Graphillion: ZDD-based Software Library for Very Large Sets of Graphs

Takeru Inoue\textsuperscript{1} Hiroaki Iwashita\textsuperscript{1,3} Jun Kawahara\textsuperscript{2} Shin-ichi Minato\textsuperscript{3,1}

\textsuperscript{1} JST ERATO Minato Discrete Structure Manipulation System Project, Sapporo 060-0814, Japan
\textsuperscript{2} Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0192, Japan
\textsuperscript{3} Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

Abstract—Graphillion is a library for manipulating very large sets of graphs, based on zero-suppressed binary decision diagrams (ZDDs) with advanced graph enumeration algorithms. Graphillion is implemented as a Python extension in C++, to encourage easy development of its applications without introducing significant performance overhead. Experimental results show that Graphillion allows us to manage an astronomical number of graphs with very low development effort.

I. INTRODUCTION

A graph is a representation of a set of edges, each of which connects a pair of vertices. It is often used as a mathematical model for a variety of problems. Researchers have developed many sophisticated graph libraries, but the design was focused on handling a small number of graphs. Thus they cannot work with very large sets of graphs, even though the set can grow exponentially with graph size since a graph with $N$ edges induces $2^N$ subgraphs. A graph library that could efficiently manage very large and complex sets of graphs within a small amount of memory would provide a novel way for powerful graph operations; e.g., an optimizer that efficiently finds the best graph from a non-convex graph set, and a graph database that can select all matched graphs from a very large set.

In this paper, we introduce Graphillion, a software library optimized for very large sets of graphs. Traditional graph libraries maintain each graph individually, which causes poor scalability, while Graphillion handles a set of graphs collectively without considering individual graph. Graphillion concentrates on edge-induced subgraphs of a given graph $G = (V,E)$, and a set of graphs is reduced into a set of edge collections\textsuperscript{1}, or a family of sets of edges more formally; i.e., a set of graphs, $\{G_1 = (V,E_1), G_2 = (V,E_2)\}$, is regarded as a set of edge collections, $\{G_1 = E_1, G_2 = E_2\}$. This reduction loses the properties of each vertex, but allows programmers to apply a powerful theory on the family [1]. A set of collections can be represented in a compressed form by sharing common parts of similar collections, so a huge number of graphs can be stored in a small amount of memory. We also employ efficient algebra called family algebra [2], in order to perform optimization, selection, and modification on very large graph sets; the efficiency is due to the fact that they can be executed without decompressing the data.

This family theory, of course, is uncerned about graph structure like a tree or a path, since it considers a graph to be just an edge collection with no structure. We rectify this omission by employing the graph enumeration algorithm called frontier-based search [3, 4, 5]. The algorithm lists all graphs that have a specified structure, and then the listed graphs (edge collections) are handled by family algebra. The number of graphs listed, of course, can be very enormous, but a recent development in enumeration algorithms allows us to output the graphs in compressed form without enumerating them one by one. This compressed form is easily converted into the compressed form of the family theory [6], and so there is no difficulty to adopt family algebra.

There are several graph libraries like NetworkX [7] and Boost Graph Library [8]. These libraries are widely used for graph analysis. They are, however, designed for a small number of graphs or a simple power set of edges; i.e., they can find a shortest path just from a power set of edges without constraints. We often use general optimizers like CPLEX [9] for graph optimization. However, they require us to describe the constraints in simple formuleae, but many practical problems are too complicated to permit this. In addition, general optimizers are not designed to search for multiple solutions. Graph databases [10] store multiple graphs and provide selection methods on graph structure. However, they do not employ efficient graph set representation. VSOP [11] employs family algebra as does Graphillion, but it provides an abstraction for combinatorial item sets, not graph sets. Frontier search is, of course, not implemented in VSOP, and so it does not create graph sets of a given structure efficiently.

The rest of this paper is organized as follows. Sections II and III discuss the theoretical aspects of Graphillion. Section IV describes implementation, Section V reports experimental results, and Section VI concludes the paper.

