THE MINIMUM SPANNING TREE PROBLEM IN ARCHAEOLOGY

Per Hage, Frank Harary, and Brent James

The minimum spanning tree problem is a well-known problem of combinatorial optimization. It was independently discovered in archaeology by Renfrew and Sterud in their method of close proximity analysis. Unlike traditional methods of seriation, this method permits branching structures that reveal clustering in archaeological data. Identifying close proximity analysis as the minimum spanning tree problem permits a more efficient means of computation, an explicit rule of clustering, and a recognition of problems of indeterminacy in the analysis of network data. These points are illustrated with reference to Irwin's recent study of voyaging and cultural similarity in Polynesia.

El problema del árbol de expansión mínima es bien conocido en optimización combinatorial. Este problema fue descubierto independientemente en la arqueología cuando Renfrew y Sterud aplicaron su análisis de "close proximity." Este método es diferente de métodos tradicionales de seriación porque trabaja con estructuras dendríticas que muestran racemización en datos arqueológicos. El identificar el análisis de "close proximity" como el problema del árbol de expansión mínima permite (1) la aplicación de técnicas de computación más eficientes, (2) el uso de una regla explícita de racemización, y (3) un reconocimento de los problemas indeterminados en el análisis de datos de red. Comparamos este método con la obra reciente de Irwin sobre la navegación en barco y la similitud de culturas en Polinesia.

The minimum spanning tree problem (MSTP) is a well-known problem of combinatorial optimization (Graham and Hell 1985). We illustrate the MSTP with an instance of how to determine the monthly telephone charge to a generic large corporation G with offices in many cities, $v_1, \ldots, v_n$. All the distances $d(v_i, v_j)$ are known and are distinct. The corporation G does not wish to pay the phone company an amount proportional to the sum $\Sigma d(v_i, v_j)$ of all the distances since not all pairwise connections are needed for each office to be able to communicate with every other office. What is needed is a tree on these n nodes having minimum total distance. The problem has applications to the design of all kinds of communication networks and to numerous other problems including classification and clustering. The MSTP was independently discovered by Renfrew and Sterud (1969) who developed a seriation technique called the "double link method of close proximity analysis." They were not aware of the spanning tree feature implicit in their method. Our purpose is to make this feature explicit so that it can be utilized in close proximity analysis. Close proximity analysis is described as a "graphical method" for ordering archaeological assemblages or types on the basis of their similarity. Unlike traditional methods of seriation (Brainerd 1951; Robinson 1951), which "compress archaeological data into a single linear series," close proximity analysis permits branching structures which reveal clustering. For example it would take account of spatial as well as temporal variation in pottery design. The practical advantages of the method are its speed and its ability to handle large data structures without the use of a computer.

There are three reasons for identifying close proximity analysis as an MSTP. The first is to provide a more efficient means of computation. The algorithm of Renfrew and Sterud works but it is unnecessarily complicated and unwieldy: it creates cycles that must subsequently be deleted and requires a special provision for connecting com-

