Distributed Edge Partitioning for Trillion-edge Graphs

Masatoshi Hanai [1*]
mhanai@acm.org

Toyotaro Suzumura [2]
suzumura@acm.org

Wen Jun Tan [3]
wtan047@e.ntu.edu.sg

Elvis Liu [1]
esyliu@sustc.edu.cn

Georgios Theodoropoulos [1†]
georgios@sustc.edu.cn

Wentong Cai [3]
aswtcai@ntu.edu.sg

[1]Southern University of Science and Technology, Shenzhen, China
[2]IBM T.J. Watson Research Center, New York, USA
[3]Nanyang Technological University, Singapore

ABSTRACT

We propose Distributed Neighbor Expansion (Distributed NE), a parallel and distributed graph partitioning method that can scale to trillion-edge graphs while providing high partitioning quality. Distributed NE is based on a new heuristic, called parallel expansion, where each partition is constructed in parallel by greedily expanding its edge set from a single vertex in such a way that the increase of the vertex cuts becomes local minimal. We theoretically prove that the proposed method has the upper bound in the partitioning quality. The empirical evaluation with various graphs shows that the proposed method produces higher-quality partitions than the state-of-the-art distributed graph partitioning algorithms. The performance evaluation shows that the space efficiency of the proposed method is an order-of-magnitude better than the existing algorithms, keeping its time efficiency comparable. As a result, Distributed NE can partition a trillion-edge graph using only 256 machines within 70 minutes.

1. INTRODUCTION

Graph partitioning plays a critical role to efficiently analyze large-scale graphs on distributed graph-processing systems, such as Pregel, Giraph, GraphX, GPS, PowerGraph, Pregel+, PGX.D, X-Stream, ScaleGraph, GraM, PowerLyra, Gemini, and LA3 [34, 1, 17, 32, 16, 40, 24, 41, 43, 50, 20,

---

*This work is initiated when Dr. Hanai was in NTU.
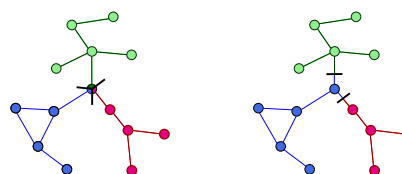†Corresponding author.

Figure 1: (a) Edge Partition (Vertex-Cut Partition) vs (b) Vertex Partition (Edge-Cut Partition).

48, 13, 55, 6]. Graph partitioning aims to divide the input graph into parts in such a way that the communication cost among the distributed processes becomes minimal while keeping the distributed load balanced.

In recent research [16, 40, 17], *edge partitioning* is shown to be more effective than the traditional vertex partitioning in large-scale real-world graphs, such as web or social graphs. The edge partitioning divides the edges of the entire graph into disjoint parts as shown in Figure 1(a), whereas the vertex partitioning divides the vertices disjointly as shown in Figure 1(b). The edge partitioning provides the better workload balance because the computational cost of the graph processing essentially depends on the number of edges rather than the number of vertices. The vertex partitioning causes serious workload imbalance between high-degree vertices and low-degree vertices in the large-scale real-world graph, which mostly has skewed-degree distribution, namely, there are a few high-degree vertices, whereas the rest of the vertices have low degree [29] (e.g., web [25] and social graphs [38]). In this paper, we focus on the edge partitioning of such skewed graphs.

The research challenges of the edge partitioning fall into two different key issues: quality and scalability. First, high *partitioning quality*, which is measured by the total number of vertex cuts, is difficult to obtain since the minimization of the vertex cuts is proved to be an NP-hard problem [54]. Second, the edge partitioning algorithm is required to scale to deal with trillion-edge graphs since the real-world graphs have been growing larger and larger, e.g., the social graph in Facebook consists of over one trillion edges [14]. The targeting scale of the recent graph analysis has been sifting from the billion-edge to the trillion-edge scale [14, 48, 33].

However, no existing edge partitioning algorithms satisfy both the quality and scalability requirements for the trillion-edge graph. On one hand, the scalable hash-based methods, such as 1D hash, 2D hash, DBH [49], Hybrid Hash [13], pro-
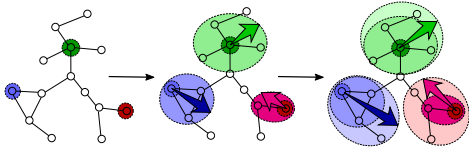
Figure 2: Parallel Expansion. 3 parts greedily expand in parallel to partition a graph.

duce the low-quality partitions because the edges are generally allocated at random by a hash operation. Although some iterative local refinements of the random allocation are proposed to improve the quality, such as Oblivious [16] and Hybrid Ginger [13], their improvement is still limited. On the other hand, the high-quality edge partitioning algorithms have a limit in their scalability. Sequential algorithms [39, 54] are not applicable for the trillion-edge graph because of the performance. Even in the state-of-art high-quality distributed method, Sheep [35] cannot scale to the trillion-edge graph because the speed up as the increase of machines is very limited.

In this paper, we propose *Distributed Neighbor Expansion (Distributed NE)*, a novel distributed edge partitioning algorithm to solve the quality and scalability issues. Distributed NE is based on a new simple greedy heuristic, called *parallel expansion* as illustrated in Figure 2. The main idea is that the algorithm starts from multiple random vertices and greedily expands each edge set in parallel such that the increase of the vertex cuts becomes minimal. Such a greedy approach provides higher-quality edge partitions because it allocates most edges in a locally optimal way and seldom includes the random allocation.

The key contributions of this paper are as follows:

A Scalable and Efficient Algorithm. Distributed NE is scalable and efficient. We carefully design the distributed algorithm (§ 3) and propose a novel distributed edge allocation to solve the scalability and performance issues (§ 4). We then propose a further optimization method which speeds up Distributed NE while keeping the quality (§ 5). Due to the scalable and efficient design, our implementation can handle the trillion-edge graph using only a few hundreds of machines (§ 7).

High-Quality Partition with Theoretical Bound. Distributed NE provides high-quality partitioning, which has a theoretical upper bound. Our theoretical analysis proves that the number of the vertex cuts computed by Distributed NE is bounded from above. Its upper bound is shown to be better than that of the existing distributed methods (§ 6). Moreover, we empirically show that Distributed NE produces significantly higher-quality partitions than the existing distributed methods in various skewed graphs (§ 7).

2. BACKGROUND

This section provides a formulation of the problem and an outline of related work.

2.1 Notation and Problem Definition

Let $G(V, E)$ be an undirected and unweighted graph which consists of a set of vertices, V, and edges, E. The vertex set involved in E is defined as $V(E)$. An edge, e ($\in E$), connecting vertices, v and u, is represented by $e_{v,u}$. Let p be a partition id and P be the set of partition ids. The number of elements in a set is represented by $|\cdot|$, e.g., $|V|$ and $|P|$.

The objective of the edge partitioning is to divide E into the disjoint subsets E_p ($\bigcup_{p \in P} E_p = E$) such that the number of vertex cuts becomes as minimal as possible while keeping the balance of the subsets. We usually measure the number of *vertex replications*, which is defined as $\sum_{p \in P} |V(E_p)|$, instead of vertex cuts. The number of vertex replications is normalized as follows:

$$\frac{1}{|V|} \sum_{p \in P} |V(E_p)|, \tag{1}$$

which is called *replication factor* [16]. Based on the replication factor, the objective of *a balanced $|P|$-way edge partitioning of G* [16] is formalized as follows:

$$\min_{f \in \mathcal{F}} \frac{1}{|V|} \sum_{E_p \in f(E)} |V(E_p)|, \quad \text{s.t.} \quad \max_{p \in P} |E_p| < \alpha \frac{|E|}{|P|}, \tag{2}$$

where $f : E \mapsto \{E_p : p \in P\}$ is a partitioning method, and $\mathcal{F}$ is the set of all partitioning methods. The imbalance factor, $\alpha \geq 1.0$, is a constant parameter.
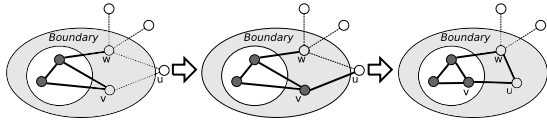
2.2 Related Work

The graph partitioning problem has been extensively investigated. We refer the reader to the comprehensive surveys by Buluç et al. [11]. Here, we provide a review of the edge partitioning methods and the vertex partitioning methods that have been proposed for trillion-edge graphs.

Edge Partitioning. One of the major approaches to edge partitioning is based on the random hash. The most straightforward approach is 1D-hash partitioning, where the edge is randomly assigned to a one-dimensional partitioning space. Another approach uses 2D-hash partitioning, where the edge is assigned to two-dimensional partitioning space by hashing the adjacent vertices separately [53, 9, 4, 17]. The latest hash-based approaches utilize the degree of vertices, where the edge is randomly assigned so that high-degree vertices are divided into more partitions than low-degree ones. The examples are Hybrid Hashing [13] and DBH [49]. Oblivious [16] and Hybrid Ginger [13] are a heuristic to iteratively refine the assignment after the hash partitioning for the improvement of partitioning quality. Since the hash operation is lightweight, this approach efficiently partitions the trillion-edge graph, but the random allocation causes the serious quality loss.

A more recent approach for the large-scale graph is based on streaming methods, such as FENNEL-based edge partitioning [45, 10], HDRF [39], and SNE [54], where the input graph is represented as a sequence of edges and processed one-by-one without the entire information. Since only a part of the entire graph is deployed on the main memory, it can handle a large graph whose size is more than the main memory. However, since the algorithm does not fully make use of the information of the entire graph, the partitioning quality is limited. Moreover, these methods basically assume the sequential processing, and thus, their performance is limited.

Sheep [35] is the state-of-the-art distributed edge partition method, where the graph is parallelly translated into the elimination tree before applying tree partitioning. Although the algorithm can be executed on the distributed memory, it does not speed up well as the increase of the machines as reported in the paper. Moreover, the algorithm does not provide any theoretical guarantee of the partitioning quality, such as the upper bound.

Figure 3: Expansion Approach. First heuristic selects v from the boundary and allocates $e_{v,u}$. Then, second heuristic allocates $e_{u,w}$. This figure is modified from [54].

Most related work to ours is a greedy approach [10, 54]. NE [54] is the state-of-the-art greedy algorithm based on the expansion of the edge set. It currently provides the best quality in practice, but the scalability is limited since it is an offline sequential algorithm, where the entire graph is deployed on the main memory on a single machine.

