On Small Universal Data Structures and Related Combinatorial Problems

Richard A. DeMillo
Stanley Eisenstat
Richard J. Lipton

July 1978

This document has been approved for public release and sale; its distribution is unlimited.

* Georgia Institute of Technology
** Yale University
ON SMALL UNIVERSAL DATA STRUCTURES
AND RELATED COMBINATORIAL PROBLEMS †
(Preliminary Report)

Richard DeMillo
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, GA 30332

Stanley Eisenstat
Richard J. Lipton
Department of Computer Science
Yale University
New Haven, CT 06520

† Work supported in part by U.S. Army Research Office, Grant No. DAAG29-76-G-0338, and by the Mathematics Research Center, University of Wisconsin-Madison.
INTRODUCTION

One of the most significant changes in theoretical computer science has been the recent infusion of the methods and problems from combinatorial analysis. Among the most powerful combinatorial theorems which have been imported to computer science are those of extremal graph theory [1]: in extremal graph theory, one is interested in the largest (or in complementary problems, the smallest) graph which avoids (or contains) a given structure. Purely combinatorial results (which have significance, e.g., for the design of circuit boards) have been obtained by Chung and Graham [2] and by Chung, Graham, and Pippenger [3]. In this paper, we extend this theory to encompass results concerning data structures.

As motivation for the results to the described, note that many of the large data structures manipulated by the programs described in [4,5] have two characteristics

(i) they are sequentially accessed, and

(ii) many distinct structures convolve in the same physical memory.

For applications of this sort, it would obviously be desirable to have available a universal data structure in which all data structures from a given class may gracefully reside. In view of (i), by "graceful" we mean that the sequential accessing characteristics of the embedded data structures are not too drastically altered. Let us measure such alterations by the dilation of logical adjacencies [6,7] needed to embed all structures from a given class into a universal structure; this is then a complementary extremal graph theory problem: what is the size (number of edges) of the smallest universal graph for a given dilation factor.
The main results contained in this paper address such problems from a number of points of view.

1. We give several asymptotically optimal universal data structures for graphs of \( n \) vertices when average dilation [7] is used as a measure.

2. We discuss a universal data structure for graphs of \( n \) vertices where worst-case dilation is used as a measure [6].

3. We consider variations of the average dilation measure which gives favorable comparisons between data structures studied in [6,7].

4. We consider the kinds of "sharing" that can take place between "almost linear" and "almost complete tree-like" structures.

5. Finally, we propose a data structure embedding model which recovers some aspects of random accessing of data items, and prove a space-time tradeoff which seems to indicate that no savings is possible in RAM models which assess accessings costs uniformly [8].

**PRELIMINARIES**

A graph, \( G \), is defined by its vertices, \( V(G) \), and edges, \( E(G) \subseteq V(G) \times V(G) \). Edges are assumed to be undirected: a pair of vertices \( x, y \) are connected if either \( (x, y) \in E(G) \) or \( (y, x) \in E(G) \). A path between \( x_0, x_n \) is said to be of length \( n \). The distance metric \( d_G(x_0, x_n) \) is defined to be \( n \) if there is no shorter path than \( x_0, \ldots, x_n \).

A graph represents a **data structure** in the obvious way: vertices represent nodes or records and connectedness models logical adjacency. The following relations and their significance for data structures can be found in [6,7]. Let \( G, G^* \) be graphs. We say that \( G \) is **\( T \)-worst case embeddable in \( G^* \)** (\( G \subseteq_T G^* \)) if there is a one-one \( \phi: V(G) \rightarrow V(G^*) \) such that \( (x, y) \in E(G) \) implies

\[
d_{G^*}(\phi(x), \phi(y)) \leq T. \tag{1}
\]
Similarly, G is A-average case embeddable in G* (G \leq_{AVG} G*) if there is a one-one \phi as above such that
\[ \sum_{x,y} d_{G*}(\phi(x), \phi(y)) \leq A \cdot |E(G)|. \] (2)

In [4,5], comparisons between several natural classes of graphs give asymptotic bounds on T, A in (1), (2) as functions of |V(G)|. Shortly after the announcement of the results of [6], R. M. Karp suggested to us the following class of problems connected with extremal graph theory: what are the characteristics of \leq_T - universal data structures; i.e., those structures which T-worst case embed all graphs in a given class. This paper grew out of considering these problems.

**UNIVERSAL GRAPHS**

Let \( \zeta^n \) be a given class of graphs G, |V(G)| = n. Let us ask about a data structure which is \leq_T or \leq_{AVG} universal for \( \zeta^n \). In particular, let us define
\[ w(\zeta^n, T) = \min \{ |E(G)| : G^n \in \zeta^n, G^n \leq_T G \} \] (3)
and
\[ a(\zeta^n, A) = \min \{ |E(G)| : G^n \in \zeta^n, G^n \leq_{AVG} G \}. \]

For T = 1, (3) becomes the complementary extremal graph problem studied in [2,3].