II. REPRESENTATIONS OF A GRAPH AND THE SET

This section formulates a graph set as a set of edge collections. Fig. 1 shows an example of the representation used in this section.

---

\textsuperscript{1}In order to describe a set of sets without confusion, the word \textit{collection} is used to indicate an “inner” set like an edge set, while \textit{set} is used for an “outer” set like a graph set.
A. Representation of a Graph

We first introduce a special graph that defines our universe (Fig. 1 (A)),

\[ U = (V_u, E_u) = \begin{array}{ccc} v_1 & v_2 \\ v_3 & v_4 \end{array} \]

where only edge collection \( E \) defines the graph, while vertices \( V \) are ignored. This simplification puts a limitation on vertices; vertices without edges cannot be recognized. However, graphs are mainly characterized by edge structures in many applications, making this limitation not a serious concern in most cases.

B. Representation of a Set of Graphs

A set of graphs, \( \mathcal{G} \), is represented by a set of collections of \( E_u \) (Fig. 1 (B, C)),

\[ \mathcal{G} \subseteq 2^{E_u}, \quad \text{e.g., } \mathcal{G} = \{ \begin{array}{ccc} \\ & \bullet & \\ & \bullet & \bullet \end{array} \} \]

where \( 2^{E_u} \) is the power set of \( E_u \). A graph used in Graphillion is defined by \( G \in \mathcal{G} \).

III. Creation and Manipulation of a Set of Graphs

This section describes the creation of a graph set using frontier search and the use of family algebra to manipulate set contents.

A. Creation of a Set of Graphs

We build a ZDD representing a set of graphs by using a graph enumeration algorithm called frontier-based search [3, 12], which integrates several advanced techniques. Frontier search finds all graphs that have a specified structure based on dynamic programming. It outputs the enumerated graphs in a compressed form that is easily converted into a ZDD [6]. The time complexity is ruled by the size of the compressed form (slightly larger than that of ZDD), not the number of graphs being output.

Frontier search was originally limited to trivial structures like trees, but it has been generalized to support various structures [4]. Table II shows the structures supported by Graphillion.

The search space can be limited within a given graph set; graphs not included in the given set are not enumerated by frontier search [5].

Simple graph sets can be created by ZDD’s primitives without frontier search; an empty set and a power set are given by
(A) membership query: e.g., \( \mathcal{G} \) is found in \( \{ \mathcal{G}, \mathcal{H} \} \)
\[
\{ \mathcal{G}, \mathcal{H} \} \cap \{ \mathcal{G}, \mathcal{I} \} = \{ \mathcal{G} \} \neq \emptyset
\]
(B) search: e.g., structure \( \mathcal{G} \) is found in \( \mathcal{H} \)
\[
\{ \mathcal{G}, \mathcal{H} \} \ni \{ \mathcal{G} \} = \{ \mathcal{G} \}
\]
(C) graft: e.g., edges \( \mathcal{I} \) are added to \( \{ \mathcal{G}, \mathcal{H} \} \)
\[
\{ \mathcal{G}, \mathcal{H} \} \cup \mathcal{I} = \{ \mathcal{G}, \mathcal{H} \}
\]

Fig. 2. Examples of graph set manipulation via family algebra.

the ZDD’s primitives, and small graph sets can be created by explicitly specifying the graphs (edge collections).

B. Manipulation of a Set of Graphs

Family algebra defines several operations on sets of collections, and the operations can be efficiently performed over ZDDs [2]. Surprisingly, these operations can be executed on the compressed data without decompression, so they are highly efficient. In this subsection, we describe the operations for optimization, selection, and modification, in the context of graph sets.

Several selection operations are defined for a set of collections, and their semantics make sense for graph sets without change. They include ordinary set operations, namely, union (\( \cup \)), intersection (\( \cap \)), difference (\( \setminus \)), and symmetric difference (\( \oplus \)). The intersection operation can be used for a membership query; to test if graph \( G \) is in set \( \mathcal{G} \) by checking \( \{ G \} \cap \mathcal{G} \neq \emptyset \) (Fig. 2 (A)).

