SCALABLE LAYOUT OF LARGE GRAPHS ON DISK

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Abdurrahman Yaşar

June, 2015

Scalable Layout of Large Graphs on Disk
By Abdurrahman Yaşar
June, 2015

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Buğra Gedik (Advisor)

Assoc. Prof Dr. Hakan Ferhatosmanoğlu

Asst. Prof. Dr. Gültekin Kuyzu

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

SCALABLE LAYOUT OF LARGE GRAPHS ON DISK

Abdurrahman Yaşar
M.S. in Computer Engineering
Advisor: Assoc. Prof. Dr. Buğra Gedik
June, 2015

We are witnessing an enormous growth in social networks as well as in the volume of data generated by them. As a consequence, processing this massive amount of data has become a major problem. An important portion of this data is in the form of graphs. In recent years, several graph processing and management systems emerged to handle large-scale graphs. The primary goal of these systems is to run graph algorithms in an efficient and scalable manner. Unlike relational data, graphs are semi-structured in nature. Thus, storing and accessing graph data using secondary storage requires new solutions that can provide locality of access for graph processing workloads. In this work, we propose a novel scalable disk layout technique for graphs, which aims at reducing the I/O cost of disk-based graph processing algorithms. To achieve this goal, we designed a scalable Map/Reduce-style method called ICBP, which can divide the graph into a series of disk blocks that contain sub-graphs with high locality. Furthermore, ICBP can order the resulting blocks on the disk to further reduce non-local accesses. We experimentally evaluated ICBP to showcase its scalability, layout quality, as well as the effectiveness of automatic parameter tuning for ICBP. We also deployed the graph layouts generated by ICBP to the Neo4j [1] graph database management system. Our experimental results show that the default layout results in 1.5 to 2.5 times higher running times compared to ICBP.

Keywords: Block formation, disk layout, graph.

ÖZET

BÜYÜK ÇİZGELER İÇİN ÖLÇEKLENEBİLİR DİSK YERLEŞİMİ

Abdurrahman Yaşar
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Assoc. Prof. Dr. Buğra Gedik
Haziran, 2015

Son yıllarda sosyal ağ kullanımının hızlı bir şekilde yaygınlaşmasına tanıklık etmekteyiz. Bu yaygınlaşmanın neticesindeyse sosyal ağlar tarafından oluşturulan veri büyüklüğü devasa boyutlara geldi ve mevcut bu verinin işlenip anlamlandırılması gerek akademi gerekse sanayii için mühim bir konu haline dönüştü. Bu verinin büyük bir kısmıysa çizgeler halinde saklanmaktadır. Bu nedenledir ki son birkaç yılda büyük ölçekli çizgeleri işleyebilmek amacıyla pek çok sistem geliştirilmiştir. Bu sistemlerin öncelikli hedefleri ise mevcut çizge algoritmalarını büyük ölçekli çizgelerde etkin bir şekilde uygulanmasını saglamaktır. Fakat ilişkisel verilerin aksine çizgeler yarı yapısal temeldedir. Bu nedenle ikincil depolama alanları üzerinden çizgelere ulaşmak ve işlemek çizge içerisindeki bezerlikleri göz önüne alan farklı çözümlere ihtiyaç duymaktadır. Bu yüzden bu çalışmada disk üzerinde rastgele gerçekleştirilen okuma yazmaları indirgemek amacıyla çizgelerin disk üzerindeki yerleşimlerini ölçeklenebilir bir şekilde gerçekleyen bir metot önermekteyiz. Bu amaçla, ICBP adını verdiğimiz, çizgeleri dağıtık ve ölçeklenebilir bir şekilde öbeklere bölebilen bir metodu Hadoop yapısını baz alarak hayata geçirdik. Önerdiğimiz bu metot öbek oluşturmanın yanında oluşturulan bu öbeklerin disk üzerinde yerleşimini de sağlamaktadır. Bu çalışmada bu metodun detaylı açıklamasının beraberinde metodun etkinliğini, kalitesini ve ölçeklenebilirliğini deneysel olarak sunacağız.

Anahtar sözcükler: Öbek oluşturma, disk yerleşimi, çizge.

Acknowledgement

First of all, I would like to thank to my supervisor, Assoc. Prof. Dr. Buğra Gedik for his exceptional inspiration, guidance, support and always being available to me when I needed help during my graduate study. I have already learned a lot from him in these two years.

I owe special thanks to Assoc. Prof. Dr. Hakan Ferhatosmanoğlu, who contributed continuously through the design and development of the studies we explain in this thesis, for his valuable suggestions and patience throughout this study.

I am grateful to my jury member, Asst. Prof. Dr. Gültekin Kuyzu for reading and reviewing this thesis.

I thank my fellow labmates Semih Şahin, Dogukan Çağatay, Fuat Basik, Kaan Bingol, Elif Eser and Mehmet Güvercin for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last two years

I would like to thank to my family for supporting me with all my decisions and for their endless love, especially my mother and sister for their support and faith during my thesis work.

A very special acknowledgement goes to my girlfriend Nida Tangün, who loved and supported me all the time, and made me feel like anything was possible. I love you, Nida.

Contents

List of Figures

Chapter 1

Introduction

We are witnessing an enormous growth in social networks and the volume of data generated by them. An important portion of this data is in the form of graphs, which are popular data structures used to represent relationships between entities. For instance, the graph structure may represent the relationships in a social network, where finding communities in the graph [2] can facilitate targeted advertising. In the telco (telecommunications) domain, CDRs (call details reports) can be used to capture the call relationships between people [3], and locating closely connected groups of people can be used for generating promotions.

With the rise in the availability and volume of graph data, several graph processing and management systems have been introduced to handle large-scale graphs [4, 5, 6, 7, 8, 9, 10, 11]. The primary goal of these systems is to manage large graphs and execute graph algorithms on them in an efficient and scalable manner. In this work, we focus on disk-based graph management systems [1, 6], and propose the first parallel and scalable Map/Reduce based disk layout technique. Unlike relational data, graphs are semi-structured in nature. Thus, storing and accessing graph data using secondary storage requires new solutions that can provide locality of access for graph processing workloads.

Many graph algorithms rely on the fundamental operation of *graph traversal* and exhibit high access locality [12]. Given that a vertex is visited during a traversal, it is quite likely that the neighbors of this vertex will be visited shortly after. For instance, an n-hop breadth first search around a vertex exhibits high locality. This observation has motivated block-based disk layouts where the neighborlists of vertices that are highly connected (e.g., form a community) are placed into the same disk block [13]. This minimizes the number of blocks read, which reduces I/O. It also avoids the costly disk seeks, since chasing blocks often requires seeking to different areas of the disk.

In this work, we propose a novel scalable disk layout technique for graphs, which aims at reducing the I/O cost of disk-based graph processing algorithms. To achieve this goal, we designed a scalable Map/Reduce style method called ICBP, which can divide the graph into a series of disk blocks that contain sub-graphs with high locality, as well as order these blocks on disk to reduce non-local accesses. In this work, we describe the ICBP method, including the challenges that arose in applying ICBP in practice, the solutions applied, and an experimental evaluation showcasing its effectiveness.

Identifying vertices that are ‘close’ with respect to locality of access during execution of graph algorithms is a challenging problem. Although neighbor lists of vertices give some information about locality, it is not sufficient, as it is possible for close vertices to have very few common neighbors. To illustrate, we can think two hop neighbors of a vertex. Although the neighbor lists of these vertices may have very few common neighbors, in a large graph we can certainly define them as ‘close’ vertices. Accordingly, there should be a diffusion factor for each vertex, which can vary based on the graph size. In this work, we use random walks to produce *diffusion sets* of vertices. The idea behind building diffusion sets is simple: for each vertex, do some number of random walks and assign weights to vertices visited during the random walks. The resulting weighted sets of vertices can be used to define closeness between the originating vertices. At this point, we run into another challenge, namely defining the number of random walks and their lengths, based on the graph characteristics. We address this challenge by automatically tuning all ICBP parameters.

Once the closeness between vertices is defined, we can use it to form disk blocks by co-locating close vertices within the same blocks. This could be achieved by using bottom-up methods from the literature, such as hierarchical clustering. Yet, these methods have high computational complexity, leading to prohibitive costs for large-scale graphs. Thus, forming the disk blocks in a scalable manner is a challenging problem. In this work, we use a coarse partitioning algorithm to divide the large graph to in-memory processable sub-graphs. This coarse partitioning gives us the ability to apply a computationally heavier block formation algorithm on these sub-graphs, in parallel.

Since the size of the disk blocks are expected to be relatively small compared to the graph size, the generated blocks are expected to contain many connections to other blocks. Therefore, to benefit completely from the locality of blocks, they need to be ordered on disk by taking into account the inter block dependencies. In this work, we solve the problem of graph block ranking using a packing algorithm which is a label based packing that follows the process of formation of blocks. Packing algorithm simply orders the blocks based on their labels that were generated as part of the block formation phase. We have integrated this packing algorithm inside the block formation algorithm to avoid an additional stage of computation.

In the literature, block formation for graphs has been considered [13], yet the solutions are not parallel or scalable. When considering the size of social media graphs and Big Data workloads, performing the block formation in a scalable manner is an important task. In this work, we achieve scalability by implementing all parts of our proposed solution as Map/Reduce (M/R) jobs, executed on the Hadoop framework.

In summary, we make the following contributions:

- We propose an effective disk layout technique, ICBP, for large-scale graphs. ICBP is aimed at increasing the performance of disk-based graph management systems by increasing the locality of access of disk blocks.

