J. Univ. Kuwait (Sci.) **12,** 1985

A path entropy function for rooted acyclic digraphs

CHRISTOPHER D. GREEN* AND FRANCIS SURAWEERA†

Department of Mathematics, University of Dundee, Dundee, Scotland

ABSTRACT

The concept of a rooted acyclic digraph is formalized and an entropy function measuring the complexity of the paths from the root to the terminal nodes is defined. Many results obtained by C. D. Green for the class of rooted trees are extended to encompass the class of rooted acyclic digraphs.

1. INTRODUCTION

In a previous paper Green (1973) developed a path entropy function for rooted, directed trees. This function was intended to provide a measure of the average information cost, in the sense of Watanabe (1969), when traversing such a tree from root to terminal vertex. The primary application was to hierarchical classification and indexing schemes, although clearly the measure could also be used for other systems invovling sequences of decisions. The path entropy can also be regarded as a measure of the complexity of the tree, although it should be noted that the definition is based upon the paths leading to the terminal vertices, and the other paths (and vertices) have only a secondary importance. It follows that our path entropy differs radically from many other measures of graph complexity, such as are described by Marshall (1971).

In the present paper we have extended the definition of path entropy to include the class of rooted acyclic digraphs as a natural generalisation of rooted directed trees. Such digraphs may be formed from a tree by the insertion of additional arcs directed away from the root, so that a path or decision sequence leading to a terminal vertex need no longer be unique. An obvious example is that of a classification scheme with 'see-also' type cross-references, and indeed the present study was motivated by this application. In what follows we develop a definition of path entropy for this situation, such that the new definition reduces to the previous function in the case of tree structures. It is shown also that our results extend the previous work of Green (1973).

The path entropy H, so defined has the following interpretation; it involves the average of the maximum entropies of the paths to each terminal vertex. Consequently, H may be regarded as a measure of the 'worst case' situation. Based on our

* Present address: Department of Accountancy and Business Finance, University of Dundee, Dundee, Scotland, UK.

† Present address: Department of Mathematics, University of Kuwait, P.O. Box 5969, Kuwait.

experience, we feel that the H function is the most logical one to represent the effect of insertion or deletion of 'see-also' type cross-reference arcs.

The rest of the paper is organized as follows. In Section 2 we develop some definitions and notations which will be used later on. In Section 3 we characterize the path entropy for rooted acyclic digraphs (RAD's) and also give a non-linear lower bound for the path entropy of a RAD with n terminal vertices. In Section 4 we formalise the product of two RAD's. The key idea presented in this section is the strict additivity of the path entropy function with respect to the graph product that preserves the path structure. In Section 5 we extend the results to terminal weighted RAD's. Finally, Section 6 contains the summary and conclusions.

2. DEFINITIONS AND NOTATIONS

The central notion is that of a rooted acyclic digraph or RAD, denoted by D_n, with unique root u_0 and terminal vertices $v_1, v_2, \ldots, v_n$. The arcs are directed away from the root so that u_0 is the unique vertex with indegree zero. The notation generally follows that of Green (1973) for directed trees, but in a RAD there may of course be more than one directed path from u_0 to a general vertex u. The length of such a path defines a general *rank* of u, denoted by $G(u)$. The supremum of $G(u)$, noted by $r(u)$, is called the (strict) rank of u. The vertices of a RAD may be assigned to levels according to the ranks, and we denote by $L(k)$ the subset of vertices u such that $r(u) = k$. The height h of D_n is the maximum such value of k. Unless otherwise stated we shall assume that D_n contains no symmetric vertices such that $od(u) = id(u) = 1$, where *od* and *id* denote out- and indegrees respectively.

In level $L(k)$ there are α_k terminal vertices and β_k interior vertices, where

$$\alpha_1 + \alpha_2 + \ldots + \alpha_h = n \qquad \text{and} \qquad \beta_0 = 1, \quad \beta_h = 0 \ .$$

A RAD is called *complete* if every vertex is adjacent to all vertices of higher rank. D_n is called (transitively) closed if D_n is isomorphic to its transitive closure in the usual sense, so that a RAD may be closed but not complete. A general RAD may be partially completed by the addition of arcs linking vertices on different levels, and if such an arc is the transitive closure of two existing arcs we call it a closure arc. Otherwise such an added arc is termed a completion arc.