Other selection operations are based on graph structures, which include subgraphs (\( \subseteq \)), supergraphs (\( \supseteq \)), maximal graphs (\( \uparrow \)), and minimal graphs (\( \downarrow \)). They do what their names suggest (they are originally called subsets or maximal sets in family algebra). The supergraphs operation can be used for search; to explore \( \mathcal{G} \) for graphs that include given structure \( G \) by \( \mathcal{G} \supseteq \{ G \} \) (Fig. 2 (B)).

All graphs in a set \( \mathcal{G} \) can be modified at once by slightly modifying family algebra. To graft edge(s) \( E \) to all graphs in set \( \mathcal{G} \), we utilize join (\( \cup \)) operation defined in the family algebra (Fig. 2 (C)). Similarly, edge(s) \( E \) can be removed by performing meet (\( \cap \)) operation against the complement edge set \( E^c = E_\cup \setminus E \) (i.e., \( E^c \) are edges not to be removed in this context). The flip (\( \equiv \)) operation flips edge status in all graphs.

Optimization is provided by a search algorithm of family algebra. To find a maximum or minimum weighted edge collection (graph) in the set. Since this search algorithm returns just a single best graph, we employ the difference operation to obtain multiple graphs in descending (or ascending) order of weight; the search algorithm is applied repeatedly while removing the previous best graph from the set by the difference operation as follows.

for \( i = 1 \to k \) do // find top-k graphs from \( \mathcal{G} \)
\[
G = \text{find_max}(\mathcal{G}) // \text{get best} \ G \text{ from } \mathcal{G}
\]
// do something with \( G \)
\[
\mathcal{G} = \mathcal{G} \setminus \{ G \} // \text{remove} \ G \text{ for the next iteration}
\]
end for

Graphillion defines other operations like hitting sets [13], random sampling, and counting graphs in a set, but we do not describe them due to space limits.

IV. IMPLEMENTATION

Frontier search and family algebra are implemented in C++, while the programming interface is written in Python. This interface is based on Python’s \texttt{set}; e.g., the size query (\texttt{len} function in Python), membership query (\texttt{in} operation), iterators (\texttt{for} operation), and general set operations (\texttt{union}, etc.). We add graph-specific operations to this interface like supergraphs, graft, and the graph-weight optimizers. Our implementation requires 14,965 lines of code in C++ and 2,251 lines in Python.

A graph set object in Python maintains a reference to the corresponding ZDD object of C++. The graph set object is very lightweight, since it has no attribute other than the reference. The selection methods return a new graph set object that refers to the associated ZDD object. The modification methods just replace their reference with a new reference to the new ZDD object. The optimizers are implemented as a Python iterator, which runs a loop step by step and yields the best graphs one by one instead of extracting all of them at once.

In order to enhance productivity further, any type of graph object (e.g., NetworkX graph) can be used in Graphillion. A graph object is transparently converted into the Graphillion’s internal representation (an edge collection) by user-defined converters. Programmers can use Graphillion as an enhancement tool for their favorite graph modules simply by registering the converters.

V. EXPERIMENTS

In this section, we first show the performance of Graphillion’s operations. We then discuss two case studies, a puzzle solver and a power network optimizer, to examine the tradeoff between performance and productivity. All experiments were conducted with Python 2.7 and GCC 4.7 on Linux 2.6 using a single core in Intel Xeon E31290 (3.60 GHz) with 32 GB of RAM.

A. Basic Performance

We evaluate the performance using a set of trees rooted at a corner on a grid graph. The set size is shown in Table I. The performance of creation is measured by building a set of the trees. The selection performance is evaluated by calculating the union of two sets of trees; trees in one set are rooted at a corner while those in the other set are rooted at the diagonally opposite corner. The modification performance is evaluated by
grafting an edge to all trees. The optimization performance is measured by finding the top-3 weighted trees with the maximizing operation.