- We develop Map/Reduce-based algorithms to implement ICBP, making the disk layout generation scalable, so that large-scale graphs can be divided into disk blocks using distributed processing.
- We propose evaluation metrics for measuring the efficacy of the ICBP disk layout technique and present an experimental evaluation showcasing its disk layout quality and running time scalability.
- We deployed the graph layouts generated by ICBP to the Neo4j [1] graph database management system to understand the impact of the layouts generated by ICBP on the performance of query evaluation in a graph database.

The rest of this thesis is organized as follows. In Chapter 2, we formalize our problem and evaluation metrics. Chapter 3 provides an high-level overview of our solution. Chapter 4 describes the ICBP technique in detail, explaining how our block formation algorithm works. Chapter 5 gives a detailed experimental evaluation of our work and Chapter 6 presents the related work. Chapter 7 concludes the thesis.

Chapter 2

Problem Definition

Most graph analytics require graph traversals, where vertex access patterns follow the connectivity structure of the graph. If the graph is laid out on the disk without considering these patterns, the traversal operations may cause too many I/O operations. This can create a bottleneck for graph processing and management systems. Therefore, storing and accessing graph data using secondary storage requires new solutions that can provide locality of access for graph processing workloads.

Locality of access for graph analytics executing on disk-based graph processing systems can be increased by locating graph vertices that are ‘close’ with respect to connectivity structure of the graph close on the disk as well. Figure 2.1 illustrates this. In the figure, we have a graph with 18 vertices stored on 6 blocks. Storing vertices in blocks aims to put close vertices together and increasing the locality of access. However, after generating locality-aware blocks, we still need to order these blocks on disk because there are inter-block dependencies between each other. In summary, our problem is composed by two sub-problems: First one is locality-aware block generation. The second one is ranking and ordering these block on disk.

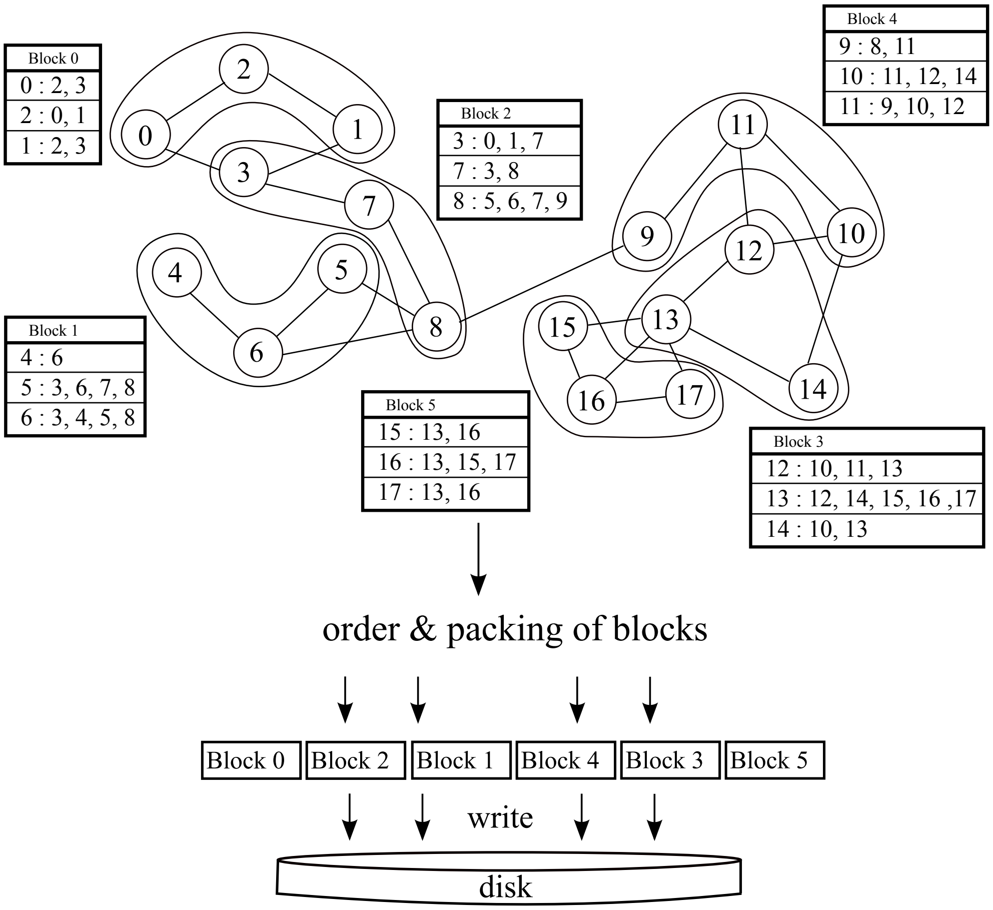

Figure 2.1: Toy graph illustrating block formation and ranking.

Illustrative Example. Assume that as part of a graph analytic task we need to access all vertices that are in 2-hop distance to vertex 0. 2-hop neighborhood of vertex 0 contains 4 vertices, which are: 1, 2, 3, and 7. In the first scenario, we consider that the assignment of vertices to blocks is being done randomly. In this case, all of the four vertices could have been assigned to different blocks, which would have resulted in 4 block accesses with a total of 12 vertex reads, resulting in 42% success rate (number of vertices used per vertex read). However, if we consider the block structure that is given in the Figure 2.1, we end up with 2 block accesses with a total of 4 vertex reads, resulting in 83% percent success rate. As we can observe in this example, locality-aware block generation decreases the number of block accesses and increases I/O performance.

Locality-aware block generation is highly critical to decrease the number of reads from disk, and ultimately, to optimize the efficiency of the system. However, if our secondary storage is an hard disk, seek time becomes important as well. In our running example, we need to access a number of blocks and if these blocks are randomly scattered on the disk, then to read a relatively small number of blocks, we would spend too much seek time. For instance, let us assume that blocks are ordered randomly on the disk as follows: 5, 2, 3, 4, 0, and 1. We need to access all vertices that are in 2-hop distance from vertex 0. To start, we need to access block 0, which is in the 5th position. Later, we must access block 2, which is in the 2nd position. This means that the disk needs to first seek to position 5 and then seek around back to position 2. However, if we use the layout that we defined in Figure 2.1, that is 0, 2, 1, 4, 3, and 5, we would avoid the additional seek. Since blocks 0 and 2 are sequential, accessing these two blocks requires only a single seek. In conclusion, with a smart ordering seek time can be decreased to improve I/O efficiency.

2.1 Notation

An undirected graph $G = (V, E)$ consists of a set of vertices V and a set of edges E. An edge is denoted as $e = (u, v) = (v, u) \in E$, where $u \neq v$ and

$u, v \in V$. The neighbor list of a vertex $u \in V$ is denoted as N_u, and defined as $N_u = \{v \in E \mid (u, v) \in E\}$. $\mathcal{N}$ represents the set of all neighbor lists, that is $\mathcal{N} = \{N_v \mid v \in V\}$. For instance, if we consider Figure 2.1, the neighbor list of vertex 0 is $N_0 = \{2, 3\}$ and $\mathcal{N}$ is $\{N_0, N_1, \cdots, N_6\}$.

Given a graph, we generate a set of blocks, denoted by $\mathcal{B}$. Each block $B \in \mathcal{B}$ contains at least one vertex and its neighbor list. Thus we can view a block as a non-empty subset of the set of all vertex-neighborlist pairs. Formally, $\forall B \in \mathcal{B},\ B \subset \{(u, N_u) \mid u \in V\}$ and $|B| > 0$. Blocks do not share their elements, that is $\forall_{\{B,B'\} \subset \mathcal{B}}, B \cap B' = \emptyset$. We denote the set of vertices in a block B as $V_B = \{u \mid (u, N_u) \in B\}$ and the set of neighbor lists as $\mathcal{N}_B = \{N_u \mid (u, N_u) \in B\}$. The set of blocks cover the entire graph G, that is $V = \bigcup_{B \in \mathcal{B}} V_B$. Finally, each block is limited in size by a block size threshold denoted by t. Let $s : \mathcal{B} \to \mathbb{N}$ be a function that assigns a size to a block, then we have $\forall B \in \mathcal{B},\ s(B) \leq t$.

We assume that blocks are laid out on the disk sequentially. The place of a block B on the disk is determined by its rank, denoted by $r(B)$. The rank of a block is simply the number of blocks that have been written before it. We have $0 \leq r(B) < |\mathcal{B}|$, and $\forall_{\{B,B'\} \subset \mathcal{B}}, r(B) \neq r(B')$. Finally, we define a function $d : \mathcal{B} \times \mathcal{B} \to \mathbb{N}$ that represents the distance between two blocks on the disk. We have $d(B, B') = |r(B) - r(B')|$.

2.2 Problem Formulation

Our problem has two aspects, namely *block formation* and *block ranking*. In the block formation problem, the aim is to generate blocks with high locality. We define the locality of a block B using a metric that measures how well connected the vertices within the block are and how well separated they are from the vertices in other blocks, denoted by $L(B)$. Thus, the goal is to maximize the total locality over all blocks, denoted by $L = \sum_{B \in \mathcal{B}} L(B)$.

In the block ranking problem, the aim is to assign close ranks to blocks that have many edges connecting them, so that they are close on the disk. We define the ranking locality of a block B using a metric that measures the on-disk distance of B to other blocks it has edges into, denoted by $R(B)$. Thus, the goal is to maximize the total locality over all blocks, denoted by $R = \sum_{B \in \mathcal{B}} R(B)$.

2.3 Metrics

Evaluation of our proposed system depends on the definition of block and block ranking localities. We now formally define these localities.

