DIMACS Technical Report 98-50
November 1998

Extremal Phylogeny Numbers

by

Fred S. Roberts[1]
Department of Mathematics, RUTCOR, and DIMACS
Rutgers University
New Brunswick, New Jersey 08903

Li Sheng[2]
Department of Mathematics and Computer Science
Drexel University, Philadelphia, PA 19014

[1]The author gratefully acknowledges the support of the National Science Foundation under grants BIR 94-12594 and SBR 97-09134 to Rutgers University.
[2]The author gratefully acknowledges the support of the National Science Foundation under grants BIR 94-12594 and SBR 97-09134 to Rutgers University.

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, Bellcore and NEC Research Institute.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

ABSTRACT

Given an acyclic digraph $D = (V, A)$, define its *phylogeny graph* $G = P(D)$ by taking the same vertex set as D and, for $x \neq y$ in V, letting $xy \in E(G)$ if and only if $(x, y) \in A$ or $(y, x) \in A$ or $(x, a), (y, a) \in A$ for some vertex $a \in V$. Given a graph $G = (V, E)$, we shall call the acyclic digraph D a *phylogeny digraph* for G if G is an induced subgraph of $P(D)$ and D has no arcs from vertices outside of G to G. The *phylogeny number* $p(G)$ is defined to be the smallest r such that G has a phylogeny digraph D with $|V(D)| - |V(G)| = r$. In this paper, we apply the famous theorem of Turán to show that for graph G with n vertices, $p(G) \leq \lfloor \frac{n^2}{4} \rfloor - n + 1$ and that $p(G)$ achieves this upper bound if and only if $n \leq 3$ or $G = K(\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil)$.

1 Introduction

Motivated by problems in ecology, Cohen [1] and Roberts [6], respectively, introduced the notions of competition graph and competition number, and motivated in part by the phylogenetic tree reconstruction problem, Roberts and Sheng [10] introduced analogous notions of phylogeny graph and phylogeny number. Competition graphs and competition numbers have been widely studied and one can find extensive references to the literature in Kim [3], Lundgren [4] and Roberts [8]. The analogies between competition graphs and phylogeny graphs, competition numbers and phylogeny numbers have been studied in Roberts [8] and Roberts and Sheng [9, 10, 11]. In this paper, we present one more analogous result by finding the largest phylogeny number of a graph with n vertices and determining the graphs that attain the optimal.

Suppose $D = (V, A)$ is a digraph, where V is the vertex set and A the set of directed arcs. We adopt the graph-theoretical notation and terminology of Roberts [7], except where noted otherwise. Throughout, $n(G) = n$ will represent the number of vertices of G and $e(G) = e$ the number of edges. We will use the notation $G_1 \cup G_2$ for the disjoint union of the two graphs G_1 and G_2. Unless explicitly mentioned, digraphs in this paper will have no loops, i.e., arcs from a vertex to itself.

The *competition graph* $K(D)$ is the undirected graph (V, E) with the same vertex set as D and with an edge $xy \in E$ if and only if there is a vertex a so that $(x, a) \in A$ and $(y, a) \in A$. Food webs are usually acyclic, and the study of competition graphs has often been restricted to the competition graphs of acyclic digraphs. Roberts [6] noted that for every graph G, G together with sufficiently many isolated vertices is a competition graph of some acyclic digraph. We shall denote by $G \cup I_k$ the graph consisting of G plus k isolated vertices. The smallest k so that $G \cup I_k$ is a competition graph of an acyclic digraph is called the *competition number* of G and is denoted by $k(G)$. Opsut [5] showed that the problem of calculating $k(G)$ is an NP-complete problem.

