{\phi\}$  and all vertices in  $U_i$  have tip numbers less than  $\theta(i+1)$ .  $\square$

LEMMA 4. *There cannot exist a vertex  $u$  such that  $u \in U_i$  and  $\theta_u < \theta(i)$ .*

PROOF. Let  $i$  be the smallest integer for which there exists a set  $S_w \neq \{\phi\}$  such that  $\theta(i) \leq \theta_u < \theta(i+1) \forall u \in S_w$ , but  $u \in U_p$ , where  $p > i$ . Let  $j$  be the last iteration that peels vertices in  $U_i$ . Clearly,  $\varkappa_u(j) \geq \theta(i+1) \forall u \in S_w$  otherwise  $u$  would be peeled in or before iteration  $j$ , and will not be added to  $U_p$ .

From lemma 2,  $\varkappa_u(j)$  correctly represents the butterfly count of  $u$  after vertices in  $U_1 \cup U_2 \dots \cup U_i$  are deleted. In other words, every vertex in  $S_w$  participates in at least  $\theta(i+1)$  butterflies with vertices

in  $U_{i+1} \cup U_{i+2} \dots \cup U_p$  and hence, is a part of  $\theta(i+1)$ -tip (def.1). Therefore, by the definition of tip number,  $\theta_u \geq \theta(i+1) \forall u \in S_w$  which is a contradiction.  $\square$

THEOREM 1. *RECEIPT CD (alg.3) correctly computes the vertex-subsets corresponding to every tip number range.*

PROOF. Follows directly from lemmas 3 and 4.  $\square$

THEOREM 2. *RECEIPT correctly computes the tip numbers for all  $u \in U$ .*

PROOF. Consider an arbitrary vertex  $u \in U_i$ . From theorem 1,  $\theta(i) \leq \theta_u < \theta(i+1)$ . Let  $S = U_1 \cup U_2 \dots U_{i-1}$  denote the set of vertices peeled before  $U_i$  in RECEIPT CD. For all vertices  $u' \in S$ ,  $\theta_{u'} < \theta_u$  and hence,  $S$  will be completely peeled before  $u$  in BUP as well. We now compare peeling  $U_i$  in RECEIPT FD (alg.4) to peeling  $U_i$  in sequential algorithm BUP, and show that the two output identical tip numbers. It is well known that BUP correctly computes tip numbers [50].

Note that initial support of  $u$  in RECEIPT FD i.e.  $\varkappa_u^{\text{init}}$  is the support of  $u$  in RECEIPT CD after  $S$  is peeled (sec.3.1, lines 6-7 of alg.3). From lemmas 1 and 2, this is equal to the support of  $u$  in BUP after  $S$  is peeled. Further, by theorem 1, any vertex in  $U_{i+1} \cup U_{i+2} \dots \cup U_p$  has tip number strictly greater than all vertices in  $U_i$ , and will be peeled after  $U_i$  in BUP. Next, we note that subgraph  $G_i$  is induced on subset  $U_i$  and entire set  $V$ . Thus, every butterfly shared between any two vertices in  $U_i$  is present in  $G_i$ . Therefore, for any vertex  $u' \in U_i$  peeled before  $u$ , the update  $\varkappa_{u',u}$  computed by BUP and RECEIPT FD will be same. Hence, BUP and sequential peeling of  $U_i$  in RECEIPT FD apply the same support updates to vertices in  $U_i$  and therefore, follow the same order of vertex peeling. Thus, the final tip number  $\theta_u$  computed by RECEIPT FD will be the same as that computed by BUP.  $\square$

It is important for a parallel algorithm to be not only scalable, but also computationally efficient. The following theorem shows that for a reasonable upper bound<sup>3</sup> on  $P$ , RECEIPT is at least as efficient as the best sequential tip decomposition algorithm BUP.

THEOREM 3. *For  $P = O\left(\frac{\sum_{u \in U} \sum_{v \in N_u} d_v}{n \log n}\right)$  vertex subsets, RECEIPT is work-efficient with computational complexity of  $O\left(\sum_{u \in U} \sum_{v \in N_u} d_v\right)$ .*

PROOF. **RECEIPT CD** – It initializes the vertex support using  $O\left(\sum_{(u,v) \in E} \min(d_u, d_v)\right)$  complexity per-vertex butterfly counting. Range computation for each subset requires constructing a  $O(|U|)$  size hashmap (*work*) whose keys are the unique support values in  $\varkappa_U$ . In this hashmap, wedge counts of all vertices with support  $\varkappa$  are accumulated in value  $\text{work}[\varkappa]$ . Next, *work* is sorted on the keys and a parallel prefix sum is computed over the values so that the final value  $\text{work}[\varkappa]$  represents cumulative wedge count of all vertices with support less than or equal to  $\varkappa$ . Parallel implementations of hashmap generation, sorting and prefix scan perform  $O(|U| \log |U|) = O(n \log n)$  work. Computing scaling factor  $s_i$  for adaptive range determination requires aggregating wedge count of vertices in  $U_i$ , contributing  $O(|U|) = O(n)$  work over all subsets.

<sup>3</sup>In practice, we use  $P \ll \frac{\sum_{u \in U} \sum_{v \in N_u} d_v}{n \log n}$  for large graphs.

Constructing *activeSet* for first peeling iteration of each subset requires an  $O(n)$  complexity parallel filtering on  $\succ_U$ . Subsequent iterations construct *activeSet* by tracking support updates doing  $O(1)$  work per update. For every vertex  $u \in U$ , the update routine is called once when  $u$  is peeled, and it traverses at most  $\sum_{v \in N_u} d_v$  wedges. At most one support update is generated per wedge, resulting in  $O(1)$  work per wedge (RECEIPT CD stores vertex supports in an array  $\succ_U$ ). Thus, the complexity of RECEIPT CD is  $O\left(\sum_{u \in U} \sum_{v \in N_u} d_v + Pn \log n\right)$ .

**RECEIPT FD** – For every subset  $U_i$ , alg.4 creates an induced subgraph  $G_i(U_i, V_i, E_i)$  in parallel. This requires  $O(n + |E_i|)$  work for subset  $U_i$  and  $O(Pn + m)$  work over all  $P$  subsets. When  $u \in U_i$  is peeled, the corresponding call to update explores wedges in the induced subgraph  $G_i$  and generates support updates to other vertices in  $U_i$ . These are a subset of the wedges and support updates generated when  $u$  was peeled in RECEIPT CD. A fibonacci heap can be used to extract minimum support vertex ( $O(\log n)$  work per vertex), and enable constant complexity updates ( $O(\sum_{v \in N_u} d_v)$  work in update call for  $u$ ). Thus, the complexity of RECEIPT FD is  $O\left(\sum_{u \in U} \sum_{v \in N_u} d_v + Pn + n \log n\right)$ .