Vertex Partitioning of Trillion-edge Graphs. Recently, a few distributed vertex partitioning methods which can handle the trillion-edge graph have been proposed. XtraPuLP [42] can scale to the trillion-edge graph, but it requires over 8K machines, which are excessive because PageRank of Facebook's trillion-edge graph can be calculated on only a few hundreds of machines [14]. Spinner [36] is reported to partition the trillion-edge graph by using a few hundreds of machines, but it generates low-quality partitions because it involves the initial random allocation in the same way as the hash-based edge partitioning methods. Moreover, these methods do not give the theoretical guarantee of the partitioning quality.

3. DISTRIBUTED NEIGHBOR EXPANSION

In this section, we first illustrate the greedy expansion approach, which forms the basis of our distributed algorithm. Second, we summarize challenging issues towards the scalable distributed algorithm. Finally, we give an overview of our proposal.

3.1 Edge Partitioning Based on Expansion

We first provide a short overview of the expansion of an edge set for partitioning. After that, we review two greedy heuristics for effective expansion.

Edge Partitioning Using Expansion of Edge Set. For $G(V, E)$, let X be an edge set ($X \subset E$). We refer to the vertex set $B(X) := \{v \mid v \in V(X) \land \exists e_{v,u} \in E \backslash X\}$ as *boundary* of X. The *expansion of* X is an operation defined as follows:

Select $v \in B(X)$, and $X \leftarrow X \cup \{e_{v,u} \mid e_{v,u} \in E \setminus X\}$,

where we also suppose that if $B(X)$ is empty, v is selected at random from $V(E \setminus X)$.

Based on the expansion, we can simply generate $|P|$ edge partitions of $G(V, E)$ as the following algorithm:

(i) The computation for each partition begins with a single boundary vertex and an empty edge set.

(ii) The edge set is expanded until its size reaches to the limit (i.e., $\alpha|E|/|P|$ as discussed in Section 2) or there is no edge for expansion. The generated edge set is one of the edge partitions, and it is removed from E.

(iii) The next partition is computed for the remaining edges E. If E is empty, the algorithm terminates.

There are two basic heuristics [16, 10, 39, 13, 54] to obtain high-quality edge partitions by using the expansion. One

aims to select a vertex from the boundary in such a way that the increase of vertex replications becomes minimal. The other is used during expansion to allocate additional edges which do not increase the replication vertices anymore. Figure 3 shows the overview of how the edge set is expanded by using these heuristics.

Heuristic for Vertex Selection from Boundary. In the above algorithm, let $E_p(t)$ be the set of p's edge set allocated so far until t, where t is a counter of the expansions. Based on Formula (1), the number of p's vertex replications at t is represented as $|V(E_p(t))|$.

The increase of the vertex replications by the new boundary vertex v from t to $t + 1$ is equal to the v's degree in the remaining edges.

$$|V(E_p(t+1))| - |V(E_p(t))| = D_{rest}(v), \qquad (3)$$

where $D_{rest}(v) := |N(v) \cap E|$; $N(v)$ is a set of v's neighbor edges; and E is the remaining edges. Thus, new vertex which minimizes the increase of the vertex replications, v_{min}, is represented as follows:

$$v_{min} := \operatorname*{argmin}_{x \in B} D_{rest}(x). \qquad (4)$$

For example in Figure 3, $D_{rest}(w)$ is 3 , and thus the number of vertex replications would increase by 3 if w is selected; whereas $D_{rest}(v)$ is 1, and the number of replications increases by 1 if v is selected. Finally, v is selected in this case (i.e., $v_{min} = v$).

Heuristic for Edge Allocation. In the above algorithm, two-hop neighbors of the selected vertex may include some edges which never increase the replication vertices. Such edges can be easily found with a simple yet important fact, that is commonly used in the existing algorithms [16, 10, 39, 13, 54], as follows:

If both vertex v and u are involved in $V(E_p(t))$, then

allocation of $e_{v,u}$ to p does not increase replications. (5)

By this fact, when allocating an edge to p, if a two-hop-neighbor vertex from the selected vertex is already allocated to p, then the edge from the one-hop-neighbor vertex to the two-hop-neighbor vertex is allocated.

For example in Figure 3, $e_{u,w}$ is additionally allocated during the expansion. The allocation of $e_{u,w}$ does not increase the number of the vertex replications because one of the edges to w and one of the edges to v are already allocated to the same partition.

3.2 Challenges to Distributed Algorithm

The main idea behind our proposed approach is to execute the expansion for each partition in parallel. Even though the core idea of the parallel expansion is simple, to achieve a highly scalable distributed algorithm which enables us to partition trillion-edge graphs is a challenging endeavour. The challenges are summarized as follows:

Scalability. The primal issue is constraints on the scalability of the partitioning algorithm imposed by the limited memory size. The parallel expansion is an offline algorithm, where the entire input graph is necessary to be accessible from the beginning of the algorithm. The effective initial deployment of the input graph to the distributed main memories is required for scalability and efficiency.
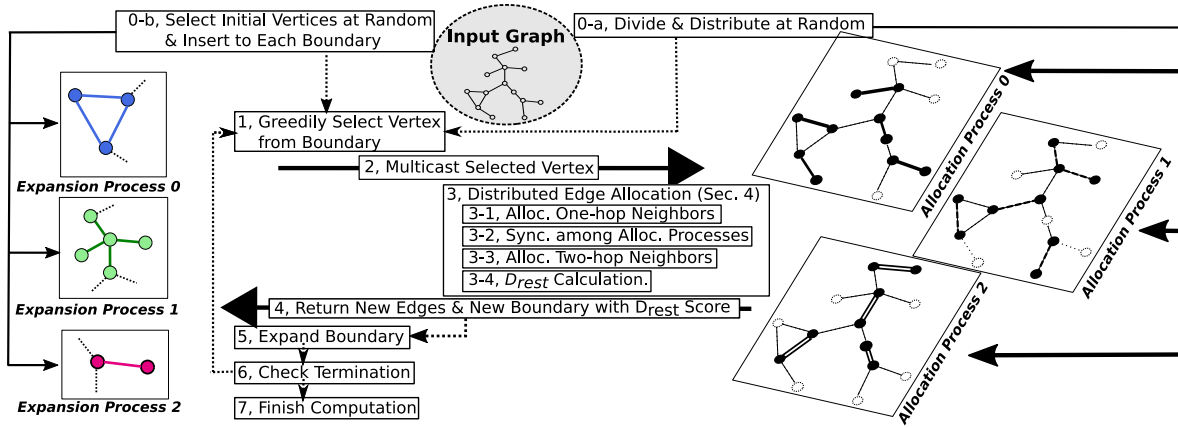
Figure 4: Overview of Distributed NE. A dotted and solid arrow represent a work and data flow, respectively.

Concurrent Allocation. Concurrency issues appear during the parallel expansion when multiple expanded parts simultaneously try to allocate the same graph element (i.e., vertex or edge). If the element is replicated among the distributed processes, the global synchronization of the allocated information is necessary among the replications to maintain consistency.

Distributed Two-hop-neighbor Search. Another issue is introduced by a two-hop-neighbor search in the greedy expansion. The search of the two-hop neighbors from the current expanded part is necessary to greedily decide which edges should be allocated next. However, in the general distributed graph, each vertex has basically only its one-hop-neighbor vertices and edges (i.e., adjacent vertices and edges), and the two-hop-neighbor search cannot be fully localized. It must include the communication among the distributed processes. The efficient two-hop-neighbor search is a key challenge, which differentiates Distributed NE from the other typical distributed graph algorithms based on the one-hop-neighbor accesses.

3.3 Proposed Distributed Approach

In this subsection, Distributed NE is presented. Suppose there are $|P|$ machines available. Since the graph partitioning is used for the pre-processing of the distributed graph applications (such as shortest path and PageRank), the number of partitions generated by Distributed NE should be equal to the most effective number for the distributed graph processing systems, which we assume are the most efficient when each partition is assigned to one machine. Thus, the goal of the task is specified to be:

To compute $|P|$-way edge partitioning of the input graph and distribute the $|P|$ partitions into the $|P|$ machines.

For the computation of the parallel expansion on the distributed environment, there are two basic requirements. First, the entire input graph needs to be initially stored in the main memories on the multiple machines for the graph traversal during the expansion. Second, the boundary vertices of the edge set in each partition need to be managed in a single machine because the selection of the local-minimal vertex from the boundary is executed locally for efficiency.

To meet the two requirements, two types of a distributed process are used: an *expansion process* and an *allocation process*. The expansion processes manage the boundary vertices to compute local-optimal vertex selection. The alloca-

tion processes manage the input graph in the distributed way and manage the allocation of vertices and edges.

The flow of the computation is illustrated in Figure 4. The overall computation begins with launching $|P|$ expansion processes and $|P|$ allocation processes (i.e., Expansion Process 0,1,2 and Allocation Process 0,1,2 in Figure 4). Each expansion process and allocation process is deployed to one of the $|P|$ machines. First, the input graph is randomly distributed to the allocation processes. The expansion in each expansion process begins with a random-selected vertex. The algorithm is executed iteratively as illustrated by Step 1–6 in Figure 4. During the iterations, the allocated edges are copied and sent from the allocation processes to the expansion processes, and at the end of the computation, the entire edges are distributed to the $|P|$ expansion processes.

Expansion Process. Each expansion process computes in parallel the selection of a new vertex from the boundary, B_p to expand the edge set, E_p.

Algorithm 1 shows the details of the computation in each expansion process. Each partition is computed iteratively, including the vertex selection (Line $3 - 7$), the allocation request (Line 8), and the update of the boundary and the edge set (Line $10 - 13$). In the vertex selection, a vertex, v_{min}, which has the minimal degree for the remaining edges as discussed in Equation (4), is selected from B_p if it is not empty; otherwise, v_{min} is chosen at random from vertices with non-allocated edges. The random vertex is basically taken from the allocation process in the same machine. It is from the other machines only if necessary. After that, in the allocation request, each expansion process multicasts v_{min} to allocation processes in charge, and waits for the edge allocation by the allocation processes. When the edge allocation is done, each expansion process receives new boundary, B_{new}, which also includes D_{rest} score of each vertex as discussed in Equation (3), and newly allocated edges, E_{new}. B_{new} and E_{new} are inserted to B_p and E_p respectively. Finally, the expansion process checks for the termination condition and stops accordingly.

Allocation Process. The allocation processes take the selected vertices from the expansion processes and return newly allocated edges and new boundary vertices with their D_{rest} scores. For each selected vertex v of partition p, two types of edges are allocated as discussed in Section 3.1.

(i) v's one-hop neighbor edges, $\{e_{v,u}\}$.

