A Satisfiability-based Approach for Generalized Tanglegrams on Level Graphs*

Andreas Wotzlaw[1] Ewald Speckenmeyer[1] Stefan Porschen[2]

[1]Department of Computer Science
University of Cologne, Germany

[2]Department 4
HTW-Berlin, Germany

Dagstuhl Seminar ”SAT Interactions”, Nov. 18-23, 2012

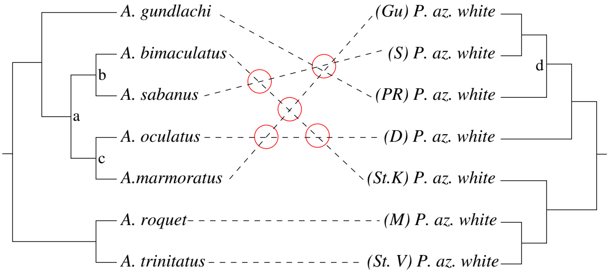
*This talk has been presented at EURO 2012, Vilnius, Lithuania.

Outline

Binary tanglegrams

Binary tanglegram

An embedding (drawing) in the plane of a pair of rooted binary trees which leaf sets are in one-to-one correspondence (perfect matching).



Important questions:

1. Is there an embedding inducing no crossings? $\rightarrow$ **planarity test**
2. If not, find an embedding with as few crossings as possible? $\rightarrow$ **crossing minimization**

Generalized tanglegrams

Motivation

- good display of hierarchical structure, e.g., in software engineering, database design, project management [di Battista, 1998]
- matching and aligning phylogenetic trees in computational biology [DasGupta et al., 1999; Dufayard et al., 2005]

Complexity results for binary tanglegrams

- planarity test decidable in linear time [Fernau et al., '10]
- crossing minimization is NP-complete (MAX-CUT) [Fernau et al., '10]

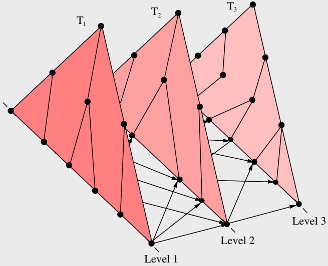
Generalized tanglegram [Bansal et al., 2009]

- the number of leaves in the two binary trees may be different
- no perfect matching required
- can address more problems in bioinformatics

Generalized tanglegrams on level graphs

Generalized tanglegram (G, F) on a level graph

- forest F of k-ary trees $T_1, T_2, \ldots$
- level graph G with n nodes and inter-tree edges E
- **Question**: Does there simultaneously exist a planar embedding of G (horizontal plane) with planar embeddings for F (vertical planes)?



Observation 1

- crossing minimization in level graphs is NP-hard [Eades/Wormald '94]
- level graphs with $|E| > 2|V| - 4$ are not planar [Randerath et al., '01]

Generalized tanglegram as a satisfiability problem

Given an instance (G, F) of a generalized tanglegram on a level graph G with n nodes and k-ary trees F defined on the levels of G, for some fixed $k > 1$

Goal a satisfiability-based formulation of crossing minimization for (G, F)

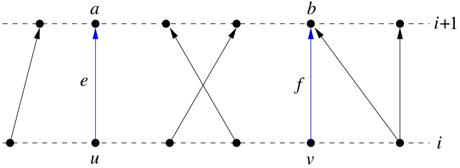
Transformation procedure

- Step 1: Construction of a CNF-formula C_G for the level graph G
- Step 2: Construction of a CNF-formula C_F for the forest F

Result: a CNF-formula $C_{GF} := C_G \wedge C_F$ for (G, F) such that C_{GF} is satisfiable iff (G, F) has planar embedding (no crossings).
Crossing minimization as an instance of PARTIAL MAX-SAT on C_{GF}.