Combining both the steps, the total work done by RECEIPT is  $O\left(\sum_{u \in U} \sum_{v \in N_u} d_v + Pn \log n\right) = O\left(\sum_{u \in U} \sum_{v \in N_u} d_v\right)$ , if  $P = O\left(\frac{\sum_{u \in U} \sum_{v \in N_u} d_v}{n \log n}\right)$ . This is the same as sequential BUP and hence, RECEIPT is work-efficient.  $\square$

## 4 OPTIMIZATIONS

Despite the parallelism potential, RECEIPT may take hours or days to process large graphs such as *TrU*, that contain hundreds of trillions of wedges (sec.1). To this purpose, we develop novel optimizations that *exploit the properties of RECEIPT to dramatically improve its computational efficiency* in practice, making it feasible to decompose graphs like *TrU* in minutes (objective 2, sec.2.2.1).

### 4.1 Hybrid Update Computation (HUC)

The complexity of RECEIPT is dominated by wedge traversal done during peeling in RECEIPT CD. In order to reduce this traversal, we exploit the following insights about the behavior of counting and peeling algorithms:

- Butterfly counting is computationally efficient (low complexity) and easily parallelizable (sec.2.1).
- Some peeling iterations in RECEIPT CD may peel a large number of vertices.

Given a vertex set (*activeSet*) to peel, we compute the cost of peeling  $C_{peel}$  as  $\sum_{u \in activeSet} \sum_{v \in N_u} d_v$ , which is the total wedge count of vertices in *activeSet*. However, the cost of re-counting butterflies  $C_{rcnt}$ , is computed as  $\sum_{(u,v) \in E} \min(d_u, d_v)$  which represents the bound on wedge traversal of counting (sec.2). Thus, if  $C_{peel}$  exceeds  $C_{rcnt}$ , we *re-compute butterflies* for all remaining vertices in  $U$  instead of computing support updates obtained by peeling *activeSet*. This optimization is denoted Hybrid Update Computation (HUC).

We note that compared to RECEIPT CD, HUC is relatively less beneficial for RECEIPT FD because: (a) the induced subgraphs have

significantly less wedges than the original graph, and (b) few vertices are typically deleted per peeling iteration in RECEIPT FD. Further, re-counting for subset  $U_i$  in RECEIPT FD must account for butterflies with vertices in  $\cup_{j>i}\{U_j\}$ . This external contribution for a vertex  $u$  can be computed by deducting the butterfly count of  $u$  within  $G_i$ , from  $\succ_u^{init}$ .

## 4.2 Dynamic Graph Maintenance (DGM)

We note that after a vertex  $u$  is peeled in RECEIPT CD or RECEIPT FD, it is excluded from future computation in the respective step. However, the graph data structure (adjacency list/Compressed Sparse Row) still contains edges to  $u$  interleaved with other edges. Consequently, wedges incident on  $u$  are still explored after  $u$  is peeled. To prevent such wasteful exploration, we *periodically update the data structures* to remove edges incident on peeled vertices. We denote this optimization as Dynamic Graph Maintenance (DGM).

The cost of DGM is determined by the work done in parallel compaction of all adjacency lists, which grows linearly with the number of edges in the graph. Therefore, if the adjacency lists are updated only after  $m$  wedges have been traversed since previous update, DGM will not alter the theoretical complexity of RECEIPT and pose negligible practical overhead.

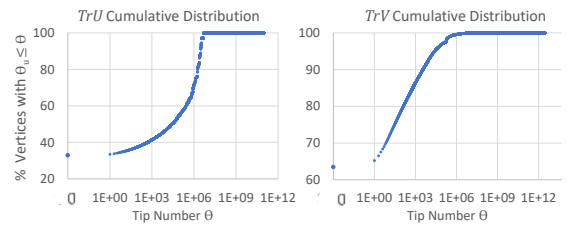
## 5 EXPERIMENTS

### 5.1 Setup

We conduct the experiments on a 36 core dual-socket linux server with two Intel Xeon E5-2695 v4 processors@ 2.1GHz and 1TB DRAM. All algorithms are implemented in C++-14 and are compiled using G++ 9.1.0 with the -O3 optimization flag. We use OpenMP v4.5 for multithreading.

**Baselines:** We compare RECEIPT against the sequential BUP algorithm (alg.2) and its parallel variant ParB [54]. ParB resembles PARBUTTERFLY with BATCH mode peeling<sup>4</sup>. ParB uses the bucketing structure of Julienne [13] with 128 buckets as given in [54].

**Datasets:** We use six unweighted bipartite graphs obtained from the KOBLENZ collection [28]. To the best of our knowledge, these are some of the largest publicly available bipartite datasets. Within each graph, number of wedges with endpoints in one vertex set can be significantly different than the other, as can be seen from  $\wedge^{BUP}$  in table 3. We label the vertex set with higher number of wedges as  $U$  and the other as  $V$ , and accordingly suffix "U" or "V" to the dataset name to identify which vertex set is decomposed.



**Figure 4: Tip number distribution in Trackers graph – y-axis shows percentage vertices with tip number less than abscissa.**

<sup>4</sup>We were unable to verify the correctness of tip numbers generated by public code for PARBUTTERFLY and hence, implemented it ourselves for comparison.



