GrapH: Heterogeneity-Aware Graph Computation with Adaptive Partitioning

Christian Mayer, Muhammad Adnan Tariq, Chen Li, Kurt Rothermel
Institute of Parallel and Distributed Systems, University of Stuttgart, Germany
Email: {christian.mayer, adnan.tariq, chen.li, kurt.rothermel}@ipvs.uni-stuttgart.de

***Abstract*—Vertex-centric graph processing systems such as Pregel, PowerGraph, or GraphX recently gained popularity due to their superior performance of data analytics on graph-structured data. These systems exploit the graph structure to improve data access locality during computation, making use of specialized graph partitioning algorithms. Recent partitioning techniques assume a uniform and constant amount of data exchanged between graph vertices (i.e., uniform vertex traffic) and homogeneous underlying network costs. However, in real-world scenarios vertex traffic and network costs are heterogeneous. This leads to suboptimal partitioning decisions and inefficient graph processing. To this end, we designed GraphH, the first graph processing system using vertex-cut graph partitioning that considers both, diverse vertex traffic and heterogeneous network, to minimize overall communication costs. The main idea is to avoid frequent communication over expensive network links using an adaptive edge migration strategy. Our evaluations show an improvement of 60% in communication costs compared to state-of-the-art partitioning approaches.**

I. Introduction

In recent years, a strong demand to perform complex data analytics on graph-structured data sets, such as the web graph, simulation grids, Bayesian networks, and social networks [1]–[4] has lead to the advent of distributed graph processing systems, such as Pregel, PowerGraph, and GraphX [5]–[7]. These systems adopt a user-friendly programming paradigm, where users specify vertex functions to be executed in parallel on vertices distributed across machines ("think-like-a-vertex"). During execution, vertices iteratively compute their local state based on the state of neighboring vertices, therefore efficient communication across vertices is vital in building highly-efficient graph processing systems. In fact, recent work on data analytics frameworks suggests that network-related costs are often the bottleneck for overall computation [8]–[11].

To overcome these inefficiencies, graph processing systems require suitable partitioning methods improving the locality of vertex communication. Mainly, there are two types of partitioning strategies: edge-cut and vertex-cut. These strategies minimize the number of times an edge or vertex spans multiple machines (*cut-size*). The idea is that a decreased cut-size leads to lower communication costs due to less inter-machine traffic [6], [7]. But this holds only under two assumptions: *vertex traffic homogeneity*, i.e., processing each vertex involves the same amount of communication overhead, and *network homogeneity*, i.e., the underlying network links between each pair of machines have the same usage costs (e.g., [6], [11]). However, these assumptions oversimplify the target objective, i.e., **minimize overall communication costs**, for two reasons.

First, real-world vertex traffic is rarely homogeneous. This is due to computational hotspots causing processing to be unevenly distributed across graph areas and vertices. Hotspots arise mainly in three cases: i) the vertices process different amounts of data, ii) the graph system executes vertices a different number of times, and iii) the graph analytic algorithms concentrate on specific graph areas. Examples of the first case are large-scale simulations of heart cells [12], liquids or gases in motion [13], and car traffic in cities [1], where each vertex is responsible for a small part of the overall simulation. Vertices simulating real-world hotspots (e.g., the Times Square in NY) have to process more data. The second case consists of algorithms defining a convergence criteria for vertices. The graph system skips execution of converged vertices (*dynamic scheduling* [6]) leading to inactive graph areas and therefore different frequencies of vertex execution. Concerning this, a popular example is the PageRank algorithm [6]. The third case includes user-centric graph analytic algorithms such as k-hop random walk and graph pattern matching. A prominent example is Facebook Graph Search, where users pose search queries to the system ("find friends who tried this restaurant"). In general, our evaluations show that vertex traffic often resembles a Pareto distribution, whereby a higher percentage of the total traffic is contributed by a much lower percentage of the vertices (cf. Fig. 2). We argue, that the one-size-fits-all approaches for vertex traffic misfit real-world, heterogeneous and dynamic traffic conditions in modern graph processing systems.

Second, network-related costs, such as bandwidth, latency, or monetary costs, are subject to large variations. Today, it is common to run graph analytics in the cloud, because of low deployment costs and high scalability [6], [14], [15]. Network heterogeneity exists even in a single data center because the machines are connected via a tree-structured switch topology, where machines connected to the same switch experience high-speed communication, while distant machines suffer from degraded performance due to multi-hop forwarding of network packages [16]. Nevertheless, modern cloud infrastructures are *geo-distributed* [17]. Cloud providers are deploying data centers globally to provide low latency user-access. For instance, Amazon, Google, and Microsoft maintain dozens of data centers world-wide. Global services, such as Twitter and Facebook, are deployed on these data centers and

Published in Proceedings of the 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), 118-128, Nara / Japan, 2016.
© IEEE 2016
http://dx.doi.org/10.1109/ICDCS.2016.92

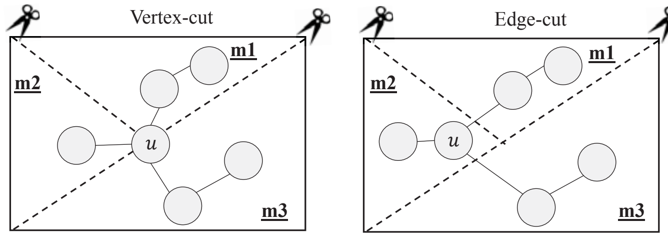


Fig. 1. Vertex-cut and Edge-cut.

produce large amounts of data (e.g., user friendship relations) that need to be analyzed. Data analytics spanning multiple data centers is often the only option. For instance, data should be stored close to the geo-distributed users to reduce access latency, but replication may be prohibitive for legal reasons or efficiency considerations [18], [19]. But here, network link costs can differ by orders of magnitudes (cf. Fig. 2d, Tab. I). These heterogeneous network costs should be considered when partitioning the graph.

To overcome these limitations, we developed GrapH, a graph processing system for distributed, in-memory data analytics on graph-structured data. GrapH is aware of both dynamic vertex traffic during execution and underlying network link costs. Considering this information, it adaptively partitions the graph at runtime to minimize overall communication costs by systematically *avoiding frequent communication over expensive network links*. In particular, the contributions of this paper are as follows:

- A fast partitioning algorithm, named H-load, and a fully distributed edge migration strategy for runtime refinement, named H-move, solving the dynamic vertex traffic- and network-aware partitioning problem.
- A graph processing system named GrapH enabling network-aware, geo-distributed execution of graph algorithms. In contrast to most state-of-the-art graph processing systems, we improve efficiency of multi-algorithm execution by keeping the graph in memory across graph algorithms.
- Evaluations on PageRank, and two important classes of graph algorithms: subgraph isomorphism to find arbitrary subgraphs in the graph, and cellular automaton to simulate social movement patterns of people in Beijing. We show, that GrapH reduces communication costs by up to 60% compared to state-of-the-art partitioning methods.

Outline: we formulate the heterogeneity-aware partitioning problem in Sec. II and present our novel partitioning algorithms H-load and H-move in Sec. III. Afterwards, we evaluate our system in Sec. IV, discuss related work in Sec. V, and conclude in Sec. VI.

II. Preliminaries and Problem Formulation

In this section, we provide background information about the graph execution model and standard vertex-cut partitioning. Then, we present the network- and traffic-aware dynamic partitioning problem to be addressed in this paper.

A. Preliminaries

We assume a widely-used distributed vertex computation model similar to PowerGraph [6], where computation is organized in *iterations*. In each iteration, the system executes the user-defined vertex function for all *active* vertices. It proceeds with the next iteration only after all active vertices have finished vertex function execution (synchronized model). The vertex function operates on user-defined vertex data and consists of three phases, **G**ather, **A**pply and **S**catter (GAS). In the gather phase, each vertex aggregates data from its neighbors into a *gathered sum* σ (e.g., a union of all neighboring vertex data). In the apply phase, a vertex changes its local data using σ. In the scatter phase, a vertex activates neighboring vertices for future execution in the next iteration. For example, in PageRank each vertex has vertex data $rank \in \mathbb{R}$. The gathered sum σ is the sum over all neighboring vertices' $rank$ values. A vertex changes its vertex data $rank$ using σ (i.e., $rank = 0.15 + 0.85 * \sigma$) and activates all neighbors, if $rank$ has changed more than a certain threshold.

