The Link and Node Biased Encoding Revisited: Bias and Adjustment of Parameters

Thomas Gaube and **Franz Rothlauf**

Working Paper 10/2000
November 2000

Working Papers in Information Systems
Editor: Prof. Dr. Armin Heinzl

University of Bayreuth
Department of Information Systems
Universitaetsstrasse 30
D-95440 Bayreuth, Germany
Phone +49 921 552807, Fax +49 921 552216
E-Mail: wi@uni-bayreuth.de
Internet: *h*ttp://wi.oec.uni-bayreuth.de

The Link and Node Biased Encoding Revisited: Bias and Adjustment of Parameters

Thomas Gaube
Department of Information Systems
University of Bayreuth
Universitaetsstr. 30
D-95440 Bayreuth/Germany
gaube@stud.uni-bayreuth.de

Franz Rothlauf
Chair of Information Systems
University of Bayreuth
Universitaetsstr. 30
D-95440 Bayreuth/Germany
rothlauf@uni-bayreuth.de

November 27, 2000

Abstract

When using genetic and evolutionary algorithms (GEAs) for the optimal communication spanning tree problem, the design of a suitable tree network encoding is crucial for finding good solutions. The link and node biased (LNB) encoding represents the structure of a tree network using a weighted vector and allows the GEA to distinguish between the importance of the nodes and links in the network. This paper investigates whether the encoding is unbiased in the sense that all trees are equally represented, and how the parameters of the encoding influence the bias. If the optimal solution is underrepresented in the population, a reduction in the GEA performance is unavoidable. The investigation reveals that the commonly used simpler version of the encoding is biased towards star networks, and that the initial population is dominated by only a few individuals. The more costly link-and-node-biased encoding uses not only a node-specific bias, but also a link-specific bias. Similarly to the node-biased encoding, the link-and-node-biased encoding is also biased towards star networks, especially when using a low weighting for the link-specific bias. The results show that by increasing the link-specific bias, that the overall bias of the encoding is reduced. If researchers want to use the LNB encoding, and they are interested in having an unbiased representation, they should use higher values for the weight of the link-specific bias. Nevertheless, they should also be aware of the limitations of the LNB encoding when using it for encoding tree problems. The encoding could be a good choice for the optimal communication spanning tree problem as the optimal solutions tend to be more star-like. However, for general tree problems the encoding should be used carefully.

1 Introduction

The optimal communication spanning tree (OCST) problem (Hu, 1974) is defined to find a tree-structured communication network that connects all given nodes and satisfies their communication requirements for their minimum total cost. The number and positions of the network nodes are given a priori and the cost of the network is determined by the cost of the links.

Like other constrained spanning tree problems, the OCST problem is NP-hard (Garey & Johnson, 1979, p. 207). Thus, several genetic and evolutionary based algorithms (GEA) were proposed for solving the problem (Davis, Orvosh, Cox, & Qiu, 1993; Berry, Murtagh, & Sugden, 1994; Kim & Gen, 1999). When using GEAs for the OCST problem and not applying genetic operators directly to the phenotypes (Li & Bouchebaba, 1999), the design of a proper representation is demanding as there is a semantic gap between the structure of tree networks and common integer or bitstring representations. Recently researchers proposed encodings such as characteristic vectors (Davis, Orvosh, Cox, & Qiu, 1993; Tang, Man, & Ko, 1997; Sinclair, 1995; Berry, Murtagh, & Sugden, 1994), predecessor encodings (Krishnamoorthy, Ernst, & Sharaiha, 1999), Prüfer numbers (Kim & Gen, 1999), random keys (Rothlauf, Goldberg, & Heinzl, 2000) and weighted encodings (Raidl & Julstrom, 2000) for encoding the problem.

The characteristic vector indicates by a bitstring vector of length $n(n-1)/2$ if a link is used. This representation has high locality as small changes of the bitstring also result in small changes of the encoded tree. However, there are a lot of invalid solutions which makes using either repair operators, or specific mutation and recombination operators necessary.