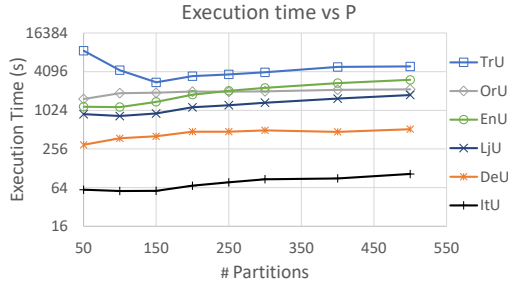
**Table 2: Bipartite Datasets for evaluation with the corresponding number of butterflies ( $\bowtie_G$ ) and wedges ( $\wedge_G$ ) in billions, and maximum tip numbers for  $U$  ( $\theta_U^{\max}$ ) and  $V$  ( $\theta_V^{\max}$ ).  $d_U$  and  $d_V$  denote the average degree of vertices in  $U$  and  $V$ , respectively.**

Dataset	Description	U	V	E	$d_U / d_V$	$\bowtie_G$ (in B)	$\wedge_G$ (in B)	$\theta_U^{\max}$	$\theta_V^{\max}$
ItU, ItV	Pages and editors from Italian Wikipedia	2,255,875	137,693	12,644,802	5.6 / 91.8	298	361	1,555,462	5,328,302,365
DeU, DeV	Users and tags from www.delicious.com	4,512,099	833,081	81,989,133	18.2 / 98.4	26,683	1,446	936,468,800	91,968,444,615
OrU, OrV	Users' group memberships in Orkut	2,783,196	8,730,857	327,037,487	117.5 / 37.5	22,131	2,528	88,812,453	29,285,249,823
LjU, LjV	Users' group memberships in Livejournal	3,201,203	7,489,073	112,307,385	35.1 / 15	3,297	2,703	4,670,317	82,785,273,931
EnU, EnV	Pages and editors from English Wikipedia	21,504,191	3,819,691	122,075,170	5.7 / 32	2,036	6,299	37,217,466	96,241,348,356
TrU, TrV	Internet domains and trackers in them	27,665,730	12,756,244	140,613,762	5.1 / 11	20,068	106,441	18,667,660,476	3,030,765,085,153

Note that the maximum tip numbers are extremely high because of few high degree vertices sharing a large number of common neighbors. However, most of the vertex' tip numbers lie in a much smaller range as shown in fig.4. For example, 99.98% vertices in  $TrU$  have tip number  $< 5M$  (0.027% of maximum).

**Implementation Details:** Unless otherwise mentioned, the parallel algorithms use all 36 threads for execution<sup>5</sup> and RECEIPT includes all workload optimizations discussed in sec.4. In RECEIPT FD and sequential BUP, we use a  $k$ -way min-heap for efficient retrieval of minimum support vertices. We found it to be faster in practice than the bucketing structure of [51] or fibonacci heaps.

To analyze the impact of number of vertex subsets on RECEIPT's performance, we vary  $P$  from 50 to 500. Fig.5 reports the effect of  $P$  on some large datasets that showed significant performance variation. Typically, RECEIPT CD dominates the overall execution time because of the larger number of wedges traversed compared to RECEIPT FD. The performance of RECEIPT CD improves with a decrease in  $P$  because of reduced synchronizations. However, a small value of  $P$  reduces parallelism in RECEIPT FD and makes the induced subgraphs larger. Consequently, for  $P \leq 100$ , we observed that RECEIPT FD became the bottleneck for some large graphs such as  $TrU$  and  $LjU$ . For all the datasets shown in fig.5, execution slows down when  $P$  is increased beyond 150. Based on these observations, we use  $P = 150$  for all datasets, in rest of the experiments.



**Figure 5: Execution time of RECEIPT vs  $P$ .**

## 5.2 Results

**5.2.1 Comparison with Baselines.** Table 3 shows a detailed comparison of various tip decomposition algorithms. Note that  $\rho$  represents the number of synchronization rounds. Each round consists of one iteration of peeling all vertices with minimum support (or support in minimum range for RECEIPT)<sup>6</sup>. Each round involves multiple

<sup>5</sup>We did not observe any benefits from enabling the 2-way hyperthreading.

<sup>6</sup>Wedge traversal by BUP can be computed without executing alg.2, by simply aggregating the 2-hop neighborhood size of vertices in  $U$  or  $V$ . A given wedge can be traversed

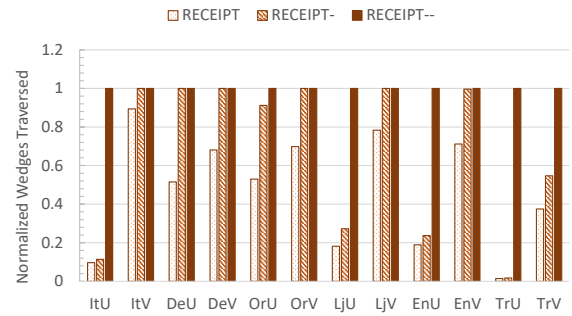
(but constant) thread synchronizations, for example, once after computing the list of vertices to peel, once after computing the updates etc. RECEIPT FD does not incur any synchronization as the threads in alg.4 only synchronize once at the end of computation. For the rest of this section, we will use the term *large datasets* to refer to a graph having large number of wedges with endpoints on the vertex set being peeled such as  $ItU$ ,  $OrU$  etc.

**Execution Time ( $t$ ):** With up to **80.8 $\times$**  and **64.7 $\times$**  speedup over BUP and ParB, respectively, RECEIPT is radically faster than both baselines, for *all* datasets. The speedups are typically higher for large datasets that offer large amount of computation to parallelize and bigger benefits from workload optimizations(sec.4). Only RECEIPT is able to successfully tip decompose  $TrU$ . Contrarily, ParB achieves a maximum 1.6 $\times$  speedup compared to sequential BUP for  $TrV$ .

**Wedges Traversed ( $\wedge$ ):** For *all* datasets, RECEIPT traverses fewer wedges than the existing baselines. RECEIPT's built-in optimizations achieve up to **64 $\times$**  reduction in wedges traversed. This enables RECEIPT to decompose large datasets  $EnU$  and  $TrU$  in few minutes, unlike baselines that take few days for the same.

**Synchronization ( $\rho$ ):** In comparison with ParB, RECEIPT reduces the amount of thread synchronization by up to **1105 $\times$** . This is primarily because RECEIPT CD peels vertices with a broad range of support in every iteration. This drastic reduction in  $\rho$  is the primary contributor to RECEIPT's parallel efficiency.

**5.2.2 Effect of Workload Optimizations.** Fig.6 and 7 show the effect of HUC and DGM optimizations (sec.4.1 and 4.2) on the performance of RECEIPT. The execution time closely follows the trend in wedge workload.