Large real-world graphs have billions of vertices and the sequential execution of all vertex functions on a single machine is not scalable. In order to parallelize execution, the graph has to be distributed onto multiple machines by cutting it through edges or vertices (*edge-cut* or *vertex-cut*). In Fig. 1, we divided the graph into three parts using both strategies. Edge-cuts distribute vertices to machines, therefore an edge can connect vertices on different machines. Vertex-cuts distribute edges across machines and make vertices span multiple machines, each having its own *vertex replica*. The set of machines, where vertex u is replicated, is denoted as *replica set* R_u (e.g., the set $\{m1, m2, m3\}$ for vertex u). Note, that we do not differentiate between replicas and machines, because there is a one-to-one mapping given a vertex v. Now, we can define the *replication degree* as the total number of vertex replicas. Vertex-cut has superior partitioning properties for real-world graphs with power-law degree distribution such as the Twitter or Facebook graph [6]. Thus, we use vertex-cut in this paper.

Inter-machine communication happens only in the form of *vertex traffic* between replicas. For instance, if all neighbors of vertex v are on the same machine, no inter-machine communication is needed because all neighboring data can be accessed locally. However, if vertex v is distributed, replicas have to communicate to access neighboring data by exchanging the gather, apply, and scatter messages. One dedicated replica, the *master* $\mathcal{M}_v$, initiates the distributed vertex function execution and keeps vertex data consistent on other replicas denoted as *mirrors*. More precisely, there are three types of communication.

First, a master sends a *gather request* to each mirror; in reply each mirror sends back a *gather response* containing an aggregation of local neighboring data (e.g., a sum of all local $rank$s for PageRank). We denote the number of bytes, exchanged in the gather phase between the master of vertex v and a mirror r in iteration i as $g_r^v(i)$. Second, after computing the new vertex data in the apply phase, the master sends a

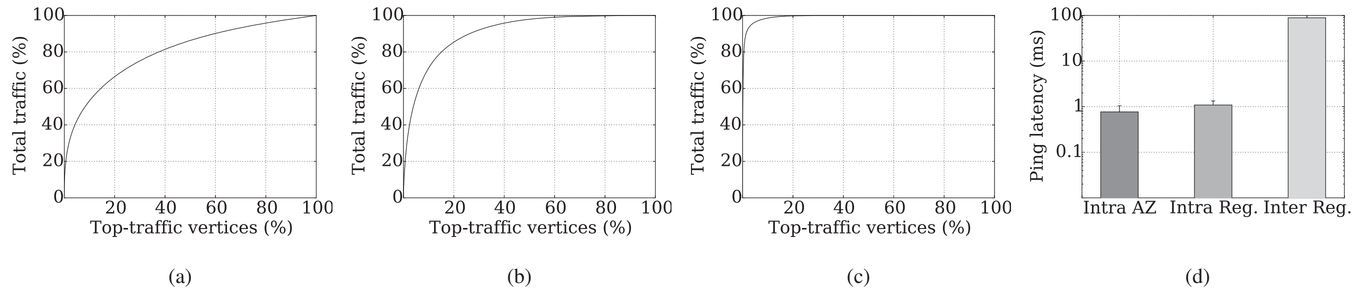


Fig. 2. (a)-(c) Distribution functions of vertex traffic for PageRank, subgraph isomorphism, and cellular automaton. (d) Latency between machines intra-Availability Zone (AZ), inter-AZ and intra-region, and inter-region.

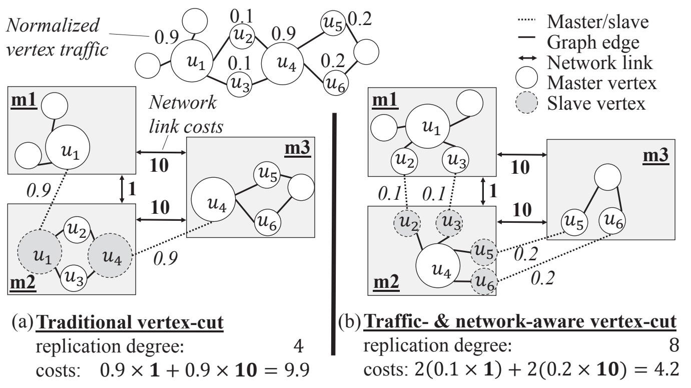

Normalized vertex traffic
Network link costs
Master/slave
Graph edge
Network link
Master vertex
Slave vertex
m1
m2
m3
(a) Traditional vertex-cut
replication degree: 4
costs: $0.9 \times \mathbf{1} + 0.9 \times \mathbf{10} = 9.9$
(b) Traffic- & network-aware vertex-cut
replication degree: 8
costs: $2(0.1 \times \mathbf{1}) + 2(0.2 \times \mathbf{10}) = 4.2$

Fig. 3. (a) Vertex-cut minimizing replication degree. (b) Network- and traffic-aware vertex-cut minimizing costs.

vertex data update to all mirrors (e.g., the new $rank$). The size of this message, $a^v(i)$, depends on the local vertex data on the master and can vary significantly. Third, in order to schedule neighbors of v for future execution, *scatter requests* of constant size s are exchanged between master and mirrors. With this, we can define vertex traffic $t^v(i)$ of vertex v in iteration i (averaged over all replicas in the replica set $R_v(i)$).

$$t^v(i) = \frac{1}{|R_v(i)|} \sum_{r \in R_v(i)} (g_r^v(i) + a^v(i) + s) \qquad (1)$$

B. Network- and Traffic-aware Vertex-cut

We now show heterogeneity of vertex traffic and the network and formulate the heterogeneity-aware vertex-cut partitioning problem.

In general, vertex traffic and network costs are heterogeneous. In Fig. 2(a)-(c), we show vertex traffic heterogeneity for three algorithms: PageRank, subgraph isomorphism and cellular automaton (cf. Sec. IV for details about the algorithms). The graph shows the x/y distribution: x% of the top-traffic vertices are responsible for y% of overall traffic. In our evaluations, PageRank is 20/65 distributed, because of different convergence behaviors of vertices as mentioned in Sec. I. Subgraph isomorphism is more extreme with a 20/84 distribution, because some vertices match more subgraphs than others. Cellular automaton is highly imbalanced (20/100), because vertices simulating unpopular regions in Beijing have almost zero traffic (cf. Sec. IV). Besides vertex traffic, network communication is also subject to significant variations in terms of bandwidth and latency, even in a single data center [9], [14], [20]. For Amazon EC2 machines, we show orders-of-magnitude variations of latency (cf. Fig. 2d). Likewise, many cloud providers charge variable prices for intra and inter data center communication. For instance, Amazon charges nothing for communication within the same availability zone (AZ), while communication across different AZs and regions costs respectively 0.01$/GB and 0.02$/GB of outgoing traffic (cf. Tab. I).

Therefore, efficient graph partitioning should utilize these diverse costs. For example in Fig. 3(a), vertices are annotated with their (normalized) vertex traffic, also indicated by the vertex size. Machines $m1 - m3$ communicate via network links with different costs, given by the weights in bold. The vertex-cut leads to distributed vertices u_1 and u_4, both having traffic 0.9. We define *communication costs* via a network link as the costs of the link multiplied by the traffic sent over this link. The summed communication costs over all network links are the *total communication costs*. In the example, total communication costs are $(0.9 * 1) + (0.9 * 10) = 9.9$. Here, traditional vertex-cut leads to minimal replication degree, but high communication costs, because high-traffic vertices u_1 and u_4 send many messages over the network. To this end, we introduce the network- and traffic-aware dynamic vertex-cut partitioning. The idea is to cut the graph on the low-traffic vertices to decrease inter-machine communication. In Fig. 3(b), we minimize communication by cutting the graph on vertices u_2,u_3 and u_5,u_6. This increases the replication degree, but decreases overall communication costs. Note that this partitioning could be improved even further by exploiting heterogeneous network link costs. Suppose, the subgraph assignments of m1 and m3 were swapped. Then, the (relative) high traffic vertices u_5,u_6 communicate over the inexpensive link (m1,m2), decreasing overall communication costs to $2(0.2 * 1) + 2(0.1 * 10) = 2.4$.