When using the predecessor encoding, one node must be assigned to be the root of the network. The immediate predecessor on the path in the direction to the root is stored in a chromosome of length n. Therefore, invalid solution candidates can exist. Each element of the vector is of base n. As there are n choices for the root node, the encoding is redundant.

The Prüfer number encoding described in Prüfer (1918) encodes a tree with a string of length $n-2$ and each element of the string is of base n. However, although the encoding is compact and elegant, it suffers due to its weak locality. As a result, slight mutations of genes are followed by totally different network structures, and recombined offspring do not resemble the trees of their parents (Rothlauf & Goldberg, 2000).

Rothlauf, Goldberg, and Heinzl (2000) extended the random key encoding (Bean, 1994) to tree network design problems. A vector of length $n(n-1)/2$ consisting of floating numbers ranging from zero to one was used as a chromosome. A decoding algorithm builds the tree based on the order of the values in the vector and skips edges which form cycles until the tree is complete.

The link and node biased (LNB) encoding is a representation from the class of weighted encodings and was developed by Palmer (1994). Additional encoding parameters are necessary to balance the importance of link and node weights. The encoding was proposed to overcome the problems of characteristic vectors, predecessor representations and Prüfer numbers. Later Abuali, Wainwright, and Schoenefeld (1995) compared different representations for probabilistic minimum spanning tree (PMST) problems and in some cases found the best solutions by using the LNB encoding. Raidl and Julstrom (2000) observed for a similar weighted encoding which was used for the degree-constrained minimum spanning tree (d-MST) problem, solutions superior to those of several other optimization methods.

In this paper we want to investigate if there are any solution candidates preferred in the initial population, and how the setting of the two encoding parameters affects the bias of an arbitrary initial population. This is important because if the encoding prefers some solution candidates, a degradation of a GEA is often inescapable. To get rid of adjusting the encoding parameters, Palmer (1994) presented in the original paper, results using only one of the two possible parameters. We want to investigate the limitations of this approach and show the dependency of the solution quality of the initial population when using both parameters.

The paper is structured as follows. In the following section we give a short description of the LNB encoding. In section 3 we give a theoretical reason for the problems with biased encoding by introducing the notion of building blocks (BB) and reviewing results from Harik, Cantú-Paz, Goldberg, and Miller (1999). This is followed by an investigation into whether the LNB encoding using node weights only is biased, and how the individuals are represented in an initial population. In section 5 we investigate how the setting of the encoding parameters affects the bias of the encoding. The paper ends with concluding remarks.

2 A short description of the LNB encoding

In this section we want to review the motivation and the resulting properties of the link-and-node-biased encoding as described in Palmer (1994), and Palmer and Kershenbaum (1994).

As the costs of a communication network strongly depend on the length of the links, network structures that prefer short distance links tend to have in general a higher fitness. It is useful to run more traffic over the nodes near the gravity center of an area, than over nodes at the edge of this area (Kershenbaum, 1993). Thus, it is desirable to be able to characterize nodes to either be interior (some traffic only transits), or leaf nodes (all traffic terminates). As a result, the more important a link is, and the more transit traffic that crosses the node, the higher in general is the degree of the node. Nodes near the gravity center tend to have a higher degree than nodes at the edge. So the basic idea of the encoding is to encode the importance of a node. The more important the node is, the more traffic that should transit over this node.

When applying this idea to the OCST problem, the given distance matrix that defines the distances between any two nodes, is biased according to the importance of the nodes a link is connected to. If a node is not important, the modified distance matrix should increase the length of all links that are connected to this node.

The chromosome $\boldsymbol{b}$ holds the biases for each node, and has length n for an n node network. The values in the distance matrix d_{ij} are modified according to $\boldsymbol{b}$ using the weighting function

$$d'_{ij} = d_{ij} + p(b_i + b_j)d_{max}.$$

The bias b_i is a floating number between zero and one, d_{max} is the largest value in the distance matrix and p controls the influence of the biases. In the following we want to denote this approach as the node-biased encoding.