Given an acyclic digraph D, let us say its *phylogeny graph* $G = P(D)$ has vertex set $V(D)$ and has an edge between $x \neq y$ if and only if either $(x, a) \in A(D)$ and $(y, a) \in A(D)$ for some vertex $a \in V(D)$ or $(x, y) \in A(D)$ or $(y, x) \in A(D)$. Given G, we can always realize G as an induced subgraph of $P(D)$ for D an acyclic digraph. Simply add a new vertex $\alpha(x, y)$ for each edge xy in G, with arcs from x and y to $\alpha(x, y)$. Note that in D we do not have arcs from added vertices to vertices in G. Any acyclic digraph D for which G is an induced subgraph of $P(D)$ and such that D has no arcs from vertices outside of G to G is called a *phylogeny digraph* for G. The *phylogeny number* $p(G)$ of G is the smallest r so that G has a phylogeny digraph D with $|V(D)| - |V(G)| = r$. A phylogeny digraph D for G for which $|V(D)| - |V(G)| = p(G)$ is called an *optimal phylogeny digraph* for G. Roberts and Sheng [10] showed that the problem of calculating $p(G)$ is an NP-complete problem.

Harary, Kim and Roberts [2] introduced the study of extremal competition numbers by giving an upper bound on the competition number for graphs with n vertices.

Theorem 1 (Harary, Kim and Roberts [2]) *Let $G = (V, E)$ be a graph with n vertices. Then $k(G) \leq \lfloor \frac{n^2}{4} \rfloor - n + 2$. Moreover, this bound is achieved if and only if G is $K(\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil)$ or*

K_3.

In this paper, we study the analogous extremal problem for phylogeny numbers. That is, we seek the maximum phylogeny number of a graph with n vertices. As for the extremal competition number, we will be utilizing the famous theorem of Turán [12] in extremal graph theory.

Theorem 2 (Turán [12]) *A graph of n vertices and no triangles can have no more that $\lfloor \frac{n^2}{4} \rfloor$ edges, and this bound is achieved only by complete bipartite graph $K(\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil)$.*

2 The Maximum Phylogeny Number for Graphs with n Vertices

We shall need the following results of [10].

Theorem 3 (Roberts and Sheng [10]) *If $G = (V, E)$ is a triangulated graph, then $p(G) = 0$.*

Theorem 4 (Roberts and Sheng [10]) *If $G = (V, E)$ is a connected graph with no triangles, then $p(G) = e - n + 1$*

Theorem 5 (Roberts and Sheng [10]) *Given a graph $G = (V, E)$, let $G_1, G_2, \ldots, G_m$ be the connected components of G. Then $p(G) = p(G_1) + p(G_2) + \cdots + p(G_m)$.*

Let $f(n) = \lfloor \frac{n^2}{4} \rfloor - n + 1$. Then,

$$f(n) \geq 1 \text{ if } n \geq 4 \text{ and } f(n) > f(n-1) \text{ if } n \geq 4 \tag{1}$$

Lemma 6 *Let $G = (V, E)$ be a graph with n vertices.*

(1) *If G has no triangles and is connected, then $p(G) \leq f(n)$.*

(2) *If $n \geq 4$, G has no triangles and is disconnected, then $p(G) < f(n)$.*

Proof. If G has no triangles and is connected, then $p(G) = e - n + 1 \leq f(n)$ by Theorem 4 and Turán's Theorem. So (1) is true.

Suppose G has no triangles and is disconnected with connected components $G_1, G_2, \cdots, G_m$. Let n_i be the number of vertices in G_i for $i = 1, 2, \cdots, m$. If G has no edges, then $p(G) = 0 < f(n)$. So, we can assume without loss of generality that $n_1 \geq 2$.

Suppose $m = 2$. If $n_2 = 1$, then $p(G) = p(G_1) + p(G_2) = p(G_1) \leq f(n-1) < f(n)$ by Theorem 5, part (1) of this Lemma and Equation (1). If $n_2 \geq 2$, then

$$\lfloor \frac{(n_1 + n_2)^2}{4} \rfloor - 1 = \lfloor \frac{n_1^2}{4} + \frac{n_2^2}{4} + \frac{n_1 n_2}{2} - 1 \rfloor > \lfloor \frac{n_1^2}{4} + \frac{n_2^2}{4} \rfloor \geq \lfloor \frac{n_1^2}{4} \rfloor + \lfloor \frac{n_2^2}{4} \rfloor$$

So by Theorem 5 and part (1) of this Lemma, we have that

$$\begin{aligned}
&p(G) = p(G_1) + p(G_2) \\
&\leq f(n_1) + f(n_2) \\
&= \lfloor \tfrac{n_1^2}{4} \rfloor + \lfloor \tfrac{n_2^2}{4} \rfloor - (n_1 + n_2) + 2 \\
&< \lfloor \tfrac{(n_1+n_2)^2}{4} \rfloor - (n_1 + n_2) + 1 \\
&= f(n)
\end{aligned}$$