Per Hage ■ Department of Anthropology, University of Utah, Salt Lake City, UT 84112
Frank Harary ■ Department of Computer Science, New Mexico State University, Las Cruces, NM 88003-0001
Brent James ■ Department of Biostatistics, Harvard School of Public Health, Cambridge, MA 02138
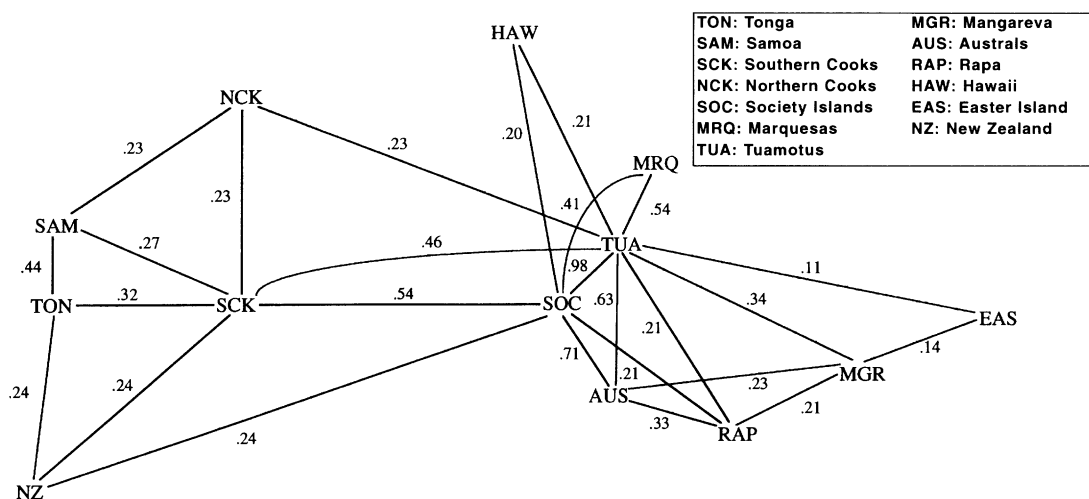
149

Figure 1. Irwin's (1992) close proximity analysis of the mutual accessibility network of Polynesian islands.

ponents of a network, which would otherwise remain separate. The second reason is to define an explicit rule of clustering in an MST. The third reason is to note a problem of indeterminacy which may arise when the values of the edges are not distinct. We illustrate these points with a recent application of close proximity analysis to Oceanic archaeology.

In a challenging and innovative study Irwin (1992) proposes that the prehistoric exploration and colonization of the Pacific Islands was rapid, deliberate, and systematic, based on continually improving navigational techniques and an expanding body of geographical knowledge. Drawing on computer simulations and practical sailing experience, Irwin argues that early voyagers followed a conservative strategy in order to ensure a safe return to their point of departure in the event that they did not find land. They sailed first into the wind, returning with the wind at their back, then, with accumulating geographical knowledge, across the wind, and finally, riskiest of all, downwind. In general, the archaeological evidence supports the hypothesis that islands to which it was easiest to return were settled first. Irwin also hypothesizes that island communities did not necessarily become isolated after settlement but remained in communication and, depending on their degree of mutual accessibility, continued to influence each other. Accessibility

would account, in part, for cultural, linguistic, and biological similarities between islands or island groups. Mutual accessibility is defined as a product of closeness and angle of target size between island pairs. From a matrix of interisland accessibility, analogous to the Renfrew-Sterud "similarity coefficient matrix," Irwin generates a close proximity graph (network) like the one in Figure 1. The higher the values of the edges (lines) the greater the accessibility.[1]

In support of the hypothesis that accessibility is predictive of patterns of cultural similarity, Irwin notes parallels between his close proximity network and the subgroupings in Polynesia identified by Burrows (1938). Unfortunately, the network in Figure 1 contains many superfluous edges and hence cycles which tend to obscure such parallels.[2] We can clarify and simplify matters of computation, presentation, and analysis by using an MST algorithm. Some definitions will be helpful.