Using the bias-vector for encoding tree networks we get the encoded network structure by calculating the minimum spanning tree for the modified distance matrix. In the original work Prim's algorithm (Prim, 1957) was used. By running Prim's MST algorithm, nodes that are situated near other nodes will probably be interior nodes of high degree in the network. Nodes that are far away from the other nodes will probably be leaf nodes. Thus, the higher the bias of a node, the higher is the probability that it will be a leaf node. To finally get the tree's fitness the encoded network is evaluated by using the original distance matrix.

Palmer noticed in the original work that each bias b_i modifies a whole row and a whole column in the distance matrix. Thus, not all possible solution candidates can be encoded by this representation (Palmer, 1994, pp. 66-67).

To overcome this problem he introduced in the second, extended version of the representation an additional link-bias. The chromosome holds biases not only for the n nodes but also for all possible $n(n-1)/2$ links and has overall length $l = n(n+1)/2$. Therefore, the weighting function for the elements in the distance matrix was extended to

$$d'_{ij} = d_{ij} + P_1 b_{ij} d_{max} + P_2(b_i + b_j)d_{max}$$

with the link-specific bias b_{ij}. Using this representation the encoding could represent all possible trees. However, the string length is increased from $l = n$ to $l = n(n+1)/2$. In the following we want to denote this representation as the link-and-node-biased encoding.

Using the simple node-biased, or the more general link-and-node-biased encoding, makes it necessary to determine the value of one, respectively two, additional encoding parameters. In the original work from Palmer only results for the node-biased encoding and $p = 1$ are presented.

If the setting of the parameters could result in a biased representation of the individuals, a degradation of a GA is sometimes unavoidable as illustrated in the following section.

3 Unbiased initial populations and building block supply

In this section we want to review the requirement for representations to be unbiased, and strengthen our investigation using work from Harik, Cantú-Paz, Goldberg, and Miller (1999). Their model could be used for explaining why and how the quality of genetic search is affected by biased representations.

The equal distribution of the initial population is a desirable property of effective encodings (Ronald, 1997). Palmer formulated in his thesis necessary criteria for tree representations (Palmer, 1994, pp. 39):

> "It should be unbiased in the sense that all trees are equally represented; i.e., all trees should be represented by the same number of encodings. This property allows us to effectively select an unbiased starting population for the GA and gives the GA a fair chance of reaching all parts of the solution space."

Palmer recognized correctly that a widely usable encoding should be unbiased. For the LNB encoding he drew the conclusion at the end of his thesis that the

> "... new Link and Node Bias (LNB) encoding was shown to have all the desirable properties ..." (Palmer, 1994, pp. 90)

including those to be unbiased. However, we will illustrate in the following sections that this claim is not true.

Using the notion of building blocks (BB) in the context of representations means that we want unbiased encodings to represent all BBs uniformly. The encoding should uniformly represent individuals containing high- and low-quality BBs. If the individuals are represented unbiased, then the BBs are represented unbiased, too. Biased encodings, however, overrepresent some specific building blocks in a randomly generated population.

The theoretical results from Harik, Cantú-Paz, Goldberg, and Miller (1999) could be used to explain why and how the quality of genetic search is affected by biased representations. They calculated the probability of failure for a GA as

$$\alpha = \exp\left(-x_0 \frac{2d}{\sigma_{BB}\sqrt{\pi m'}}\right),$$

where d is the signal difference between the best and second best BB, $m' = m - 1$ with m is the number of BB in the problem, σ_{BB}^2 is the BB variance and x_0 is the expected number of copies of the best BB in the randomly initialized population. The probability of GA failure goes with $O(\exp(-x_0))$.

The result from Harik et. al. tells us that when the number of copies of the best BB in the randomly initialized population is reduced, the probability of GA failure grows exponentially. Transferring these results to biased encodings, we recognize that encodings that overrepresent individuals, which consist of mainly low-quality BBs, result in an exponential decrease of GA solution quality.