Algorithm 1: Expansion Process for Partition p

Input: p – This Partition ID, P – Set of Partition IDs
Output: E_p – Partitioned Edges for p

Var: B_p – Priority Queue of $\langle D_{rest}(v), v \rangle$
  : $|E|$ – Total # of Allocated Edges
  : $|E_{init}|$ – Total # of Initial Edges

1 $E_p \leftarrow \varnothing$, $B_p \leftarrow \varnothing$
2 **while** *true* **do**
      /* Selection of Expansion Vertex */
3    $v_{min} \leftarrow \varnothing$
4   **if** $B_p \neq \varnothing$ **then**
5      $v_{min} \leftarrow B_p.\texttt{popMinDrestVertex()}$
6   **else**
7      $v_{min} \leftarrow \texttt{getRandomVertex()}$
      /* Request Allocation */
8    $\texttt{MulticastVertexToAllocators}(v_{min}, p)$
9    $\texttt{Barrier()}$ /* Wait for Allocation */
      /* Update Boundary and Edge Sets */
10    $B_{new} \leftarrow \texttt{ReceiveNewBoundaryWithDrest()}$
11    $B_p.\texttt{insert}(B_{new})$
12    $E_{new} \leftarrow \texttt{ReceiveNewAllocatedEdges()}$
13    $E_p \leftarrow E_p \cup E_{new}$
      /* Check Termination */
14    $|E| \leftarrow \texttt{AllGatherSum}(|E_p|)$
15   **if** $E_p > \alpha \cdot |E_{init}|/|P| \ \lor \ |E| = |E_{init}|$ **then break**

Algorithm 2: Overview of Edge Allocation

Args: VP_{rec} – Received $\{\langle v_{min}, p \rangle\}$ via Line 8 in Alg. 1

Var : BP_{new} – New Boundary-Partition Pairs $\{\langle v, p \rangle\}$
  EP_{1hop} – Alloc. 1Hop-Partition Pairs $\{\langle e, p \rangle\}$
  EP_{2hop} – Alloc. 2Hop-Partition Pairs $\{\langle e, p \rangle\}$
  LD_{rest} – Local D_{rest} Scores $\{\langle v, D_{rest}(v) \rangle\}$

/* This function is called after all the selected vertices
 * are received. (After Line 8 in Algorithm 1) */
1 $\texttt{EdgeAllocation}(VP_{rec})$
2    $\langle BP_{new}^{local}, EP_{1hop} \rangle \leftarrow$
      $\texttt{AllocteOneHopNeighbors}(VP_{rec})$
3    $BP_{new} \leftarrow \texttt{SyncVertexAllocations}(BP_{new}^{local})$
4    $EP_{2hop} \leftarrow \texttt{AllocateTwoHopNeighbors}(BP_{new})$
5    $LD_{rest} \leftarrow \texttt{ComputeLocalDrest}(BP_{new})$
6    $\texttt{SendNewBoundaryWithLocalDrest}(BP_{new}, LD_{rest})$
7    $\texttt{SendNewAllocatedEdges}(EP_{1hop} \cup EP_{2hop})$

  (ii) v's two-hop neighbor edges which do not increase the vertex replications, namely, which satisfy Condition (5).

Unlike the single machine environment, where these edges are straightforwardly searched and allocated, the allocation on the distributed environment causes two issues: the management of concurrent allocation and the two-hop neighbor search. In the next section, we propose a new technique to solve the issues.

4. DISTRIBUTED EDGE ALLOCATION

In this section, an edge allocation algorithm to utilize in the distributed allocation processes is proposed. We first present a basic data structure which is used to divide the input graph and to store the divided graphs. The input graph needs to be evenly divided for workload balancing and to be stored in a space-efficient way. Second, we present an algorithm which efficiently allocates one-hop and two-hop neighbors while solving concurrency problems.

Data Structure. In the allocation processes, the input graph is distributed by the edge partitioning algorithm, which uniquely divides the edges and replicates the vertices among the allocation processes. Specifically, we use 2D-hash partitioning and compressed sparse row (CSR) [53, 9, 4, 17] for the initial distribution of the input graph. In 2D-hash partitioning, the metadata of replicated vertices (i.e., process ids of each vertex) can be calculated from vertex id instead of storing pairs of vertex and process id, which suppresses memory space in the case of trillion-edges graphs.

On the edge-partitioned graph, global synchronization is executed only in the vertex replications, whereas the edge's allocation information does not need to synchronize since it is unique. For example in Figure 4, the input edges are uniquely divided into three processes (solid line, dotted line and double line). The vertices are replicated to the three processes and synchronize their allocation information during the computation.

There are two main advantages of this vertex-replicated approach compared to other approaches, such as replicating edges only or replicating both edges and vertices hybridly.

First, this approach can locally solve the allocation conflict, which occurs only in the edge and not in the vertex. On one hand, the edge allocation may result in conflicts because each edge is allocated by the unique partition. Thus, if the multiple partitions concurrently try to allocate the same edge, then conflict resolution is necessary to decide which partition actually allocates the edge. On the other hand, the vertex allocation does not cause any conflict because a vertex may be allocated by the multiple partitions. If the multiple partitions concurrently allocate the same vertex, these partitions can be simply assigned to the vertex.

Second, the replication of vertices is space efficient compared to the replication of edges because skewed graphs have much more edges than vertices (e.g., Facebook graph [14] contains 1.45B vertices and 1T edges). Replications of vertices generally require less memory than that of edges.

Algorithm. The key idea of the algorithm is to first allocate the one-hop-neighbor edges and then allocate two-hop-neighbor edges which satisfy Condition (5). Between the one-hop-neighbor and two-hop-neighbor allocation, synchronization is executed to detect two-hop neighbors on the other processes and to make the allocation information consistent among the replications. The conflicts of the edge allocation are solved locally. Algorithms 2 and 3 provide a high level and a detailed description of the algorithm respectively. Figure 5 illustrates the edge allocation for the graph in Figure 1.

Specifically, `EdgeAllocation()` in Algorithm 2 is called after receiving all the selected vertices with their partition id, VP_{rec}, from the expansion processes (i.e., after all the expansion processes call `MulticastVertexToAllocators()` at Line 8 in Algorithm 1).

The function consists of 4 phases as follows.

(1) Expansion and One-hop-neighbor Allocation.
In `AllocteOneHopNeighbors()`, for each received vertex and its associated partition $\langle v, p \rangle$, v's one-hop-neighbor edge and
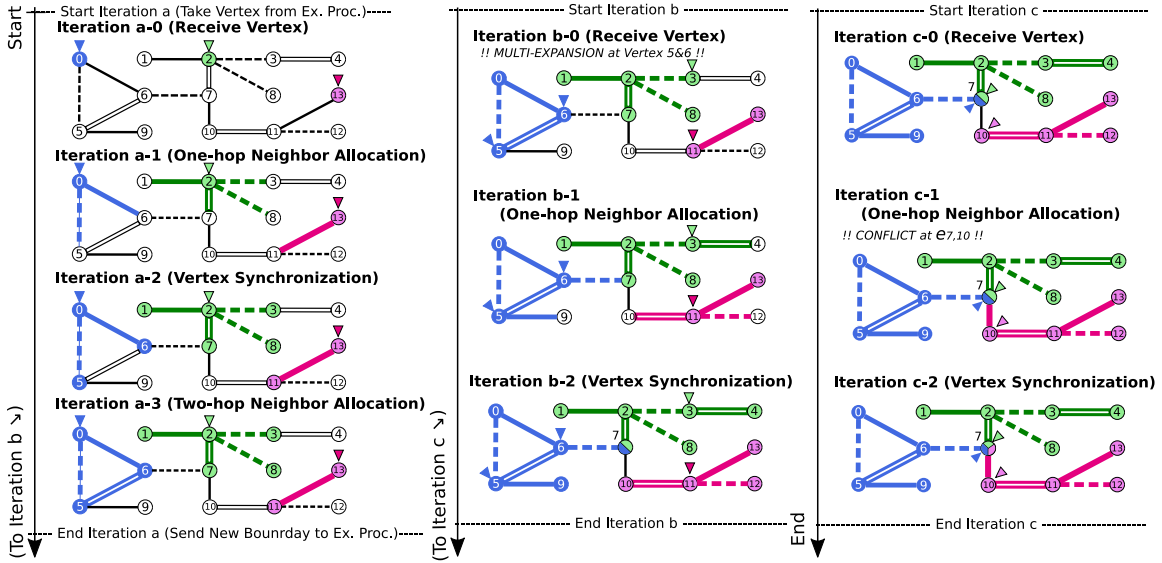
Figure 5: Distributed Edge Allocation. Each line type (solid, dotted, double) represents the initial partition for allocation process. Each color (blue, red, green) is the final partitioning result.

Algorithm 3: Details of Algorithm 2

SubGraph: $SubG$ – Initial Distributed Subgraph

1 AllocteOneHopNeighbors(VP_{rec})
2 $BP_{new}^{local} \leftarrow \varnothing;\ EP_{1hop} \leftarrow \varnothing$
3   **for** $\langle v, p \rangle \in VP_{rec}$ **do in parallel**
4   **for** $e_{v,u} \in SubG.NonAllocNeighbors(v)$ **do**
5 $SubG.\texttt{AppendVertexPartition}(u,\ p)$
6 $SubG.\texttt{AllocateEdge}(e_{v,u},\ p)$
7 $BP_{new}^{local} \leftarrow BP_{new}^{local} \cup \{\langle u, p \rangle\}$
8 $EP_{1hop} \leftarrow EP_{1hop} \cup \{\langle e_{v,u}, p \rangle\}$

9   **return** $\langle BP_{new}^{local},\ EP_{1hop} \rangle$

10 AllocateTwoHopNeighbors(BP_{new})
11 $EP_{2hop} \leftarrow \varnothing$
12   **for** $\langle u, . \rangle \in BP_{new}$ **do in parallel**
13   **for** $e_{u,w} \in SubG.NonAllocNeighbors(u)$ **do**
14 $P_{new} \leftarrow SubG.\texttt{Parti}(u) \cap SubG.\texttt{Parti}(w)$
15   **if** $P_{new} \neq \varnothing$ **then**
16 $p_{new} \leftarrow \underset{x \in P_{new}}{\text{argmin}}\ SubG.\texttt{NumEdges}(x)$
17 $SubG.\texttt{AllocateEdge}(e_{u,w},\ p_{new})$
18 $EP_{2hop} \leftarrow EP_{2hop} \cup \{\langle e_{u,w}, p_{new} \rangle\}$

19   **return** EP_{2hop}

---

vertex, $e_{v,u}$ and u, are allocated to p in parallel as shown in Algorithm 3. $\langle e_{v,u}, p \rangle$ and $\langle u, p \rangle$ are inserted to the new one-hop-neighbor edges, EP_{1hop}, and the new boundary, BP_{new}^{local}, respectively. The conflict among different threads to allocate the same edge is solved by a CAS operation for performance.