As defined in Buckley and Harary (1990) and Hage and Harary (1991) a graph G consists of a finite nonempty set $V = V(G)$ of *nodes* together with a set $E = E(G)$ of *edges*[3] joining certain pairs of distinct nodes of G. A *path* in G is an alternating sequence $v_0, e_1, v_1, e_2, v_2, \ldots , v_{n-1}, v_n$ of distinct nodes and edges. A *cycle* is obtained from a path when the initial and terminal nodes v_0 and v_n are joined by an edge. Graph G is connected if every pair of nodes are joined by a path.
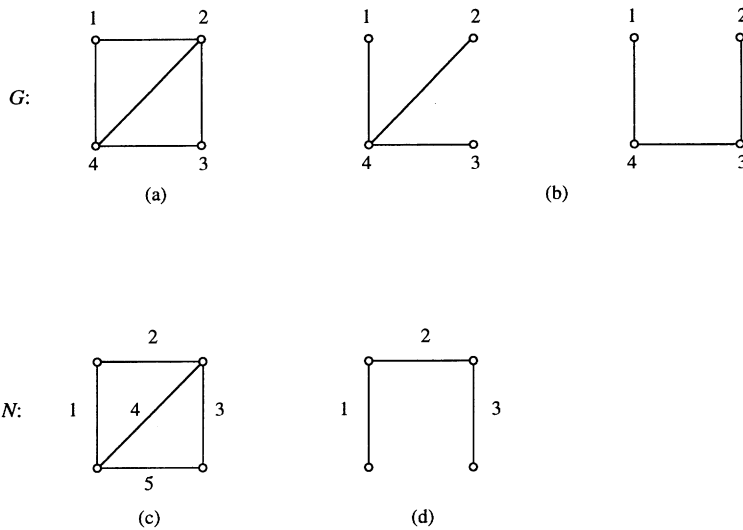
Figure 2. Illustrations of (a) a graph *G*, (b) two of the spanning trees of *G*, (c) a network *N*, and (d) the minimum spanning tree of *N*.

Figure 2a shows a connected graph with four nodes and five edges. It contains three cycles: (1,2,4,1), (2,3,4,2), and (1,2,3,4,1).

A *tree* is a connected graph with no cycles (acyclic). A *spanning tree* of a connected graph *G* is a subtree of *G* which contains all the nodes of *G*. Figure 2b shows two trees. They are two of the eight spanning trees of *G*.

A *network N* (or weighted graph) is a graph with a positive value $f(e)$ assigned to each edge *e*. A *minimum spanning tree* (MST) of *N* is a spanning tree with minimum value sum. Figure 2c illustrates a network *N* and Figure 2d its MST.

Table 1 shows the mutual accessibility of 13 islands (or island groups) in Polynesia. Based on Irwin's Table 22, it conforms with the standard procedure for constructing an MST—the lower the value assigned to an edge, the greater the accessibility.

Algorithm 1–Kruskal's Algorithm for Constructing an MST

Given: An undirected network *N* with distinct positive integer values (equivalently positive real numbers) $f(e)$ on each edge *e* of *N*.

Wanted: A spanning tree *T* of *N* with minimum edge-value sum. We will specify such a tree by building its edge set E_T.

Step 1. Label the edges of *N* by $e_1, e_2, \ldots, e_q$ such that whenever $i < j$, we have $f(e_i) < f(e_j)$.

Call this sequence of edges σ.

Step 2. Place e_1 in E_T.

Step 3. On arriving in σ at a generic edge e_i, place e_i in E_T if and only if e_i together with the edges of *N* already in E_T do not contain a cycle. If e_i is not placed in E_T, go to Step 4.

Step 4. If $| E_T | = p - 1$, stop. Otherwise repeat Step 3.

Theorem 1. Given a connected network *N* in which all the edge-values are distinct, Kruskal's algorithm will terminate with *N*'s unique minimum spanning tree *T*.

From the upper (or lower) half of Table 1 we list the edges of *N* in order of their value (the sequence σ). We begin Kruskal's algorithm by adding the first edge SOC-TUA (.02) to E_T followed by the second edge SOC-AUS (.29). We cannot add the third edge TUA-AUS (.37) to E_T as it would create a cycle, so we proceed to the next edge SCK-SOC (.46). Continuing in this way we add a last edge EAS-MGR (.86) which gives $p - 1$ edges in all. An MST of mutual accessibility of Polynesian islands is shown in Figure 3.

The MST of Figure 3 is identical to a close proximity graph but Kruskal's algorithm is much simpler than the Renfrew-Sterud procedure. We just keep adding edges of least value as long as they do not form a cycle until we have an MST. Two comments can be made on Figure 3.

1. Clustering. By definition an MST connects

Table 1. A Minimum Accessibility Matrix for Polynesia.