By an n-tree G, we mean a connected acyclic graph G, with |V(G)| = n. It is also convenient to think of trees as rooted in the following sense: accompanying G, there is an ancestor-descendent relation that assigns direct ancestors and direct descendants to vertices in the obvious way so that a vertex with no ancestors can be designated as the root of the tree.
(Obviously this choice is not going to be unique, but we assume that $G$ is not characterized until such a choice is made). A $d$-ary $n$-tree is an $n$-tree in which each vertex has at most $d$ direct descendents. We denote, respectively, the classes of $n$-trees and $d$-ary $n$-trees by $\Gamma^n$ and $\Gamma^n_d$.

By [2] it is known that $\frac{1}{4}n \log n < w(\Gamma^n, 1) < n^{1+\kappa(n)}$, $k(n) = \frac{1}{\log \log n}$.  

The upper bound was improved in [3] to

$$w(\Gamma^n, 1) = O(n \log n \log \log n^2)$$

The bounds on $w(\Gamma^n, 1)$ are apparently not elsewhere considered.

Superficially, at least, all interest in further characterization of (3) is destroyed by the following obvious

Theorem. For $T \geq 2$

$$w(\Gamma^n, T) = n$$

Of course, in (3), the "target" graph $G$ may have unbounded degree. Therefore, it is natural to consider $w(\Gamma^n, T, S)$ and $a(\Gamma^n, T, S)$ where in both cases the target graph $G$ is restricted to be in the set $S$. Note that now the theorem just cited is no longer obviously true.

* Thus $\Gamma^n_2$ = binary trees on $n$ vertices.

† In the sequel, we use $\log x$ for $\log_2 x$ and $\ln x$ for $\log_e x$. 

4
The best that is known is the upper bound of [3] \( S = \) all cubic graphs

\[
 w(\Gamma^n, 1, S) \leq \frac{2\sqrt{2}}{n} \exp \left( \log^2 n / 2 \log 2 \right). \tag{4}
\]

It is not obvious that when (i) "targets" are restricted to binary trees and (ii) \( w(\Gamma^n, T, \Gamma^n_2) \) is considered, that it is possible to do any better than the union of all trees in \( \Gamma^n_2 \), giving a structure of size \( 4^n/2^{n^2} \).

But, we have the following

**Theorem.** For each \( T \geq 1 \), there is a binary tree \( H \), such that \( G \subseteq H \) for all \( G \in \Gamma^n_2 \), and

\[
 \ln |E(G)| \leq \frac{2n^2}{\ln 4};
\]

or in other words

\[
 w(\Gamma^n_2, T, \Gamma^n_2) = \exp \left( \frac{1}{\ln 4} (\ln n)^2 + O((\ln n)^2) \right).
\]

A key step in the proof of this theorem hinges on the solution to the fascinating "almost linear" recurrence

\[
 u_n = u_{n-1} + \left\lfloor \frac{u_n}{2} \right\rfloor, \tag{5}
\]

first considered by Knuth [9]. This also establishes a connection between the theorem and ineq. (4): \( u_n \) is also the number of partitions of \( 2n \) of the form \( \sum_{i=1}^{\alpha_i} 2^i \), \( \alpha_i = 0, 1 \). Knuth [9] bounds the partition function

\[
 P(m) = \frac{1}{4\sqrt{3m}} \exp \left( \pi \frac{2}{3} m \right). 
\]
There are two possibilities for improving the bounds in \( w(\Gamma_2^n, T, \Gamma_2^n) \). The first possibility is to introduce circuits to the target graph of the previous theorem, but this does not appear to give an asymptotically better bound than (4). The second possibility is to prove that balanced trees and unbalanced trees are \( \leq_T \) - equivalent. This seems unlikely since combining such a result with the proof method of the previous theorem gives a polynomial sized universal tree. However, in trying to improve the bounds on \( w(\Gamma_2^n, T, \Gamma_2^n) \) it may be desirable to ignore irregular trees, letting only very balanced or very unbalanced trees reside in the same universal data structure.

In any case, it seems unlikely that polynomial structures are possible. We are, however, far from proving this; indeed, the best known lower bound is the following

**Theorem.** For all \( n > N \)

\[
w(\Gamma_2^n, T, \Gamma_2^n) > c(T) \cdot n \log n,
\]

where \( c(T) > 0 \) is a constant for fixed \( T \geq 1 \).

Certain other subcases are also of interest. Erdős, Chung, and Graham, consider \( w(S,1) \) and obtain

\[
w(S,1) \leq \frac{4}{11} \cdot n^2.
\]

The following theorem is an improvement, but is surely not the best possible bound.

