Discrete Applied Mathematics 160 (2012) 2349–2363

Contents lists available at SciVerse ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

# Generalized *k*-ary tanglegrams on level graphs: A satisfiability-based approach and its evaluation☆

Andreas Wotzlaw [a,*], Ewald Speckenmeyer [a], Stefan Porschen [b]

[a] *Institut für Informatik, Universität zu Köln, D-50969 Köln, Germany*

[b] *Fachgruppe Mathematik, Fachbereich 4, HTW-Berlin, D-10318 Berlin, Germany*

ARTICLE INFO

*Article history:*
Received 30 August 2011
Received in revised form 22 May 2012
Accepted 24 May 2012
Available online 22 June 2012

*Keywords:*
Satisfiability
Mixed Horn formula
2-CNF
Level graph
Planar embedding
Tanglegram
Crossing minimization
Graph drawing
Computational biology
Combinatorial optimization

ABSTRACT

A tanglegram is a pair of (not necessarily binary) trees on the same set of leaves with matching leaves in the two trees joined by an edge. Tanglegrams are widely used in computational biology to compare evolutionary histories of species. In this work we present a formulation of two related combinatorial embedding problems concerning tanglegrams in terms of CNF-formulas. The first problem is known as the planar embedding and the second as the crossing minimization problem. We show that our satisfiability-based encoding of these problems can handle a much more general case with more than two, not necessarily binary or complete, trees defined on arbitrary sets of leaves and allowed to vary their layouts. Furthermore, we present an experimental comparison of our technique and several known heuristics for solving generalized binary tanglegrams, showing its competitive performance and efficiency and thus proving its practical usability.



## 1. Introduction

In recent times the interest in designing exact algorithms and efficient heuristics providing a better performance ratio for various variants of the tanglegram problem has increased. This is primarily motivated by their broad applications in computational biology, especially in phylogenetics.

In this work we introduce *generalized k-ary tanglegrams on level graphs*, a generalization of the well-known binary tanglegrams, and study two combinatorial embedding problems connected with them. A *binary tanglegram* [29] is an embedding (drawing) in the plane of a pair of rooted binary trees whose leaf sets are in one-to-one correspondence (perfect matching), such that the matching leaves are connected by *inter-tree edges*, called sometimes *tangles* or *tangle edges* and drawn as straight-line segments. A *crossing* occurs when two inter-tree edges intersect in a drawing. Clearly, the number of crossings between the inter-tree edges depends on the layout of the trees. From a practical point of view, an embedding with many crossings can hardly be analyzed. Fig. 1 shows an example of a binary tanglegram coming from phylogenetic studies done by Charleston and Perkins [9]. Taking this into account, the first problem one can consider here consists of determining an embedding of one or both trees such that the inter-tree edges do not cross, if such an embedding exists. This

☆ A preliminary version of this paper appeared in Speckenmeyer (2011) [35].

* Corresponding author. Fax: +49 2214705387.
*E-mail addresses:* wotzlaw@informatik.uni-koeln.de (A. Wotzlaw), esp@informatik.uni-koeln.de (E. Speckenmeyer), porschen@htw-berlin.de (S. Porschen).

0166-218X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2012.05.028

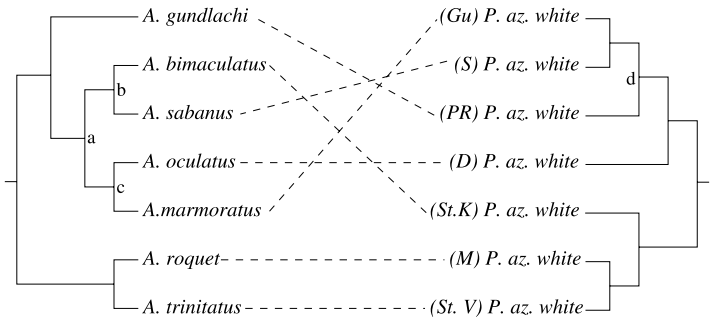


**Fig. 1.** A binary tanglegram from [9] showing phylogenetic trees for lizards (left tree) and strains of malaria (right tree) found in the Caribbean tropics. The dashed lines represent the host–parasite relationship. Here, the number of crossings is 7. This can be reduced to 1 by interchanging the children of nodes a–d.

problem is known as the *planar embedding* problem. If such a planar embedding is not possible, then we may want to find an embedding with as few crossing inter-tree edges as possible. This second problem, *crossing minimization*, is known in the literature also as the *tanglegram layout* problem [4,7,37].