However, in general, biased encodings do not always lead to a decrease in solution quality. If an encoding is biased towards individuals that are similar to the good solutions, the solution quality of a GA is increased exponentially. Therefore, researchers and practitioners should be careful with using biased encodings. If the good solutions are similar to the individuals the encoding is biased towards, it could be a good choice. However, if it is biased towards low quality solutions, a failure of the GA is inescapable.

4 The node-biased encoding

It is known that the node-biased encoding is not capable of representing all possible network structures (Palmer, 1994). We want to investigate if the represented networks are encoded unbiased. We start with a distance matrix where all elements have the same value. This is followed by an investigation where the position of the nodes is chosen randomly.

4.1 All links have the same length

We assume that all values d_{ij} in the unbiased distance matrix are equal. Thus, the values in the biased distance matrix are determined by $\boldsymbol{b}$. We denote with b_l the lowest bias in $\boldsymbol{b}$. It is the bias for the lth node and all other biases are larger. Using this definition the modified length of each link connecting node i and j is always higher than the length of the link connecting either i and l or j and l:

$$d_{i,l} < d_{i,j} \text{ for } b_l = \min\{b_1, \dots b_n\},$$

where $i, j, l \in \{0, \dots, n\}$, $i \neq l$, $i \neq j$ and $l \neq j$. As the decoding algorithm chooses the shortest $n-1$ links that do not create a cycle for creating the encoded network, the only structure that could be represented by the node-biased encoding is a star network with center l.

For an n node network the number of possible star networks is n, whereas the number of all possible networks is n^{n-2}. Thus, only a small fraction of networks could be represented by the node-biased encoding. However, at least the represented star networks are unbiased, as the elements of $\boldsymbol{b}$ are uniformly distributed in the initial population.

We present an empirical verification of these results for a small 4 node problem in table 1. There are 16 possible networks, and 4 of them are stars with center l. For the experiments we created 1000 initial populations of size 1000. We see that it is not possible to create non-stars, and that the stars are represented uniformly. As a result, the node-biased representation is biased towards star networks if the distances between the nodes have the same value.

non-star	star with center			
	$l=1$	$l=2$	$l=3$	$l=4$
0%	25.01%	24.97%	24.92%	25.10%

Table 1: Average percentage of represented network types for a 4 node problem

4.2 Random length of links

Now we assume that the nodes are placed randomly on a two-dimensional plane.

For our investigation we randomly placed four nodes on a quadratic plane of size 1000 x 1000 and created randomly 1000 node-biased vectors. The distances between the nodes were calculated using the Euclidean distance. We performed 1000 experiments with different randomly located nodes.

If the representation would be unbiased each individual is created with probability $p = 1/n^{n-2} = 1/16 = 6.25\%$. However, our experiments revealed that star networks are strongly overrepresented. An average of 50.8% of all randomly created individuals are a star, whereas the portion of stars for all individuals is only $4/16 = 25\%$. Furthermore, the results show that all four stars are created uniformly. This means each of the four star networks is created with an average probability of about 12.7%

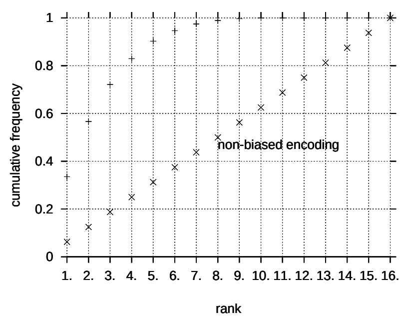


Figure 1: Distribution of network types

To investigate how the individuals are distributed in a randomly created population, we ordered the represented networks according to their frequency. In figure 1, we plot the cumulative frequency of the ordered number of copies an individual has in a randomly created population for a 4 node problem. If all individuals would be created with the same probability, the cumulative frequency would be linear over all possible individuals. However, for the node-biased encoding some individuals are created much more often. For our 4 node problem more than 90% of all individuals encode only five possible networks. In contrast to the overrepresentation of some individuals, some of the individuals are not represented at all. On average, 6 out of 16 possible networks are not represented at all when randomly creating 1000 individuals.