**Theorem**

\[
w(S,1) \leq \frac{2}{9} \cdot n^2
\]
A non-trivial lower bound would clearly be desirable. Another class of interest are graphs of high genus. We conjecture that for graphs of fixed genus $\gamma$, it is possible to do better than the naive $\binom{n}{2}$ bound obtained by embedding in the complete graph.

Our next series of results show impressive improvements by passing to average dilations. We now get optimal constructions, even in a variety of limited settings.

We have, for instance, the

**Theorem.** For $\alpha > 0$,

$$a(\Gamma^a_2, \frac{1}{\alpha}, S) = o(n^{\log(2+\alpha)}) .$$

Since there is a linear lower bound on $a(\gamma, \gamma, \gamma)$, this construction is optimal. By a slight modification of the construction, this gives $a(\Gamma^a_2, A, S) = o(n)$, for all $A \geq 1$, but this result may be superceeded by the following

**Theorem.** For each $A \geq 1$, there is a binary tree $H$, such that

$$G \leq \frac{\text{avg}}{A} H$$

for all $G \in \Gamma^a_2$, and

$$|E(G)| = o(n) ;$$

or, in other words

$$a(\Gamma^a_2, A, \Gamma^a_2) = o(n) .$$

†† A graph is of genus $\gamma$ if it can be embedded in a sphere with $\gamma$ handles [10].
These results are related to the ability to "cut" graphs in advantageous ways. For example, a generalization of the planar separator theorem [11] to graphs of high genus, obtained by Lipton and Tarjan, gives us the following

**Theorem.** Let $L^n_\gamma$ be the class of graphs $G$ with genus $\gamma$ and $|V(G)| = n$. Then, for all $n > N$, 

$$a(L^n_\gamma, A, \Gamma_2^n) \leq c(A) \cdot n,$$

where $c(A)$ does not depend on $n$.

**EXTENDED MODEL**

In comparing classes of data structures (see, e.g., [6,7], the measures of "efficiency" have implicitly assumed that only sequential accessing is important. Thus, when in [6], we bound the efficiency, $T$, of an embedding of $n \times n$ array into binary trees by

$$T \geq c \log n,$$

the function $T(n)$ captures the dilation factor in an embedding. We now describe a generalization of this concept which recovers a certain kind of random accessing. Since the precise definitions are quite complex, we will settle for a less exact — but more picturesque — rendering. Let us assume that we have in front of us an illustration of a graph $G$, and also a number of friends who agree to lend us their forefingers for use in tracing the paths of the graph. Our friends oblige us as follows: We may start traversing at any vertex already visited. The traversal rule is, then, that we must either traverse graph edges or "jump" to a vertex pointed to
by a friend. The time required to traverse a sequence of vertices is then simply the number of applications of traversal rules. Notice that the result of a traversal is not necessarily a path of G. The connection between fingers and random accessing is that traversals requiring k-fingers also require k-"addresses" for the vertices pointed to.

We then say that $G \preceq_{k,T} G^*$ if there is a one-one $\phi: V(G) \rightarrow V(G^*)$, so that for every $x, y \in V(G)$ with $d_G(x, y) = m$, there is a k-finger traversal from $\phi(x) = x^*$ to $\phi(y) = y^*$ with time at most $\Delta$, and $\Delta \leq t_{G^*}(x^*, y^*)$.

We have the following

**Theorem.** If $G_n$ is the $n \times n$ array [7], H is a binary tree and

$$G_n \preceq_{k,T(n)} H,$$

then

$$k + T(n) \geq c \log n,$$

where $c$ is a constant independent of $n$.

**OTHER TYPES OF AVERAGE EMBEDDING**

The relation $\preceq_{\text{avg}}$ may be thought of as averaging - with relative frequencies uniformly distributed to the edges $E(G)$ - over the edges of G. We now make a more global definition which may be used to recover our intuitions about path lengths in binary trees [7]. We will essentially average our shortest paths:

$G \leq_{\text{paths}} G^*$ if there is an embedding $\phi: V(G) \rightarrow V(G^*)$ such that

$$\sum_{\phi(x), \phi(y)} d_{G^*}(\phi(x), \phi(y)) \leq A \cdot \sum_{x,y} d_G(x, y).$$
We then have the following

**Theorem.** For each $n \geq 0$, let $A_n$ be the least real number such that

$$G_n \leq \frac{\text{paths}}{A},$$

for a binary tree $H$. Then

$$\lim_{n \to \infty} A_n = 0$$

Thus, we see that if the average embedding is required to work well on all shortest paths, then the embedding cost goes to zero. In a sense, then $\frac{\text{avg}}{A}$ "charges" more heavily than $\frac{\text{paths}}{A}$ for any bottlenecks.

**REFERENCES**


On Small Universal Data Structures and Related Combinatorial Problems

Richard A DeMillo, Stanley Eisenstat, Richard J. Lipton

School of Information and Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30332

Approved for Published Releases; Distribution Unlimited