We measured the CPU time and the memory usage of these operations with and without Graphillion. In the implementation without Graphillion, graphs are created as NetworkX objects, and are stored in Python’s built-in set object (the union operation is provided by the built-in set, but the other operations were added by us). In order to evaluate Python’s overhead, we developed pure C++ implementation of the operations just for the experiments.

The results are shown in Fig. 3. The implementation without Graphillion could not finish any operation for a $5 \times 5$ grid within an hour due to the very large number of trees. Graphillion performs a little poorly on the small grids due to the overhead of object mapping and conversion, but the overhead is negligible in the larger grids. It finished all operations in less than 10 seconds with 500 MB of memory even for the $9 \times 9$ grid, which has $10^{37}$ trees. Creation and optimization are slower than the other operations, because they involve complicated search algorithms. Selection requires twice as much memory than the others since it uses two sets of trees\(^2\), but it is the fastest due to its simple operation.

B. Puzzle Solver

The first case study is the Slitherlink puzzle\(^3\), which is a logic puzzle to find a cycle that satisfies given hints (Fig. 4). The Slitherlink solver in [14] was the fastest one that could list all solution cycles. The solver employs frontier search redesigned for Slitherlink; it has special algorithms to process hints. It is written in 2,116 lines of C++ code.

We developed another solver with Graphillion, without frontier search dedicated for Slitherlink. This new solver, first, enumerates subgraphs that satisfy the hints, and then runs frontier search over the hint-satisfying subgraphs to select solution cycles. Thanks to the generality of Graphillion, the new solver is written just in 153 lines of Python. This is a 93 \% reduction in code line number, and it is, in addition, written in easy Python, not in complicated C++.

We measure the CPU time and memory usage on three problems found in a Slitherlink book [15], all of which have just a single solution. We also conduct an experiment against a modified problem in which ten hints are randomly removed to permit multiple solutions. Fig. 5 shows the results. Both solvers scaled similarly with problem size, and their memory usages were roughly comparable. The Graphillion solver is slightly outperformed in CPU time due to the special algorithms in the dedicated solver, but the tradeoff between performance and productivity is acceptable.

---

\(^2\)Selection requires 500 MB of memory, which is slightly larger than double the theoretical value (207 MB), shown in Table I, because of the unused slots in the hash table used to maintain ZDDs. The flat regions seen in the memory usage for smaller grids are also due to the pre-allocated slots of the hash table.

\(^3\)http://www.nikoli.com/en/puzzles/slitherlink/
We can obtain top-k longest or shortest cycles with Graphillion’s iterators, when the problem has multiple solution cycles. It took just another 0.24 seconds to find the three longest cycles from among the 117059496 solutions in the modified problem.

C. Power Network Optimizer

The second case study is power loss minimization in a distribution network; this is a discrete non-convex optimization problem involving hundreds of variables [16]. A power distribution network can be represented by a graph in which a vertex corresponds to a town block or a power substation while an edge is a power line with a switch. The power flow is configured by changing the open/closed status of switches. It must be cycle-free to avoid short circuits, and must cover all blocks to avoid blackouts; the power flow, as a consequence, forms a spanning forest, in which each tree is rooted at a power substation. The flow also must satisfy complicated electrical constraints on line capacity and voltage drop; roughly speaking, very large or tall trees are forbidden. The network is operated to minimize resistive line losses while satisfying these constraints.

In [17], a power loss optimizer is developed using frontier search and family algebra in an ad-hoc manner without the unified concept discussed in this paper. The loss optimizer first enumerates all spanning forests rooted at substations by frontier search (1st line of Fig. 7). It then enumerates all spanning forests rooted at substations by conducting complicated power calculations (2nd line of Fig. 7). Family algebra selects forests that do not include the infeasible trees (3rd line of Fig. 7). Finally, the minimum-loss forest is found from the selected feasible forests; since the search space consists of only the feasible forests, the search algorithm does not need to consider the complicated constraints. To handle the nonlinear nature of the power loss, a dedicated search algorithm had been developed (that of family algebra was not used). Our past work implemented a part of frontier search and of family algebra in 6,856 lines of C++ code, while the complicated power calculations, including nonlinear optimization, was written in 1,221 lines of Python code. Intermediate data are serialized into a file, which is exchanged between the C++ program and the Python program.