3. PATH ENTROPY FOR ROOTED ACYCLIC DIGRAPHS

Let $P_i(j) = P_i[u_0, v_i, j] = (u_0, u_1, \ldots, u_k, v_i)$ denote the jth path from the root to a terminal vertex v_i, the collection of such paths being indexed in an arbitrary way. We define the entropy of this path to be the sum of the logarithms (to base 2) of the outdegrees of the vertices composing the path, thus

$$\eta[P_i(j)] = \sum_{r=0}^{k} \log od\,(u_r) \ .$$

The path entropy of the terminal vertex v_i is then defined to be the maximum of the entropies of the paths leading from the root to v_i, and the path entropy $\eta(D_n)$ of the RAD to be the sum of the entropies of the n terminal vertices, so that

$$\eta(D_n) = \sum_{i=1}^{n} \max_j \eta[P_i(j)] \quad .$$

The normalised path entropy is then

$$H(D_n) = (1/n)\,\eta(D_n) \quad .$$

It is clear that this definition of path entropy reduces to that introduced in Green (1973) in the case that D_n is a tree. $H(D_n)$ may be regarded as a measure of the average 'decision cost' in the worst case, i.e. when the most complex path from root to terminal vertex is used in every case. We remark here that the presence of symmetric vertices does not substantially affect the value of η, since such vertices do not contribute to the relevant summations.

Theorem 1 of Green (1973), which gives a lower bound to $\eta(T_n)$ in the tree case can be extended to cover the new definition.

Theorem 1. Let D_n be a RAD with n (> 1) terminal vertices. Then, $\eta(D_n) \geqslant n \log n$ for all $n > 1$.

Proof. Suppose that we take one of the terminal vertices v_i of D_n at the highest level, and select a maximal entropy path, say $P_{\max}[u_0, v_i]$. Starting at the top we move down towards the root, at each level deleting any in-arcs that do not belong to $P_{\max}[u_0, v_i]$. Repeating this process for each terminal vertex in turn, level by level, we ultimately obtain a tree $T_{n'}$, where $n' \geqslant n$, since it is possible for some interior vertices of D_n to appear ultimately as terminal vertices of $T_{n'}$. Note that this tree may contain symmetric vertices. It is clear that none of the arcs which have been deleted can be a disconnecting arc, and that

$$\eta(D_n) \geqslant \eta(T_n) \geqslant n \log n,$$

where T_n is the tree obtained by successively deleting any 'spurious' terminal vertices from $T_{n'}$. The last inequality follows from Theorem 1 of Green (1973).

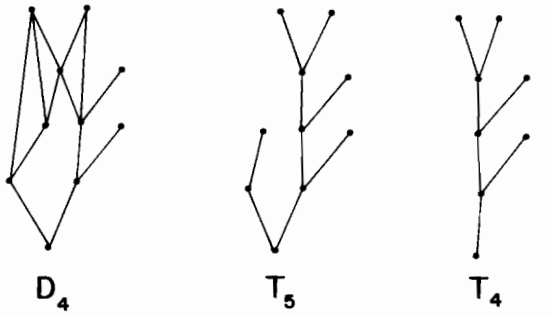
Fig. 1 shows the sequence of graphs obtained from D_n by the construction used in the proof of Theorem 1.

4. PRODUCTS OF ROOTED ACYCLIC DIGRAPHS

The tree product introduced in Green (1973) can be generalised without difficulty to the present case, although it should be remarked that now the assumption concerning the absence of symmetric vertices becomes vital since the ranking of the vertices is an essential feature of the construction.

Definition Let D_m, D'_n be rooted acyclic digraphs, with vertex sets $V(D_m)$, $V(D'_n)$ respectively. Then the product $D_m \cdot D'_n$ (having mn terminal vertices) is the RAD with vertex set contained in $V(D_m) \times V(D'_n)$ defined inductively by

(1) (u_0, u'_0) is the root of $D_m \cdot D'_n$ where u_0, u'_0 are the roots of D_m, D'_n respectively;
(2) $(b, b') \in V(D_m \cdot D'_n)$ and is an immediate successor of (a, a') if
   (i) both a and a' have the same general rank in D_m and D'_n respectively and a immediately precedes b (in D_m) and a' immediately precedes b' (in D'_n), or



$\eta(D_4)g + 3\log 3 = 14.755$ $\eta(T_5) = 14$ $\eta(T_4) = 9$

Fig. 1.

(ii) $a = b$ is a terminal vertex of D_m and a' immediately precedes b' in D'_n, or
(iii) a immediately precedes b in D_m and $a' = b'$ is a terminal vertex of D'_n.

The resulting RAD has height equal to the maximum of the heights of D_m and D'_n. It is easily shown that the graph product '.' is closed with respect to the class of RAD's, is commutative and is associative.

Corresponding to Theorem 3 of Green (1973) we can prove that the normalised entropy H is a strictly additive function with respect to this product.

Theorem 2. $H(D_m \cdot D'_n) = H(D_m) + H(D'_n)$.