Both problems belong to the area of graph drawing [11] and are motivated by the desire to find a good display of hierarchical structures, e.g., in software engineering, project management, or database design. For instance, tanglegrams occur when analyzing software projects in which trees are used to represent package, class, or method hierarchies. Changes in hierarchies can be analyzed over time, or automatically generated decompositions can be compared with human-made ones. This application yields tanglegrams on trees that are not binary in general [28].

Matching and aligning trees is also a recurrent problem in computational biology [29]. Embeddings with fewer crossings or with matching leaves close together are useful in biological analysis [37]. Here, prominent applications are in particular the comparisons of phylogenetic trees [9,10,12], which are used to represent a hypothesis of the evolutionary history (phylogeny) of a set of species. These species are drawn as the leaves of the tree whereas their ancestors are represented by the inner nodes. Different hypotheses may lead to a set of different candidate trees. An embedding imposes an order among the leaves of the trees. Therefore, comparing the drawings of the trees is equivalent to comparing the permutations of the leaves. In general, the simultaneous examination of a species phylogeny and a gene phylogeny can offer biologists more insights into evolutionary processes, e.g., gene duplication, gene extinction, or cross-species gene transfer, that the inspection of either tree alone cannot provide [27,36].

Bansal et al. [4] analyzed *generalized tanglegrams* where the number of leaves in the two binary trees may be different and a leaf in one tree may match multiple leaves in the other tree, thus no perfect matching is required here. They pointed out that such a generalization of the problem makes it possible to address not only the gene tree and species tree embedding problem, but also those problems in which the inter-tree edges between the trees can be completely arbitrary. Such general instances arise in several settings, e.g., in the analysis of host–parasite cospeciation [29,18].

*Related work.* Crossing minimization in tanglegrams has parallels to crossing minimization in graphs. Computing the minimum number of crossings in a graph is NP-hard [16]. However, it can be verified in linear time whether a graph has a planar embedding [20]. The last assertion holds also for a more special case of level graphs [25,31]. Computing the minimum number of crossings is fixed-parameter tractable [7,22]. Analogously, crossing minimization in tanglegrams is NP-hard, as shown by Fernau et al. [14] by a reduction from the MAX-CUT problem [15], while the special case of the planarity test can be executed in linear time [14].

Furthermore, the problem of minimizing the number of crossings where one tree is fixed and the layout of the other tree is allowed to vary can be solved efficiently. For binary trees with arbitrary topology, Fernau et al. [14] showed a $O(n \log^2(n))$ solution, further improved to $O(n \log^2(n)/\log\log(n))$ by Bansal et al. [4]. Here, $n$ denotes the number of leaves in each tree. Venkatachalam et al. [37] provided recently an algorithm working on the integer linear programming (ILP) formulation of the problem with the so far best-known time bound of $O(n \log(n))$.

Recently Buchin et al. [7] have proved that under the widely accepted Unique Games Conjecture [23] there is no constant factor approximation algorithm for minimizing the number of crossings of binary tanglegrams. Nöllenburg et al. [28] gave an extensive experimental evaluation of some heuristics, an exact branch-and-bound algorithm, and an ILP-based approach for binary tanglegrams. Finally, Baumann et al. [5], by exploiting the fact that crossing minimization in (not necessarily binary) tanglegrams can be seen as a generalization of bipartite crossing minimization, formulated it as a quadratic linear ordering problem with some additional side constraints and evaluated it by using semidefinite optimization and the mathematical programming software CPLEX [21].

For the case of generalized tanglegrams where the layout of one tree is fixed, Bansal et al. [4] presented two algorithms with running times $O(mh)$ and $O(m \log^2(m)/\log\log(m))$, where $m$ is the number of edges between the two trees and $h$ is the height of the tree whose layout can change. To obtain those polynomial running times, the layout of the other tree must be fixed. Based on the result of Fernau et al. [14], they also showed that the existence of a planar embedding can be verified in $O(m)$ time.