Problem Formulation: Let $G = (V, E)$ be a directed graph with the vertex set V and edge set $E \in V \times V$. Let $M = \{m_1, ..., m_k\}$ be the set of all participating machines. The network cost matrix $T \in \mathbb{R}^{k \times k}$ assigns a cost value to each pair of machines (e.g., monetary costs for sending one

Machine placement	Incoming traffic	Outgoing traffic
Same AZ	0.00-0.01 \$/GB	0.00-0.01 \$/GB
Different AZ, same region	0.01 \$/GB	0.01 \$/GB
Different region	0.00 \$/GB	0.02 \$/GB
Internet	0.00 \$/GB	0.00-0.09 \$/GB

TABLE I
HETEROGENEOUS COMMUNICATION COSTS FOR AMAZON EC2 CLOUD INSTANCES (AUGUST 2015).

byte of data). Hence, $T_{m,m'}$ represents the network costs between machines m and m'. The set of all iterations needed for the graph processing task be $I = \{0, 1, 2, ...\}$. Vertex traffic for all iterations $i \in I$ and vertices $v \in V$ is denoted as $t^v(i)$. The assignment function $a : E, I \to M$ specifies the mapping of edges to machines in a given iteration. The *replica set* of vertex v in iteration i based on assignment function a is denoted as $R^a_v(i)$. It represents the set of machines maintaining a replica of v: $R^a_v(i) = \{m | a((u, v), i) = m \lor a((v, u), i) = m\}$. In the following, we denote R_v to be v's replica set under the assignment in the present context. One dedicated replica $\mathcal{M}_v \in R_v$ is the *master replica* of vertex v.

Our goal is to *find an optimal dynamic assignment of edges to machines minimizing overall communication costs*:

$$a_{opt} = \operatorname*{argmin}_a \sum_i \sum_{v \in V} \sum_{m \in R^a_v(i)} t^v(i) \ T_{m,\mathcal{M}_v} \tag{2}$$

The load $L_m(i)$ of machine m in iteration i is defined as the summed vertex traffic over all vertices replicated on m in iteration i. To balance machine load, we require for each iteration i and machine m (having vertices V_m) that load deviation is bounded by a small balancing factor $\lambda > 1$:

$$L_m(i) = \sum_{v \in V_m} t^v(i) < \lambda \frac{\sum_{v \in V} t^v(i)}{|M|}. \tag{3}$$

Hardness: The dynamic network- and traffic-aware partitioning problem is NP-hard.

Proof sketch: Reduce the NP-hard balanced vertex-cut problem to Eq. 2. Set input: $I = \{1\}$, $t^v(i) = 1$, $T_{m_1,m_2} = 1$. By, $a_{opt} = \operatorname{argmin}_a \sum_i \sum_{v \in V} \sum_{m \in R^a_v(i)} 1{*}1 = \operatorname{argmin}_a \sum_{v \in V} |R^a_v|$, Eq. 2 becomes the network- and traffic-*unaware* vertex-cut problem, which is NP-hard (e.g., [21]). $\square$

III. ALGORITHMS FOR NETWORK- AND TRAFFIC-AWARE PARTITIONING

In this section, we present our novel algorithms addressing the network- and traffic-aware partitioning problem. We developed two methods: a partitioning algorithm called **H-load** for pre-partitioning the graph, and a dynamic algorithm called **H-move** for runtime refinement using migration of edges. In Tab. II, we summarize notation used in the paper.

A. H-load: Initial Partitioning

Graph processing systems have to pre-partition the graph, so that each machine can load its partition into local memory.

M	The set of machines.
k	The number of machines.
a	Function mapping edges to machines.
R_v	Replica set of vertex v.
$\mathcal{M}_v$	Master of replica set of vertex v.
$t^v(i)$	Vertex traffic of vertex v in iteration i.
$\hat{t}^v(i)$	Vertex traffic estimation of vertex v for iteration i.
$L_m(i)$	Load (i.e., summed vertex traffic) of machine m in iteration i.
C	Capacity of exchange partner machine.
$\beta(v)$	Byte size of serialized vertex u.
μ	Aggressiveness parameter specifying *willingness-to-move*.
c_+	Investment costs of migrating edges.
c_-	Payback costs in terms of saved future traffic.

TABLE II
NOTATION OVERVIEW.

To this end, we developed H-load, a fast pre-partitioning algorithm that consists of two phases. First, it partitions the graph using a vertex-cut algorithm utilizing network heterogeneity. Second, it determines a cost-efficient mapping from partitions to machines. We describe these two phases in the following.

1) Initially, our goal is to find a reasonable partitioning of the graph into k balanced parts, ignoring concrete mapping of partitions to machines. In order to improve partitioning performance of billion-scale graphs, we assume a streaming setting: the graph is given as a stream of edges $e_1, e_2, ..., e_{|E|}$ with $e_i \in E$ and we consecutively read and assign one edge at a time to a partition until there are no more edges to read. Hence, computational complexity is linear to the number of edges. Our method is similar to the vertex-cut algorithm of PowerGraph [6], which greedily reduces replication degree of vertices. However, the PowerGraph pre-partitioning leads to relatively homogeneous total traffic between each pair of partitions. But to exploit heterogeneity of network costs, we also require heterogeneity of inter-partition traffic: partitions exchanging more traffic should be mapped to machines with low-cost network links. More precisely, the network cost matrix T often consists of several *clusters* of machines that have low intra-cluster and high inter-cluster costs (e.g., EC2 machines running in different availability zones). The number of clusters c can be determined from the matrix T using well-established clustering methods (e.g., [22]). Next, we group partitions into c clusters and map edges to partitions such that replicas preferentially lie in the same cluster.

Each edge (u, v) is assigned to a partition p as follows. If there exists no replica of u or v on any partition, assign (u, v) to the least loaded partition. If there are partitions containing replicas of u *and* v, assign (u, v) to the least loaded of those partitions. Otherwise, a new replica has to be created, because there is no partition containing both replicas of u and v. For example, if we place the edge on a partition p that already has a replica of u, we have to create a new replica of v. We choose partition p, such that the new replica preferentially lies in the same cluster as already existing replicas. With this method, our algorithm ensures a clustered traffic behavior: partitions in the same cluster share the same replicas and thus are expected to exchange more traffic than partitions in different clusters. In the next phase, we try to find a good mapping of partitions to machines.

2) Now, we try to find a mapping of the $|M|$ partitions to $|M|$ machines while minimizing overall communication costs. In order to minimize these costs, an optimal mapping of partitions to machines would assign two partitions with higher inter-partition traffic to machines connected via a low-cost network link. This is an instance of the well-known quadratic assignment problem: map $|M|$ factories (i.e., partitions) to $|M|$ locations (i.e., machines), so that the mapping has minimal costs of factories sending their goods to other factories (i.e., communication costs). We used the iterated local search algorithm of Stützle et al. [23] which greedily minimizes (communication) costs. Initially, partitions are randomly mapped to machines. Then the algorithm iteratively improves the total costs using the following method. Find two machines, such that an exchange of partition assignments would result in lower total communication costs. For example in Fig. 3, exchanging partition assignments of machine $m1$ and $m3$ results in lower total communication costs. If an improvement is found, it is applied immediately. In order to address convergence to local minima, we perturb a local optimal solution by randomly exchanging two assignments. Note, that this algorithm is computationally feasible, because it runs on a relatively small problem set with size $|M| << |V|$. Clearly, the above method assumes that the traffic exchanged between each pair of partitions is known (i.e., cumulative traffic exchanged between vertex replicas shared by each pair of partitions). This information can be determined from the previous executions of the GAS algorithm. Otherwise, homogeneous traffic between vertex replicas is assumed.