Proof. Consider an arbitrary terminal vertex v_i of D_m and similarly v'_j of D'_n. Select in each case a path of maximal entropy from the respective root, and denote these by $P_{\max}[u_0, v_i]$ and $P_{\max}[u'_0, v'_j]$ respectively. It follows from the definition of the product that every path $P[(u_0, u'_0), (v_i, v'_j)]$ in $D_m \cdot D'_n$ is the product of two paths of the form of $P[u_0, v_i]$ and $P[u'_0, v'_j]$. Further, from Theorem 3 of Green (1973) it is true that the entropy of the product of two such paths (regarded as sub-trees) is the sum of the entropies of the paths. It follows that $P_{\max}[u_0, v_i]$. $P_{\max}[u'_0, v'_j]$ will be a maximal path from (u_0, u'_0) to (v_i, v'_j). Hence, summing over all the terminal vertices of the product graph we get

$$\eta(D_m \cdot D'_n) = n \sum_{i=1}^{m} \eta(P_{\max}[u_0 v_i]) + m \sum_{j=1}^{n} \eta(P_{\max}[u'_0, v'_j])$$

so that, after division by mn

$$H(D_m \cdot D'_n) = H(D_m) + H(D'_n) \quad .$$

The results obtained by Green (1973) concerning terminal weighted trees can easily be extended. In the next section, we give the definition of a terminal weighted rooted acyclic digraph and the corresponding results.

5. TERMINAL WEIGHTED RAD's

A digraph D_n is said to be terminal-weighted with W if the set $W = \{w_1, w_2, \ldots, w_n\}$ of non-negative numbers is associated with the terminal nodes $v_1, v_2, \ldots, v_n$ in the obvious way. We can assume without loss of generality that the weights are normal-

ised so that $\sum_{i=1}^{n} w_i = 1$. We denote such a weighted digraph by $(D_n;W)$. The path entropy of a weighted digraph is defined as follows:

$$H(D_n;W) = \sum_{i=1}^{n} w_i \max_j \eta(P[u_0, v_i; j]).$$

If the uniform weighting for which each w_i has the value $1/n$ is denoted by I, then the following theorem is immediate.

Theorem 3. For a uniform weighting I of a terminal weighted RAD, $H(D_n;I) = H(D_n)$.

The proof of Theorem 3 is elementary.

Our next result gives a lower bound for the terminal weighted RAD's. We shall state this as a theorem.

Theorem 4. For fixed n and $W = \{w_1, w_2, \ldots, w_n\}$, the path entropy of a weighted RAD $(D_n;W)$ satisfies the inequality $H(D_n;W) \geqslant n \log n \cdot \min_i w_i$.

Proof. The proof of the above theorem follows from our Theorem 1 and the Lemma in Green (1973).

Finally we extend Theorem 3 of Green (1973) to cover the terminal weighted RAD's.

Theorem 5. Given the RAD's D_m and F_n and weight sets $Y = (y_1, y_2, \ldots, y_m)$ and $Z = (z_i, z_2, \ldots, z_n)$ we construct $D_m \cdot F_n$ as before and assign weights to the mn terminal vertices from the elements of the normalised product set $Y \otimes Z$; i.e. $w_{ij} = y_i z_j$. Calling this product set W, we have

$$H(D_m \cdot F_n;W) = H(D_m;\ Y) + H(F_n;\ Z) \quad .$$

Proof. Immediate from Theorem 2.

6. SUMMARY AND CONCLUSIONS

In this paper we have formalised the concept of rooted acyclic diagraphs (RAD's). The main characteristics of RAD's are (i) RAD's have a finite number of terminal nodes (ii) all nodes are reachable from some specified node designated as the root (iii) all nodes are ranked, and (iv) RAD's contain no cycles. A path entropy function is defined on a RAD which may be considered as a measure of the choices to be made when traversing the path from the root to a given vertex. A lower bound for the path entropy of a RAD is given. In addition, a graph product is defined for RAD's which is a generalisation of Green's graph product for trees. We have shown that the path entropy function is additive with respect to this graph product. Many of the results obtained were generalised to the case when the terminal nodes were assigned weights. One possible application of these results is that of obtaining bounds for the H value of a particular classification scheme. Furthermore, RAD models appear quite naturally in connection with communication and decision networks and in these cases also such values can be of importance.

ACKNOWLEDGEMENT

We would like to thank the referees for their exceptionally thorough job in catching and correcting a number of minor errors in the original manuscript.

REFERENCES

Green, C. D. 1973. A path entropy function for rooted trees. J. Assoc. Comp. Machinery **20:** 378–84.
Marshall, C. W. 1971. Applied graph theory. Wiley-Interscience, New York.
Watanabe, S. 1969. Knowing and guessing. Wiley, New York.

(Received 12 December 1983, revised 28 October 1984)

مسار دالة الانتروبي للبيانات المتجهة الجذرية غير الدائرية

كريستوفر د . جرين وفرانسيس سوراويرا*
قسم الرياضيات بجامعة دندي ، دندي ، سكوتلندا ، المملكة المتحدة

خلاصة

لقد أمكن تشكيل فكرة البيانية الجذرية غير الدائرية كما أمكن تعريف قياس تعقيد المسارات ، من الجذور الى العقد النهائية ، لدالة الانتروبي . وتم تعميم نتائج عديدة ، سبق أن حصل عليها ك . د . جرين ، لصنف أشجار جذرية لتشمل صنف البيانات المتجهة غير الدائرية .

*العنوان الحالي : قسم الرياضيات بجامعة الكويت