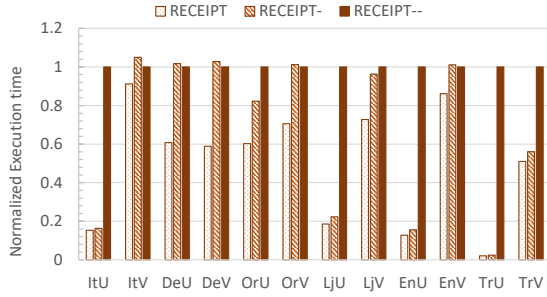


**Figure 6: Effect of workload optimizations on wedge traversal (normalized with respect to RECEIPT--). RECEIPT- is RECEIPT without DGM. RECEIPT-- is RECEIPT without DGM and HUC.**

twice. Similarly,  $\rho$  for ParB can be computed from a modified version of RECEIPT FD where we retrieve all vertices with minimum support in a single iteration.

**Table 3: Comparing execution time ( $t$ ), # wedges traversed ( $\wedge$ ) and # synchronization rounds ( $\rho$ ) of RECEIPT and baseline algorithms. ParB traverses the same # wedges as BUP and has missing entries due to out-of-memory error. pvBcnt denotes per-vertex butterfly counting and  $t = \infty$  denotes execution did not finish in 10 days.**

		ItU	ItV	DeU	DeV	OrU	OrV	LjU	LjV	EnU	EnV	TrU	TrV
$t$ (s)	pvBcnt	0.3	0.3	8.3	8.3	45.6	45.6	5.1	5.1	6.9	6.9	7.8	7.8
	BUP	3,849	8.4	12,260	428	39,079	2,297	67,588	200	111,777	281	$\infty$	5,711
	ParB	3,677	8.1	-	377.7	-	1,510	-	132.5	-	198	-	3,524
	RECEIPT	<b>56.8</b>	<b>3.1</b>	<b>402.4</b>	<b>32.4</b>	<b>1,865</b>	<b>136</b>	<b>911.1</b>	<b>23.7</b>	<b>1,383</b>	<b>31.1</b>	<b>2,784</b>	<b>530.6</b>
$\wedge$ (billions)	pvBcnt	0.19	0.19	20.3	20.3	74.8	74.8	4.7	4.7	6.3	6.3	6.3	6.3
	BUP	723	0.57	2,861	70.1	4,975	231.4	5,403	14.3	12,583	29.6	211,156	1,740
	RECEIPT	<b>71</b>	<b>0.56</b>	<b>1,503</b>	<b>51.3</b>	<b>2,728</b>	<b>170.4</b>	<b>1,003</b>	<b>11.7</b>	<b>2,414</b>	<b>22.2</b>	<b>3,298</b>	<b>658.1</b>
$\rho$	ParB	377,904	10,054	670,189	127,328	1,136,129	334,064	1,479,495	83,423	1,512,922	83,800	1,476,015	342,672
	RECEIPT	<b>967</b>	<b>280</b>	<b>1113</b>	<b>406</b>	<b>1,160</b>	<b>639</b>	<b>1,477</b>	<b>456</b>	<b>1,724</b>	<b>453</b>	<b>1,335</b>	<b>1,381</b>

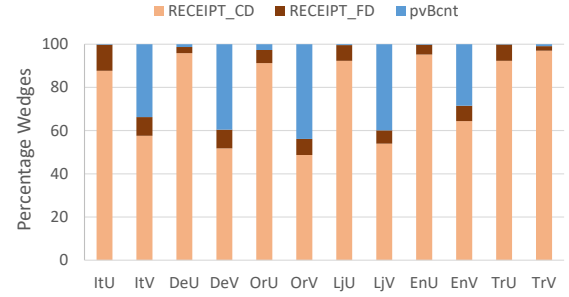


**Figure 7: Effect of workload optimizations on execution time (normalized with respect to RECEIPT-). RECEIPT- is RECEIPT without DGM. RECEIPT-- is RECEIPT without DGM and HUC.**

HUC reduces wedge traversal by opting to re-count butterflies instead of peeling, in selective iterations. A comparison of the work required for peeling vertices in  $U$  versus counting all butterflies can be an indicator of the benefits of HUC. Therefore, we define a ratio  $r = \frac{\wedge^{peel}}{\wedge^{cnt}}$ , where  $\wedge^{peel}$  and  $\wedge^{cnt}$  denote the number of wedges traversed for peeling all vertices in  $U$  ( $\wedge^{BUP} - \wedge^{pvBcnt}$  in table 3) and for butterfly counting ( $\wedge^{pvBcnt}$  in table 3), respectively. Datasets with large value of  $r$  (for example  $ItU$ ,  $LjU$ ,  $EnU$  and  $TrU$  with  $r > 1000$ ) benefit heavily from HUC optimization, since peeling in high workload iterations is replaced by re-counting, which dramatically reduces wedge traversal. Especially for  $TrU$  ( $r > 33,500$ ), HUC enables  $57\times$  and  $42\times$  reduction in wedge traversal and execution time, respectively. Contrarily, in datasets with small value of  $r$  ( $ItV$ ,  $DeV$ ,  $OrV$ ,  $LjV$  and  $EnV$  with  $r < 5$  due to low  $\wedge^{peel}$ ), none of the iterations utilize re-counting. Consequently, performance of RECEIPT- and RECEIPT-- is similar for these datasets.

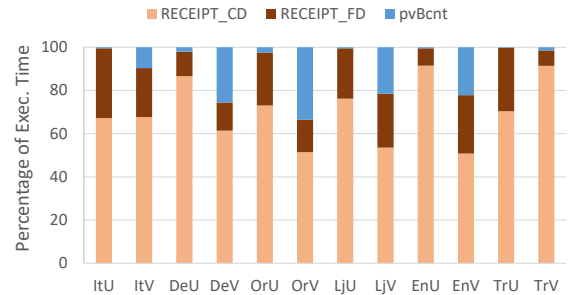
DGM can potentially half the wedge workload since each wedge has two endpoints, but needs to be traversed only by the vertex that gets peeled first. However, the reduction is less than  $2\times$  in practice, because for many wedges, both endpoints get peeled between consecutive DGM data structure updates. It achieves  $1.41\times$  and  $1.29\times$  average reduction in wedges and execution time, respectively.

**5.2.3 RECEIPT Breakup.** In this section, we analyze the work and execution time contribution of each step of RECEIPT individually. We further split RECEIPT CD (alg.2) into a peeling step (denoted as RECEIPT CD), and a per-vertex butterfly counting (pvBcnt) step used for support initialization.