B. H-move: Distributed Migration of Edges

The H-load algorithm is suitable for a static network-aware and traffic-aware partitioning. However, often the vertex traffic changes dynamically at runtime. To this end, we developed the distributed edge-migration algorithm *H-move* solving the dynamic heterogeneity-aware partitioning problem. The idea is that each machine locally reduces the communication costs by migrating edges to distant machines. To this end, we define the term *bag-of-edges* as the set of edges to be migrated. Machines migrate bag-of-edges in parallel after each GAS iteration. Finally, if no further improvements can be performed, migration is switched off.

Approach overview: The overall migration strategy is given in Alg. 1. After activation of the migration algorithm (line 1), machine m first selects partner machine m' (line 2) and then calculates the bag-of-edges to be send to m' (line 3). In order to prevent inconsistencies due to parallel updates on the distributed graph, machine m requests locks for all vertices in the bag-of-edges (line 4). Afterwards, m updates the bag-of-edges to contain only those edges, whose endpoint vertices could be locked (line 5) and determines, whether sending the updated bag-of-edges results in lower total communication costs (line 6). When sending the bag-of-edges, communication

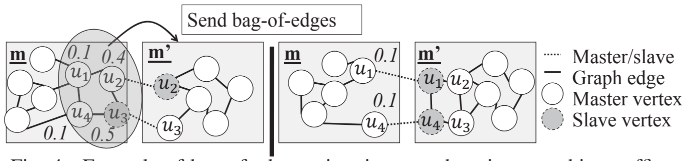


Fig. 4. Example of bag-of-edges migration to reduce inter-machine traffic.

costs change due to modifications of the vertex replica sets. To calculate Δc in line 6, machine m considers both: the migration overhead c_+ of sending the bag-of-edges, as well as the decrease of communication costs c_- when improving the partitioning. If Δc is negative, the bag-of-edges is migrated to m'. Finally, machine m releases all held locks in line 9.

Algorithm 1 Migration algorithm on machine m.

```
1: $waitForActivation()$
2: $m' \leftarrow selectPartner()$
3: $b \leftarrow bagOfEdges(m')$
4: $lock(b)$
5: $b \leftarrow updateLocked(b)$
6: $\Delta c \leftarrow c_+ - c_-$
7: if $\Delta c < 0$ then
8:     $migrateBag(b)$
9: $releaseLocks(b)$
```

We give an example of this procedure in Fig. 4. Two machines m and m' have replicas of high-traffic vertices u_2 and u_3. In order to reduce communication costs, m decides to send the bag-of-edges $b = \{(u_1, u_2), (u_2, u_3), (u_3, u_4), (u_4, u_1)\}$ to m'. Machine m' receives b and adds all edges in b to the local subgraph. The right side of the Fig. 4 shows the final state after migration of b. Now, low-traffic vertices u_1 and u_4 are cut leading to less inter machine traffic. In the following, we describe the proposed migration approach (cf. Alg. 1) in more details.

1) Selection of partner and bag-of-edges: Which machine is suitable as exchange partner? Intuitively, two machines sharing high-traffic replicas are strong candidates for exchanging bag-of-edges, because improving their partitioning can potentially reduce the overall communication costs. To this end, each machine m maintains a list of potential exchange partners (with decreasing priority). This list is computed by sorting neighboring machines w.r.t. the total amount of exchanged traffic. On each round of Alg. 1, the top-most machine m' is selected as an exchange partner and removed from the list. Once the list is empty, it is recomputed as mentioned above.

Now, we determine the maximal size of the bag-of-edges to be sent to m' in order to ensure balanced machine load (cf. Eq. 3). Therefore, we introduce the notion of *capacity* of a machine m', i.e., the maximum amount of additional load, machine m' can carry. Capacity is defined as half the difference of loads L_m and $L_{m'}$ of the sending and the receiving machine: $C = (L_{m'} - L_m)/2$. To learn about the current load $L_{m'}$ of machine m', machine m sends a request to m'. Using the capacity, machine m can balance the loads by only sending edges, such that the deviation of the two machines traffic values is still bounded. For example, if sending the bag-of-edges results in a new replica of vertex v on m', this increases load of m' by the vertex traffic of v. If this violates load balancing between m and m', machine m will not include v into the bag-of-edges.

Once the exchange partner is selected and we know its capacity, we determine a bag-of-edges (bag) to be send. Selecting a suitable bag is crucial for optimizing communication costs and migration overhead. Theoretically, the perfect bag could be any subset out of p edges on a machine (i.e., 2^p subsets). In order to keep the migration phase lean, we developed a fast heuristic to find a bag improving communication costs (cf. Alg. 2). Initially, machine m determines the set of candidate vertices, those replicated on both machines, because they are responsible for all the traffic between m and m'. Machine m sorts the candidates by descending vertex traffic in order to focus on the high-traffic vertices first (line 3). Then, m iterates the following steps until m' has no more capacity. It checks for the top-most candidate vertex (line 5), whether sending all adjacent edges results in lower total communication costs of the overall graph processing (line 6-7, cf Sec. III-B2). If the total communication costs would decrease when sending the edges, machine m adds them to the bag (line 8-9).

Algorithm 2 Determining the bag-of-edges to exchange.

```
1: function $bagOfEdges(m')$:
2: $bag \leftarrow []$
3: $candidates \leftarrow sort(adjacent(m'))$
4: while $hasCapacity(m', bag)$ do
5:     $v \leftarrow candidates.removeFirst()$
6:     $b \leftarrow \{(u,v) | u \neq v\}$
7:     $\Delta c \leftarrow c_+ - c_-$
8:     if $\Delta c < 0$ then
9:         $bag \leftarrow bag + b$
10: return $bag$
```

2) Calculation of costs: Clearly, migrating bag b from one machine to another is only beneficial, if it results in lower overall costs (i.e., line 6 in Alg. 1, and line 7 in Alg. 2). In general, two types of costs have to be considered in calculating the resulting overall costs: *investment costs* and *payback costs*. Investment costs represents the overhead for migrating the bag and should be avoided. Payback costs are the saved costs after migrating bag b in the form of less future inter-machine traffic. In the following, we formulate both costs.

Investment costs: After sending b to m', m can remove isolated replicas that have no local edges anymore (Fig. 4 vertices u_2, u_3). If machine m is the master $\mathcal{M}_v$ of a vertex v to be removed, i.e., $\mathcal{M}_v = m$, we have to select a new master after removing v from m. We set the partner machine m' to be the new master of v: $\mathcal{M}'_v = m'$. On the other hand, some vertices may not exist on m' leading to creation of new replicas (Fig. 4 vertices u_1, u_4). In both cases, the replica set R_u of a vertex u might have changed (i.e., remove m or add m' to R_u). Then m has to send an update to all machines in R_u with the new vertex replica set, denoted as R'_u. Additionally, when creating a new replica on m', machine m has to send the state of v, i.e., vertex data and meta information such as the vertex id. This can be very expensive for large vertex data, and should be taken into account when deciding whether to migrate a bag. Together, the investment costs are the sum of three terms. The first term calculates the costs of sending the bag b to m'. The second term calculates the costs of sending new replicas to m', if needed. The third term calculates the costs of updating machines in all replica sets that have changed.

$$c_+ = \sum_{e=(u,v)\in b} \beta(e) T_{m,m'} + \sum_{u \in V_b} \delta(u)\beta(u) T_{m,m'} + \sum_{u\in V'_b, r\in R_u \cup R'_u} \beta(R_u) T_{m,r}, \tag{4}$$

where (i) the function $\beta(x)$ returns the number of bytes needed to encode x (to be sent over the network), (ii) the indication function $\delta(u)$ returns 1, if machine m' has no local replica of u, otherwise 0, (iii) V_b is the set of all vertices in bag b, and (iv) V'_b is the set of all vertices whose replica sets will change when sending the bag b to m'.