Suppose that $m \geq 3$. Then since $n_1 \geq 2$,

$$\begin{aligned}
&\lfloor \tfrac{(n_1+n_2+\cdots+n_m)^2}{4} \rfloor - (m-1) \\
&= \lfloor \tfrac{n_1^2}{4} + \tfrac{n_2^2}{4} + \cdots + \tfrac{n_m^2}{4} + \tfrac{n_1(n_2+n_3+\cdots+n_m)+n_2n_3+\cdots+n_{m-1}n_m}{2} - (m-1) \rfloor \\
&> \lfloor \tfrac{n_1^2}{4} + \tfrac{n_2^2}{4} + \cdots + \tfrac{n_m^2}{4} \rfloor \\
&\geq \lfloor \tfrac{n_1^2}{4} \rfloor + \lfloor \tfrac{n_2^2}{4} \rfloor + \cdots + \lfloor \tfrac{n_m^2}{4} \rfloor
\end{aligned}$$

So again by Theorem 5 and part (1) of this Lemma, we have that

$$\begin{aligned}
&p(G) = p(G_1) + p(G_2) + \cdots + p(G_m) \\
&\leq f(n_1) + f(n_2) + \cdots + f(n_m) \\
&= \lfloor \tfrac{n_1^2}{4} \rfloor + \lfloor \tfrac{n_2^2}{4} \rfloor + \cdots + \lfloor \tfrac{n_m^2}{4} \rfloor - (n_1 + n_2 + \cdots + n_m) + m \\
&< \lfloor \tfrac{(n_1+n_2+\cdots+n_m)^2}{4} \rfloor - (n_1 + n_2 + \cdots + n_m) + 1 \\
&= f(n)
\end{aligned}$$

□

Lemma 7 *Let $G = (V, E)$ be a graph of n vertices. If G has a triangle and $n \geq 4$, then $p(G) < f(n)$.*

Proof. We will proceed by induction. Notice that $f(n) \geq 1$ since $n \geq 4$. First assume that $n \leq 5$. If G is triangulated, then by Theorem 3, $p(G) = 0 < f(n)$. If G is not triangulated and G has a triangle, then G must be one of three possible graphs of five vertices, and it is easy to show that $p(G) \leq 1 < f(5)$ for each of these graphs. So, again $p(G) < f(n)$.

Suppose that the lemma is true for any graph with a triangle and with less than n vertices and let G be a graph with a triangle and with $n \geq 6$ vertices. If G is a complete graph, then $p(G) = 0 < f(n)$. If G is not a complete graph, then it is easy to see that there is a triangle v_1, v_2, v_3 such that either there is a vertex u with u not not adjacent to any vertex in $\{v_1, v_2, v_3\}$, or there is a vertex $v \neq v_i, i = 1, 2, 3$, so that v is adjacent to at least one and at most two vertices in $\{v_1, v_2, v_3\}$. If the former case occurs, then let $v_4 = u$. Otherwise, we may assume without loss of generality that $vv_1 \in E, vv_2 \notin E$, and let $v_4 = v$. Let the rest of the vertices be labeled by $v_5, v_6, \cdots, v_n$. Let $G' = G - \{v_1\} - \{v_2\}$. Then by the induction hypothesis,

$$p(G') \leq f(n-2) = \lfloor \frac{(n-2)^2}{4} \rfloor - (n-2) + 1 = \lfloor \frac{n^2}{4} \rfloor - 2n + 4$$

Let $D' = (V', A')$ be an optimal phylogeny digraph for G'. We construct a phylogeny digraph D from D' as follows.

Add arcs $(v_1, v_2), (v_3, v_2)$ to A', and add arc (v_4, v_1) to A' if $v_4 = v$. Now, consider $i = 5, 6, \cdots, n$. If $v_i v_1 \notin E$ and $v_i v_2 \notin E$, the construction is complete. Otherwise, add a vertex w_i to V'. If v_i is adjacent to both v_1 and v_2, then add arcs $(v_1, w_i), (v_2, w_i), (v_i, w_i)$ to A'; if v_i is adjacent to one of v_1, v_2 but not both, say $v_i v_1 \in E$, then add arcs $(v_1, w_i), (v_i, w_i)$ to A'.