The simple node-biased encoding is biased towards star networks. A few individuals dominate a randomly created population and some solution candidates are impossible to be created. In the following section we want to investigate how the bias is influenced by the setting of its two parameters when using the link-and-node-biased encoding.

5 The link-and-node-biased encoding

Palmer (1994) proposed the link-and-node-biased encoding using node and link biases as a way to overcome some of the problems with the node-biased (NB) encoding. In the following we investigate how the bias which we have noticed for the NB encoding, is influenced by the choice of the two parameters P_1 and P_2.

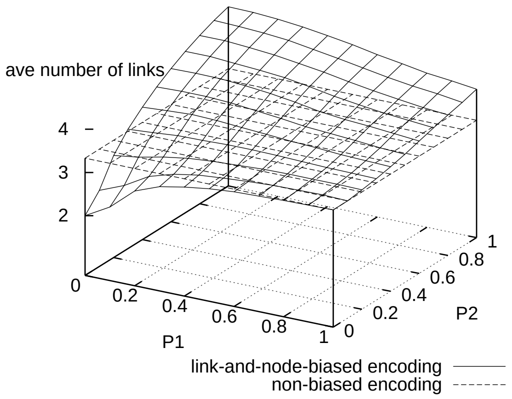
To investigate the bias of the LNB encoding, we randomly create a link-and-node-biased vector and measure the maximum number of links the individual has in common with one of the n stars. The more links an individual has in common with one of the stars, the more star-like it is. In figure 2 we present results for a randomly created 8, 16 and 32 node-problem. The average maximum number of links a randomly created link-and-node-biased individual has in common with a star is plotted over P_1 and P_2 and compared to an unbiased encoding. The number of links could vary from 2 (the individual represents a list network) to $n-1$ (the individual represents a star network). The parameters P_1 and P_2 vary between 0 and 1, and we generated 2000 individuals for each parameter setting.

The results show that for all three problem instances that the bias of the LNB encoding strongly depends on P_2. For small values of P_2 the individuals are slightly biased towards non-star structures, whereas for large values of P_2 the individuals are strongly biased towards star networks. An increase of P_1 reduces the strong influence of P_2 on the bias of the encoding. For small P_2, the individuals are only slightly biased, and for large P_2 they are less biased towards stars.

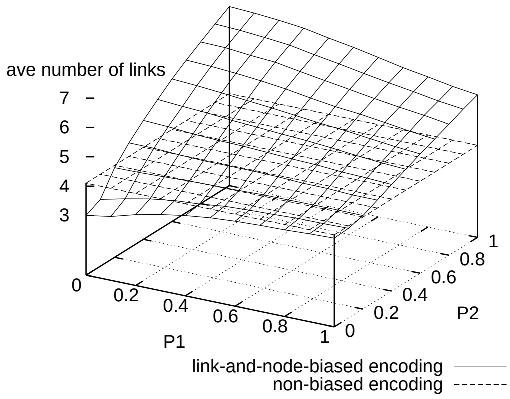
The results show that with increasing P_2, a randomly created link-and-node-biased individual is more biased towards star networks. With increasing P_1, however, the population becomes less biased. In agreement with section 4 we see, that using only a node bias ($P_1 = 0$) leads to a strongly biased representation. To get a more unbiased representation large values should be used for P_1.