Payback costs: we can also save costs when sending bag b from m to m'. Suppose the replication degree decreases because of sending (u, v), i.e., $|R'_u| < |R_u|$ or $|R'_v| < |R_v|$. Then, we save *for each iteration* (starting from the current iteration i_0) the costs of exchanging gather, apply, and scatter messages across replicas, i.e., the vertex traffic $t^v(i)$ of vertex v in iteration i. Theoretically, exact payback costs are given by the following formula that calculates for all future iterations and each vertex in the bag the difference of the new costs and the old costs of v's replica set.

$$c^*_- = \sum_{i>i_0} \sum_{v\in V_b} (\sum_{r\in R'_v} t^v(i) T_{r,\mathcal{M}'_v} - \sum_{r\in R_v} t^v(i) T_{r,\mathcal{M}_v}) \tag{5}$$

Here, it is assumed that vertex traffic is known for all future iterations. This is not the case in real systems. Therefore, we describe next, how to estimate the payback costs in the presence of uncertainty about future vertex traffic.

In order to estimate payback costs, we have to predict vertex traffic for future iterations. More formally, given vertex traffic $t^v(0), t^v(1), ..., t^v(i)$, we estimate traffic values $t^v(i+1), ..., t^v(|I|)$. The prediction should be quick, have low computational overhead, and low memory requirements, because we have to predict vertex traffic in each migration phase for millions of vertices. We investigate three well-known methods for time series prediction of the next traffic value $t^v(i+1)$, that fit to our requirements [24]. The first method is *most recent value* (Last) taking the last traffic value as prediction for the next traffic value: $\hat{t}^v(i+1) = t^v(i)$. The second method is *incremental moving average* (MA) with the idea to use the moving average of the last w observations, while not storing the values in the window: $\hat{t}^v(i+1) = \frac{\hat{t}^v(i)(w-1)+t^v(i)}{w}$. The third method is *incremental exponential average* (EA) that calculates the estimation based on the old estimation and the last traffic value: $\hat{t}^v(i+1) = \alpha t^v(i) + (1-\alpha)\hat{t}^v(i)$. The parameter $\alpha \in [0, 1]$ specifies the amount of decaying

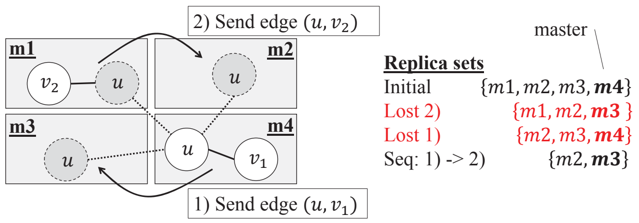

2) Send edge (u, v_2)
master
m1
m2
v_2
u
u
Replica sets
Initial $\{m1, m2, m3, \mathbf{m4}\}$
Lost 2) $\{m1, m2, \mathbf{m3}\}$
Lost 1) $\{m2, m3, \mathbf{m4}\}$
Seq: 1) -> 2) $\{m2, \mathbf{m3}\}$
m3
m4
u
u
v_1
1) Send edge (u, v_1)

Fig. 5. Lost update problem for parallel edge migration.

older traffic values and thus, the importance of recent traffic. We assume $t^v(0) = 0$ for all methods. In Sec. IV, we have compared overall system performance for these methods with different parameter choices, i.e., window sizes w and decay parameters α.

With the above mentioned methods, we can determine the vertex traffic estimation for the next iterations. However, Eq. 5 expects a vertex traffic value for all future iterations. In general, accuracy of the predicted vertex traffic $\hat{t}^v(i)$ can decrease with increasing i, because vertex traffic patterns may change over time. Therefore, we introduce a factor μ, which represents the minimum number of iterations, we expect to save communication costs as a result of migrating bag b. This parameter can be used to specify the aggressiveness with which migration should be performed. Together, our estimated payback costs are the following.

$$c_{-} = \mu \sum_{v \in V_b} \Big(\sum_{r \in R'_v} \hat{t}^v(i+1) T_{r,\mathcal{M}'_v} - \sum_{r \in R_v} \hat{t}^v(i+1) T_{r,\mathcal{M}_v} \Big) \quad (6)$$

3) Graph consistency: When two machines independently send edges and change replica sets of vertices, inconsistencies of the data graph can arise. In Fig. 5, we give an example. Machines $m1$-$m4$ maintain a replica of vertex u. Suppose machine $m4$ wants to send edge (u, v_1) to machine $m3$ and machine $m1$ edge (u, v_2) to $m2$. After sending the edge, machines $m4$ ($m1$) removes the vertex replicas of u with no further local edges. Therefore, machine $m4$ ($m1$) has to update all other machines having a replica of v with the new replica set. Machine $m4$ sends the new replica set $\{m1, m2, m3\}$ to all machines, while machine $m1$ sends $\{m2, m3, m4\}$. The machines can receive these updates in different orders leading to inconsistent views on the replica sets (lost update problems). Only a sequential update would result in a consistent state (machine $m4$ changes the replica set *before* machine $m1$). To guarantee sequential updates during edge migration, we lock endpoint vertices in the bag to be sent to the partner machine. We use a simple master locking scheme, i.e., each machine intending to change vertex u sends a locking request to $\mathcal{M}_u$. If vertex u is already locked, the master machine returns $false$, otherwise it locks u and returns $true$. However, this also implies that not all locks may be acquired. Therefore, we exclude edges from the bag for which not both endpoints have been locked successfully (see line 5 in Alg. 1). Only after the locking machine has successfully implemented the bag-of-edge exchange, it releases all locks.

Name	$\lvert V\rvert$	$\lvert E\rvert$
Gnutella	8,114	26,013
Facebook	4,039	88,234
WikiVote	7,115	103,689
Twitter	81,308	1,768,149
GoogleWeb	875,713	5,105,039
TwitterLarge	41,652,230	1,468,365,182

TABLE III
REAL-WORLD GRAPHS FOR EVALUATIONS.

IV. EVALUATIONS

In the following, we present evaluations for GrapH on two computing clusters for three different algorithms with high practical relevance, i.e., PageRank, subgraph isomorphism, and cellular automaton, on several real-world graphs[1] with up to 1.4 billion edges given in Tab. III. We compared our migration strategies with state-of-the-art static vertex-cut partitioning approaches: hashing of edges (Hash) and PowerGraph (PG) [6], [7]. We show the efficiency and effectiveness of the partitioning strategies H-load and H-move.

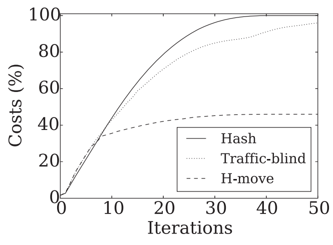
A. Graph Algorithms

For our experiments, we have implemented the three graph algorithms: PageRank (cf. [6]), subgraph isomorphism, and social simulations via agent-based cellular automaton, denoted as PR, SI, and CA, respectively. For SI and CA there is, to the best of our knowledge, no implementation using the GAS API, so we have designed our own algorithms. For implementation details on these algorithms, we refer the reader to [25].

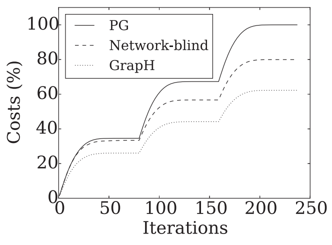
Subgraph Isomorphism is an NP-complete graph problem. It can be used to query a graph for certain patterns (*subgraphs*). Examples are simple queries for Facebook Graph Search ("Has a friend of mine been in that restaurant?") or complex queries for community detection ("Is there a k-clique in the graph?"). More precisely, given a graph $G = (V, E)$ and a graph pattern $P = (V_P, E_P)$, subgraph isomorphism is the problem of finding subgraphs $G_{sub} = (V_{sub}, E_{sub})$, with $V_{sub} \subseteq V, E_{sub} \subseteq E$, that are isomorphic to the graph pattern P. Each graph vertex can have an optional label (e.g., an importance weight such as a PageRank value). In this case, the labeled SI additionally requires both vertices in V_{sub} and V_P to have matching labels (cf. [4]).