We developed another power loss optimizer that implements the same algorithms but employs Graphillion for frontier search and family algebra; we are allowed to focus on the power calculation and the nonlinear optimization. Since this optimizer is implemented as a single program, it does not need to exchange intermediate data. It is written in 1,164 lines of Python code without C++. This Python code is shorter than the original, because it does not require serialization and object mapping. In total, we achieved a 86 % reduction in code line number, and it is, in addition, written in easy Python, not complicated C++.

The two optimizers are compared for power distribution networks used in [17]; the largest network has 432 blocks (vertices) and 468 power lines (edges). The results are shown in Fig. 8. Both implementations demonstrate comparable performance in the CPU time and memory usage (the memory usage includes both C++ and Python programs for the implementation without Graphillion). The Graphillion optimizer was slightly faster due to its omission of data exchange, while it required a bit more memory because of the full Python implementation. This memory overhead is negligible compared to the productivity improvement, which allows programmers to focus on their own problems without considering complicated graph operations. Surprisingly, more than $10^{58}$ feasible forests were handled with only 1.5 GB memory in the largest network. Graphillion needed just one thousand code lines to find an optimal solution from a non-convex set of $10^{58}$ graphs in just one minute.

Graphillion also can be used as a graph database of feasible forests. We issued queries specifying an open/closed switch to select all the forests matched to the queries, like Fig. 2 (B). Graphillion processed the queries within just 1.5 seconds for a closed switch and within 0.5 seconds for an open switch in the largest network.

<table>
<thead>
<tr>
<th>Grid size</th>
<th>CPU time [sec.]</th>
<th>Memory usage [M Byte]</th>
</tr>
</thead>
<tbody>
<tr>
<td>11x11</td>
<td>w/ Graphillion</td>
<td>0.01</td>
</tr>
<tr>
<td>19x11</td>
<td>w/ Graphillion</td>
<td>0.1</td>
</tr>
<tr>
<td>37x21</td>
<td>w/ Graphillion</td>
<td>10000</td>
</tr>
<tr>
<td>37x21 (mod.)</td>
<td>w/ Graphillion</td>
<td>100000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grid size</th>
<th>CPU time [sec.]</th>
<th>Memory usage [M Byte]</th>
</tr>
</thead>
<tbody>
<tr>
<td>11x11</td>
<td>w/o Graphillion</td>
<td>0.01</td>
</tr>
<tr>
<td>19x11</td>
<td>w/o Graphillion</td>
<td>0.1</td>
</tr>
<tr>
<td>37x21</td>
<td>w/o Graphillion</td>
<td>10000</td>
</tr>
<tr>
<td>37x21 (mod.)</td>
<td>w/o Graphillion</td>
<td>100000</td>
</tr>
</tbody>
</table>

Fig. 5. CPU time and memory usage of the dedicated solver and the Graphillion solver on Slitherlink problems.

Fig. 8. CPU time and memory usage of power network optimizers with and without Graphillion.
VI. CONCLUSIONS

In this paper, we have introduced Graphillion, which is a software library designed for very large sets of graphs. Our representation of a graph set allows us to utilize the theory of the “family of sets”, which can compress graph sets and manipulate them efficiently. Graphillion is implemented in Python and provides a sophisticated but easy to use interface. Experiments showed the excellent performance of Graphillion. Two case studies revealed that programmers can handle very large graph sets with just a small number of lines of code.

Future work includes a plug-in mechanism for operation customization, generalized design for directed graphs and hyper graphs, and analysis of compression ratio on graph set characteristics. Since we would like to find out more applications for which Graphillion works well, we make it publicly available online.

REFERENCES