6 Summary and Conclusions

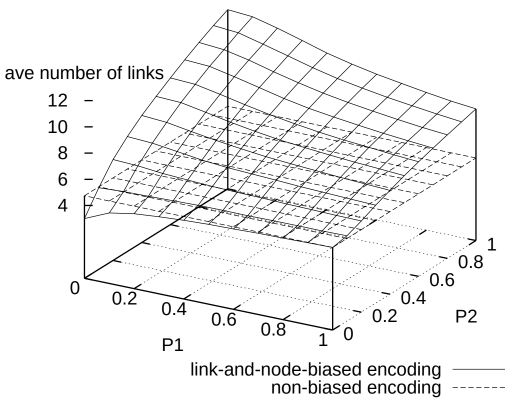
After a short review of tree network encodings we focus on the link and node biased encoding as described in Palmer (1994). We describe the encoding and illustrate that for setting the link-specific bias to zero we get the simplified node-biased encoding. Palmer stated as a necessary condition for encodings that it has to be unbiased in the sense that all individuals are equally represented. In section 3 we use the results from Harik, Cantú-Paz, Goldberg, and Miller (1999) to explain the effects of biased encodings on the solution quality of GAs. This is followed by an investigation into



(a) 8 node problem



(b) 16 node problem



(c) 32 node problem

Figure 2: Average maximum number of links a randomly generated individual has in common with a star. With increasing P_2 an individual is strongly biased towards star structures. Higher values for P_1 result in a more unbiased representation.

the bias of the node-biased encoding. Finally, we study how the bias of the link-and-node-biased encoding depends on the setting of both parameters.

We have seen that the simple node-biased encoding is biased towards star networks. Furthermore, a randomly created population is dominated only by a few individuals and some individuals could not be represented by the encoding at all. The investigation in the link-and-node-biased encoding reveals that the bias of the representation depends strongly on the used parameter setting. Using only a node-specific bias results, similar to the node-biased encoding, in a strong bias towards star networks. But fortunately the bias of the representation is decreased when using higher weights for the link-specific bias. Therefore, we strongly encourage users not to use only a node-specific, but also a link-specific bias, if they want to use the link-and-node-biased encoding, and if they are interested in having a more unbiased encoding.

Because optimal solutions for the optimal communication spanning tree problem often tend to be star-like, the LNB encoding could be a good choice for this problem. In general, however, the encoding has some serious problems, especially when using the simplified node-biased encoding.

Researchers should therefore be careful when using this encoding for other problems because some network structures are not encoded at all, and a randomly generated individual is biased.

Acknowledgments

The authors would like to thank David E. Goldberg, Martin Pelikan, Kumara Sastri, Armin Heinzl, Clarissa von Hoyweghen and the anonymous reviewers of the paper for useful comments and help with the project.

Support for this work was provided by the Deutschen Forschungsgemeinschaft under SFB-DFG 1083. This work was also sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant F49620-00-1-0163. Research funding for this work was also provided by the National Science Foundation under grant DMI-9908252. Support was also provided by a grant from the U. S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0003. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the National Science Foundation, the U. S. Army, or the U. S. Government.

References

Abuali, F. N., Wainwright, R. L., & Schoenefeld, D. A. (1995). Determinant factorization: A new encoding scheme for spanning trees applied to the probabilistic minimum spanning tree problem. In Eschelman, L. (Ed.), *Proceedings of the Sixth International Conference on Genetic Algorithms* (pp. 470–477). San Francisco, CA: Morgan Kaufmann.

Bean, J. C. (1994). Genetic algorithms and random keys for sequencing and optimization. *ORSA Journal on Computing*, *6*(2), 154–160.

Berry, L. T. M., Murtagh, B. A., & Sugden, S. J. (1994). A genetic-based approach to tree network synthesis with cost constraints. In Zimmermann, H. J. (Ed.), *Second European Congress on Intelligent Techniques and Soft Computing - EUFIT'94*, Volume 2 (pp. 626–629). Promenade 9, D-52076 Aachen: Verlag der Augustinus Buchhandlung.

Davis, L., Orvosh, D., Cox, A., & Qiu, Y. (1993). A genetic algorithm for survivable network design. In Forrest, S. (Ed.), *Proceedings of the Fifth International Conference on Genetic Algorithms* (pp. 408–415). San Mateo, CA: Morgan Kaufmann.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York: W. H. Freeman.