2.3.1 Block Locality

Locality of a block can be defined using two concepts: conductance and cohesiveness. Conductance is commonly used for graph partitioning. In our context it is defined as the ratio of the number of edge cuts to the total number of edges in a block. Formal definition of conductance is as follows:

$$C^d(B) = \frac{|\{(u, v) \in E \mid |\{u, v\} \cap V_B| = 1\}|}{|\{(u, v) \in E \mid |\{u, v\} \cap V_B| > 0\}|} \tag{2.1}$$

For example conductance of Block 0 in Figure 2.1 is $C^d(B_0) = \frac{2}{4} = 0.5$. Because, in the block, there are four edges, two of which are going out, that is $(0, 3)$ and $(1, 3)$.

Conductance of a block is not sufficient to determine the locality of a block. What is missing is the cohesiveness of the block. Cohesiveness is generally used for finding highly connected regions or communities in graphs. In this work we define cohesiveness of a block as the number of vertex pairs that are connected to each other via an edge in the block, divided by the total number of vertex pairs. Denoted by C^h, cohesiveness is formally defined as follows:

$$C^h(B) = \frac{|\{(u,v) \in E \mid u,v \in V_B\}|}{|B| \cdot (|B|-1)/2} \tag{2.2}$$

Again, if we consider Block 0 in Figure 2.1, cohesiveness of the block becomes $C^h(B_0) = \frac{2}{3} = 0.66$. Because in block there are 2 connected pairs of vertices, out of 3 possible connections.

These two metrics are complementary. Impact of dangling edges is captured by conductance and connectivity within a block is captured by cohesiveness. To obtain a high locality block, we need to increase cohesiveness, while decreasing conductance.

As a result, we define the locality of a block B, denoted by $L(B)$, as the geometric mean of cohesiveness and one minus the conductance. That is:

$$L(B) = \sqrt{C^h(B) \times (1 - C^d(B))} \tag{2.3}$$

Finally, if we apply this formula to Block 0, we obtain: $L(B_0) = \sqrt{0.33 \times (1-0.5)} = 0.41$.

2.3.2 Ranking Locality

We define ranking locality in terms of the distance between blocks of neighboring vertices. Let us denote the ranking distance a vertex $u \in V$ has to its neighbor vertices by $R(u)$. Formally, we have:

$$R(u) = \sum_{v \in N_u} d(r(u), r(v)) \tag{2.4}$$

Then the ranking locality for a block B is defined as:

$$R(B) = 1 - \frac{\sum_{u \in V_B} R(u)}{d_{max} \times \sum_{u \in V_B} |N_u|} \tag{2.5}$$

In this formula, d_{max} represents the maximum possible distance in the layout such that $d_{max} = max_{u,v \in V} d(r(v), r(u)$. When there are no edges going outside of a block, the ranking locality is 1. This is the ideal scenario. The ranking locality could be negative.

Chapter 3

Solution Overview

In this chapter, we give an overview of our solution to scalable layout of large-scale graphs. Our approach, named ICBP[1], consists of a multi-stage process, where each stage can be implemented in a scalable manner using map/reduce style parallelism.

3.1 General Approach

ICBP has three major stages. The first stage identifies the diffusion sets of vertices. The second stage performs coarse partitioning of the graph based on locality. It uses the diffusion sets from the first stage to guide the partitioning. The last two stages are used to form blocks and rank them. The forming of blocks and their ranking are implemented in an integrated manner to reduce the overhead of having an extra stage in the map/reduce flow. Figure 3.1 illustrates these stages.

[1] ICBP acronym is formed by the first letters of the four stages in our solution.

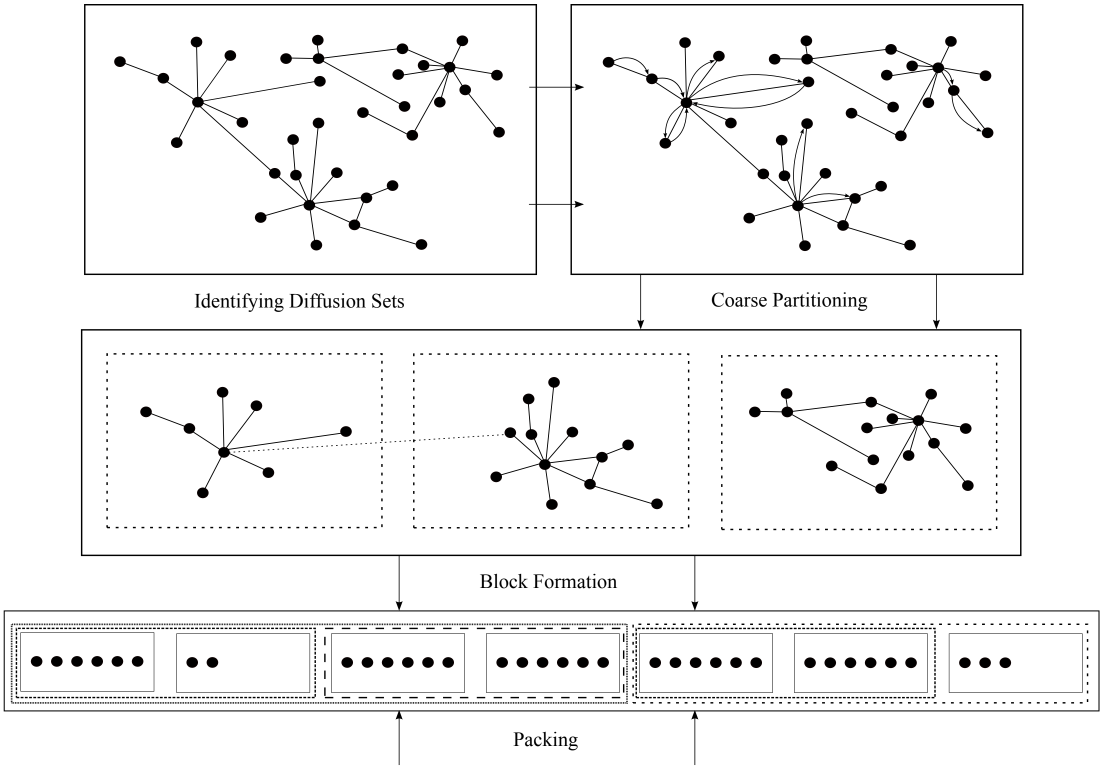

Figure 3.1: Solution overview.

3.1.1 Identifying Diffusion Sets

Diffusion set of a vertex is a summarized representation of its neighborhood in the graph, not limited to single-hop neighbors. It can be used to define closeness between vertices. To identify the diffusion set of a vertex, we perform random walks starting from the vertex and record the vertices visited, together with the number of times they have been visited during the random walks. The end result is a weighted set of vertices. We perform t random walks, each of length l. If we choose small values for l and t, then the neighborhoods will be sparse and thus similarities among neighborhoods of close vertices will be low. Conversely, if we choose large values for l and t, then many neighborhoods will end up looking similar, even if the vertices are not close. Also, large values will increase the computation time significantly, as diffusion sets are computed for each vertex. We address tuning of l and t in Section 4.1.

3.1.2 Coarse Partitioning

After identifying diffusion sets for each vertex in the graph, we divide the graph into k vertex-disjoint sub-graphs. Vertices that are close based on the similarity of their diffusion sets are co-located on the same sub-graphs, as much as possible. The goal of the coarse partitioning is to create sub-graphs that can fit into the memory available on a single machine. Furthermore, coarse partitioning also helps us create sufficiently small sub-graphs that are suitable for executing computationally more expensive block formation algorithms inspired by hierarchical clustering. Naturally, as the input graph becomes larger in size, the number of partitions we need tp create, that is k, increases as well. We address the tuning of k in Section 4.2.

3.1.3 Block Formation

Block formation is performed in a bottom-up fashion. Initially, each vertex is in a partition by itself. Then we successively merge pairs of partitions to create bigger partitions. Among the possible pairs, we pick the one that minimizes the distance between the diffusion sets of the vertices in the partitions. We further detail this in Section 4.3. If a partition exceeds the maximum block size, a block is formed from it. This block is output and removed from the partition. The block formation completes when all vertices are assigned to a block.

3.1.4 Packing

Packing is performed in an integrated manner as part of the block formation. When the block formation algorithm finalizes a block, the packing algorithm assigns a *rank label* to the block. This rank label is a multi-segment string that approximates the location of the block within the hierarchical merge-tree of the vertices. Ordering the resulting blocks based on their rank labels gives their rank. The base packing algorithm only orders blocks within the same coarse partition, as the packing is performed independently for different partitions. A post-packing algorithm applies the same logic to order the coarse partitions, to achieve the final global ranking.

3.2 Scalability

Since we aim to perform locality-aware block formation and ordering for large-scale graphs, scalability is a primary concern. Therefore, our entire solution is designed to be run as a series of map/reduce (MR) tasks.

First, an MR task transforms the input graph given in the form of an edge list into an adjacency list formatted graph. This step is not needed if the input graph is already in the adjacency list format.

Second, we use two MR tasks to form the diffusion sets. The first task is responsible for performing random walks and forming the vertex visit lists. The second job uses these lists to assign weights to vertices and form the final diffusion sets.

Third, we run a series of MR tasks to perform the coarse partitioning. The coarse partitioning is implemented as variation of iterative k-means clustering. An initial MR task is used to form initial partition centroids and the remaining tasks are used to perform a single iteration of a k-means clustering algorithm.

Last, we use an MR task to run the block formation and packing for each one of the coarse partitions we have created in the earlier stage.

Chapter 4

Scalable Block Formation & Ranking

In this chapter, we discuss the details of the four stages comprising ICBP. For each stage, we describe parameter tuning and scalable implementation techniques.