Cellular Automaton is a powerful and well-established model of simulation, where the problem space is expressed as a grid of cells with each cell having a finite number of states. A cell iteratively calculates its own state based on the states of neighboring cells. Many complex simulation problems can be modeled as cellular automata and computation is easily parallelizable using the GAS programming abstraction for recent graph processing systems. We implemented an agent-based variant for simulating movements of people in Beijing using real-world movement data[2] [1].

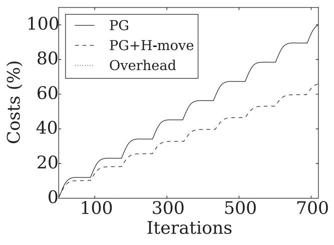
[1]http://konect.uni-koblenz.de/networks/twitter, http://snap.stanford.edu/data
[2]http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/



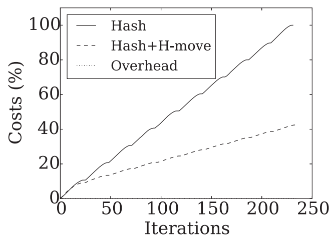
(a) PR on *Twitter*(Traffic-awareness)



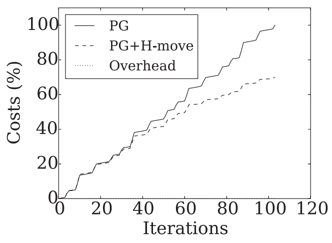
(b) PR on *GoogleWeb*



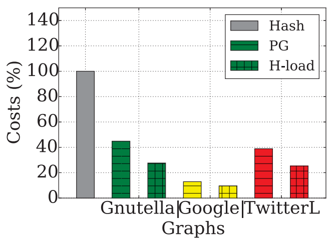
(c) PR on *GoogleWeb*.



(d) PR on *WikiVote*.



(e) SI on *Facebook*.



(f) Pre-partitionings

Fig. 6. (a) Traffic- and (b) network-awareness reduce communication costs. (c)-(e) H-move reduces communication costs with low overhead. (f) Pre-partitioning with H-load reduces communication costs.

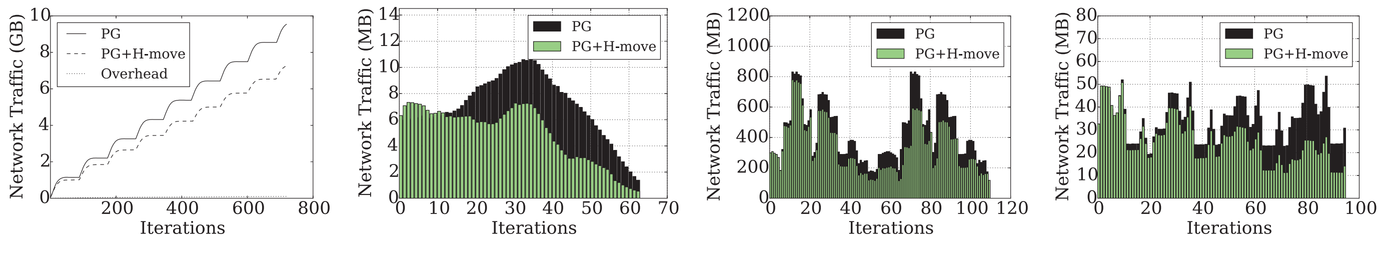
B. Evaluation Setup and Experiments

We have implemented GrapH in the Java programming language (10,000 lines of code). GrapH consists of a master machine and multiple client machines performing graph analytics. The master receives a sequence of graph processing queries $q_1, q_2, q_3, ...$ consisting of user specified GAS algorithms. All machines communicate with each other directly via TCP/IP. We used two computing clusters with homogeneous, and heterogeneous network costs. The homogeneous computing cluster (ComputeC) consists of 12 machines, each with 8 cores (3.0GHZ) and 32GB RAM, interconnected with 1 Gbps ethernet. The heterogeneous computing cluster (CloudC) is deployed in the Amazon cloud using 8 geographically distributed EC2 instances (1 virtual CPU with 3.3 GHz and 1 GB RAM) that are distributed across two regions, US East (Virginia) and EU (Frankfurt), and four different availability zones. If not mentioned otherwise, the experiments are performed on CloudC. As network costs between these instances, we used the real monetary costs charged by Amazon (cf. Tab. I).

1) Communication costs: The main idea of this paper is to consider network- and traffic-heterogeneity while constantly repartitioning the graph during computation. In the following, we evaluate the effect of traffic- and network-awareness on total communication costs. We show the improvement of communication costs as defined in Eq. 2: *total traffic sent via each network link, weighted by the costs of the network link*. In our first experiment (cf. Fig 6a), we compared three different partitioning methods: (i) hashing of edges to machines without dynamic migration (Hash), (ii) our dynamic migration strategy assuming homogeneous vertex traffic on the hash partitioned graph (Traffic-blind), and (iii) our dynamic migration strategy on the hash partitioned graph (H-move). We accumulated communication costs over 50 iterations of PR on *Twitter*. As we can see, taking heterogeneous vertex traffic into account greatly improves total communication costs by up to 50% compared to Traffic-blind. The reason is that GrapH implicitly prioritizes high-traffic vertices, that are the major sources of total traffic (cf. Sec. II, Fig. 2).

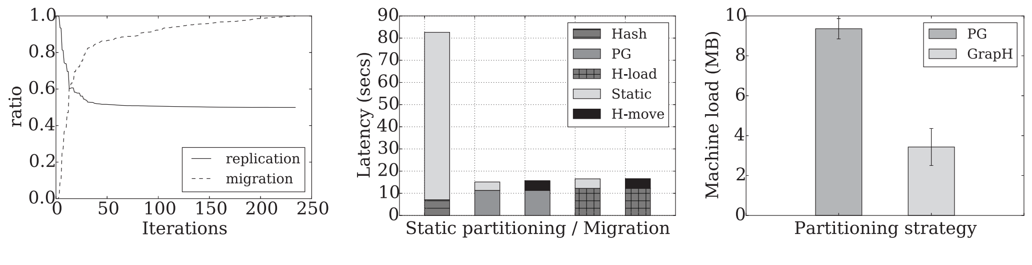
How does network-awareness improve total communication costs? In Fig. 6b, we compare three different partitioning methods: (i) PowerGraph partitioning without dynamic migration (PG), (ii) our dynamic migration strategy on the partitioned graph using our static partitioning strategy, while both strategies assume homogeneous network (Network-blind), and (iii) our dynamic migration strategy H-move on the H-load partitioned graph (GrapH). We performed 250 iterations of PageRank on *GoogleWeb* (and restarted computation after termination of one PageRank instance). It can be seen, that Network-blind already reduces communication costs by 20% compared to PG. However, taking heterogeneous network into account reduces total costs by additional 20%. Our experiments therefore indicate, that the awareness of traffic and network heterogeneity improves partitioning quality significantly.

To learn about the overhead of migrating edges in terms of additional communication costs, we have performed experiments on both computing infrastructures, for different graph algorithms and different real-world graphs (cf. Fig 6c-e). Our findings are, that H-move consistently improves communication costs while keeping migration overhead extremely low. For instance in Fig. 6c, we show total communication costs and migration overhead of 700 iterations of PageRank on *GoogleWeb*. While communication costs decreased by 33% compared to PG, the costs for migration (investment costs, cf Eq. 4) are very low compared to the saved costs (in fact, the migration overhead can not be differentiated from the x-axis in the figure). In Fig. 6d, we repeated this experiment for the hash prepartitioned *WikiVote* graph. Here, communication costs decrease by 60%, when migration is switched on. Additionally,



(a) PR on *GoogleWeb*. (b) CA on grid (10^5 vertices) (c) SI on *Twitter*. (d) SI on *Facebook*.