**Figure 8: Breakup of wedges traversed in RECEIPT**

Fig.8 shows a breakdown of the wedges traversed during different steps as a percentage of total wedges traversed by RECEIPT. As discussed in sec.3, RECEIPT CD traverses significantly more wedges than RECEIPT FD. For all the datasets, RECEIPT FD incurs  $< 15\%$  of the total wedge traversal. Note that for a given graph, the number of wedges explored by pvBcnt is independent of the vertex set being peeled. For example,  $ItU$  and  $ItV$  both traverse 188 million wedges during pvBcnt. However, its percentage contribution depends on the total wedges explored during the entire tip decomposition, which can vary significantly between  $U$  and  $V$  vertex sets (table 3).



**Figure 9: Breakup of execution time of RECEIPT**

Fig.9 shows a breakdown of the execution time of different steps as a percentage of the total execution time of RECEIPT. Intuitively, the step with most workload i.e. RECEIPT CD, has the largest contribution ( $> 50\%$  of the total execution time for all datasets). In most datasets with a small value of  $r = \frac{\wedge^{peel}}{\wedge^{cnt}}$  ( $ItV$ ,  $DeV$ ,  $OrV$ ,  $LjV$  and  $EnV$ ), even pvBcnt has a significant share of the execution

time. Note that for some datasets, contribution of RECEIPT FD to the execution time is more than its contribution to the wedge traversal. This is due to the following reasons: (a) RECEIPT FD has computational overheads other than wedge exploration, such as creating induced subgraphs and applying support updates to a heap, and (b) Lower parallel scalability compared to counting and RECEIPT CD (sec.5.2.4). Still, RECEIPT FD consumes < 25% of the overall execution time for almost all datasets.

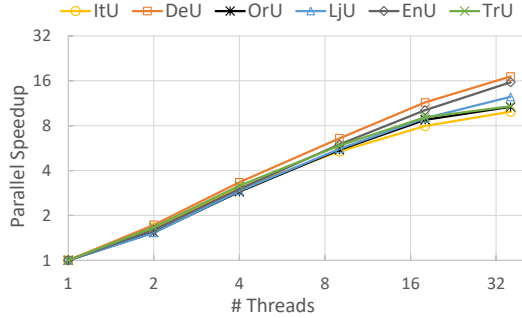


Figure 10: Parallel Speedup of RECEIPT when peeling set  $U$

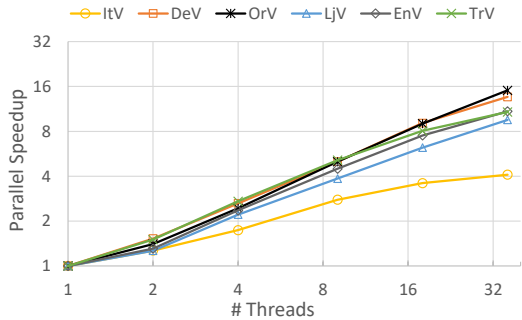


Figure 11: Parallel Speedup of RECEIPT when peeling set  $V$

5.2.4 *Parallel Scalability.* We measure RECEIPT’s performance with 1, 2, 4, 9, 18 and 36 threads. Fig.10 and 11 show the parallel speedup obtained by RECEIPT over single-threaded execution<sup>7</sup>.

For most of the datasets, RECEIPT exhibits almost linear scalability. With 36 threads, RECEIPT achieves up to 17.1× self-relative speedup. In comparison, we observed that ParB exhibits at most 2.3× parallel speedup (self-relative) with 36 threads. RECEIPT’s speedup is poor for *ItV* because it is a very small dataset that gets peeled in < 4s. It requires very little computation (0.56B wedges) and hence, employing a large number of threads is not useful.

Typically, datasets with small amount of wedges (*ItV*, *LjV*, *EnV*) exhibit lower scalability, because compared to larger datasets, they experience lower workload per synchronization round on average. For example, *LjV* traverses 86× fewer wedges than *LjU* but incurs only 3.2× fewer synchronizations. This increases the relative overheads of parallelization and restricts the parallel scalability of

<sup>7</sup>We also developed a sequential version of RECEIPT with no synchronization primitives (atomics) and sequential implementations of basic kernels such as prefix scan, scatter etc. However, the observed performance difference between sequential implementation and single-threaded execution of parallel implementation was negligible.

RECEIPT CD, which is the highest workload step in RECEIPT (fig.8). For example, parallel speedup of RECEIPT CD with 36 threads is 15.1× for *LjU* but only 7.1× for *LjV*.

In RECEIPT FD, parallel speedup is restricted by workload imbalance across the subgraphs. This is because RECEIPT CD tries to balance total wedge counts of vertex subsets as seen in original graph, whereas work done in RECEIPT FD is determined by wedges in induced subgraphs. Consequently, we observed that for some datasets, parallel scalability of RECEIPT FD is poorer than RECEIPT CD. For example, for *TrU* with 36 threads, parallel speedup of RECEIPT FD was only 5.3× compared to 13.1× of RECEIPT CD, 12.5× of counting (pvBcnt) and 10.7× of RECEIPT overall. Note that even sequential RECEIPT is much faster than BUP because of the following:

- (1) *Fewer support updates* – updates to  $\mathfrak{A}_u$  from all vertices in a peeling iteration are aggregated into a single update.
- (2) *Lesser work* – reduced wedge traversal due to HUC and DGM optimizations (sec.4).

We also observe that slope of the speedup curve decreases from 18 to 36 threads. This could potentially be due to the NUMA effects as RECEIPT does not currently have NUMA specific optimizations. Up to 18 threads, the execution is limited to single socket but 36 threads are spawned across two different sockets.

## 6 RELATED WORK

Dense subgraph discovery is a widely used analytic in several applications [9, 14, 15, 20, 44, 56]. Researchers have developed many techniques for mining dense regions in unipartite graphs [4, 16, 21, 27, 31, 36, 49, 52, 60, 62]. Particularly, motif-based techniques have gained tremendous interest [3, 19, 51, 57, 58, 60, 65]. Motifs like triangles represent a quantum of cohesion in graphs. Hence, the number of motifs incident on a vertex or an edge is an indicator of their involvement in dense subgraphs. Several recent works are focused on efficiently finding such motifs in the graphs [1, 18, 25, 35, 54, 63].