4.1 Identifying Diffusion Sets

Diffusion set of a vertex v, denoted by $\mathcal{D}_v$, is used to capture the close vertices around v based on the vertices visited during random walks that start from v. To find $\mathcal{D}_v$, we apply t random walks around v, each of length l. These random walks aim to locate vertices that are encountered close to each other during a graph traversal. We compute the diffusion sets for all vertices in the graph and implement it in a scalable manner using Map/Reduce. The more challenging aspect of identifying diffusion sets is tuning the parameters k and l based on the graph size and structure, which we discuss next.

Choosing t**.** Number of random walks (t) is critically important because if we set a too small t value, then the diffusion sets of vertices become very sparse and defining similarity of vertices using these sets becomes ineffective. Otherwise, if we set a too large t value, then the computation cost significantly increases without any benefit in terms of creating a diffusion set that can capture vertex similarity.

For a given graph, we define f as a cumulative distribution function of degrees, such that for $x \in \mathbb{N}$ $f(x) = P(d \leq x)$. In other words, $f(x)$ is the fraction of vertices that have a degree less than equal to x. Then we choose t as follows:

$$t = min\{x : f'(x) \leq \epsilon\} \tag{4.1}$$

Here, f' is the derivative of the cumulative degree distribution function f. In effect, we pick the smallest degree for which the distribution function's slope reaches ϵ. Our experimental evaluation has shown that choosing $\epsilon = 1.0$ gives robust results for varying graph sizes.

Choosing l. Vertex similarities are directly related to the setting of l. With large l values, the number of unique vertices that appear in diffusion sets increase and all vertices becomes similar. On the other hand, with small l values, the effectiveness of diffusion sets decreases as they become dissimilar even for close vertices.

In order to decide l, the first thing we should know is the diameter of the graph. Since social graphs exhibit small world phenomenon, their diameter can be estimated as the natural logarithm of the number of vertices they have, that is $ln(|V|)$. Accordingly, l should be at most $ln(|V|)$. Recall that after finding diffusion sets, we apply a coarse partitioning algorithm to divide the graph into k sub-graphs. Therefore we choose l so as to cover the space with a sub-graph, as follows:

$$l = 1 + \left\lceil \frac{ln(|V|)}{k} \right\rceil \tag{4.2}$$

M/R Implementation. t-l random walks are implemented via l repeated M/R jobs, each one producing the vertices visited during the next hop of the random

Algorithm 1: Random Walk Mapper

```
Param : $t$, the number of random walks; isFirst, whether this is the first job
Input : $\langle key, value \rangle$
if isFirst then
    let $\langle v, N_v \rangle = \langle key, value \rangle$
    for $t$ times do
        $u \leftarrow N_v[rand()]$
        output $\langle v, u \rangle$
else
    if value is a neighbor list then
        let $\langle u, N_u \rangle = \langle key, value \rangle$
        output $\langle u, N_u \rangle$
    else
        let $\langle v, u \rangle = \langle key, value \rangle$
        output $\langle u, v \rangle$
```

walks, followed by a final M/R job for creating the diffusion sets. During the first iteration, the mapper takes the input graph as the input as a list of vertex to neighbot list mappings. For each vertex, it chooses t random nodes from the neighbor list and sends each vertex, neighbor pair to the reducer. The reducer is an indentity reducer in the first iteration. The result is a file that contains the *initiator* vertex as the key, and the *visited* vertex as the value. This MR job is run for $l - 1$ more times after the first iteration. In the following iterations, the mapper takes the original graph and the output from the previous step as input. If a key/value pair comes from the original graph, then the mapper sends this pair directly to the reducer. If not, it switches the initiator with the visitor and sends the resulting pair to the reducer. This swapping enables joining the visited vertex with its neighbor list, so that the next vertex to visit can be determined at the reducer side. For each visited vertex, the reducer collects the initiators vertices plus the neighborlist of the visited vertex. For each initiator, it determines the next visited vertex using the neighborlist of the current one, and outputs an initiator, next visited vertex pair. Algorithms 1 and 2 give the pseudo-codes for the mapper and the reducer for the iterative steps of the random walks, respectively.

When l iterations are completed, the final M/R job combines all intermediate files and outputs the diffusion sets. Assigning weights to vertices in the diffusion

Algorithm 2: Random Walk Reducer

```
Param : $isFirst$, whether this is the first job
Input : $\langle key, values \rangle$
$N \leftarrow nil$                          ▷ neighbor list of last visited vertex
$V \leftarrow []$                           ▷ initiator vertices for last visited vertex
if $isFirst$ then
    let $\langle v, U \rangle = \langle key, values \rangle$
    foreach $u \in U$ do
        output $\langle v, u \rangle$
else
    let $u = key$
    foreach $value \in values$ do
        if $value$ is a neighbor list then
            let $N_u = value$
            $N \leftarrow N_u$
        else
            let $v = value$
            $V \leftarrow V + [v]$
    foreach $v \in V$ do
        output $\langle v, N[rand()] \rangle$
```

sets is an important task performed by this last task, because it identifies the vertices that are commonly visited (closer). We tested our system with several alternatives for the weighting:

- non-weighted diffusion paths,
- occurrence count based weighted diffusion sets, and
- tf-idf based weighted diffusion sets.

Tf-idf based weights are computed by treating each diffusion set as a document and using the traditional term frequency times inverse document frequency formulation from information retrieval [14]. In our context, the term frequency is the weight of a vertex in the diffusion set. The inverse document frequency for a vertex is the logarithm of the ratio of the total number of vertices to the number of diffusion sets that contain the vertex.

These weight assignment policies are compared in the experimental evaluation section in terms of their running times and quality of locality they provide.

4.2 Coarse Partitioning

After identifying diffusion sets for each vertex in the graph, we divide the graph into k vertex-disjoint sub-graphs as part of the coarse partitioning stage. The goal of the coarse partitioning is to create sub-graphs that can fit into the memory available on a single machine. Furthermore, coarse partitioning also helps us create sufficiently small sub-graphs that are suitable for executing computationally more expensive block formation algorithms inspired by hierarchical clustering.

Our coarse partitioning algorithm is based on k-means [15]. As such, we first choose a set of k initial centers, denoted by $\mathcal{C}$, from the graph. Then, for each vertex $v \in V$, we find the closest center $c \in \mathcal{C}$ and assign v to the cluster of c. After all vertices are assigned, we obtain a list of vertices for each cluster, denoted as as V_c for center c. We then calculate the new centers, that is we update $\mathcal{C}$, by reducing V_c into a new center value replacing the old one. The process is repeated until convergence, detected based on comparing the difference between the new and old clusters to a threshold.

We now describe the various details of the algorithm, such as the distance metric we use, setting the value of k, and determining the initial centers. We then provide a brief description of the M/R implementation.

Distance Metric. To determine closeness of vertex pairs we need to define a distance metric. Since diffusion sets are just weighted sets of vertices, we use a weighted Jaccard distance for this purpose. *Jaccard similarity* of two sets S and T is the ratio of the size of their intersection to the size of their union, that is $\frac{|S \cap T|}{|S \cup T|}$. If we apply this in our context for two vertices u, $v \in V$, we get $JS(u, v) = \frac{|\mathcal{D}_u \cap \mathcal{D}_v|}{|\mathcal{D}_u \cup \mathcal{D}_v|}$. As we mentioned before, the vertices in diffusion paths could be weighted, in which case we have a weighted Jaccard similarity, defined as $JS_w(u, v) = \frac{\sum_{x \in \mathcal{D}_v \cap \mathcal{D}_u} min\{w(x,\mathcal{D}v), w(x,\mathcal{D}_u)\}}{\sum_{x \in \mathcal{D}_v \cup \mathcal{D}_u} max\{w(x,\mathcal{D}_v), w(x,\mathcal{D}_u)\}}$. Here, $w(x, \mathcal{D})$ represents the weight of vertex x in diffusion set $\mathcal{D}$, After defining the similarity between two vertices, the Jaccard Distance between them is simply: $JD(u, v) = 1 - JS_w(u, v)$.

Choosing k. Tuning the k parameter is crucial because coarse partitioning aims to divide the graph into in-memory processable sub-graphs for the following block formation stage. Therefore, if we choose a too small k value, then we can run out of memory in the block formation stage. On the other hand, if we choose a too large k value, then we increase the processing time for the coarse partitioning stage and we also lose the locality effect that will is needed for the block formation stage to form blocks with high locality. Assume that all cores in our cluster has M byte of memory and a vertex's size is s byte. Then we choose k as follows:

$$k = \left\lceil \frac{s \times |V|}{\sqrt{0.8 * M}} \right\rceil \tag{4.3}$$

Is summary, we make k as small as possible without utilizing more than 80% of the main memory on a node.

Initial Centers. One option to decide on the initial centers is to choose them randomly. However, this has caused unstable performance both in terms of convergence of the coarse partitioning stage as well as the locality of the resulting blocks for the ICBP method. Instead, we came up with a more effective way of setting the initial centers. The idea is to pick k vertices that are distant to each other and have high degrees. These can be considered as influence centers in the graph. To compute them, we added an M/R job to the system to sort the vertices by degree. We then process this list, starting form the highest degree vertex. If a vertex has a distance 0.9 or more to all of the the previously selected ones, we select it as a center vertex. We stop when k vertices are selected.

Deciding Center Size. Cluster centers are weighted sets, just like the diffusion sets. Recall that at the end of each iteration of k-means, we have to form new centers. The size of these centers is also an important factor. If we choose a too small size, then coarse partitioning converges too fast and the resulting clustering have poor locality. If the size is too large, then the this delays convergence. We set the center size to the average length of the diffusion paths within a cluster. In our empirical study, this setting has resulted in good quality sub-graphs and has shows good convergence behavior.