Level embedding by a Boolean formula

Consider two adjacent levels i and $i + 1$ of G



Observation 2

Two arcs $e = (u, a)$ and $f = (v, b)$ connecting levels i and $i + 1$ with different tails $u \neq v$ and heads $a \neq b$ do not cross wrt. linear orders on i and $i + 1$ iff

$$u < v \Leftrightarrow a < b$$

Step 1: Construction of C_G for level graph G

1. For each pair $\{u, v\}$ of distinct nodes from each level $i \in L$, create a Boolean variable uv such that

$$uv = \mathit{true} \text{ iff } u < v \text{ in a linear order on level } i.$$

2. Create the following Boolean subformulas:

   (I) non-crossing conditions C_I: for every two arcs $e = (u, a)$ and $f = (v, b)$ connecting levels i and $i + 1$ with $u \neq v$ and $a \neq b$

   $$uv \leftrightarrow ab$$

   (II) antisymmerty conditions C_{II}: for each node pair $\{u, v\}$ from each level in L

   $$uv \leftrightarrow \overline{vu}$$

   (III) transitivity conditions C_{III}: for each node triple $\{u, v, w\}$ from each level in L

   $$uv \wedge vw \rightarrow uw$$

Result: $C_G = C_I \wedge C_{II} \wedge C_{III}$, where $C_I \wedge C_{II} \in$ 2-CNF and $C_{III} \in$ 3-CNF.

Preliminary results on C_G

It holds:

- C_G has $O(n^2)$ Boolean variables
- C_G has $O(n^3 + |E|^2)$ clauses
- by Observation 1, for the planarity test only $O(n^2)$ 2-clauses in C_I
- for the planarity test C_{III} can be dropped [Randerath et al., 2001]
  $\Rightarrow C_G \setminus C_{III} \in$ 2-CNF

Proposition 1

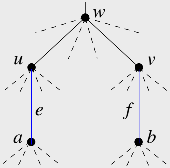
A level graph G with n nodes has a planar embedding iff $C_G \setminus C_{III}$ is satisfiable.
The test can be done in time $O(n^2)$.

Plane embedding of tree $T_i \in F$

Observation 3

Let T_i be a complete k-ary tree of height d on a level i with fixed linear order:

- for $d = 1$ the edges of T_i never cross in any drawing T_i
- let $w \in T_i$ such that the height of subtree $T_i(w)$ is at least 2



- the edges leaving w never cross in any drawing of $T_i(w)$
- let $e = \{u, a\}$ and $f = \{v, b\}$ be two edges from $T_i(w)$ with $u \neq v$ having both depth 1. In a drawing of $T_i(w)$, e and f do not cross iff

$$u < v \quad \Leftrightarrow \quad a < b.$$

Step 2: Construction of C_F for forest F

Repeat for each $T_i \in F$

1. For each level $j = 1, ..., d$ of T_i and each pair $\{u, v\}$ of distinct nodes from j, create a Boolean variable uv such that

$$uv = true \ \text{ iff } \ u < v \text{ in a linear order on level } i.$$

2. Create the following Boolean subformulas:

   (IV) non-crossing conditions $C_{IV}^{T_i}$: for each level j and two edges $e = \{u, a\}$, $f = \{v, b\}$ of T_i such that $u \neq v$ have depth j and a, b have depth $j + 1$

$$(uv \leftrightarrow ab) \wedge (vu \leftrightarrow ba)$$

   (V) antisymmerty conditions $C_V^{T_i}$: for each node pair $\{u, v\}$ from each level in T_i

$$uv \leftrightarrow \overline{vu}$$

Result: $C_{T_i} = C_{IV}^{T_i} \wedge C_V^{T_i} \in$ 2-CNF

Satisfiability-based formulation of (G, F)