For example in Figure 5, at Iteration a-0, v_0 is received in the solid-line allocation process and the dotted-line one. Then, $e_{0,6}$ is allocated in the solid-line process, while $e_{0,5}$ is allocated in the dotted-line process. v_6 and v_5 become new boundaries for the next iteration. In Iteration c-1, a conflict in the edge allocation occurs at $e_{7,10}$ if multiple threads simultaneously try to allocate $e_{7,10}$ to different partitions. In this case, the edge is finally allocated to the red partition.

(2) Synchronization of Vertex's Allocation Ids.
SyncVertexAllocations() synchronizes vertices which are newly allocated in the previous one-hop-neighbor allocation, so that all vertex replications among the different machines have the same allocation ids. Specifically, the local data of the new boundaries, BP_{new}^{local}, are sent to their replications on the other machines. Then, in the other machines, the allocation ids are appended to the associated vertices. The synchronized allocation ids, BP_{new}, are used for the two-hop-neighbor allocation. For example in Figure 5, at Iteration a-2, v_6 in the solid-line process is synchronized with its replications in the dotted-line and the double-line ones.

(3) Two-hop-neighbor Allocation.
In AllocateTwoHopNeighbors(), the neighbors of BP_{new} (thus, two-hop neighbors of the received vertex, v_{min}) are processed in parallel as shown in Algorithm 3. For a two-hop-neighbor edge, $e_{u,w}$, if both u and w have been already allocated to the same partitions, P_{new}, then the algorithm selects the minimal edge partition, $p_{new} \in P_{new}$, and allocates $e_{u,w}$ to p_{new}. The new allocated edges are stored to EP_{2hop} and sent back to the expansion processes. For example in Figure 5, $e_{5,6}$ at Iteration a-3 is allocated in this phase because both v_5 and v_6 are allocated to *blue*.

(4) Calculation of Local D_{rest} of New Boundaries.
Finally, in ComputeLocalDrest(), the local score of D_{rest} in each processes (as discussed in Equation (3)) is calculated for each vertex in BP_{new}. At a later stage, the calculated local scores will be communicated to and gathered in the corresponding expansion process in order to compute the global D_{rest} score.

After completing the 4 phases, the new boundary vertices, BP_{new}, with their local D_{rest} score, LD_{rest}, and the new allocated edges, $EP_{1hop} \cup EP_{2hop}$, are sent back to the expansion processes. In the expansion processes, they are received via ReceiveNewBoundaryWithDrest() at Line 10 in Algorithm 1 and ReceiveNewAllocatedEdges() at Line 12, respectively. Before calling ReceiveNewBoundaryWithDrest(), the local D_{rest} scores for each vertex are summed up to calculate the global score in each expansion process.

5. ITERATION REDUCTION BY MULTI-EXPANSION

In this section, we propose further performance optimization of Distributed NE by reducing the number of iterations in the expansion. Even though our algorithm uses the efficient allocation method, performance bottleneck still appears in communications among the allocation processes because a large number of iterations are executed. As a result, a large number of barrier operations are conducted, and it causes a performance problem.

Suppose each process evenly allocates n edges per iteration, the number of total iterations for computing the partition of $G = (V, E)$ is represented as $|E| / (n \cdot |P|)$. Usually, the number of edges $|E|$ (e.g., 100K – 1T) is much larger than the number of $|P|$ (e.g., up to 1K). Therefore, the number of allocating edges per iteration n is required to be large enough to reduce the number of iterations.

Algorithm 4: Multi-Expansion (Corresponding to Line 3–8 in Algorithm 1)

Variable: $V_{k\text{-}min}$ – Expanding Vertices
**Parameter:** λ – Expansion Factor.

/* Multi-Expansion */
3   $V_{k\text{-}min} \leftarrow \varnothing$
4   **if** $B_p \neq \varnothing$ **then**
5      $k \leftarrow \lambda \cdot |B_p|$
6      $V_{k\text{-}min} \leftarrow B_p.\texttt{popK-MinDrestVertices}(k)$
7   **else**
8      $V_{k\text{-}min} \leftarrow$ Random Vertex

/* Request Allocation */
9   $\texttt{MulticastVerticesToAllocators}(V_{k\text{-}min}, p)$

Algorithm 4 is the modified algorithm from Algorithm 1 in Section 3. Line 3–8 in Algorithm 1 are replaced into Line 3–9 in Algorithm 4. To increase the number of allocating edges per iteration, Algorithm 4 expands multiple boundary vertices in one iteration. Instead of selecting a single minimum scored vertex from the boundary, the algorithm selects k-minimum vertices per iteration. The number of expanding vertices k is determined by an *expansion factor* λ $(0 < \lambda \leq 1)$. In the expansion, $k = \lambda \cdot |B_p|$ vertices are expanded for allocating new edges. As λ approaches 1, the number of expanded edges per iteration increases, and the overall execution time is reduced. On the other hand, the oversize λ causes quality loss since the greedy heuristic may not work well.
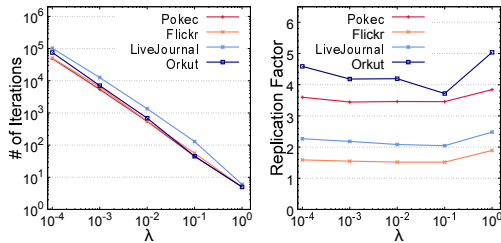
Figure 6: Trend of # of Iterations and Replication Factor on Different λ (32 Partitions).

Preliminary experiments have been conducted to select the value for λ. Figure 6 shows the relationship between the number of iterations and the replication factor (as discussed in Equation (1)) under different λ on 32 partitions. The

number of iterations linearly decreases as the increase of λ. In $\lambda = 1.0$, which means all boundary vertices are chosen every iteration, the total number of iterations is less than 10 in all graph data. On the other hand, the replication factor slightly decreases from $\lambda = 10^{-4}$ to 10^{-1}. Then, in $\lambda = 1.0$, it becomes worse in all graph data. We find the similar characteristics in the other numbers of partitions such as 4, 8, 16, and 64 partitions. Based on these results, we choose $\lambda = 0.1$ in our implementation to maximize the performance and quality.

6. THEORETICAL ANALYSIS

In this section, we first analyze a theoretical upper bound of the partitioning quality provided by Distributed NE (Theorem 1, 2). We then provide the efficiency analysis of Distributed NE (Theorem 3).

The partitioning quality given by Distributed NE (without the muti-expansion in Section 5) has the upper bound and satisfies the following two theorems. The upper bound of partitioning quality by the sequential expansion has been obtained in [54], but it does not satisfy the parallel case because the bound strongly assumes the sequential processing. We provide new upper bound for the parallel case.

Theorem 1 (Upper Bound). *Edge partitions of a graph* $G(V, E)$ *computed by Distributed NE have replication factor* (*RF*), *such that*

$$RF \leq \frac{|E| + |V| + |P|}{|V|},$$

where $|E|$, $|V|$ *and* $|P|$ *are the number of edges, vertices, and partitions , respectively.*

Proof. We will prove the formula using a potential function as follows:

$$\Phi(t) := |E_{rest}(t)| + |V_{rest}(t)| + |P_{rest}(t)| + \sum_{p \in P} |V(E_p(t))|,$$

where t is the counter of the iteration in Algorithm 1; $E_{rest}(t)$ is the set of non-allocated edges at t; $V_{rest}(t)$ is the set of vertices which has non-allocated adjacent edges at t; $P_{rest}(t)$ is the set of parallely computed partitions whose edges are less than the limit; $E_p(t)$ is the set of allocated edges; and $V(E_p(t))$ is the set of vertices covered by $E_p(t)$.

Suppose Algorithm 1 is terminated at T, we will prove $\Phi_T \leq \Phi_0$. Let $\Delta(\cdot)$ be the differentiation operator from t to $t + 1$ as $\Delta(\cdot) := (\cdot)(t + 1) - (\cdot)(t)$. Then, $\Delta\Phi$ is represented as follows:

$$\Delta\Phi = \Delta|E_{rest}| + \Delta|V_{rest}| + \Delta|P_{rest}| + \sum_{p \in P} \Delta|V(E_p)|$$

Suppose a vertex x_p is selected from the boundary set B_p for each partition $p \in P_{rest}(t)$ in parallel at t as shown in Algorithm 1. Let n_p be the number of x_p's one-hop-neighbor edges added to $E_p(t)$, and n'_p be the number of the two-hop-neighbor edges also added to $E_p(t)$. Obviously, n_p and n'_p is equal to zero if p's expansion has finished already. Then, $\Delta|E_{rest}|$ is represented as follow.

$$\Delta|E_{rest}| = -\sum_{p \in P}(n_p + n'_p)$$

$|V_{rest}(t)|$ is a non-increasing function over t, and $\Delta|V_{rest}|$ can be represented using an indicator function $\chi(p)$ as follow.

$$\Delta|V_{rest}| \leq -\sum_{p \in P} \chi(p), \quad \chi(p) = \begin{cases} 0 & (p \text{ ends at } t \text{ or earlier}) \\ 1 & (\text{Other}) \end{cases}$$

This means that if p's computation is terminated at t, then x_p's one-hop neighbors may not fully allocated, and $|V_{rest}(t+1)|$ at p may be equal to $|V_{rest}(t)|$. Otherwise, x_p's one-hop neighbors are fully allocated, and $|V_{rest}(t)|$ at p is always decreased by up to one (x_p and the other vertices due to the allocation of the two-hop-neighbor edges).

$|P_{rest}(t)|$ is also a non-increasing function over t.

$$\Delta|P_{rest}| = -\sum_{p \in P}(1 - \chi(p))$$

If p's computation is terminated at t, then the number of computed partition $|P_{rest}(t)|$ is decreased ($\because \chi(p) = 0$); otherwise it is constant ($\because \chi(p) = 1$).

$|V(E_p(t))|$ is a increasing function over t, and $\Delta|V(E_p)|$ can be represented as $\Delta|V(E_p)| \leq n_p + 1$, where $n_p + 1$ means the number of x_p's allocated neighbors plus x_p itself.

To sum up above, the difference of the potential function from t to $t + 1$ is represented as follow:

$$
\begin{aligned}
\Delta\Phi &= \Delta|E_{rest}| + \Delta|V_{rest}| + \Delta|P_{rest}| + \sum_{p \in P_{rest}(t)}\Delta|V(E_p)| \\
&\leq -\sum_{p \in P_{rest}(t)}\{(n_p + n'_p) + \chi(p) + (1 - \chi(p)) - (n_p + 1)\} \\
&= -\sum_{p \in P_{rest}(t)} n'_p \quad \leq 0
\end{aligned}
$$

Therefore, $\Delta\Phi \leq 0$ for all t, and thus $\Phi_T \leq \Phi_0$. The following equation is established.

$$RF := \sum_{p \in P}\frac{|V(E_p)|}{|V|} = \frac{\Phi_T}{|V|} \leq \frac{\Phi_0}{|V|} = \frac{|E| + |V| + |P|}{|V|}$$

Thus, Theorem 1 is proved. $\qquad\square$

In addition, the tightness of the upper bound (called UB) is provided by the following theorem.