Algorithm 3: Coarse Partitioning Mapper

```
Param : $\mathcal{C}$, set of centers, where for $c \in C$, $c.id$ is the center id and $c.S$ is the
        diffusion set for the center.
Input : $\langle key, value \rangle$

let $\langle v, \mathcal{D}_v \rangle = \langle key, value \rangle$
$c \leftarrow \text{argmin}_{c \in \mathcal{C}} JD(\mathcal{D}_v, c.S)$
output $\langle c.id, \mathcal{D}_v \rangle$
```

Algorithm 4: Coarse Partitioning Reducer

```
Param : $isLast$, whether this is the last job
Input : $\langle key, values \rangle$

$O \leftarrow \{\}$                        ▷ Map from vertex to in-cluster occurrence count
$size \leftarrow 0$                        ▷ Average diffusion set size in cluster
let $cId = key$                            ▷ key is the cluster id
if not $isLast$ then
    foreach $value \in values$ do
        let $\mathcal{D} = value$          ▷ each value is a diffusion set
        foreach $v \in \mathcal{D}$ do
            $O[v] \leftarrow O[v] + 1$
        $size \leftarrow size + |\mathcal{D}|$
    $size \leftarrow size / |values|$
    $\mathcal{D} \leftarrow \text{argtop-}k_{v \in O} O[v]$, where $k = size$
    $c \leftarrow tuple(id{=}cId, S{=}\mathcal{D})$
    output $\langle cId, c \rangle$
else
    foreach $value \in values$ do
        let $\mathcal{D} = value$
        $c \leftarrow tuple(id{=}cId, S{=}\mathcal{D})$
        output $\langle cId, c \rangle$
```

M/R Implementation. Coarse partitioning implemented via repeated sequential M/R jobs. The first iteration, takes a set of initial centers denoted by $\mathcal{C}$. Other sequential jobs produce the new centers for following iterations until the final M/R job. We produce new centers simply by counting number of occurances of vertices in that cluster and getting the most frequent ones. In the final job we generate clusters. Algorithms 3 and 4 give the pseudo-codes for the mapper and the reducer for the coarse partitioning stage, respectively.

4.3 Block Formation

During block formation, vertices are placed into partitions in a bottom-up fashion. Each vertex starts in its own partition and partitions are successively merged by picking the closest pair of partitions at each step. We define the closeness of two partitions as the minimum Jaccard distance between the diffusion sets of the vertices contained within. For partitions P and P', this is given as $\min\{JD(\mathcal{D}_u, \mathcal{D}_v) : u \in P \wedge v \in P'\}$. When the size of a potential block that would be formed from vertices in the partition without a block assigned so far exceeds the maximum block size, then a full block is formed and output. The block formation completes when all vertices are assigned to a block.

Super blocks. In large graphs that exhibit power law [16] degree distribution, popular nodes require special treatment. If we take the Twitter graph as an example, a user with millions of followers becomes an exceptional case because the size of his/her neighbor list exceeds the block size. In such exceptional cases, we divide the neigbor list of the vertex into multiple block sized partitions. We define a block that points to multiple such partitioned blocks a *super block*.

Block labeling. We assign labels to blocks for helping with the last stage of the ICBP solution, that is packing. For this purpose, during the execution of the block formation algorithm, each partition maintains a label. This partition label is used to derive the block label later. It captures the merge history of partitions with respect to blocks. Initially, each partition has its vertex id as its label. When two partitions merge, this label is updated as follows: If the two partitions have not produced a block before, the new label is taken as the label of the larger partition. If only one of them has formed a block before, then its label is taken as the partition label. Finally, if both of the partitions have produced a block before, then the label is taken as the concatenation (using `":"` as a delimiter) of the two labels, label of the bigger partition appearing on the left. When a block is produced, it gets the label of its partition, with an additional suffix (using `"."` as a separator) representing the index among blocks generated with the same partition label. Figure 4.1) shows an example block formation process, where numbers represent the order in which the partitions are merged. The partition

Algorithm 5: Block Formation Algorithm

```
Param : $S$, block size; $V$: the set of vertices in the sub-graph
$\mathcal{B} \leftarrow \emptyset$                                        ▷ Blocks to be generated
$\mathcal{P} \leftarrow \bigcup_{v \in V} \{tuple(l{=}str(v), i{=}false, V{=}[v], U{=}\{v\})\}$
while $|\mathcal{P}| > 1$ do
    $\{P, P'\} \leftarrow \text{argmin}_{\{P,P'\} \subseteq \mathcal{P}}$
            $\min\{JD(\mathcal{D}_u, \mathcal{D}_v) : u \in P.U \wedge v \in P'.U\}$
    ▷ Setup the partition label
    let $P_n = \text{argmin}_{P'' \in \{P,P'\}} |P''.U|$                 ▷ Small partition
    let $P_x = P''\ s.t.\ P'' \neq P_n \wedge P'' \in \{P, P'\}$           ▷ Large part.
    if $P_n.i \wedge P_x.i$ then $P_x.l \leftarrow P_x.l + \text{“ : ”} + P_n.l$
    else if $\neg P_x.i \wedge P_n.i$ then $P_x.l \leftarrow P_n.l$
    ▷ Merge the partitions
    $\mathcal{P} \leftarrow \mathcal{P} \setminus \{P_n\}$
    $P_x.U \leftarrow P_x.U \cup P_n.U$
    $P_x.V \leftarrow P_x.V \cup P_n.V$
    if $blockSize(P_x.V) \geq S$ then
        $P_x.i \leftarrow true$                              ▷ Remember generation of block
        $k \leftarrow \max\{k : blockSize(P_x.V[0{:}k]) \leq S\}$
        $V' \leftarrow P_x.V[0{:}k]$                              ▷ Vertices to form a block
        $B \leftarrow \{(v, N_v) : v \in V'\}$                          ▷ Form the block
        $\mathcal{B} \leftarrow \mathcal{B} \cup B$
        $P_x.V \leftarrow P_x.V \setminus V'$                     ▷ Update unassigned vertices
return $\mathcal{B}$
```

labels are indicated on tree edges representing the merges. Blocks are marked with dotted boxes and their block lables are indicated next to the boxes.

M/R implementation. Block formation is implemented with a single M/R job, making use of only the map operation. Each map performs block formation on one of the sub-graphs generated by the coarse partitioning stage and produce blocks with their associated labels. Algorithm 5 gives the pseudo-code for this process.

4.4 Packing

Social graphs exhibit small-world behavior, and thus most vertices are reachable from each other via a small number of hops. Therefore, even with locality-aware

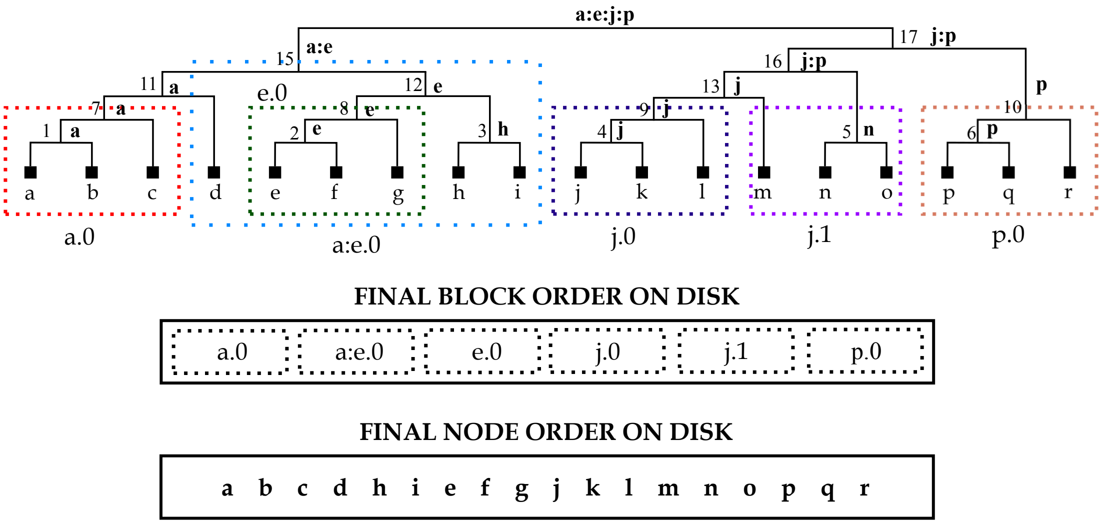
a:e:j:p
17 j:p
15 a:e
16 j:p
11 a
12 e
13 j
p
e.0
7 a
8 e
9 j
10
1 a
2 e
3 h
4 j
5 n
6 p
a b c d e f g h i j k l m n o p q r
a.0
a:e.0
j.0
j.1
p.0
FINAL BLOCK ORDER ON DISK
a.0 a:e.0 e.0 j.0 j.1 p.0
FINAL NODE ORDER ON DISK
a b c d h i e f g j k l m n o p q r

Figure 4.1: Illustration of packing.

block formation, we will have many edges crossing between blocks. With the packing algorithm, we aim to provide a locality-aware disk layout for graphs by considering inter block similarities. Primary goal of packing process is to store similar blocks close on disk.

The packing algorithm simply orders the blocks based on their labels that were generated as part of the block formation phase. Before the sort, we replace the vertex names that appear in the block labels with their order in the leaves of the hierarchical merge tree. Then sorting the blocks by their labels locate blocks that were close in the merge tree close on the disk as well.