Fig. 7. (a) H-move reduces accumulated network traffic. (b)-(d) H-move reduces current network traffic.



(a) PR on *WikiVote*. (b) PR on *GoogleWeb*. (c) PR on *GoogleWeb*.

Fig. 8. (a) H-move reduces replication degree. (b) Latency remains relatively stable using H-move. (c) GrapH reduces total workload.

we tested H-move on the complex SI algorithm (Fig. 6e). We have sequentially executed SI algorithms on *Facebook*, mainly searching for relatively simple graph patterns such as triangles. Here, H-move decreases communication costs of the PG pre-partitioned graph by 30%, although subsequent executions will benefit even more from the improved partitioning.

Finally, we evaluated communication costs improvement of H-load compared with PowerGraph and Hashing for three different graphs: *Gnutella*, *GoogleWeb*, and *TwitterLarge* (Fig 6f). We assume homogeneous vertex traffic and heterogeneous network costs. H-load greatly reduces total graph processing costs by 70-90% compared to Hash and by 25-38% compared to PowerGraph partitioning.

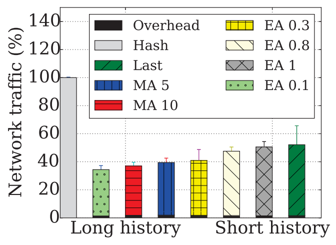
Overall, graph processing using H-move reduces total communication costs at runtime, independently from the concrete pre-partitioning. GrapH automatically decides, whether migrating edges is worth the additional overhead, so that migration overhead is kept minimal.

2) Network traffic: Often network costs are unknown or relatively homogeneous. In this case, H-move dynamically reduces the total amount of machine communication, i.e., *network traffic*. In the following, we show the extent of this improvement for PR, SI, and CA. In Fig. 7a, we have executed multiple PR algorithms on *GoogleWeb*. We plotted the *accumulated* network traffic for PG pre-partitioned graph with H-move switched off and on. As we can see, H-move decreases total traffic by 12%. In order to learn about the *current* network traffic, we have performed experiments for CA and SI (cf. Fig. 7b-d). We show the current network traffic (averaged over a sliding window of 10 iterations) for these algorithms on ComputeC. Our migration strategy reduces total network traffic by up to 50% compared to PowerGraph partitioning.

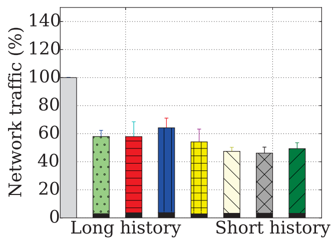
3) Replication degree: A standard metric to measure the partitioning quality of vertex-cut algorithms is the replication degree, i.e., the number of vertex replicas in the system. In Fig. 8a, we can see the improvement of the replication degree using H-move during PR on *WikiVote*. For each iteration, we plot the current replication degree divided by the replication degree of a hash partitioned graph, as well as the current migration overhead divided by the total migration overhead (CDF). During the first 50 iterations, there is an improvement of more than 50% in replication degree, followed by saturation. Therefore, migration is switched off after saturation.

4) Latency: The partitioning problem is computationally hard and solving it during execution can be extremely expensive. However, our heuristic performs well as can be seen in Fig. 8b. We measured the average graph processing latency for one PR iteration, with and without migration, as well as the latency for pre-partitioning the graph using Hash, PowerGraph, and H-load. We denote an execution as *Static*, if dynamic migration is switched off. Due to huge memory overhead, the hash-based approach leads to a much higher latency in our setting (note that memory of our EC2 instances is relatively small). We can also see that H-move induces relatively little latency overhead compared to migration switched off ($0.01 - 0.13$ times). Experiments for SI on a PG partitioned graph show similar results of 0.04 times increased latency compared to static partitioning. We believe that the latency punishment of H-move can be reduced further by *automatically* switching the migration on or off at runtime.

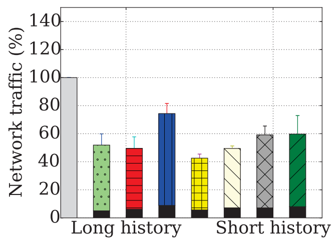
5) Load balancing: GrapH balances the *summed vertex traffic over all vertices on a machine* (cf. Eq. 3). In Fig. 8c, we show the average machine workload after one PR execution (78 iterations) for PG and GrapH as well as the deviation of the machines from this average workload. GrapH leads



(a) PR on *Twitter*.



(b) SI on *Twitter*.



(c) CA on grid (2,500 vertices).

Fig. 9. (a)-(c) Prediction methods influence migration efficiency.

to a slightly higher workload imbalance because we balance for (more volatile) vertex traffic. In contrast, PG balances the number of edges. However, GrapH's most loaded machine still has less workload than the least loaded machine in PG. Thus, GrapH reduces total workload by more than 60%.

6) Prediction methods: We expected that choosing the right prediction method for future vertex traffic is important for overall performance of our system. Therefore, we compared three methods (cf. Sec. III): last value (Last), moving average (MA) with window sizes 5 and 10, and exponential averaging (EA) with decay parameter $\alpha = 0.1, 0.3, 0.8, 1.0$ in Fig. 7a-c. We evaluated the reduction of total network traffic for PR, SI, and CA, compared to Hash. The position of the bars from left to right reflects the size of the considered history for prediction in descending order. For example, since *Last* considers only the last value, we plotted it on the right. As it can be seen, all prediction methods meet the goal of minimizing overall network traffic. However, no method leads to consistently better results for all algorithms. We attribute this to the different stability of vertex traffic patterns. For example in PR, considering a longer history shows better results, because vertex traffic patterns remain stable over time. Nevertheless, considering a large history in SI actually harms performance, because the subsequent short-lived queries lead to diverse vertex traffic patterns. For CA, apparently, there is no such clear trend. This is because of the sudden changes of vertex traffic, when agents move to neighboring cells. However, exponential averaging with parameter $\alpha = 0.3$ showed good results for all three algorithms, so we use this configuration as default. A thorough study in this direction is left for future work.

V. Related Work

Recently, many systems for distributed graph processing have emerged that inspired our work. PowerGraph [6] suggests the GAS programming model and uses a distributed vertex computation strategy based on vertex-cut partitioning. They provide a greedy streaming heuristic (*Coordinated*) for static vertex-cut partitioning, which we used for comparison. Petroni et al. (*HDRF* [21]) consider the vertex degree in order to find a minimum vertex-cut in the streaming setting. They argue that real-world graphs with skewed degree distribution have a high clustering coefficient for low-degree vertices, and therefore *HDRF* minimizes the cut size only for low-degree vertices. A similar argumentation is used in degree-based hashing (*DBH* [26]), where an edge is assigned to a partition based on the hash value of the edge's low-degree vertex (cf. [27]). PowerLyra [28] extends PowerGraph with hybrid-cuts: cutting vertices with high degree and edges of low-degree vertices decreasing expensive replica communication overhead. These strategies minimize the replication degree, but ignore diverse and dynamic vertex traffic.

On the other hand, the graph systems Mizan [29] and GPS [11] propose adaptive edge-cut partitioning using vertex migration. Mizan considers the real traffic sent via each edge to balance workload at runtime. Vaquero et al. [30] apply edge-cut to changing graphs using a decentralized algorithm for iterative vertex migration to avoid costly re-execution of static partitioning algorithms. Shang et al. [31] nicely identified and categorized three types of vertex activation patterns for graph processing workload: always-active-, traversal-, and multi-phase-style. They exploit these workload patterns to dynamically adjust the partitioning during computation. Yang et al. [32] (*Sedge*) improve localization of processing small-sized queries by introducing a two-level complimentary partitioning scheme using vertex replication. While these edge-cut systems adapt to changing graphs or traffic behaviors, they do not consider network topology and migration costs. Furthermore, these approaches focus on edge-cut and optimal edge-cuts can not be transformed into close-to-optimal vertex-cuts for graphs with high-degree vertices (cf. [33]). For instance, a star graph with $|E|$ edges has an edge-cut size of $\Omega(|E|)$ but a vertex-cut size of $O(1)$. Since real-world graphs often resemble a star-like degree distribution (e.g. *President Obama* in the Twitter social network), vertex-cut strategies should not be based simply on transformed edge-cuts, but be tailored specifically to the vertex-cut problem.