Theorem 2 (Tightness).

$$\exists G(V, E), \exists P \ s.t. \ RF \sim UB \ as \ |V| \to \infty.$$

Proof. Suppose a graph consisting of two isolated sub-graphs: a complete graph and a ring. We assume that the complete graph includes n vertices and $\frac{1}{2}n(n-1)$ edges, and that the ring has $\frac{1}{2}n(n-1)$ vertices and $\frac{1}{2}n(n-1)$ edges. We will show the graph satisfies the condition of the theorem in a certain number of partitions. Specifically, we will prove that for the graph,

$$|P| = 2^{-1}n(n-1) \Rightarrow RF \sim UB \ as \ |V| \to \infty.$$

Figure 7 shows the example of the case $n = 4$. In general, the total number of vertices $|V|$ is equal to $\frac{1}{2}n(n-1)+n$. The total number of edges $|E|$ is $\frac{1}{2}n(n-1)+\frac{1}{2}n(n-1) = n(n-1)$.
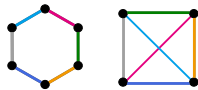
Figure 7: Partitions of Ring + Complete Graph.

Consider a case where the initial random vertex selection for each partition is from the ring, and each partition allocates one of the edges. Then, in the next iteration, the random vertex selection occurs again in the complete graph since no partitions have boundary vertices. Each partition

allocates one of its one-hop-neighbor edges and then terminates the expansion because the number of allocated edges reaches to the limit ($\frac{|E|}{|P|} = \frac{n(n-1)}{2^{-1}n(n-1)} = 2$). Finally, each edge in the complete graph is allocated to one of the partitions. Figure 7 shows the final condition, where different colors represent different partitions. In each partition, the allocated edges are disconnected.

In this case, both the complete graph and the ring have $n(n-1)$ vertex replications. The replication factor is represented as follows:

$$RF = |V|^{-1}\{n(n-1) + n(n-1))\} = |V|^{-1}\{2n(n-1)\}$$

The upper bound is represented as follows:

$$
\begin{aligned}
UB &:= |V|^{-1}(|E| + |V| + |P|) \\
&= |V|^{-1}\{n(n-1) + 2^{-1}n(n-1) + n + 2^{-1}n(n-1)\} \\
&= |V|^{-1}\{2n(n-1) + n\}
\end{aligned}
$$

Therefore,

$$\lim_{|V| \to \infty}\frac{RF}{UB} = \lim_{n \to \infty}\frac{2n(n-1)}{2n(n-1) + n} = 1 \quad \therefore RF \sim UB \quad \square$$

Comparison with the Other Distributed Methods. We compare the theoretical upper bound of Distributed NE with the other distributed edge partitioning methods discussed in [49]. In the paper, the theoretical upper bounds of three hash-based methods (*Random*, *Grid*, and *DBH*) are formulated in the case of power-law graphs.

Since our upper bound is for the general graph, we will apply it to the power-law graph. Based on the formulation of the power-law distribution by Clauset et al. [15], we use a graph, $G_\zeta(V, E)$, satisfying the following condition.

$$Pr[d] = d^{-\alpha} \cdot \zeta(d, d_{min})^{-1}, \tag{6}$$

where $Pr[d]$ is the probability that vertex's degree becomes d; α is the scaling parameter (Typically, $2 < \alpha < 3$); $\zeta(d, d_{min})$ is the generalized (or Hurwitz) zeta function; and d_{min} is the minimum degree. We assume that $d_{min} = 1$ (thus, $\zeta(d, d_{min})$ becomes Riemann zeta function) and that $|V|$ is much bigger than $|P|$ such that $|P|/|V| \approx 0$.

In G_ζ, the expected upper bound of Distributed NE is represented as

$$\mathbb{E}[UB] \approx \mathbb{E}\left[\frac{|E|}{|V|}\right] + 1 = \frac{1}{2} \cdot \frac{\zeta(\alpha - 1, 1)}{\zeta(\alpha, 1)} + 1.$$

This is because the mean of (6) is equal to $\zeta(\alpha-1, 1)/\zeta(\alpha, 1)$ when $d_{min} = 1$, while the mean degree of a graph is generally $2 \times \mathbb{E}[|E|/|V|]$.

Table 1 shows the calculation result in various α under $|P| = 256$. We calculate the existing upper bound based on the formulas provided in [49]. Distributed NE always provides the better upper bound than the existing distributed methods. Especially, in the smaller α, the difference is more significant.

Table 1: Theoretical Upper Bound of Replication Factor in Power-law Graph (256 Partition).

Partitioner	$\alpha = 2.2$	2.4	2.6	2.8
Random (1D-hash)	5.88	3.46	2.64	2.23
Grid (2D-hash)	4.82	3.13	2.47	2.13
DBH [49]	5.54	3.19	2.42	2.05
Distributed NE	**2.88**	**2.12**	**1.88**	**1.75**

We provide the time complexity for each computing unit since Distributed NE is executed in parallel. Let n be the number of the computing units (i.e., the cores in practice).

Theorem 3 (Efficiency). *In Distributed NE, suppose the initial graph is evenly distributed to each allocation process; the workload in each machine is evenly assigned to its computing units; and $|V|, |E| \gg |P|, n$. The worst-case time complexity of the local computation per unit is $\mathcal{O}\left(\frac{d|E|(|P|+d)}{n|P|}\right)$, where d is the maximum degree.*

Proof. The dominant part in the local computation is `AllocateTwoHopNeighbors()` in Algorithm 3. In the worst case, the number of messages in BP_{new} is $\mathcal{O}(|E|/|P|)$. In Line 12, $|BP_{new}|/n = \mathcal{O}(|E|/n|P|)$ vertices per unit are processed. For each vertex u, $\mathcal{O}(d)$ neighbors are processed. The complexity from Line 13 to 18 is $\mathcal{O}(|P|+d)$ since the time complexity is $\mathcal{O}(|P|)$ at Line 14,16 and $\mathcal{O}(d)$ at Line 17. Therefore, the total time complexity of the function is $\mathcal{O}\left(\frac{|E|}{n|P|}\right) \times \mathcal{O}(d) \times \mathcal{O}(|P|+d) = \mathcal{O}\left(\frac{d|E|(|P|+d)}{n|P|}\right)$ □

7. EMPIRICAL EVALUATION

In this section, we discuss the empirical analysis of the quality, scalability, and efficiency. Our claims are as follows:

Highest Quality. In various types of skewed graphs, Distributed NE always generates higher-quality partitions than any other state-of-the-art distributed methods.

Highest Scalability. Memory usage of Distributed NE is an order of magnitude smaller than that of the state-of-the-art high-quality distributed methods. It can handle a larger graph with fewer machines than the existing methods.

Comparable Efficiency. The elapsed time of Distributed NE is comparable to or better than the state-of-the-art high-quality distributed methods.

Trillion-edge Graph. Due to its scalability and efficiency, Distributed NE can generate high-quality partitions of the trillion-edge graph using only a few hundreds of machines.

7.1 Benchmarks and Setup

Real-world Datasets. For the evaluation, we use different graph datasets with over a billion of vertices, provided by SNAP [30], KONECT [26, 27] and LWA [3]. We use 7 real-world skewed graphs as summarized in Table 2.

Table 2: Real-world Skewed Graphs

Dataset	Vertices	Edges	Dataset	Vertices	Edges
Pokec [44]	1.63M	30.62M	Twitter [28]	41.65M	1.46B
Flickr [37]	2.30M	33.14M	Friendster [51]	65.60M	1.80B
LiveJ. [7]	4.84M	68.47M	WebUK [8]	105.15M	3.72B
Orkut [52]	3.07M	117.18M			

Synthetic Datasets and Trillion-edge Graph. Since each real-world dataset has the entirely different graph feature, the results would not provide any common performance characteristics. Thus, further experiments are conducted using various sizes of synthetically generated graphs which have the similar graph features. We use RMAT [12, 29] graphs whose vertex size are from Scale20 to Scale30, where *ScaleN* is referred to as a graph with 2^N vertices. Their average number of edges per vertex, called *edge factor* (*EF*), ranges from 2^4 to 2^{10} according to Graph500 setting [2] (the edge factor is 2^4) and Facebook's trillion-edge graph [14], which is reported to have 1.45 billions vertices and 1 trillion edges (thus, the edge factor is around 2^{10}).

The trillion-edge graph is simulated using RMAT with the same scale as Facebook's trillion-edge graph because no real-world trillion-edge graph is publicly available. The graph is Scale30 with Edge factor 2^{10}, namely, the graph consists of $1,073,741,824$ ($= 2^{30}$) vertices and $1,099,511,627,776$ ($= 2^{30} \times 2^{10}$) edges.

Benchmark Partitioning Algorithms. We compare *Distributed NE* with 8 distributed partitioning methods. *Random* is the simple one-dimensional hashing. *2D-Random* (or referred to as *Grid*) is the two-dimensional hashing, which is used for the initial assignment of *Distributed NE*. *Oblivious* [16] and *Hybrid Ginger* [13] are the state-of-the-art hash-based edge partitioning methods including iterative refinements. *Spinner* [36] is the state-of-the-art hash-based vertex partitioning method, where vertices are assigned randomly followed by the iterative refinements based on Label Propagation. *ParMETIS* [23] is the standard multi-level vertex partitioning. *Sheep* [35] is the state-of-the-art high-quality distributed edge partitioning method based on the elimination-tree conversion. *XtraPuLP* [42] is the state-of-the-art high-quality distributed vertex partitioning method, where vertices are directly assigned based on Label Propagation without initial random allocation. For the comparison of the partitioning quality with the vertex partitioning methods, such as *ParMETIS*, *Spinner*, and *XtraPuLP*, we convert their vertex-partitioned graphs into the edge-partitioned ones as demonstrated in [10], that is, each edge is randomly assigned to one of its adjacent vertices' partitions.

Parameter Setting. The imbalance factor α discussed in Section 2 is set to 1.1. Moreover, we set the expansion factor λ equal to 0.1 to balance the quality and performance as discussed in Section 5.

7.2 Quality Evaluation

The quality of partitioning is evaluated by means of the replication factor (RF) as Equation (1) in Section 2. We execute *Distributed NE* with five different random seeds and show the median value, where the relative standard error of the result is less than 5%.