For instance, in Figure 4.1, first nodes a and b are merged. then e and f, and so on. As you can see, we construct a tree in a bottom up manner. In this toy example, for brevity and ease of exposition, we assume that all vertices have the same degree d and size limit for a block is $3 \times (d + 1)$, thus only 3 vertices fit in a block. We observe that in the 7^{th} iteration, the vertices $[a, b, c]$ reach the size limit and block formation algorithm generates them as a block. This block is labeled as $a.0$ by taking the partition label at the time of block generation (a in this case), and the index among the blocks that are generated with that partition label (0 in this case). This procedure continues to create blocks out of vertices $[e, f, g]$, $[j, k, l]$, $[m, n, o]$, and $[p, q, r]$.

In Figure 4.1, the block that contains vertices $[d, h, i]$ is different, because the vertices in this block are not contiguous at the leaf level. In the 11^{th} step, the partition that contains d merges with the partition that has earlier produced block $a.0$. And in the 12^{th} step, the partition that contains h and i merges with the partition that earlier produced block $e.0$. Finally, 15^{th} step, we merge these two partitions. The resulting partitions gets the label $a : e$, because the constituent partitions both have produced blocks earlier. Since the partition reaches the maximum size, a new block that contains the vertices $[d, h, i]$ and has label $a : e.0$ is generated.

Finally, when block formation is completed, we order blocks by sorting their labels. The end result is seen at the bottom of Figure 4.1.

Recall that this packing procedure is performed for each sub-graph in parallel. Once the order of blocks with each sub-graph is determined, a sequential version of the same process is applied across sub-graphs, by treating each sub-graph as a virtual vertex and pre-computing the distances among them based on the number of edges going in-between. The end result is an ordering that specifies which sub-graph blocks go first on the disk.

Chapter 5

Experimental Evaluation

In this section, we evaluate our system, with a special focus on the impact of the proposed optimizations on locality and scalability. Scalability experiments evaluate the running time of our ICBP algorithm as a function of number of cores used and the size of the graph. Locality experiments evaluate the performance of ICBP using locality metrics, as well as query running times using an industrial graph database system.

5.1 Experimental Setup

We first provide details on our implementation, evaluation environment, the datasets used, and the metrics employed in our evaluation.

5.1.1 Implementation

Our implementation was done in Java 1.7 using Hadoop v2.6 [17] framework. For evaluation of the coarse partitioning method we use Metis [18] graph partitioning tool and for evaluation of the layout we use Neo4j [1] graph database. For workload generation, we use RMAT [19] implementation of Boost Graph Library [20].

5.1.2 Environment

For running the ICBP algorithm, we used a cluster with 8 machines and a total of 96 cores. Each machine has 2 Intel Xeon E5-2620 2.00GHz processors and 32GB of memory. Each processor has 6 cores and our implementation use all of these cores. Each machine has 1TB disk space, made of up of 4 IBM Server X 5400 SATA disks using RAID-5. The operating system used was CentOS GNU/Linux with the 2.6 kernel and ext4 filesystem. It is worth noting that our evaluation heavily focuses on scalability and impacts of optimizations on locality and not on absolute performance. In the experiments where we evaluate the performance of our disk layout using the Neo4j [1] graph database, we use a Macbook-Pro with an Intel i5 processor and 4GB of memory.

5.1.3 Data Sets

We used R-MAT [19] generated graphs, as well as real-world graphs obtained from SNAP [21].

Synthetic Data: In our experiments, we use R-MAT generated power-law graphs with small world properties. The R-MAT graph generator provides an efficient way for generating very large realistic graphs. We apply our ICBP method to the graphs generated by R-MAT and analyze their locality and running time performance under different configurations. In our testes we use RMAT graphs with different sizes, where the number of edges is taken as 20 times the number of vertices.

Real Data: In addition to the RMAT graphs, we also selected several small, medium, and large size graphs from SNAP. These graphs are: *ego-Facebook* (4039 vertices, 88234 edges), *wiki-Vote* (7,115 vertices, 103,689 edges), *wiki-Talk* (2,394,385 vertices, 5,021,410 edges).

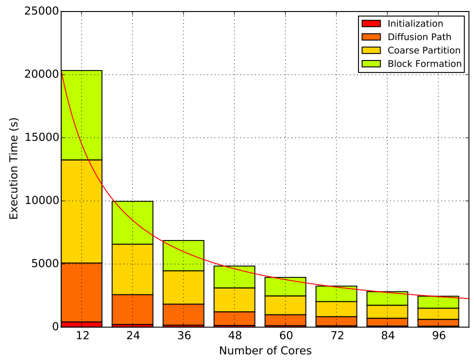

Figure 5.1: Scalability w.r.t. # of cores.

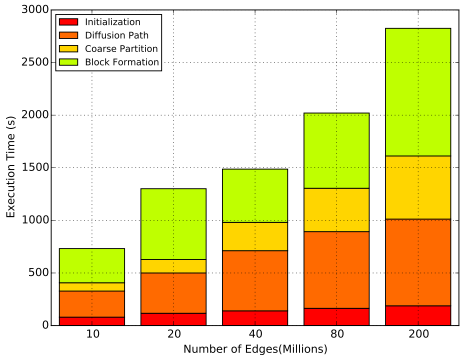

Figure 5.2: Scalability w.r.t. # of edges.

5.2 Scalability

Figure 5.1 shows the running time of the ICBP method as a function of the number of cores used. The graph used in this experiment is an 80 million edge R-MAT graph. Each bar represents the total amount of time the ICBP algorithm took to generate the disk layout. The different colored sub-bars represent the time taken by different stages on the ICBP method. The first sub-bar represent *initialization*, which is used to convert the initial graph from edge list representation to adjacency list representation. The second sub-bar represents forming the diffusion sets and the third bar represents coarse partitioning. The fourth and final sub-bar represents block formation, which also performs packing. The figure also shows an *ideal line* representing perfect scale-up. Figure 5.2 shows the running time with the same breakdown, but as a function of the number of edges. Graphs used in this experiment are 10, 20, 40, 80 and 200 million edge R-MAT graphs.

We observe from Figures 5.1 and 5.2 that initialization step takes negligible time compared to other stages, as it is very light on computation. Among the remaining stages, forming the diffusion sets is cheaper than coarse partitioning and block formation, but in general the distribution is quite balanced, especially with increasing number of cores. The most striking observation from Figure 5.1 is about scalability. We see that ICBP method's running time with increasing core sizes closely matches the running times represented by the ideal scale-up line.

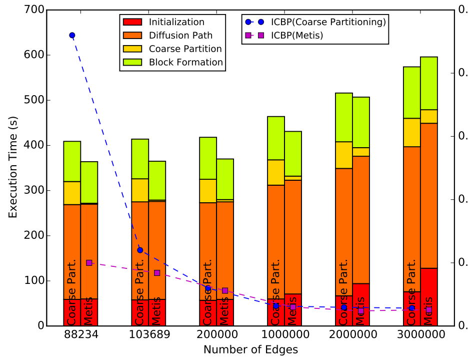

Figure 5.3: ICBP with Metis & coarse partitioning.

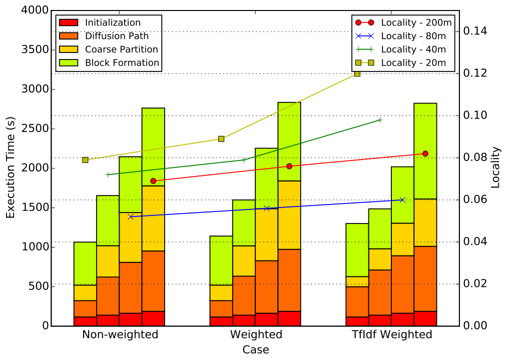

Figure 5.4: Assigning weights to diffusion sets.

We observe from Figure 5.2 that the running time is not always linear in the number of edges. t parameter is one of the key factors which determine the size of diffusion sets. In our parameter selection policy, t doesn't increase proportional to number of edges of the graph, instead it increases slowly. Therefore, the size of diffusion so the running time is not always linear in the number of edges.

5.3 Locality

In this section, we study the effectivenesses of our proposed optimizations on the locality of the layouts generated by ICBP.

5.3.1 Effectiveness of Coarse Partitioning

Coarse partitioning plays an important role in ICBP, as the localities of the generated blocks are affected by the quality of the sub-graphs generated by coarse partitioning. To understand the effectiveness of coarse partitioning, we compare it to a more traditional approach: graph partitioning.

Metis [18] is one of the popular and effective graph partitioning methods in the

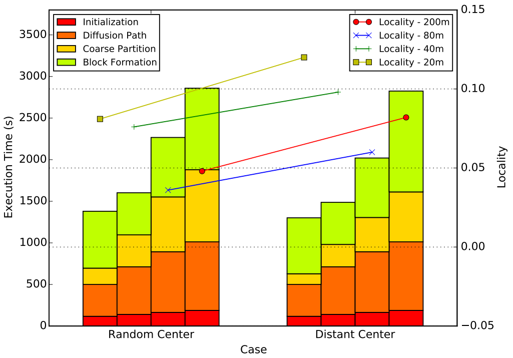

Figure 5.5: Choosing initial centers.

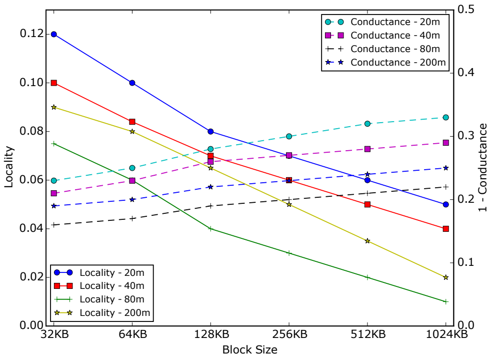

Figure 5.6: Locality vs. block size.

literature and it produces high-quality graph partitions. Therefore, in this experiment, we compared the results from ICBP with those from a variant of ICBP where the coarse partitioning is replaced by graph partitioning. The graph partitioning aims to minimize the edge cut, while balancing the number of vertices in each partition. Figure 5.3 plots the locality of the resulting blocks, as a function of graph size. We use 6 different graphs for this purpose. The first two graphs are real graphs from SNAP, namely *ego-Facebook* and *wiki-Vote*, and the last four ones are generated using R-MAT.