	TON	SAM	SCK	NCK	SOC	MRQ	TUA	MGR	AUS	RAP	HAW	EAS	NZ
Tonga (TON)		.56	.68	.82	.78	.82	.76	.91	.84	.90	.89	.97	.77
Samoa (SAM)	.56		.73	.77	.83	.85	.83	.90	.81	.89	.81	.95	.76
S. Cook (SCK)	.68	.73		.77	.46	.70	.54	.86	.62	.82	.85	.92	.76
N. Cook (NCK)	.82	.77	.77		.77	.85	.79	.89	.82	.86	.89	.95	.90
Society (SOC)	.78	.83	.46	.77		.59	.02	.78	.29	.79	.80	.91	.80
Marquesas (MRQ)	.82	.85	.70	.85	.59		.46	.80	.66	.85	.87	.90	.88
Tuamotu (TUA)	.76	.83	.54	.79	.02	.46		.66	.37	.79	.79	.89	.76
Mangareva (MGR)	.91	.90	.86	.89	.78	.80	.66		.77	.79	.92	.86	.92
Australs (AUS)	.84	.81	.62	.82	.29	.66	.37	.77		.67	.81	.94	.79
Rapa (RAP)	.90	.89	.82	.86	.79	.85	.79	.79	.67		.95	.90	.91
Hawaii (HAW)	.89	.81	.85	.89	.80	.87	.79	.92	.81	.95		.97	.91
Easter (EAS)	.97	.95	.92	.95	.91	.90	.89	.86	.94	.90	.97		1.00
New Zealand (NZ)	.77	.76	.76	.90	.80	.88	.76	.92	.79	.91	.91	1.00	

all the nodes of a network. It therefore does not define clusters, but it does contain information useful for clustering. Thus an edge with large value relative to other edges incident with a node can be "cut" (deleted) to form two clusters. An endnode is incident with a single edge which can always be cut. With this procedure in mind let us examine the relation between interisland accessibility as calculated by Irwin and areal differentiation as delineated by Burrows.

On the basis of a distributional analysis of cultural traits and complexes including artifacts (tools, canoe types, bark cloth, etc.), aspects of social organization (chiefly languages, and kinship terms and practices), and religious ideas (origin myths, concepts of the afterworld, and lunar names), Burrows identified four subgroupings in Polynesia. Restricting Burrows's set of islands to the subset in Irwin's model the subgroupings are:

Western Polynesia (Samoa, Tonga)

Central Polynesia (Society Islands, Tuamotus, Southern Cooks, Australs, Rapa, Hawaii)

Marginal Polynesia (New Zealand, Easter, Marquesas, Mangareva)

Intermediate Polynesia (Northern Cooks)

The basic contrast according to Burrows is between Western Polynesia, which centers on Tonga and Samoa, and Central Polynesia, which has as its "nucleus" the Society Islands. Archaeologists regard this dichotomy as "fundamental" (Bellwood 1987).
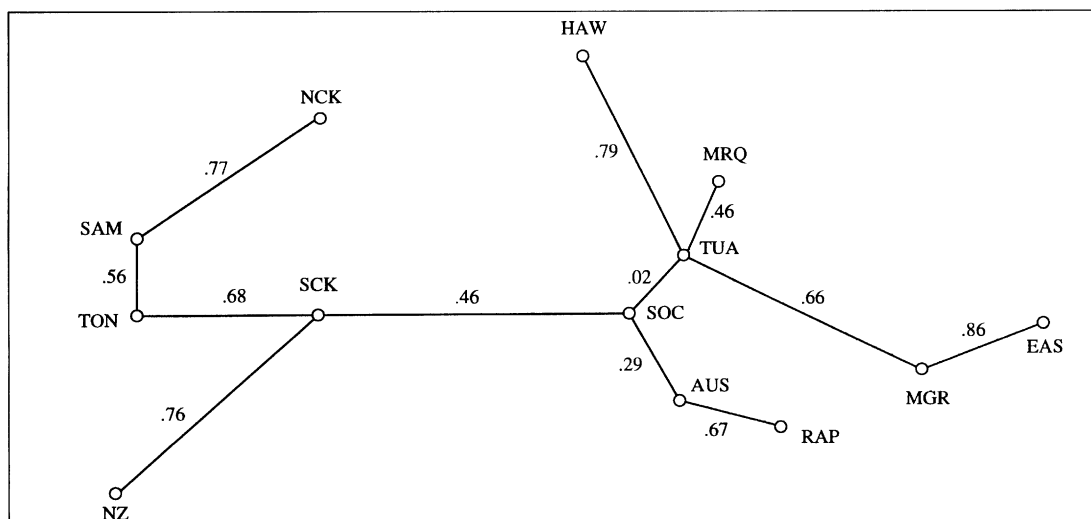
Can these subgroupings be identified as clus-