Figure 8(a) to 8(g) show the replication factor of real-world graphs on different numbers of partitions from 4 to 64. Note that even on our big memory server (around 1TB), *ParMETIS* is unable to process `Twitter`, `Friendster`, and `WebUK` due to the insufficient memory space.

Overall, *Distributed NE* outperforms the other methods in real-world graphs. Especially in the severe cases, such as, a case of more partitions or a case where replication factor is relatively high (e.g., `Orkut`, `Pokec`, and `Friendster`), the improvement from the others is much more significant. Only in the smaller number of partitions or in graphs whose replication factor is low, the other methods are comparable to *Distributed NE*. For example, in `Flickr` and `Twitter` of 4 to 16 partitions, *Sheep* is slightly better than *Distributed NE*. In `LiveJ.`, *ParMETIS* is also slightly better than *Distributed NE* in 4 to 16 partitions. In `WebUK`, replication factor of *Distributed NE* is similar quality to that of *Sheep* and *XtraPuLP*. Their scores are less than 1.1, which is a nearly ideal case, that is, there are no vertex replications, and replication factor is equal to 1.

Moreover, the partitioning quality of *Distributed NE* is stable. It always provides high-quality partitions. On the other hand, the other methods have some specific graphs which are unsuited to partition. For example, *Sheep* gener-
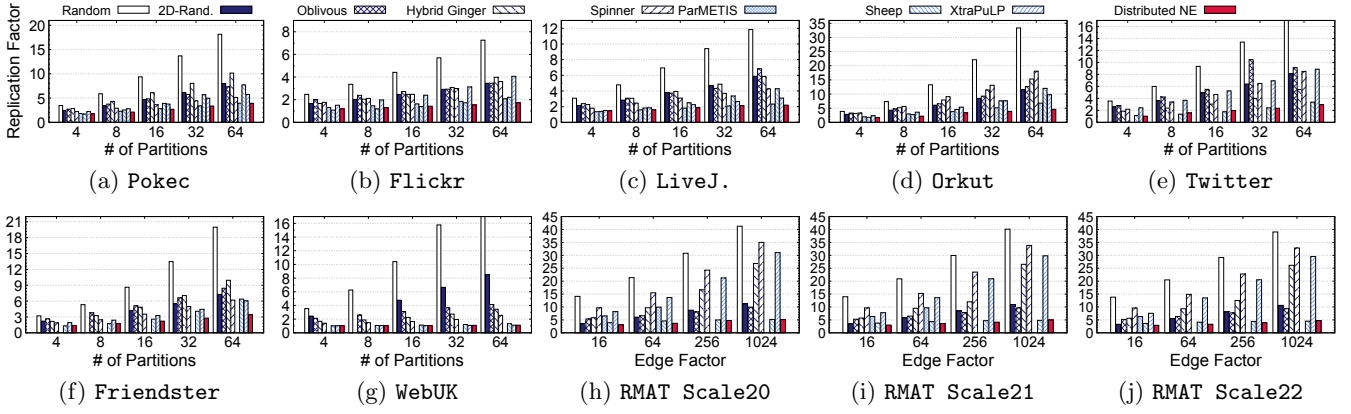
Figure 8: Replication Factor of Real-world Graphs and RMAT Graphs. $|P| = 64$ in RMAT Graphs.

ates high-quality partitions in `Twitter` but worse in `Pokec`, `LiveJ.`, `Orkut` and `Friendster`. *XtraPuLP* is significantly worse in `Twitter`, and `Friendster` and RMAT graphs discussed below.

Figure 8(h) to 8(j) show the replication factor of RMAT synthetic graphs on 64 partitions. Note that *ParMETIS* is unable to process 256 and 1024 in all scales due to the insufficient memory space.

As with the real-world graphs, *Distributed NE* outperforms the other methods. In general, the replication factor becomes higher as the increase of the edge factor. The result is intuitive since graphs become more complicated as the increase of edges, and thus it becomes more difficult to generate high-quality partitions. Moreover, in the same edge factor, the replication factor is almost the same in the different scales: Scale20, Scale21, and Scale22. This means that the difficulty in partitioning a graph depends on its complexity rather than its scale.

To summarize, we can classify all the methods into 3 categories. The first categories are based on the hashing, such as *Random, 2D-Random, Oblivious, Hybrid Ginger* and *Spinner*. These methods provide low quality due to the randomness. The methods of the second category indirectly solve the graph partitioning problem by converting it into the other similar problem, such as label propagation (*XtraPuLP*) or tree partitioning (*Sheep*). Such indirect methods provide high-quality partitions only for some cases. Finally, the methods of the last category directly solve the optimization problem based on the approximate algorithm, such as *ParMETIS* and *Distributed NE*. These methods stably generate high-quality partitions.

7.3 Performance Evaluation

The performance of the partitioning algorithms is evaluated. We compare the memory consumption and elapsed time of *Distribute NE* with the three high-quality partitioning methods: *ParMETIS, Sheep,* and *XtraPuLP*. Hash-based algorithms, such as *Random, Oblivious, Hybrid Ginger,* and *Spinner*, are omitted from the performance evaluation because the partitions by these algorithms do not reach to the sufficient quality as shown previously. Instead, these algorithms is efficient and scalable since they only include light-weight hash calculation and local refinements.

Table 3 shows the configuration of the distributed environment used to evaluate the performance. According to Facebook's work [14], where trillion-edge graph can be processed on only 200 machines, we use the similar scale of the

Table 3: Computational Environment

Service	ASPIRE 1 in NSCC Singapore
CPUs per Machine	Dual Sockets Intel E5-2690v3 (2×12 cores)
Memory / Network	98 GB per Machine / InfiniBand EDR
# of Machines	4 to 256 (2^2 to 2^8)
C++ Compiler / MPI	GCC 5.4 (–O3 flag) / IntelMPI 5.1.2
OS	Red Hat Server release 6.9

distributed environment consisting of up to 256 machines.

Memory Consumption. The scalability of the graph partitioning is evaluated with memory consumption. Figure 9(a) shows the memory consumption with the real-world graphs on 64 machines (64 processes). We take snapshots of all distributed processes every 0.5 second during execution to get their memory usage. Then, we use the snapshot, s_{max}, at which the total memory usage of the processes becomes maximum. The score is normalized by the number of edges, that is,

$$(Mem\ Score) := \frac{1}{|E|} \sum_{pr \in Pr} \{pr\text{'s Mem. Usage (byte) at } s_{max}\},$$

where pr is a process, and Pr is a set of the processes. Overall, *Distributed NE* outperform the other methods by around one order of magnitude. On average, its mem score is only 5.89 % of the other methods. The mem score slightly decreases as the increase of the graph size because the proportion of the edges to the total memory usage becomes decrease accordingly.

(a) Real-world Graphs   (b) RMAT Graphs
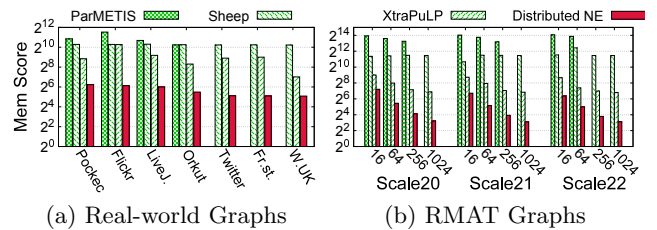
Figure 9: Memory Consumption. *Mem Score* is a total memory usage (byte) normalized by # edges.

Distributed NE is highly space-efficient because of two main reasons. First, each distributed edge is unique without any replication among machines, while vertices are replicated. In general, the replication of vertices is more space-efficient than that of edges since the vertex requires fewer bytes than the edge. In the vertex partitioning such as *ParMETIS* and *XtraPuLP*, the memory consumption may become higher since the edges are replicated among machines instead of vertices. Especially in the skewed graph,

Figure 10: Elapsed Time (sec) to Partition Real-world Graphs and RMAT Graphs.

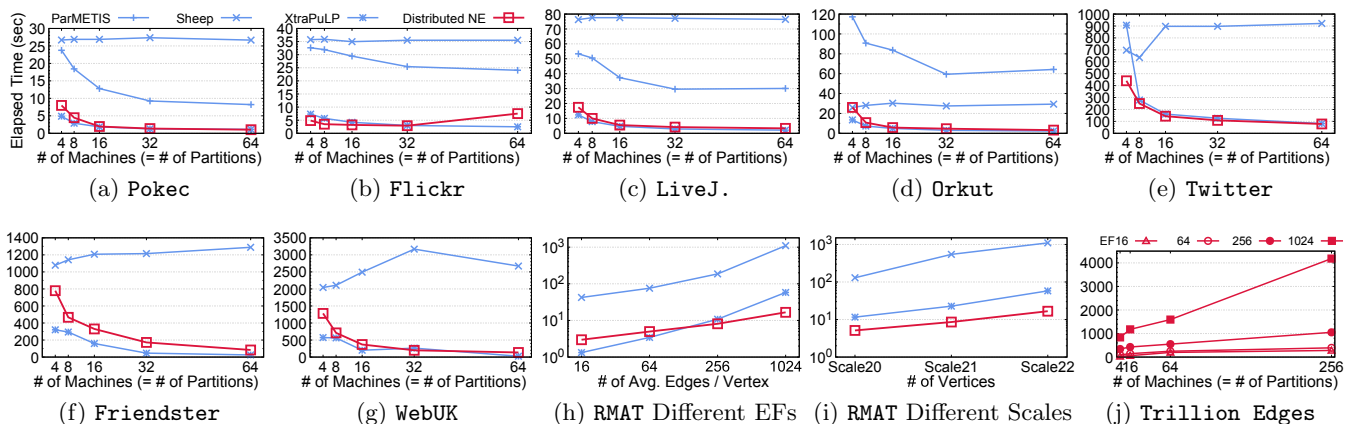the vertex replication is better since there are much more edges than vertices. Moreover, in the common multiple coarsening-refinement approaches such as *ParMETIS*, graph data are replicated multiple times for coarsening, and it requires much more memory than the others. Second, in *Distributed NE*, the graph data are stored without any memory-consuming data structure such as the hash map, which usually consumes around an order of magnitude memory space compared to the continuous array. The core components of the graph are stored in CSR, and their metadata is functionally computed instead of storing them.

Figure 9(b) shows the memory consumption of the RMAT graphs on 64 machines. *ParMETIS* is unable to run all scales with 1024 and Scale22 with 256 due to out-of-memory. As with the real-world graphs, *Distributed NE* is much better than the others. In *Distributed NE*, the mem score shows the substantial decrease as the increase of the edge factor because during computation it compacts the duplicated edges, which have the same sources and destinations. The duplication often appears in the higher edge factor.