CNF-Formula for the forest F:

$$C_F = \bigwedge_{T_i \in F} C_{T_i}$$

- C_F has $O(n^2)$ Boolean variables
- C_F has $O(n^2)$ 2-clauses

Finally, by applying the 2-step transformation procedure to (G, F), we obtain

$$C_{GF} = C_G \wedge C_F = (C_I \wedge C_{II} \wedge C_{III}) \wedge C_F$$

- C_{GF} has $O(n^2)$ Boolean variables
- C_{GF} has $O(n^3 + |E|^2)$ clauses
- only the transitivity conditions $C_{III} \in$ 3-CNF, the rest $\in$ 2-CNF

Main complexity results

By Proposition 1, for the planarity test C_{III} can be omitted
$\Rightarrow C_{GF} \setminus C_{III} \in$ 2-CNF solvable for SAT efficiently [Aspvall et al., 1979]

Theorem 1 [Wotzlaw et al., 2012]

(G, F) has a planar embedding iff $C_{GF} \setminus C_{III}$ is satisfiable. The test needs $O(n^2)$ time, for some fixed integer $k > 1$.

Crossing minimization is an instance of PARTIAL MAX-SAT $\in$ NP-hard.

Theorem 2 [Wotzlaw et al., 2012]

Let t be a truth assignment satisfying $C_{GF} \setminus C_I$ and **minimizing the number** τ **of not satisfied clauses in** C_I for some fixed integer $k > 1$. Then τ is the minimum number of arc crossings in an embedding of (G, F).

Experimental evaluation

Goals

Evaluation of **SatTG** for generalized binary tanglegrams (GBT) in terms of:

- computation of optimal layouts
- performance ratio $\rho := \frac{1+\tau}{1+\tau_{OPT}}$
- computation time t

Details on **SatTG** [Wotzlaw et al., 2012]

- performs crossing minimization as described above
- resulting PARTIAL MAX-SAT encodings contain up to one million Boolean variables and 40 millions clauses
- utilizes several complete PARTIAL MAX-SAT solvers, depending on the problem type and size, e.g., akmaxsat, clasp, QMaxSAT0.4
- computes exact or approximate solutions (with timeout set)

Experimental setup

We compare **SatTG** with

- an exact integer LP-based method **ILPTG** (using CPLEX 12.1)
- three polynomial-time heuristics **AH**, **LH**, and **LAH** [Bansal et al., 2009]
  $\rightarrow$ the fastest heuristics for GBT known so far

Test data:

- random GBTs: $n \leq 800$ and $|E| = 1.15n$ [Bansal et al., 2009]
- simulated gene/species trees: $n \leq 1200$ and $|E| \leq 2n$ [Syvanen, 1985; Arvestad et al., 2004]
- real-world GBTs: $n \leq 101$ and $|E| \leq 3n$ [Sanderson/McMahon, '07]

Evaluation results

Computation of optimal layouts:

- **SatTG** and **ILPTG** comparable for instances with $n < 200$
- instances with $n > 400$ very time and resource consuming

Average performance ratios ρ:

Category	n	**AH**	**LH**	**LAH**	**IPLTG**	**SatTG**
random	≤ 100	1.109	1.020	1.006	**1***	**1***
random	≥ 200	1.026	1.016	**1.011**	1.082	1.076
simulated	≤ 100	1.269	1.023	1.001	**1***	**1.003**
simulated	≥ 200	1.265	1.072	1.024	5.533	**1.017**
real-world	10-200	1.668	1.012	1.001	**1***	**1***

Computation time of **SatTG**:

- the fastest method for real-world GBTs with $n < 200$ and random and simulated GBTs with $n < 60$
- better than **ILPTG** and similar to **LAH** for simulated GBTs
- outperformed for random GBTs with $n \geq 60$

Summary

Conclusion:

- generalization of the tanglegram problem on level graphs
- planarity test solvable efficiently in $O(n^2)$
- crossing minimization intractable (PARTIAL MAX-SAT)
- competitive for computing optimal layouts of medium-sized instances
- very well qualified for application in interactive visualization tools

Open problems:

- bounds for the approximation ratio for generalized tanglegrams

Thank you for your attention!