PegasusN: A Scalable and Versatile Graph Mining System

Ha-Myung Park,1 Chiwan Park,2 U Kang2
1KAIST, 2Seoul National University
hamyung.park@kaist.ac.kr, chiwanpark@snu.ac.kr, ukang@snu.ac.kr

Abstract
How can we find patterns and anomalies in peta-scale graphs? Even recently proposed graph mining systems fail in processing peta-scale graphs. In this work, we propose PegasusN, a scalable and versatile graph mining system that runs on Hadoop and Spark. To handle enormous graphs, PegasusN provides and seamlessly integrates efficient algorithms for various graph mining operations: graph structure analyses, subgraph enumeration, graph generation, and graph visualization. PegasusN quickly processes extra-large graphs that other systems cannot handle.

Introduction
Graphs are everywhere: friendship networks, the World Wide Web, knowledge bases, biological networks, etc. Many machine learning tasks are directly or indirectly related to graphs; some tasks such as clustering, partitioning, and classification operate directly on graphs, while others use graph properties such as degrees, graph patterns, connected components, and centralities (Faloutsos, Faloutsos, and Faloutsos 1999; Kang et al. 2010; 2011; Jung et al. 2017). Meanwhile, with the rapid development of technologies, graphs that do not fit into the memory are now commonplace: e.g., more than one billion Facebook users form a huge friendship network, and trillions of web pages are linked to each other on the Web.

Several systems have been developed to handle such enormous graphs but they all fail in processing web-scale graphs. Single machine systems like GraphChi (Kyrola, Blelloch, and Guestrin 2012) and distributed memory systems like GraphX (Gonzalez et al. 2014) cannot process graphs exceeding the external memory space and the distributed memory space, respectively. Even Pegasus (Kang, Tsourakakis, and Faloutsos 2009), a MapReduce based system resolving the space shortage problem, cannot process extra large graphs because of heavy network and disk I/Os by reading and shuffling the entire data many times.

In this work, we introduce PegasusN, a scalable and versatile graph mining package that runs on Hadoop and Spark. To handle enormous graphs, PegasusN provides and seamlessly integrates distributed algorithms for the following four types of operations with various applications (see Figure 1).

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Pregelix
PowerGraph
Hama
GraphX
Giraph

data that will not change in the future. Gaussian significantly reduces communication cost as well as process change over time and dwindle rapidly in size. PegasusN avoids the generation of massive intermediate data and balances workloads for superior performance in distributed systems. Pre-partitioning — many iterative graph algorithms require dividing the entire graph into overlapping or non-overlapping subgraphs multiple times. Instead of generating subgraphs from scratch in every iteration, PegasusN partitions the graph into blocks that do not share any edges with each other and makes each subgraph by combining blocks. The blocks are created in consideration of the sparsity of the graph so that subgraphs are similar in size, and thus workloads are evenly distributed. By caching the blocks in memory or on disk once at the beginning, PegasusN avoids the generation of massive intermediate data in every iteration. Pruning — in iterative algorithms, data to process change over time and dwindle rapidly in size. PegasusN significantly reduces communication cost as well as computational cost by excluding from the next iteration the data that will not change in the future.

Performance
Figure 2 shows the running times of PegasusN and existing graph mining systems on three tasks: PageRank, connected components, and triangle enumeration. We use a cluster of 20 machines equipped with an Intel Xeon E5-2640v3 and 32GB RAM. We use five real-world graphs summarized in Table 1. PegasusN is the only one that succeeds in processing CW12, the largest graph used in this experiment.

Conclusion
PegasusN is a scalable graph mining system for Hadoop and Spark. The system supports four types of graph algorithms (graph structure analyses, subgraph enumeration, graph generation, and graph visualization) in a distributed manner. PegasusN shows the best performance on real world graphs, processing CW12, the largest graph used in this paper with 6.2 billion nodes and 72 billion edges. PegasusN provides extremely useful toolsets that are needed by data mining and machine learning researchers and practitioners.

Figure 2: Running times of various systems on three tasks. Missing methods for some datasets mean they failed on the datasets.

Table 1: Summary of datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Nodes</th>
<th>Edges</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW12</td>
<td>6231126594</td>
<td>71746553402</td>
<td>Webgraph by Lemur Proj. (2009)</td>
</tr>
<tr>
<td>CW09</td>
<td>1684876525</td>
<td>7939647897</td>
<td>Webgraph by Lemur Proj. (2012)</td>
</tr>
<tr>
<td>YW</td>
<td>720242173</td>
<td>6636600779</td>
<td>Webgraph by Yahoo (2002)</td>
</tr>
<tr>
<td>TWT</td>
<td>41652230</td>
<td>1468365182</td>
<td>Follow network in Twitter</td>
</tr>
<tr>
<td>LJ</td>
<td>4847571</td>
<td>6893773</td>
<td>Friendship network in LiveJournal</td>
</tr>
</tbody>
</table>

Acknowledgments
This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.R0190-15-2012, High Performance Big Data Analytics Platform Performance Acceleration Technologies Development). U Kang is the corresponding author.

References