Elapsed Time. The efficiency is evaluated with the elapsed time. We measure the time to compute partitions excluding the loading time to deploy input graph data. The conversion time from the vertex partition to the edge partition is also excluded in *XtraPuLP* and *ParMETIS*. We run the programs ten times and show the median value.

Figure 10(a) to 10(g) show the elapsed time in the real-world graphs. Overall, *Distributed NE* outperform *ParMETIS* and *Sheep*. In 64 partition, its speed up over *ParMETIS* is up to 9.1 in LiveJ.. The speed up over *Sheep* is up to 19.8 in Twitter. Its performance is basically comparable to *XtraPuLP* and slightly worse in Flickr on 64 machines and Friendster. In Flickr on 64 machines, *Distributed NE* takes lots of iterations at the final part of the computation because many isolated edges remain to be allocated. In the situation, edges are allocated by the random selection, but the number of allocated edges per iteration is few. Friendster is the special case for *XtraPuLP* to execute very fast as mentioned in the original paper [42].

Figure 10(h) shows the elapsed time to partition Scale22 into 64 partitions on the different edge factors (EFs). As the edge factor increases exponentially, so does the elapsed time in all algorithms. The rate of the increase in *Distributed NE* is lower than the others. As a result, *Distributed NE* is slower than *XtraPuLP* in 16 and 64, but it becomes faster in 256 and 1024.

Figure 10(i) shows the elapsed time of RMAT with Edge factor 1024 on the different scales using 64 machines. The result is straightforward. The elapsed time increases as the scale of RMAT graph. Their increasing rates are similar in all algorithms.

7.4 Scalability to Trillion-edge Graph

Figure 10(j) shows the scalability of *Distributed NE* to trillion-edge graph. We fix the number of vertices per machine as 2^{22} and change the number of machines (= the number of partitions). Namely, we use Scale24 on 4 machines, Scale26 on 16 machines, Scale28 on 64 machines, and Scale30 on 256 machines, respectively. As the number of machines increases, so does the elapsed time linearly. This is mainly because workload imbalance occurs in the vertex selection in expansion processes. Since the expansion rate is completely different in each partition, few partitions which have much more boundary vertices become the bottleneck of the entire processing. For example in Edge factor 1024 (EF 1024), the elapsed time of the vertex selection on 4 machines is less than 1 % of the entire processing, but it becomes 30.3 % on 256 machines. The communication cost also linearly increases as the increase of the number of machines.

Finally, due to the space-efficient design and implementation, *Distributed NE* can cope with the trillion-edge graph (Scale30, EF 1024) using only 256 machines. Its elapsed time is 69.7 minutes.

7.5 Comparison with Sequential Algorithms

Table 6 shows the comparison with the state-of-the-art sequential and streaming algorithms in the middle-scale real-world graphs. *NE* is the sequential offline algorithm, whereas *HDRF* and *SNE* are the sequential streaming algorithms. Each graph is partitioned into 64 subgraphs. Thus, *Distributed NE* are computed on 64 machines. Although *NE* provides the best replication factor, *Distributed NE* is much faster than the sequential algorithms.

Table 4: Comparison with Sequential Algorithms.

		Pokec	Flickr	LiveJ.	Orkut
RF	HDRF [39]	6.92	3.33	4.71	10.42
	NE [54]	**2.71**	**1.51**	**1.72**	**3.05**
	SNE [54]	3.89	1.78	2.12	5.66
	Distributed NE	*3.92*	*1.72*	*2.19*	*4.60*
Time (sec)	HDRF	24.310	24.370	57.228	92.479
	NE	61.890	62.910	143.690	182.288
	SNE	82.999	131.926	370.335	206.482
	Distributed NE	*1.029*	*7.523*	*3.309*	*3.224*

Table 5: Performance of Graph Applications (*SSSP*, *WCC*, *PageRank*) on 64 Partitions. `EB` is Edge Balance. `VB` is Vertex Balance. `ET` is Elapsed Time (sec). `COM` is COMmunication cost (GB). `WB` is Workload Balance.

		Flickr			Pokec			LiveJ.			Orkut			Twitter			FriendSter			WebUK		
		RF	EB	VB	RF	EB	VB	RF	EB	VB	RF	EB	VB	RF	EB	VB	RF	EB	VB	RF	EB	VB
Quality	Rand.	7.3	1.0	1.0	18.1	1.0	1.0	11.8	1.0	1.0	33.4	1.0	1.0	17.8	1.0	1.0	20.0	1.0	1.0	21.6	1.0	1.0
	2D-R.	4.4	1.0	1.0	9.1	1.0	1.0	6.8	1.0	1.0	12.7	1.0	1.0	9.1	1.0	1.0	8.3	1.0	1.0	10.1	1.0	1.0
	Obli.	6.3	1.7	1.1	13.6	1.6	1.1	9.0	1.1	1.0	20.9	1.3	1.0	13.8	1.0	1.0	14.3	1.0	1.0	4.0	1.3	1.0
	H.G.	4.0	1.2	1.0	10.2	1.2	1.1	6.0	1.1	1.1	14.3	2.5	1.1	5.5	1.3	1.1	9.6	1.3	1.0	3.4	1.0	1.0
	D.NE	*1.8*	*1.1*	*3.5*	*4.3*	*1.1*	*1.2*	*2.5*	*1.1*	*1.3*	*5.1*	*1.1*	*1.6*	*2.9*	*1.1*	*1.6*	*3.5*	*1.1*	*1.9*	*1.5*	*1.1*	*1.6*
		ET	COM	WB	ET	COM	WB	ET	COM	WB	ET	COM	WB	ET	COM	WB	ET	COM	WB	ET	COM	WB
SSSP	Rand.	2.96	1.78	1.58	2.91	3.10	1.46	4.08	6.02	1.41	4.45	11.3	1.25	22.7	87	1.15	50.3	146	1.20	88.4	254	1.27
	2D-R.	2.98	1.16	1.36	2.63	1.70	1.32	3.36	3.70	1.16	3.25	5.2	1.22	14.0	53	1.22	27.3	73	1.27	60.6	141	1.21
	Obli.	2.99	1.57	1.57	2.77	2.40	1.68	3.67	4.68	1.38	3.61	7.6	1.32	19.4	73	1.15	38.7	112	1.22	39.4	83	1.21
	H.G.	2.98	2.75	1.56	3.46	3.01	1.67	3.18	6.45	1.43	3.24	9.0	1.24	11.6	88	1.25	26.8	145	1.23	N/A	N/A	N/A
	D.NE	*2.94*	*0.63*	*1.28*	*2.63*	*1.03*	*1.42*	*3.15*	*1.83*	*1.46*	*2.48*	*3.1*	*1.71*	*7.8*	*30*	*1.34*	*17.6*	*44*	*1.42*	*28.5*	*58*	*1.43*
WCC	Rand.	4.77	3.87	1.30	6.58	8.33	1.30	10.08	14.7	1.25	17.50	31.1	1.16	89.3	156	1.18	286.0	406	1.12	396.2	733	1.16
	2D-R.	3.90	2.33	1.18	4.24	4.26	1.19	6.65	8.5	1.16	9.53	12.3	1.11	56.9	85	1.15	169.6	173	1.18	231.6	350	1.22
	Obli.	4.59	3.36	1.38	5.44	6.24	1.40	8.54	10.9	1.30	13.70	19.9	1.13	74.5	122	1.14	217.6	293	1.12	108.7	144	1.25
	H.G.	3.97	3.43	1.37	4.64	5.60	1.33	6.44	9.8	1.27	10.84	15.7	1.35	41.1	91	1.20	159.2	239	1.18	119	232	1.06
	D.NE	*3.48*	*0.74*	*1.31*	*3.55*	*1.94*	*1.30*	*4.69*	*2.7*	*1.34*	*7.09*	*5.2*	*1.24*	*31.1*	*31*	*1.28*	*115.3*	*71*	*1.26*	*61.2*	*55*	*1.25*
PageRank	Rand.	51.2	35.0	1.32	72.8	65.6	1.29	120.1	130	1.23	182.0	228	1.11	1568	1607	1.14	2820	2942	1.11	3370	3853	1.12
	2D-R.	36.2	19.8	1.14	45.4	32.6	1.13	79.1	71	1.13	93.2	91	1.05	863	798	1.11	1407	1239	1.07	1650	1826	1.09
	Obli.	45.6	28.9	1.38	63.0	51.2	1.39	100.7	96	1.28	129.2	147	1.10	1223	1252	1.14	2070	2112	1.12	769	776	1.15
	H.G.	31.1	14.9	1.23	41.3	24.4	1.26	61.8	43	1.33	87.1	74	1.14	446	462	1.19	1253	1151	1.20	682	687	1.06
	D.NE	*28.0*	*4.6*	*1.69*	*34.4*	*14.0*	*1.33*	*49.4*	*20*	*1.36*	*65.4*	*33*	*1.44*	*362*	*216*	*1.35*	*806*	*432*	*1.22*	*289*	*137*	*1.36*

7.6 Effect on Distributed Graph Applications

We briefly evaluate the effect of the partitioning methods on the distributed graph applications with the different real-world graphs (see Table 2). We use 3 common graph applications which have different communication patterns: Single Source Shortest Path (*SSSP*), Weakly Connected Component (*WCC*) and PageRank (*PR*). These are implemented on PowerLyra [13] forked from PowerGraph [16]. *SSSP* is the lightest workload and only involves a few communications; *WCC* is medium; and *PR* is the heaviest, where all the vertices send messages to their destinations in every iteration. We also refer the reader to the performance analysis of distributed graph applications, such as [19, 18, 47, 5]. Our result is consistent with the results presented in these analyses. In *SSSP*, we select Vertex 0 as the source. In *PR*, we conduct 100 times iterations.

Table 5 shows the result with regard to partitioning quality, balanceness, elapsed time, and communication cost on 64 machines. In these result, we ignore the initialization phase (e.g., system setup, data loading, and partitioning) to show only the application performance. We run each application 5 times and show the median. We compare *Distributed NE* (*D.NE*) with 4 methods in PowerLyra: *Random* (*Rand.*), *2D-Radom*(*2D-R.*), *Oblivious*(*Obli.*), and *Hybrid Ginger* (*H.G.*). N/A means that we cannot run the program correctly due to the system error by PowerLyra.

The balanceness among partitions is generally defined as $B(\{x_p|\ p \in P\}) := \frac{\max x_p}{\bar{x}}$, where $\bar{x} := \frac{\sum x_p}{|P|}$. Therefore, the edge balance (`EB`) is $B(\{|E_p|\})$; the vertex balance (`VB`) is $B(\{|V(E_p)|\})$; and the workload balance is $B(\{LT_p\})$, where LT_p is the local elapsed time in Partition p.