From Figure 5.3 we observe that for small graphs (especially the first real-world graph), ICBP with Metis can lead to improved locality compared to using ICBP with coarse partitioning. However, for larger graphs, the localities achieved by the two approaches are identical. We prefer coarse partitioning over Metis due to its scalability and integration into ICBP's Hadoop framework, as well as its good locality for large-graphs that is the focus of this work. Figure 5.3 also shows that Metis starts to take more time as the graph size is increased. Furthermore, pre-processing also starts to take more time for Metis, as the graph needs to be converted into the input format of Metis. The time taken by coarse partitioning, on the other hand, is not effected as much from the number of vertices, even though in absolute terms it takes more time than Metis for smaller graph sizes. For 300 million edges, ICBP with coarse partitioning starts to take less time compared to Metis. While there are parallel, scalable versions of Metis [22], they do not integrate well with our M/R framework.

5.3.2 Assigning weights

Having weighted diffusion sets helps us better capture similarity for vertices, which in turn is expected to improve block locality. To understand the impact of weight assignment on the locality of the generated blocks, we compared three alternatives schemes: non-weighted, occurrence counts as weights, and tf-idf weights computed over occurrence counts. For the weighted schemes, it is important to note that during random walks, the host vertex is assumed to be visited as the first vertex.

Figure 5.4 plots the execution time of ICBP (using the left y-axis) and locality (using the right y-axis), for different weighting schemes and for R-MAT generated graphs of different sizes (20, 40, 80, and 200 million edges).

We observe from Figure 5.4 that for all graphs sizes, tf-idf based weight assignment improves locality compared to non-weighted and occurence count based weighted cases, with relative improvements ranging from 20% to 50%. Since tf-idf based weights decrease the importance of very popular vertices in diffusion sets, this type of weight assignment improves the quality of sub-graphs that are generated with coarse partitioning by reducing the tendency of vertices to accumulate in one cluster.

5.3.3 Choosing Centers

During coarse partitioning, in each iteration, we assign vertices to clusters based on the similarity of their diffusion sets to cluster centers. The initial center selection for coarse partitioning impacts these iterations, and thus the locality and convergence.

In this experiment we examine two center selection strategies, namely *random* and *distant*. The first selection strategy is to choose randomly selected k host vertices and their adjacency lists as centers. The second selection approach is to choose k most distant and highest degree host vertices and their adjacency lists as initial centers, as explained earlier in Section 4.2. For this experiment, we again used RMAT-generated graphs.

Figure 5.5 plots the execution time of ICBP (using the left y-axis) and locality (using the right y-axis), for the two center selection schemes and for 4 different graph sizes (20, 40, 80, and 200 million edges).

We see that initial center selection strategies impact the convergence speed of coarse partitioning. Based on our experiments, we have observed that starting coarse partitioning with randomly selected centers from the graph sometimes requires more iterations to converge. The 40 million edge graph is a good example of this in Figure 5.5, where the coarse partitioning takes almost two times longer with random center selection.

From Figure 5.5, we also observe that initial center selection strategy impacts locality. For all graph sizes, the distant center selection strategy outperforms the random one, up to 30% in some cases.

Although distant center selection strategy improves locality and speeds up convergence, in some cases it also increases the time taken by the following stage of ICBP, that is block formation. This can be observed for the 200 million edge graph in Figure 5.5. Still, ICBP with distant center selection completes faster than random selection, for all graph sizes. The reason block formation sometimes takes longer with distant center selection is that, higher quality sub-graphs formed by it may have higher skew in their sizes, resulting in load imbalance during the block formation stage.

5.3.4 Locality and Block Size

In this experiment we examine the effect of block size on locality. We apply ICBP with blocks of size 32, 64, 128, 256, 512, and 1024 KBs. We use R-MAT graphs with differing sizes and measure locality.

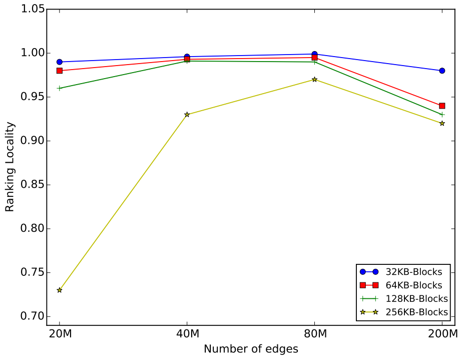

Figure 5.7: Ranking locality vs. graph size.

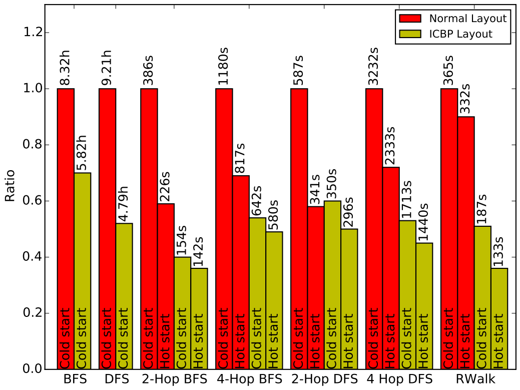

Figure 5.8: Query running times with Neo4j.

Figure 5.6 plots locality as a function of the block size, for graphs of different sizes. The overall locality is shown on the left y-axis and 1 - conductance is shown on the right y-axis. Since cohesiveness has a term that graphs quadratically with the number of vertices in a block, it brings down the overall locality significantly. Thus, we also show conductance separately in this experiment. We observe that, as the block sizes increase, the conductance decreases. This is intuitive, as if there was only a single block, then conductance would have been 1. However, the overall locality decreases as the block size increases, due to the impact of cohesiveness.

5.3.5 Ranking Locality

In this experiment, we evaluate ranking locality for different graph and block sizes. We use Equation 2.3.2 to compute ranking localities over all disk blocks. We use distant center selection and tf-idf weight assignment strategies. The graphs used are R-MAT generated.

Figure 5.7 plots ranking locality as a function of graph size, for different block sizes. Overall, ranking localities are high. An important observation from the figure is about the sensitiveness of ranking locality to graph size. Small blocks are more resilient to changes in the graph size. In fact, 32KB blocks have ranking localities almost independent of graph size. On the other hand, 256KB blocks show high variation in locality as the graph size changes, compared to smaller block sizes.

5.3.6 Query Running Times

To understand the impact of the layouts generated by ICBP on the performance of query evaluation in a graph database, we deployed the graph layouts generated by ICBP to the Neo4j [1] graph database management system. For this experiment, we used the 80 million edge R-MAT graph. To evaluate query performance, we used global BFS and DFS queries, limited hop BFS and DFS queries and random walks. The limited hop queries were run 100 times and the average results are reported. These graph algorithms were implemented using the Java API provided by Neo4j [1].

Deployment of the ICBP generated layout to Neo4j is performed in two stages. First stage is for preparation and the second one is for generation of the Neo4j specific files on the disk. Neo4j stores graphs in separate files and uses a modified version of edge list format to represent relationships between vertices. Since, Neo4j doesn't have a specific block notion and uses edge lists, our adjacency list based block structure needs to be converted. In preparation stage we do this conversion with two steps. First we merge blocks according to packing order and obtain a single file; and second we transform this file into edge list format. After the edge list file is generated, we create a second file which stores vertices in the order of their appearance in the edge list. These two files become inputs of the second stage. In the generation stage, we create Neo4j specific files using Map/Reduce jobs, consisting of consecutive join, union, and ascending sort jobs.

Figure 5.8 shows the running times of the algorithms, normalized with respect to Neo4j's default layout (labeled as Normal in the graph). We also have absolute running times as annotations in the figure. We show running times for both cold start and hot start cases, except for the global queries for which a hot cache does not make a difference (since the query touches the entire database). We observe that the default layout of Neo4j has 43% and 92% higher running times compared to ICBP for the BFS and DFS algorithms, respectively. For the cold start case using limited hop queries, the default layout results in 1.5 to 2.5 times higher running times compared to ICBP. The relative results are similar even for the hot start case, except for 2-hop DFS where the normal layout and ICBP perform similarly.

Chapter 6

Related Work

With the popularization of social networks and availability of large amounts of relationship data in the form of graphs, graph data management and mining became an important area of research and development. A survey can be found here [23].

Graph representation is used frequently in many domains, such as social media and telecommunications. For example, we can model the relationships in a social network using graphs and finding communities in the graph [2] can facilitate targeted advertising. In the telco domain, CDRs (call details reports) can be used to capture the call relationships between people [3], and locating closely connected groups of people can be used for generating promotions. To handle the graph processing and management needs of an increasing number of applications in diverse domains, several graph processing and management systems have been introduced to handle large-scale graphs [4, 5, 6, 7, 8, 9, 10, 11, 24, 25]. The primary goal of these systems is to manage large graphs and execute graph algorithms on them in an efficient and scalable manner.

In this work, we focus on disk-based graph management systems [1, 6]. Unlike relational data, graphs are semi-structured in nature. Thus, storing and accessing graph data using secondary storage requires new solutions that can provide locality of access for graph processing workloads. In the literature there are several works which try to increase efficiency of graph management systems, like [13] and [26].