Nucleus decomposition is a popular motif-based dense graph mining technique. It considers the distribution of motifs across vertices or edges within a subgraph as an indicator of its density [2], resulting in denser subgraphs compared to counting [2, 53]. For example, truss decomposition mines subgraphs called  $k$ -trusses, where every edge participates in at least  $k - 2$  triangles within the subgraph. Truss decomposition is one of the three tasks in the popular GraphChallenge [46], that has resulted in highly efficient parallel solutions scalable to billion edge graphs [11, 23, 55, 59]. However, such solutions cannot be directly applied on bipartite graphs due to the absence of triangles. Chakravarthy et al.[8] propose a distributed truss decomposition algorithm that trades off computational efficiency to reduce synchronization. This approach requires triangle enumeration and cannot be adopted for tip decomposition due to prohibitive space and computational requirements.

The simplest non-trivial motif in a bipartite graph is a Butterfly (2,2-biclique, quadrangle). Several algorithms covering various aspects of butterfly counting have been developed: in-memory or external memory [61, 63], exact or approximate counting [47, 48] and parallel counting on various platforms [54, 61, 63]. Particularly, the in-memory algorithms for exact counting are relevant to our work. Wang et al.[61] count rectangles in bipartite graphs by traversing wedges with  $O\left(\sum_{u \in W} d_u^2\right)$  complexity. Sanei-Mehri

et al.[47] reduce this complexity to  $O\left(\min \sum_{u \in U} d_u^2, \sum_{v \in V} d_v^2\right)$  by choosing the vertex set where fewer wedges have end points.

Before the aforementioned works, Chiba and Nishizeki [10] had proposed an  $O(\alpha \cdot m)$  complexity vertex-priority based quadrangle counting algorithm for generic graphs. Wang et al.[63] further propose a cache optimized variant of this algorithm and use shared-memory parallelism for acceleration. Independently, Shi et al.[54] develop provably efficient shared-memory parallel implementations of vertex-priority based counting. All of these approaches are amenable for per-vertex or per-edge counting.

Butterfly based decomposition, albeit highly effective in finding quality dense regions in bipartite graphs, is computationally much more expensive than counting. Sariyuce et al.[51] defined  $k$ -tips and  $k$ -wings as subgraphs with minimum  $k$  butterflies incident on every vertex and edge, respectively. Correspondingly, they defined the problems of Tip decomposition and Wing decomposition. Their algorithms use bottom-up peeling with respective complexities of  $O\left(\sum_{v \in V} d_v^2\right)$  and  $O\left(\sum_{(u,v) \in E} \sum_{w \in N_v} (d_u + d_w)\right)$ . Independently, Zou [69] defined the notion of bitruss similar to  $k$ -wing and showed its utility for bipartite network analysis. Shi et al.[54] propose parallel bottom-up peeling used as a baseline in our evaluation. Their wing decomposition uses hash tables to store edges and has a complexity of  $O\left(\sum_{(u,v) \in E} \sum_{w \in N_v} \min(d_u, d_w)\right)$ .

In their seminal paper, Chiba and Nishizeki [10] had remarked that butterflies can be represented by storing the  $O(\alpha \cdot m)$  triples traversed during counting. In a very recent work, Wang et al.[64] store these triples in the form of maximal *blooms* (bicliques), that enables quick retrieval of butterflies shared between edges. Using this structure, authors develop the Bit-BU algorithm for peeling bipartite networks. To the best of our knowledge, it is the most efficient sequential algorithm that can compute wing decomposition in  $O(\bowtie_G)$  time. Yet, it takes more than 15 hours to process the Livejournal (*Lj*) dataset, whereas RECEIPT can tip decompose both *LjU* and *LjV* in less than 16 minutes. Moreover, Bit-BU has a non-trivial  $O(\alpha \cdot m)$  space requirement, which in practice, amounts to hundreds of gigabytes for datasets like Orkut(*Or*). Comparatively, RECEIPT has a space-complexity of  $O(n \cdot T + m)$  (same as parallel counting [63]) and consumes only few gigabytes for all the datasets used in sec.5. Further, the fundamental ideas employed in RECEIPT and Bit-BU are complimentary. While RECEIPT attempts to exploit parallelism across  $k$ -tip hierarchy, Bit-BU tries to make peeling every edge more efficient. We believe that an amalgamation of our ideas with [64] can produce highly scalable parallel solutions for peeling large bipartite graphs.

## 7 EXTENSIONS

In this section, we list some directions for future research in the context of our work:

- *Parallel Edge peeling*: RECEIPT can be adapted for parallel wing decomposition (edge peeling) in bipartite graphs [51, 64]. It’s workload optimizations can have a large impact on edge peeling due to the higher complexity and smaller range of wing numbers. Parallel edge peeling may incur conflicts if multiple edges in a butterfly get deleted in the same iteration. However, such conflicts can be resolved by imposing a priority ordering of edges.

- *Distributed Tip Decomposition*: Distributed-memory systems (like multi-node clusters) offer large amount of extendable computational resources and are widely used to scale high complexity graph analytics [6, 8, 29]. We believe that the fundamental concept of creating independent tip number ranges and vertex subsets can be very useful in exposing parallelism for distributed-memory algorithms. In the past as well, distributed graph processing algorithms have achieved high scalability by creating independent parallel tasks [5, 29].

However, support updates generated from peeling may need to be communicated on the network. This can affect scalability of the algorithm because of high communication cost in clusters [34, 37]. Further, execution on distributed systems may exhibit load imbalance due to large number of threads and limitations of dynamic task scheduling across processes.

- *System Optimizations*: Memory access optimizations [30, 66] and SIMD parallelism [24] have significant impact on graph analytics. Enhancing memory access locality can also mitigate the NUMA effects that limit parallel speedup (sec.5.2.4). RECEIPT CD and RECEIPT FD exploit parallelism at vertex and subgraph granularity, respectively. Using fine-grained parallelism at edge or wedge granularity can further improve load balance.

## 8 CONCLUSION

In this paper, we proposed RECEIPT – a novel shared-memory parallel algorithm for tip decomposition. RECEIPT is the first algorithm that exploits the massive parallelism across different levels of  $k$ -tip hierarchy. Further, we also developed pragmatic optimizations to drastically improve the computational efficiency of RECEIPT.

We empirically evaluated our approach on a shared-memory multicore server, and showed that it can process some of the largest publicly available bipartite datasets orders of magnitude faster than the state-of-the-art algorithms, with dramatic reduction in synchronization and wedge traversal. Using 36 threads, RECEIPT can provide up to 17.1× self-relative speedup, being much more scalable than the best available parallel algorithms for tip decomposition.