Overall, *Distributed NE* outperforms the others in the elapsed time for all the cases due to the reduction of the communication cost. Especially, the improvement of the elapsed time is significant in *PR* due to its high-communication workload; on the contrary, that of *SSSP* is few due to its low-communication workload.

Distributed NE achieves the good edge balance due to the algorithmic constraint, whereas the vertex balance becomes worse in some cases (e.g., `Flickr`). However, this does not seriously affect the elapsed time. Although Distributed NE is not explicitly aware of the vertex balance, it is not sig-

nificantly worse due to its algorithmic characteristics. That is, during computation, each $|V(E_p)|$ slightly increases at a similar speed because each part greedily selects the next expanding vertex so that the increase of $|V(E_p)|$ is as minimal as possible.

7.7 Evaluation with Non-skewed Graphs

We evaluate the effect of *Distributed NE* to 3 real-world road networks (California: 1.96M vertices and 2.76M edges; Pennsylvania: 1.08M vertices and 1.54M edges; and Texas: 1.37M vertices and 1.92M edges [31]) as the representative example of the large-scale non-skewed graphs. *Distributed NE* provides the similar or slightly better partitioning quality compared to the other methods. However, our claim is that traditional vertex partitioning would be a good choice in some cases because *Distributed NE* is basically developed for skewed graphs as we already discussed in Section 1.

Table 6: Replication Factor of Road Networks [31].

	Rand.	2D-R.	Obli.	H.G.	P.M.	Sheep	X.P.	*D.NE*
Calif.	3.72	3.54	2.13	2.32	**1.002**	1.03	1.12	*1.02*
Penn.	3.74	3.55	2.14	2.40	**1.004**	1.03	1.11	*1.01*
Tex.	3.70	3.51	2.13	2.35	**1.003**	1.03	1.12	*1.02*

8. CONCLUSION

In this paper, we presented Distributed NE, a parallel and distributed edge partitioning algorithm, which produces high-quality partitions of skewed graphs fast and at scale as well as provides the theoretical upper bound of the partitioning quality. There are two directions to improve our proposal. On one hand, the further speed-up technique will be established for coping with exascale graphs [46]. On the other hand, the extension to more complicated graph structures, such as dynamic graphs [21], and hypergraphs [22], will be investigated. **Distributed NE is publicly available in** `http://www.masahanai.jp/DistributedNE/`.

Acknowledgement

9. REFERENCES

[1] Apache Giraph. http://giraph.apache.org.

[2] Graph500. https://graph500.org/.

[3] Laboratory for web algorithmics (LWA). http://law.di.unimi.it/index.php.

[4] On two-dimensional sparse matrix partitioning: Models, methods, and a recipe. *SIAM J. SCI. COMPUT*, 32(2):656–683, 2010.

[5] Z. Abbas, V. Kalavri, P. Carbone, and V. Vlassov. Streaming graph partitioning: an experimental study. *PVLDB*, 11(11):1590–1603, 2018.

[6] Y. Ahmad, O. Khattab, A. Malik, A. Musleh, M. Hammoud, M. Kutlu, M. Shehata, and T. Elsayed. LA3: A scalable link- and locality-aware linear algebra-based graph analytics system. *PVLDB*, 11(8):920–933, 2018.

[7] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *KDD*, pages 44–54, 2006.

[8] P. Boldi, M. Santini, and S. Vigna. A large time-aware graph. *SIGIR Forum*, 42(2):33–38, 2008.

[9] E. G. Boman, K. D. Devine, and S. Rajamanickam. Scalable matrix computations on large scale-free graphs using 2d graph partitioning. In *SC*, pages 1–12, 2013.

[10] F. Bourse, M. Lelarge, and M. Vojnovic. Balanced graph edge partition. In *KDD*, pages 1456–1465, 2014.

[11] A. Buluç, H. Meyerhenke, I. Safro, P. Sanders, and C. Schulz. Recent advances in graph partitioning. *Algorithm Engineering*, pages 117–158, 2016.

[12] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-MAT: A recursive model for graph mining. In *SDM*, pages 442–446, 2004.

[13] R. Chen, J. Shi, Y. Chen, and H. Chen. PowerLyra: Differentiated graph computation and partitioning on skewed graphs. In *EuroSys*, pages 1:1–1:15, 2015.

[14] A. Ching, S. Edunov, M. Kabiljo, D. Logothetis, and S. Muthukrishnan. One trillion edges: Graph processing at Facebook-scale. *PVLDB*, 8(12):1804–1815, 2015.

[15] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[16] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin. PowerGraph: Distributed graph-parallel computation on natural graphs. In *OSDI*, pages 17–30, 2012.

[17] J. E. Gonzalez, R. S. Xin, A. Dave, D. Crankshaw, M. J. Franklin, and I. Stoica. GraphX: Graph processing in a distributed dataflow framework. In *OSDI*, pages 599–613, 2014.

[18] Y. Guo, M. Biczak, A. L. Varbanescu, A. Iosup, C. Martella, and T. L. Willke. How well do graph-processing platforms perform? an empirical performance evaluation and analysis. In *IPDPS*, pages 395–404, 2014.

[19] M. Han, K. Daudjee, K. Ammar, M. T. Özsu, X. Wang, and T. Jin. An experimental comparison of pregel-like graph processing systems. *PVLDB*, 7(12):1047–1058, 2014.

[20] S. Hong, S. Depner, T. Manhardt, J. Van Der Lugt, M. Verstraaten, and H. Chafi. PGX. D: A fast distributed graph processing engine. In *SC*, pages 58:1–58:12, 2015.

[21] J. Huang and D. J. Abadi. Leopard: Lightweight edge-oriented partitioning and replication for dynamic graphs. *PVLDB*, 9(7):540–551, 2016.

[22] I. Kabiljo, B. Karrer, M. Pundir, S. Pupyrev, and A. Shalita. Social hash partitioner: a scalable distributed hypergraph partitioner. *PVLDB*, 10(11):1418–1429, 2017.

[23] G. Karypis and V. Kumar. A parallel algorithm for multilevel graph partitioning and sparse matrix ordering. *J. Parallel Distrib. Comput.*, 48(1):71–95, 1998.

[24] Z. Khayyat, K. Awara, A. Alonazi, H. Jamjoom, D. Williams, and P. Kalnis. Mizan: a system for dynamic load balancing in large-scale graph processing. In *EuroSys*, pages 169–182, 2013.

[25] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The Web As a Graph: Measurements, Models, and Methods. In *COCOON*, pages 1–17, 1999.

[26] konect network dataset KONECT. http://konect.uni-koblenz.de.

[27] J. Kunegis. KONECT: The koblenz network collection. In *WWW Companion*, pages 1343–1350, 2013.

[28] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.

[29] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*, 11:985–1042, 2010.

[30] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data.

[31] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[32] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein. Distributed GraphLab: a framework for machine learning and data mining in the cloud. *PVLDB*, 5(8):716–727, 2012.

[33] S. Maass, C. Min, S. Kashyap, W. Kang, M. Kumar, and T. Kim. MOSAIC: Processing a trillion-edge graph on a single machine. In *EuroSys*, pages 527–543, 2017.

[34] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: a system for large-scale graph processing. In *SIGMOD*, pages 135–146, 2010.

[35] D. Margo and M. Seltzer. A scalable distributed graph partitioner. *PVLDB*, 8(12):1478–1489, 2015.

[36] C. Martella, D. Logothetis, A. Loukas, and G. Siganos. Spinner: Scalable graph partitioning in the cloud. In *ICDE*, pages 1083–1094, 2017.

[37] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *WOSN*, pages 25–30, 2008.

[38] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC*, pages 29–42, 2007.

[39] F. Petroni, L. Querzoni, K. Daudjee, S. Kamali, and G. Iacoboni. HDRF: Stream-based partitioning for power-law graphs. In *CIKM*, pages 243–252, 2015.

[40] A. Roy, I. Mihailovic, and W. Zwaenepoel. X-stream: Edge-centric graph processing using streaming partitions. In *SOSP*, pages 472–488, 2013.

[41] S. Salihoglu and J. Widom. Optimizing graph algorithms on pregel-like systems. *PVLDB*, 7(7):577–588, 2014.

[42] G. M. Slota, S. Rajamanickam, K. Devine, and K. Madduri. Partitioning trillion-edge graphs in minutes. In *IPDPS*, pages 646–655, 2017.

[43] T. Suzumura and K. Ueno. ScaleGraph: A high-performance library for billion-scale graph analytics. In *BigData*, pages 76–84, 2015.

[44] L. Takac and M. Zabovsky. Data analysis in public social networks. In *DTI*, volume 1, 2012.

[45] C. Tsourakakis, C. Gkantsidis, B. Radunovic, and M. Vojnovic. FENNEL: Streaming graph partitioning for massive scale graphs. In *WSDM*, pages 333–342, 2014.

[46] K. Ueno, T. Suzumura, N. Maruyama, K. Fujisawa, and S. Matsuoka. Efficient breadth-first search on massively parallel and distributed-memory machines. *DSE*, 2(1):22–35, 2017.

[47] S. Verma, L. M. Leslie, Y. Shin, and I. Gupta. An

[48] M. Wu, F. Yang, J. Xue, W. Xiao, Y. Miao, L. Wei, H. Lin, Y. Dai, and L. Zhou. GraM: scaling graph computation to the trillions. In *SOCC*, pages 408–421, 2015.

[49] C. Xie, L. Yan, W.-J. Li, and Z. Zhang. Distributed power-law graph computing: Theoretical and empirical analysis. In *NIPS*, pages 1673–1681, 2014.

[50] D. Yan, J. Cheng, Y. Lu, and W. Ng. Effective techniques for message reduction and load balancing in distributed graph computation. In *WWW*, pages 1307–1317, 2015.

[51] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *ICDM*, pages 745–754, 2012.

[52] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *KAIS*, 42(1):181–213, 2015.

[53] A. Yoo, A. H. Baker, R. Pearce, and V. E. Henson. A scalable eigensolver for large scale-free graphs using 2d graph partitioning. In *SC*, pages 63:1–63:11, 2011.

[54] C. Zhang, F. Wei, Q. Liu, Z. G. Tang, and Z. Li. Graph edge partitioning via neighborhood heuristic. In *KDD*, pages 605–614, 2017.

[55] X. Zhu, W. Chen, W. Zheng, and X. Ma. Gemini: A computation-centric distributed graph processing system. In *OSDI*, pages 301–316, 2016.

experimental comparison of partitioning strategies in distributed graph processing. *PVLDB*, 10(5):493–504, 2017.