Surfer [14] tailors graph processing to the cloud by considering bandwidth unevenness to map graph partitions with a high number of inter-partition links to machines connected via high-bandwidth networks. They provide a straightforward method for measuring bandwidth heterogeneity that could be applied to our system. However, they assume homogeneous traffic via each edge. GraphIVE [15] strives for a minimal *unbalanced* k-way vertex-cut for machines with heterogeneous computation and communication capabilities, in order to put more work to more powerful machines. Therefore, they search the optimal number of edges for each machine. This approach is orthogonal to our proposed heterogeneity-aware partitioning algorithms. Xu et al. [20] consider network and vertex weights to find a static edge-cut with minimal communication costs. They do not consider adaptive vertex-cut partitioning, and vertex weights reflect only the number of executions, but not the real vertex traffic. Zheng et al. [34] propose an architecture-aware graph repartitioning method that also considers the amount of communication going over each edge and the costs of migrating a vertex. Unfortunately, it can not be used for vertex-cut partitioning and the GAS execution model.

General data processing in the geo-distributed setting is addressed by Pu et al. [18] and Jayalath et al. [19]. They argue that aggregating geographical distributed data into a single data center can significantly hurt overall data processing performance. Hence, while focusing on MapReduce-like computations, they share our idea of saving overall costs in geo-distributed data analytics on heterogeneous networks. LaCurts et al. [9] point out that considering network heterogeneity for an optimal task placement, improves overall end-to-end data analytics performance, even in a single data center. Therefore, they place communicating tasks in such a way that most communication flows over fast network links. However, task placement is orthogonal to graph partitioning and could be used on top of our system.

VI. Conclusion

Modern graph processing systems use vertex-cut partitioning due to its superiority of partitioning real-world graphs. These partitioning methods minimize the replication degree, which is expected to be the dominant factor for communication costs. However, the underlying assumptions of uniform vertex traffic and network costs do not hold for many real-world applications. To this end, we proposed GrapH, a graph processing system taking dynamic vertex traffic and diverse network costs into account to adaptively minimize communication costs of the vertex-cut at runtime. Our evaluation show, that GrapH outperforms PowerGraph's vertex-cut partitioning algorithm by more than 60% w.r.t. communication costs.

References

[1] T. Suzumura, C. Houngkaew, and H. Kanezashi, "Towards billion-scale social simulations," in *WSC, 2014*.

[2] A. Pascale, M. Nicoli, and U. Spagnolini, "Cooperative bayesian estimation of vehicular traffic in large-scale networks," *IEEE Transactions on Intelligent Transportation Systems, 2014*.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web." 1999.

[4] S. Ma, Y. Cao, W. Fan, J. Huai, and T. Wo, "Capturing topology in graph pattern matching," *PVLDB, 2011*.

[5] 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 *ACM SIGMOD, 2010*.

[6] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin, "Powergraph: Distributed graph-parallel computation on natural graphs." in *OSDI, 2012*.

[7] 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, 2014*.

[8] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, "Towards predictable datacenter networks," in *SIGCOMM CCR, 2011*.

[9] K. LaCurts, S. Deng, A. Goyal, and H. Balakrishnan, "Choreo: Network-aware task placement for cloud applications," in *Proceedings of the 2013 conference on Internet measurement conference (IMC), 2013*.

[10] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Vl2: A scalable and flexible data center network," *SIGCOMM CCR, 2009*.

[11] S. Salihoglu and J. Widom, "Gps: A graph processing system," in *Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM), 2013*.

[12] K. Ten Tusscher, D. Noble, P. Noble, and A. Panfilov, "A model for human ventricular tissue," *American Journal of Physiology-Heart and Circulatory Physiology, 2004*.

[13] A. Beck, T. Bolemann, H. Frank, F. Hindenlang, M. Staudenmaier, G. Gassner, and C.-D. Munz, "Discontinuous galerkin for high performance computational fluid dynamics," in *High Performance Computing in Science and Engineering (HPCSE), 2013*.

[14] R. Chen, M. Yang, X. Weng, B. Choi, B. He, and X. Li, "Improving large graph processing on partitioned graphs in the cloud," in *SoCC, 2012*.

[15] D. Kumar, A. Raj, D. Patra, and D. Janakiram, "Graphive: Heterogeneity-aware adaptive graph partitioning in graphlab," in *ICPW, 2014*.

[16] C. Peng, M. Kim, Z. Zhang, and H. Lei, "Vdn: Virtual machine image distribution network for cloud data centers," in *INFOCOM, 2012*.

[17] I. Narayanan, A. Kansal, A. Sivasubramaniam, B. Urgaonkar, and S. Govindan, "Towards a leaner geo-distributed cloud infrastructure," in *USENIX HotCloud, 2014*.

[18] Q. Pu, G. Ananthanarayanan, P. Bodik, S. Kandula, A. Akella, P. Bahl, and I. Stoica, "Low latency geo-distributed data analytics," in *SIGCOMM, 2015*.

[19] C. Jayalath, J. Stephen, and P. Eugster, "From the cloud to the atmosphere: Running mapreduce across data centers," *IEEE Transactions on Computers (TC), 2014*.

[20] N. Xu, B. Cui, L. Chen, Z. Huang, and Y. Shao, "Heterogeneous environment aware streaming graph partitioning," *IEEE Transactions on Knowledge and Data Engineering (TKDE), 2015*.

[21] F. Petroni, L. Querzoni, K. Daudjee, S. Kamali, and G. Iacoboni, "Hdrf: Stream-based partitioning for power-law graphs," in *CIKM, 2015*.

[22] G. Hamerly and C. Elkan, "Learning the k in k-means," *NIPS, 2004*.

[23] T. Stützle, "Iterated local search for the quadratic assignment problem," *European Journal of Operational Research (EJOR), 2016*.

[24] N. R. Herbst, N. Huber, S. Kounev, and E. Amrehn, "Self-adaptive workload classification and forecasting for proactive resource provisioning," *Concurrency and Computation: Practice and Experience (CCPE), 2014*.

[25] C. Li, "Distributed data analytics using graph processing frameworks," Master's thesis, University of Stuttgart, 2015.

[26] C. Xie, L. Yan, W.-J. Li, and Z. Zhang, "Distributed power-law graph computing: Theoretical and empirical analysis," in *NIPS, 2014*.

[27] C. Xie, W.-J. Li, and Z. Zhang, "S-powergraph: Streaming graph partitioning for natural graphs by vertex-cut," *arXiv preprint arXiv:1511.02586*, 2015.

[28] R. Chen, J. Shi, Y. Chen, and H. Chen, "Powerlyra: Differentiated graph computation and partitioning on skewed graphs," in *EuroSys, 2015*.

[29] 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, 2013*.

[30] L. M. Vaquero, F. Cuadrado, D. Logothetis, and C. Martella, "Adaptive partitioning for large-scale dynamic graphs," in *ICDCS, 2014*.

[31] Z. Shang and J. X. Yu, "Catch the wind: Graph workload balancing on cloud," in *ICDE, 2013*.

[32] S. Yang, X. Yan, B. Zong, and A. Khan, "Towards effective partition management for large graphs," in *ACM SIGMOD 2012*.

[33] U. Feige, M. Hajiaghayi, and J. R. Lee, "Improved approximation algorithms for minimum weight vertex separators," *SIAM Journal on Computing, 2008*.

[34] A. Zheng, A. Labrinidis, and P. K. Chrysanthis, "Architecture-aware graph repartitioning for data-intensive scientific computing," in *IEEE International Conference on Big Data, 2014*.