Figure 3. An MST of the mutual accessibility network of Polynesian islands based on data in Irwin (1992).

ters in the MST in Figure 3? Referring to the acronyms in Figure 1, if we cut the Northern Cook–Samoa (NCK-SAM) edge we have a cluster (NCK) that corresponds to Burrows's Intermediate Polynesia, a subgrouping not considered by Irwin.

Next, if we cut the NZ-SCK, EAS-MGR, and MRQ-TUA edges we have three of the four endnodes (NZ, EAS, MRQ) belonging to Marginal Polynesia. The fourth island, MGR, cannot be assigned to this cluster because it cannot be separated from TUA whose largest edge joins it to HAW. This is a minor problem. A major problem concerns the position of HAW. While Irwin places Hawaii in Marginal Polynesia, Burrows assigns it to Central Polynesia. Burrows is emphatic on this point: "Hawaii . . . is so close culturally to Central Polynesia that . . . it is grouped with this center rather than with the other geographically marginal regions" (1938:91–92).[4] If cultural similarity is related to accessibility, then Hawaii must have been much more accessible with other islands in Central Polynesia. In fact, Lewis (1972), as Irwin notes, does not consider a Tahiti (Society Islands) to Hawaii voyage navigationally difficult. Greater mutual accessibility of Hawaii and the Society Islands (either directly or via the Tuamotus) would also be consistent with Irwin's suggestion that social stratification in Eastern Polynesia might be explained in terms of Renfrew and Cherry's (1986) Peer Polity Interaction model. Basically, this model accounts for the development of social stratification, not in terms of local adaptation or outside forces but in terms of elite interaction in a network of societies. Irwin proposes this model as a modification of the current view of Polynesian prehistory, which holds, on the basis of linguistic reconstruction, that Proto-Polynesian society was stratified (Kirch and Green 1987). There is, however, no clear artifactual or architectural evidence in the archaeological record for early distinctions of rank. As applied to Eastern Polynesia the Renfrew-Cherry model would account for "broadly parallel changes in social complexity, and to take a material example, the striking elaboration of associated religious structures especially in Hawaii and the Society Islands" (Irwin 1992:202). In this light Hawaiian traditions concerning "Kahiki"

would be interpreted not as evidence of secondary settlement from Tahiti but as a record of more recent contacts.[5]

Finally, if we cut the TON-SCK edge we have two clusters (TON, SAM) and (SOC, TUA, SCK, AUS, RAP, HAW, MGR), which, with the exception of MGR, correspond to Burrows's Western and Central Polynesia subgroupings.