One of the primary contributions of our work is the scalable block formation algorithm used to generate locality-aware blocks by storing close vertices in the same blocks as much as possible. A relevant work in this area is the disk layout techniques proposed by Hoque and Gupta [13] called Bondhu. Bondhu [13] presents a strategy for storing a social graph on disk. In this work they use community structures of social graph as a placement strategy. Using this strategy they optimize the disk layout, so that graph traversals can be performed using less I/O. Unlike their work, ICBP is a distributed graph layout algorithm (based on Map/Reduce) that can scale to large graphs.

In [26], Nodine et al. studies the graph search problem for large graphs that cannot fit into the main memory by trying to use blocks on disk efficiently. In their work, they have shown that optimizing the blocking has increased the performance of searching complete d-ary trees and d-dimensional grid graphs.

In [27], Gedik et al. have proposed a system for temporal storage and querying of evolving interaction graphs. In this work they proposed several online block formation algorithms that are used to reduce the I/O required to answer queries. Besides, they have proposed and applied several locality metrics to analyze graph blocks. In contrast to our work, their graphs are not relationship graphs, but instead append-only interaction graphs with a temporal aspect. As a result, their algorithms are streaming in nature.

GBASE [10] is a disk-based graph management system. It is related to our work in the sense that, it is a Map/ Reduce [28] based large-scale graph management system. It employs a graph storage method that relies on block compression to efficiently store homogeneous regions of graphs, and a grid based technique to efficiently place blocks into files. However, the system is not optimized for locality-awareness.

In [29], Akyürek et al. describes an adaptive technique for reducing disk seek times. To achieve this goal they copy a number of frequently referenced disk blocks to a reserved area near the middle of the disk from their current locations. Block rearrangement is related with our work, because similarly we also need to arrange and order graph blocks on disk to achieve good performance. In [29], the arrangement of blocks are done based on block access frequencies and in our work we do it based on block similarities.

BORG [30] is a self-optimizing layer in the storage stack. It reorganizes data on disk by looking at access patterns. BORG aims to optimize read and write traffic dynamically by making them more sequential. This work is relevant with ours, in which we aim to organize locality-aware blocks of a graph on disk and make reads more sequential.

TurboGraph [31] is designed as a single PC graph processing system. It leverages the advantages of low latancy and random I/O capabilities of SSDs. Although TurboGraph performs really well on SSD based disks, due to its parallel random I/O dependent design, it performs poorly on conventional magnetic disks.

In [32], Xie et al. propose a novel block-oriented computation model. In their model, computations are performed by iterating over locality-aware blocks. Although their computation model is in vertex-centric programming abstraction, instead of executing one vertex at a time they execute one block at a time and achieve good cache performance.

Neo4j [1] is a commercial disk-based graph management system. Although Neo4j has implements optimizations such as indexing and caching, its on-disk graph layout can be improved to increase query performance. In this work, we have shown that locality-aware layouts generated by ICBP can be used to improve Neo4j's query performance by a factor of 2 or more.

In [33], Dominguez-Sal et al. studies the characteristics of the graphs which are essential for benchmarks, and also the characteristics of the queries that are important in graph analysis applications. This study mainly helped us to determine graph characteristics that are useful for our methodology and parameter selection.

Chapter 7

Conclusion

We have developed a scalable system that generates locality aware blocks for large graphs. The system maintains a method called ICBP, which can divide the graph into a series of disk blocks that contain sub-graphs with high locality. Furthermore, ICBP can order the resulting blocks on the disk to further reduce non-local accesses. We experimentally evaluated ICBP to showcase its scalability, layout quality, as well as the effectiveness of automatic parameter tuning for ICBP. We demonstrated that ICBP is an effective disk layout technique, for large-scale graphs and it increases the performance of disk-based graph management systems by increasing the locality of access of disk blocks. As we have shown in our first experiment; ICBP makes the disk layout generation scalable, so that large-scale graphs can be divided into disk blocks using distributed processing. Finally, we proposed evaluation metrics for measuring the efficacy of the ICBP disk layout technique and present an experimental evaluation showcasing its disk layout quality and running time scalability.

Bibliography

[1] "Neo4j open source graph database." `http://neo4j.org/`, retrieved January, 2010.

[2] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 483, no. 3-5, pp. 75–174, 2009.

[3] A. A. Nanavati, G. Siva, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi, "On the structural properties of massive telecom call graphs: findings and implications," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 435–444, 2006.

[4] 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 *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pp. 135–146, 2010.

[5] 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," *Proceedings of the VLDB Endowment*, vol. 5, no. 8, pp. 716–727, 2012.

[6] A. Kyrola, G. Blelloch, and C. Guestrin, "GraphChi: Large-scale graph computation on just a PC," in *Proceedings of the USENIX Symposium on Operating System Design and Implementation (OSDI)*, pp. 31–46, 2012.

[7] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin, "PowerGraph: Distributed graph-parallel computation on natural graphs," in *Proceedings of the USENIX Symposium on Operating System Design and Implementation (OSDI)*, pp. 17–30, 2012.

[8] B. Shao, H. Wang, and Y. Li, "Trinity: A distributed graph engine on a memory cloud," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, 2013.

[9] J. Mondal and A. Deshpande, "Managing large dynamic graphs efficiently," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pp. 145–156, 2012.

[10] U. Kang, H. Tong, J. Sun, C.-Y. Lin, and C. Faloutsos, "Gbase: A scalable and general graph management system," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data mining (SIGKDD)*, pp. 1091–1099, 2011.

[11] "Apache Giraph." `giraph.apache.org/`, retrieved June, 2013.

[12] R. Steinhaus, "G-Store: A storage manager for graph data," Master's thesis, University of Oxford, 2011.

[13] I. Hoque and I. Gupta, "Disk layout techniques for online social network data," *IEEE Computing*, vol. 16, no. 3, pp. 24–36, 2012.

[14] A. Rajaraman and J. D. Ullman, "Data mining," in *Mining of Massive Datasets*, pp. 1–17, Cambridge University Press, 2011.

[15] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 281–297, 1967.

[16] M. Newman, "Power laws, pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[17] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino,

O. O'Malley, S. Radia, B. Reed, and E. Baldeschwieler, "Apache Hadoop YARN: Yet another resource negotiator," in *Proceedings of the Annual Symposium on Cloud Computing (SOCC)*, pp. 5:1–5:16, 2013.

[18] G. Karypis and V. Kumar, "Multilevel graph partitioning schemes," in *Proceedings of the International Conference on Parallel Processing (ICPP)*, pp. 113–122, 1995.

[19] D. Chakrabarti, Y. Zhan, and C. Faloutsos, "R-MAT: A recursive model for graph mining," in *In Fourth SIAM International Conference on Data Mining*, 2004.

[20] J. G. Siek, L.-Q. Lee, and A. Lumsdaine, *Boost Graph Library, The: User Guide and Reference Manual*. Addison-Wesley, 2002.

[21] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection." `http://snap.stanford.edu/data`, retrieved June, 2014.

[22] D. Lasalle and G. Karypis, "Multi-threaded graph partitioning," in *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing (IPDPS)*, pp. 225–236, 2013.

[23] C. Aggarwal and H. Wang, "Graph data management and mining," in *A survey of algorithms and applications* (C. Aggarwal, ed.), Springer, 2010.

[24] R. S. Xin, J. E. Gonzalez, M. J. Franklin, and I. Stoica, "Graphx: A resilient distributed graph system on spark," in *First International Workshop on Graph Data Management Experiences and Systems*, pp. 2:1–2:6, 2013.

[25] V. Prabhakaran, M. Wu, X. Weng, F. McSherry, L. Zhou, and M. Haridasan, "Managing large graphs on multi-cores with graph awareness," in *Proceedings of the 2012 USENIX Conference on Annual Technical Conference*, pp. 4–4, 2012.

[26] M. H. Nodine, M. T. Goodrich, and J. S. Vitter, "Blocking for external graph searching," *Algorithmica*, vol. 16, no. 2, pp. 181–214, 1996.

[27] B. Gedik and R. Bordawekar, "Disk-based management of interaction graphs," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 26, no. 11, pp. 2689–2702, 2014.

[28] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in *Proceedings of the USENIX Symposium on Operating System Design and Implementation (OSDI)*, pp. 137–150, 2004.

[29] S. Akyurek and K. Salem, "Adaptive block rearrangement," *ACM Trans. Comput. Syst.*, vol. 13, no. 2, pp. 89–121, 1995.

[30] M. Bhadkamkar, J. Guerra, L. Useche, S. Burnett, J. Liptak, R. Rangaswami, and V. Hristidis, "Borg: Block-reorganization for self-optimizing storage systems," in *Proccedings of the 7th Conference on File and Storage Technologies*, pp. 183–196, 2009.

[31] W.-S. Han, S. Lee, K. Park, J.-H. Lee, M.-S. Kim, J. Kim, and H. Yu, "Turbograph: A fast parallel graph engine handling billion-scale graphs in a single pc," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 77–85, 2013.

[32] W. Xie, G. Wang, D. Bindel, A. Demers, and J. Gehrke, "Fast iterative graph computation with block updates," *Proceedings of the Very Large Databases Conference (PVLDB)*, vol. 6, no. 14, pp. 2014–2025, 2013.

[33] D. Dominguez-Sal, N. Martinez-Bazan, V. Muntes-Mulero, P. Baleta, and J. Larriba-Pey, "A discussion on the design of graph database benchmarks," in *Performance Evaluation, Measurement and Characterization of Complex Systems* (R. Nambiar and M. Poess, eds.), Springer Berlin Heidelberg, 2011.