We also explored the generalizability of RECEIPT for wing decomposition (edge peeling) and several interesting avenues for future work in this direction. We believe that scalable algorithms for parallel systems such as many-core CPUs, GPU or HPC clusters, can enhance the applicability of tip or wing decomposition, and our work is a step in that direction.

## ACKNOWLEDGMENTS

This work is supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Number FA8750-17-C-0086 and National Science Foundation (NSF) under Contract Number OAC-1911229. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or NSF. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. Cesar A. F. De Rose was supported by CNPq PDE grant and Fapergs PRONEX project.

## REFERENCES

- [1] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield. Efficient graphlet counting for large networks. In *2015 IEEE International Conference on Data Mining*, pages 1–10. IEEE, 2015.
- [2] N. K. Ahmed, J. Neville, R. A. Rossi, N. G. Duffield, and T. L. Willke. Graphlet decomposition: Framework, algorithms, and applications. *Knowledge and Information Systems*, 50(3):689–722, Mar. 2017.
- [3] S. G. Aksoy, T. G. Kolda, and A. Pinar. Measuring and modeling bipartite graphs with community structure. *Journal of Complex Networks*, 5(4):581–603, 2017.
- [4] A. Angel, N. Koudas, N. Sarkas, D. Srivastava, M. Svendsen, and S. Tirthapura. Dense subgraph maintenance under streaming edge weight updates for real-time story identification. *The VLDB journal*, 23(2):175–199, 2014.
- [5] S. Arifuzzaman, M. Khan, and M. Marathe. Patric: A parallel algorithm for counting triangles in massive networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 529–538, 2013.
- [6] B. Bhattarai, H. Liu, and H. H. Huang. Ceci: Compact embedding cluster index for scalable subgraph matching. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1447–1462, 2019.
- [7] F. Bonchi, A. Khan, and L. Severini. Distance-generalized core decomposition. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1006–1023, 2019.
- [8] V. T. Chakaravarthy, A. Goyal, P. Murali, S. S. Pandian, and Y. Sabharwal. Improved distributed algorithm for graph truss decomposition. In *European Conference on Parallel Processing*, pages 703–717. Springer, 2018.
- [9] J. Chen and Y. Saad. Dense subgraph extraction with application to community detection. *IEEE Transactions on knowledge and data engineering*, 24(7):1216–1230, 2010.
- [10] N. Chiba and T. Nishizeki. Arboricity and subgraph listing algorithms. *SIAM Journal on computing*, 14(1):210–223, 1985.
- [11] K. Date, K. Feng, R. Nagi, J. Xiong, N. S. Kim, and W.-M. Hwu. Collaborative (cpu+gpu) algorithms for triangle counting and truss decomposition on the minsky architecture: Static graph challenge: Subgraph isomorphism. In *2017 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7. IEEE, 2017.
- [12] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, 2001.
- [13] L. Dhulipala, G. Blelloch, and J. Shun. Julienne: A framework for parallel graph algorithms using work-efficient bucketing. In *Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures*, pages 293–304, 2017.
- [14] A. Epasto, S. Lattanzi, V. Mirrokni, I. O. Sebe, A. Taei, and S. Verma. Ego-net community mining applied to friend suggestion. *Proceedings of the VLDB Endowment*, 9(4):324–335, 2015.
- [15] Y. Fang, Y. Yang, W. Zhang, X. Lin, and X. Cao. Effective and efficient community search over large heterogeneous information networks. *Proceedings of the VLDB Endowment*, 13(6):854–867, 2020.
- [16] Y. Fang, K. Yu, R. Cheng, L. V. Lakshmanan, and X. Lin. Efficient algorithms for densest subgraph discovery. *Proceedings of the VLDB Endowment*, 12(11):1719–1732, 2019.
- [17] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting burstiness in reviews for review spammer detection. In *Seventh international AAAI conference on weblogs and social media*, 2013.
- [18] J. Fox, O. Green, K. Gabert, X. An, and D. A. Bader. Fast and adaptive list intersections on the gpu. In *2018 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7. IEEE, 2018.
- [19] E. Fratkin, B. T. Naughton, D. L. Brutlag, and S. Batzoglou. Motifcut: regulatory motifs finding with maximum density subgraphs. *Bioinformatics (Oxford, England)*, 22(14):e150–7, July 2006.
- [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*, pages 721–732, 2005.
- [21] 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*, pages 721–732, 2005.
- [22] R. L. Graham. Bounds on multiprocessing timing anomalies. *SIAM journal on Applied Mathematics*, 17(2):416–429, 1969.
- [23] O. Green, J. Fox, E. Kim, F. Busato, N. Bombieri, K. Lakhota, S. Zhou, S. Singapura, H. Zeng, R. Kannan, et al. Quickly finding a truss in a haystack. In *2017 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7. IEEE, 2017.
- [24] S. Han, L. Zou, and J. X. Yu. Speeding up set intersections in graph algorithms using simd instructions. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18*, page 1587–1602, New York, NY, USA, 2018. Association for Computing Machinery.
- [25] Y. Hu, H. Liu, and H. H. Huang. Tricore: Parallel triangle counting on gpus. In *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 171–182. IEEE, 2018.
- [26] Z. Huang, D. D. Zeng, and H. Chen. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Management science*, 53(7):1146–1164, 2007.
- [27] W. Khaouid, M. Barsky, V. Srinivasan, and A. Thomo. K-core decomposition of large networks on a single pc. *Proceedings of the VLDB Endowment*, 9(1):13–23, 2015.
- [28] J. Kunegis. Konect: the koblentz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1343–1350, 2013.
- [29] K. Lakhota, R. Kannan, Q. Dong, and V. Prasanna. Planting trees for scalable and efficient canonical hub labeling. *Proceedings of the VLDB Endowment*, 13(4), 2019.
- [30] K. Lakhota, R. Kannan, and V. Prasanna. Accelerating pagerank using partition-centric processing. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, pages 427–440, 2018.
- [31] V. E. Lee, N. Ruan, R. Jin, and C. Aggarwal. A survey of algorithms for dense subgraph discovery. In *Managing and Mining Graph Data*, pages 303–336. Springer, 2010.
- [32] E. A. Leicht, P. Holme, and M. E. Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.
- [33] W. Li, M. Qiao, L. Qin, Y. Zhang, L. Chang, and X. Lin. Scaling distance labeling on small-world networks. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD '19*, page 1060–1077, New York, NY, USA, 2019. Association for Computing Machinery.
- [34] A. Lumsdaine, D. Gregor, B. Hendrickson, and J. Berry. Challenges in parallel graph processing. *Parallel Processing Letters*, 17(01):5–20, 2007.
- [35] C. Ma, R. Cheng, L. V. Lakshmanan, T. Grubenmann, Y. Fang, and X. Li. Linc: a motif counting algorithm for uncertain graphs. *Proceedings of the VLDB Endowment*, 13(2):155–168, 2019.
- [36] F. D. Malliaros, C. Giatsidis, A. N. Papadopoulos, and M. Vazirgiannis. The core decomposition of networks: Theory, algorithms and applications. *The VLDB Journal*, pages 1–32, 2019.
- [37] F. McSherry, M. Isard, and D. G. Murray. Scalability! but at what {COST}? In *15th Workshop on Hot Topics in Operating Systems (HotOS {XV})*, 2015.
- [38] A. Mislove, M. Marcon, K. P. Gummadri, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC '07)*, San Diego, CA, October 2007.
- [39] A. Mukherjee, B. Liu, and N. Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200, 2012.
- [40] S. Navlakha, R. Rastogi, and N. Shrivastava. Graph summarization with bounded error. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 419–432, 2008.
- [41] M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.
- [42] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64:016131, Jun 2001.
- [43] H.-M. Park, S.-H. Myaeng, and U. Kang. Pte: Enumerating trillion triangles on distributed systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. ACM Press, 2016.
- [44] P. Rozenstein, A. Anagnostopoulos, A. Gionis, and N. Tatti. Event detection in activity networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1176–1185, 2014.
- [45] S. Samsi, V. Gadepally, M. Hurley, M. Jones, E. Kao, S. Mohindra, P. Monticciolo, A. Reuther, S. Smith, W. Song, et al. Static graph challenge: Subgraph isomorphism. In *2017 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2017.
- [46] S. Samsi, V. Gadepally, M. Hurley, M. Jones, E. Kao, S. Mohindra, P. Monticciolo, A. Reuther, S. Smith, W. Song, et al. Static graph challenge: Subgraph isomorphism. In *2017 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2017.
- [47] S.-V. Sane-Mehri, A. E. Sariyuce, and S. Tirthapura. Butterfly counting in bipartite networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2150–2159, 2018.
- [48] S.-V. Sane-Mehri, Y. Zhang, A. E. Sariyuce, and S. Tirthapura. Fleet: Butterfly estimation from a bipartite graph stream. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1201–1210, 2019.
- [49] A. E. Sariyuce and A. Pinar. Fast hierarchy construction for dense subgraphs. *Proceedings of the VLDB Endowment*, 10(3), 2016.
- [50] A. E. Sariyuce and A. Pinar. Peeling bipartite networks for dense subgraph discovery. *arXiv preprint arXiv:1611.02756*, 2016.
- [51] A. E. Sariyuce and A. Pinar. Peeling bipartite networks for dense subgraph discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 504–512, 2018.
- [52] A. E. Sariyuce, C. Seshadhri, and A. Pinar. Local algorithms for hierarchical dense subgraph discovery. *Proceedings of the VLDB Endowment*, 12(1):43–56, 2018.
- [53] A. E. Sariyuce, C. Seshadhri, A. Pinar, and U. V. Catalyurek. Finding the hierarchy of dense subgraphs using nucleus decompositions. In *Proceedings of the 24th International Conference on World Wide Web*, pages 927–937, 2015.