2. Indeterminacy. Theorem 1 guarantees that an MST of a network N will be unique only if the edge values of N are distinct. If they are not distinct, there may be different MSTs and hence different clusterings. Renfrew and Sterud recommended that the entries in the similarity coefficient matrix be different, but this is not always possible. Their Cycladic Cemetery matrix, for example, has many identical entries resulting in 288 different close proximity structures (MSTs), as shown by an exhaustive computer-assisted verification. The MST in Figure 3 is one of nine different MSTs generated by the matrix in Table 1. In one MST, NCK can be joined, as an endnode, to SAM in Western Polynesia, and in other MSTs, to SOC or SCK in Central Polynesia. This reflects its Intermediate position in Burrows's classification. NZ, however, presents a problem. If we consider the fundamental contrast between Western and Central Polynesia, NZ can be joined to either subgrouping (cluster) depending on which MST we use. As can be seen from Table 1, NZ can be joined to TUA in the Central group or to SAM or SCK in the Western group. The latter joining would contradict Burrows's conclusion that the geographically marginal islands, including NZ, are "all more closely related to central than to western Polynesia which strengthens the showing of two radiating centers of influence" (1938:91). It would also be inconsistent with the archaeological view that the origins of the New Zealand Maori are to be found in the islands of Eastern Polynesia—the Southern Cooks, Society or Austral Islands (Bellwood 1987). And it would contradict the linguistic evidence which assigns Maori to the Proto-Central-Eastern group of Polynesian languages (Pawley and Green 1975). We conclude from an explicit MST analysis of interisland accessibility that the Hawaiian and New Zealand cases will require some modification of Irwin's model.

Kruskal's algorithm works best for sparse networks or small networks such as this one (Figure 3) or Renfrew and Sterud's Aurignacian example. For larger, denser networks, such as that of the Cycladic Cemetery with 21 nodes and 210 edges, the most efficient algorithm is that of Prim (1957). It has the advantage of a simple easily programmed matrix method (Wilson and Watkins 1990).

Algorithm 2–Prim's Algorithm for Constructing an MST

Given: An undirected network N with distinct positive integer values $f(e)$ on each edge e of N.

Wanted: A spanning tree T of N with minimum value sum. We will specify such a tree by building its edge set E_T.

Step 1. Label the edges of N by $e_1, e_2, \ldots, e_q$ such that whenever $i < j$, we have $f(e_i) < f(e_j)$. Call this sequence of edges σ.

Step 2. Take any node u of N and place in E_T the first edge of σ incident with u.

Step 3. Add to E_T the first edge e_i of σ adjacent with at least one edge already in E_T such that the addition of e_i does not create a cycle.

Step 4. If $| E_T | = p - 1$, stop. Otherwise repeat Step 3.

Theorem 2. Given a connected network N in which all the edge values are distinct, Prim's algorithm will terminate with the unique minimum spanning tree T.

Using this algorithm we grow an MST from any initial node. For example we may take the node TON as u and place the edge TON-SAM in E_T. Next, we add the edge TON-SCK followed by SCK-SOC. We continue in this way until we have an MST. It will of course be the same MST obtained by using Algorithm 1.

There are many other potential applications of the MSTP in archaeology and anthropology. In a forthcoming work (Hage and Harary 1996), clustering in an MST is used to model the formation of dialect groups and marriage isolates in the Tuamotu Islands in Eastern Polynesia, while a third parallel processing, MST algorithm due to Boruvka (1926a, b), is used to simulate the evolution of overseas chiefdoms in the Lau Islands, Fiji.[6]

References Cited

Bellwood, P.
1987 *The Polynesians: Prehistory of an Island People.* Thames and Hudson, London.

Boruvka, O.
1926a O jistem problemu minimálním. *Acta Societatis Scientarium Naturalium Moravicae* 3:37–58.
1926b Príspevek k reseni otázky ekonomické stavby electrovodnich síti. *Eletrotechnicky Obzor* 15:153–154.

Brainerd, G. W.
1951 The Place of Chronological Ordering in Archaeological Analysis. *American Antiquity* 16:301–313.

Buckley, F., and F. Harary
1990 *Distance in Graphs.* Addison-Wesley, Reading, Massachusetts.

Burrows, E. G.
1938 *Western Polynesia: A Study of Cultural Differentiation.* Ethnologiska Studier 7, Göteborg, Sweden.