It is easy to see that the digraph obtained above is acylic, and it is a phylogeny digraph for G. Moreover, it added at most $n - 4$ vertices to D'. So,

$$p(G) \leq p(G') + (n-4) \leq \lfloor \frac{n^2}{4} \rfloor - 2n + 4 + (n-4) = \lfloor \frac{n^2}{4} \rfloor - n < f(n)$$

□

Theorem 8 *For all graphs $G = (V, E)$ with n vertices, $p(G) \leq \lfloor \frac{n^2}{4} \rfloor - n + 1$, and this bound is achieved if and only if $n \leq 3$ or $G = K(\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil)$.*

Proof. Let $f(n) = \lfloor \frac{n^2}{4} \rfloor - n + 1$. By Lemmas 6, 7, it is easy to see that if $n \geq 4$, $p(G) \leq f(n)$. Futher, if $n \leq 3$, $p(G) = f(n) = 0$. If $G = K(\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil)$, then $e = \lfloor \frac{n^2}{4} \rfloor$ by Turán's Theorem. So by Theorem 4, $p(G) = e - n + 1 = f(n)$.

Now, assume $p(G) = f(n)$ and $n \geq 4$. Then, by Lemma 7, G must have no triangles. By part (2) of Lemma 6, G must be connected. So, by Theorem 4, $p(G) = e - n + 1$. This implies that $e = \lfloor \frac{n^2}{4} \rfloor$, and so $G = K(\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil)$ by Turán's Theorem. □

References

[1] J. E. Cohen. Interval graphs and food webs: A finding and a problem. Technical Report 17696-PR, Rand Corporation Document, Monica, CA, 1968.

[2] F. Harary, S.-R. Kim, and F. S. Roberts. Extremal competition numbers as a generalization of Turán's theorem. *J. Ramanujan Math. Soc.*, **5**:33–43, 1990.

[3] S.-R. Kim. The competition number and its variants. In J. Gimbel, J.W. Kennedy, and L.V.Quintas, editors, *Quo Vadis Graph Theory?*, volume **55** of *Annals of Discrete Mathematics*, pages 313–325. 1993.

[4] J. R. Lundgren. Food webs, competition graphs, competition-common enemy graphs, and niche graphs. In F. S. Roberts, editor, *Applications of Combinatorics and Graph Theory in the Biological and Social Sciences*, volume **17** of *IMA Volumes in Mathematics and its Applications*, pages 221–243. Springer-Verlag, New York, 1989.

[5] R. J. Opsut. On the computation of the competition number of a graph. *SIAM J. Algebraic Discrete Methods*, **3**:420–428, 1982.

[6] F. S. Roberts. Food webs, competition graphs, and the boxicity of ecological phase space. In Y. Alavi and D. Lick, editors, *Theory and Applications of Graphs*, pages 477–490. Springer-Verlag, New York, 1978.

[7] F. S. Roberts. *Applied Combinatorics*. Prentice-Hall, Englewood Cliffs, N.J., 1984.

[8] F. S. Roberts. Competition graphs and phylogeny graphs. In L. Lovasz, editor, *Graph Theory and Combinatorial Biology*, Bolyai Studies. J. Bolyai Mathematical Society, Budapest, Hungary, To appear.

[9] F. S. Roberts and L. Sheng. Phylogeny graphs of arbitrary digraphs. In B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, editors, *Mathematical Hierarchies in Biology*, pages 233–238. DIMACS Series, Vol. 37, American Mathematical Society, Providence, RI, 1997.

[10] F. S. Roberts and L. Sheng. Phylogeny numbers. *Disc. App. Math.*, **87**:213–228, 1998.

[11] F. S. Roberts and L. Sheng. Phylogeny numbers for graphs with two triangles. DIMACS Tech Report, Oct. 1998, DIMACS Center, Rutgers University, submitted for publication.

[12] P. Turán. Eine extremalaufgabe Aus der Graphentheorie. *Mat. Flz. Lapok*, **48**:436–452, 1941.