- [54] J. Shi and J. Shun. Parallel algorithms for butterfly computations. *arXiv preprint arXiv:1907.08607*, 2019.
- [55] S. Smith, X. Liu, N. K. Ahmed, A. S. Tom, F. Petrini, and G. Karypis. Truss decomposition on shared-memory parallel systems. In *2017 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2017.
- [56] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.
- [57] C. E. Tsourakakis. A novel approach to finding near-cliques: The triangle-densest subgraph problem. *arXiv preprint arXiv:1405.1477*, 2014.
- [58] C. E. Tsourakakis, J. Pachocki, and M. Mitzenmacher. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1451–1460, 2017.
- [59] C. Voegelé, Y.-S. Lu, S. Pai, and K. Pingali. Parallel triangle counting and k-truss identification using graph-centric methods. In *2017 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7. IEEE, 2017.
- [60] J. Wang and J. Cheng. Truss decomposition in massive networks. *Proceedings of the VLDB Endowment*, 5(9), 2012.
- [61] J. Wang, A. W.-C. Fu, and J. Cheng. Rectangle counting in large bipartite graphs. In *2014 IEEE International Congress on Big Data*, pages 17–24. IEEE, 2014.
- [62] K. Wang, X. Cao, X. Lin, W. Zhang, and L. Qin. Efficient computing of radius-bounded k-cores. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 233–244. IEEE, 2018.
- [63] K. Wang, X. Lin, L. Qin, W. Zhang, and Y. Zhang. Vertex priority based butterfly counting for large-scale bipartite networks. *Proceedings of the VLDB Endowment*, 12(10):1139–1152, 2019.
- [64] K. Wang, X. Lin, L. Qin, W. Zhang, and Y. Zhang. Efficient bitruss decomposition for large-scale bipartite graphs. *arXiv preprint arXiv:2001.06111*, 2020.
- [65] N. Wang, J. Zhang, K.-L. Tan, and A. K. Tung. On triangulation-based dense neighborhood graph discovery. *Proceedings of the VLDB Endowment*, 4(2):58–68, 2010.
- [66] H. Wei, J. X. Yu, C. Lu, and X. Lin. Speedup graph processing by graph ordering. In *Proceedings of the 2016 International Conference on Management of Data*, pages 1813–1828. ACM, 2016.
- [67] D. Wen, L. Qin, Y. Zhang, X. Lin, and J. X. Yu. I/o efficient core graph decomposition: application to degeneracy ordering. *IEEE Transactions on Knowledge and Data Engineering*, 31(1):75–90, 2018.
- [68] Y. Yang, L. Chu, Y. Zhang, Z. Wang, J. Pei, and E. Chen. Mining density contrast subgraphs. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 221–232. IEEE, 2018.
- [69] Z. Zou. Bitruss decomposition of bipartite graphs. In *International Conference on Database Systems for Advanced Applications*, pages 218–233. Springer, 2016.