Harik, G., Cantú-Paz, E., Goldberg, D. E., & Miller, B. L. (1999). The gambler's ruin problem, genetic algorithms, and the sizing of populations. *Evolutionary Computation*, *7*(3), 231–253.

Hu, T. C. (1974, September). Optimum communication spanning trees. *SIAM Journal on Computing*, *3*(3), 188–195.

Kershenbaum, A. (1993). *Telecommunications network design algorithms*. New York: McGraw Hill.

Kim, J. R., & Gen, M. (1999). Genetic algorithm for solving bicriteria network topology design problem. In Angeline, P. J., Michalewicz, Z., Schoenauer, M., Yao, X., Zalzala, A., & Porto, W. (Eds.), *Proceedings of the 1999 IEEE Congress on Evolutionary Computation* (pp. 2272–2279). IEEE Press.

Krishnamoorthy, M., Ernst, A. T., & Sharaiha, Y. M. (1999). *Comparison of algorithms for the degree constrained minimum spanning tree* (Tech. Rep.). Clayton, Australia: CSIRO Mathematical and Information Sciences.

Li, Y., & Bouchebaba, Y. (1999). A new genetic algorithm for the optimal communication spanning tree problem. In Fonlupt, C., Hao, J.-K., Lutton, E., Ronald, E., & Schoenauer, M. (Eds.), *Proceedings of Artificial Evolution: Fifth European Conference* (pp. xx). Berlin: Springer.

Palmer, C. C. (1994). *An approach to a problem in network design using genetic algorithms.* unpublished PhD thesis, Polytechnic University, Troy, NY.

Palmer, C. C., & Kershenbaum, A. (1994). Representing trees in genetic algorithms. In *Proceedings of the First IEEE Conference on Evolutionary Computation*, Volume 1 (pp. 379–384). Piscataway, NJ: IEEE Service Center.

Prim, R. (1957). Shortest connection networks and some generalizations. *Bell System Technical Journal*, *36*, 1389–1401.

Prüfer, H. (1918). Neuer Beweis eines Satzes ueber Permutationen. *Arch. Math. Phys.*, *27*, 742–744.

Raidl, G. R., & Julstrom, B. A. (2000). A weighted coding in a genetic algorithm for the degree-constrained minimum spanning tree problem. In Carroll, J., Damiani, E., Haddad, H., & Oppenheim, D. (Eds.), *Proceedings of the 2000 ACM Symposium on Applied Computing* (pp. 440–445). ACM Press.

Ronald, S. (1997). Robust encodings in genetic algorithms: A survey of encoding issues. In *Proceedings of the Forth International Conference on Evolutionary Computation* (pp. 43–48). Piscataway, NJ: IEEE.

Rothlauf, F., & Goldberg, D. E. (2000). Pruefernumbers and genetic algorithms: A lesson on how the low locality of an encoding can harm the performance of GAs. In Deb, K., Rodolph, G., Yao, X., & Schwefel, H.-P. (Eds.), *Proceedings of the 2000 Parallel Problem Solving from Nature VI Conference* (pp. 395–404). Springer.

Rothlauf, F., Goldberg, D. E., & Heinzl, A. (2000). *Network random keys – a tree network representation scheme for genetic and evolutionary algorithms* (Technical Report No. 8/2000). University of Bayreuth, Germany.

Sinclair, M. C. (1995). Minimum cost topology optimisation of the COST 239 European optical network. In Pearson, D. W., Steele, N. C., & Albrecht, R. F. (Eds.), *Proceedings of the 1995 International Conference on Artificial Neural Nets and Genetic Algorithms* (pp. 26–29). New York: Springer-Verlag.

Tang, K. S., Man, K. F., & Ko, K. T. (1997). Wireless LAN desing using hierarchical genetic algorithm. In Bäck, T. (Ed.), *Proceedings of the Seventh International Conference on Genetic Algorithms* (pp. 629–635). San Francisco: Morgan Kaufmann.