Friedman, J.
1981 Notes on Structure and History in Oceania. *Folk* 23:275–295.

Graham, R. L., and P. Hell
1985 On the History of the Minimum Spanning Tree Problem. *Annals of the History of Computing* 7:43–57.

Grofman, B., and J. T. Landa
1983 The Development of Trading Networks among Spatially Separated Traders as a Process of Proto-Coalition Formation: The *Kula* Trade. *Social Networks* 5:347–365.

Guiart, J.
1963 Un état palatial Océanien: l'Empire maritime des Tu'i Tonga. Appendix to *Structure de la Chefferie en Mélanésie du Sud.* Travaux et Mémoires, no. 66. Institut d'Ethnologie, Paris.

Hage, P., and F. Harary
1991 *Exchange in Oceania.* Clarendon Press, Oxford.
1996 *Island Networks.* Cambridge University Press, Cambridge, in press.

Irwin, G.
1992 *The Prehistoric Exploration and Colonisation of the Pacific.* Cambridge University Press, Cambridge.

Kirch, P. V.
1984 *The Evolution of the Polynesian Chiefdoms.* Cambridge University Press, Cambridge.
1988 *Niautoputapu: The Prehistory of an Island Chiefdom.* Burke Museum, Seattle.

Kirch, P. V., and R. C. Green
1987 History, Phylogeny, and Evolution in Polynesia. *Current Anthropology* 28:431–456.

Kruskal, J. B.
1956 On the Shortest Spanning Tree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society* 7:48–50.

Lewis, D.
1972 *We the Navigators.* University of Hawaii Press, Honolulu.

Pawley, A., and R. C. Green
1975 Dating the Dispersal of the Oceanic Languages. *Oceanic Linguistics* 12:1–68.

Prim, R. C.
1957 Shortest Connection Networks and Some Generalizations. *Bell System Technical Journal* 36:1389–1401.

Renfrew, C., and J. F. Cherry
  1986 *Peer Polity Interaction and Socio-political Change.* Cambridge University Press, Cambridge.
Renfrew, C., and G. Sterud
  1969 Close-Proximity Analysis: A Rapid Method for the Ordering of Archaeological Materials. *American Antiquity* 34:265–277.
Robinson, W. S.
  1951 A Method for Chronologically Ordering Archaeological Deposits. *American Antiquity* 16:293–301.
Wilson, R. J., and J. J. Watkins
  1990 *Graphs: An Introductory Approach.* John Wiley and Sons, New York.

Notes

1. Target size is obviously a nonsymmetric relation. The Tuamotu archipelago, for example, presents a large target for the isolated island of Rapa but not conversely. To control for this, Irwin constructed three different accessibility matrices based on the minimum, maximum, and mean angle of target size between island pairs. As it turned out all three measures gave essentially the same result. His Table 22 (Irwin 1992:197) uses minimum angle.

2. The addition of superfluous edges "crowds" a network.

According to Irwin's accessibility matrix, NZ should be joined to TUA rather than SOC and to SAM rather than TON. NCK should be joined to SOC, not TUA.

3. Synonyms of node and edge are point and line, vertex and branch. Renfrew and Sterud's "link," "chain," and "loop" correspond to edge, path, and cycle in graph theory.

4. In Burrows's trait list, Hawaii and the Southern Cooks share more traits with the Society Islands (the nucleus of Central Polynesia) than any other islands.

5. The corresponding network model of social stratification in Western Polynesia, which is geographically more compact, is the prestige-good system centered on the Tongan maritime chiefdom (Guiart 1963; Kirch 1984). In effect, Irwin proposes a network alternative to Friedman's (1987) devolutionary model of Oceanic social stratification which assumes that societies in Eastern Polynesia remained relatively isolated after settlement.

6. Boruvka's algorithm was independently discovered in the social sciences by Grofman and Landa (1983) who used it to simulate the evolution of a Melanesian trade network. See the discussion in Hage and Harary (1991).