

Minimum Support Augmentation in Chordal Graphical Models

Alper Ferudun*

15 October 2026

Abstract

Given an observed support in a finite Cartesian table, we minimize the number of distinct new cells needed to make every maximal-clique projection of a chordal graphical model surjective. For a clique path, the missing clique configurations form a partially ordered set, and the minimum augmentation equals its width. This gives an exact algorithm by bipartite matching, as well as an equivalent lower-bound circulation formulation. In contrast, the decision problem is NP-complete on the fixed six-variable chordal graph with maximal cliques ABC , ABX , ACY , and BCZ , even when X, Y, Z are binary; the alphabets of A, B, C are allowed to grow. A partial Latin square supplies the reduction and a concrete support for which the three individual root-leaf optima are five but the global optimum is six. Integer separator multiplicities also give an exact dynamic program on any clique tree. The statistical interpretation uses the classical positive-MLE criterion for decomposable models.

1 Support augmentation and main results

Let V be a finite nonempty set of variables, with nonempty finite alphabets $\mathcal{I}_v$. Write $\mathcal{I}_D = \prod_{v \in D} \mathcal{I}_v$ for $D \subseteq V$, and $\mathcal{I} = \mathcal{I}_V$. The empty product has one element. For a full cell $x \in \mathcal{I}$, write x_D for its restriction to D , and $\pi_D(S) = \{x_D : x \in S\}$. Fix a chordal graph G on V , its family $\mathcal{C}$ of maximal cliques, and a nonempty support $S \subseteq \mathcal{I}$. Every cell of $\mathcal{I}$ is allowed: there are no structural zeros.

The missing configurations at a clique C and their indicators are

$$M_C = \mathcal{I}_C \setminus \pi_C(S), \quad b_C(u) = \begin{cases} 1 & u \in M_C, \\ 0 & u \notin M_C. \end{cases}$$

A *repair* is a set $T \subseteq \mathcal{I} \setminus S$ satisfying $\pi_C(S \cup T) = \mathcal{I}_C$ for every $C \in \mathcal{C}$. Define

$$\tau_G(S) = \min\{|T| : T \text{ is a repair}\}, \quad B = \sum_{C \in \mathcal{C}} |M_C|. \quad (1)$$

We count distinct new full cells, not marginal entries, multiplicities or changes to the old support. Extending each missing configuration to an arbitrary full cell gives

$$\max_{C \in \mathcal{C}} |M_C| \leq \tau_G(S) \leq B. \quad (2)$$

Such an extension cannot belong to S . In particular, repairs always exist under the Cartesian hypothesis.

*Mercury Software GmbH. Email: alper@mercurycodelab.com. GitHub: [AlperTheKing](https://github.com/AlperTheKing).

A *clique path* is an ordering $C_1, \dots, C_q$ of the maximal cliques in which the cliques containing any given variable are consecutive. A clique tree has the analogous connectedness property. Empty separators are allowed, so the definitions also cover disconnected graphs. Computational bounds below use the explicitly enumerated support, variable labels and local configuration tables $\mathcal{I}_C$ as input. They do not use a succinct encoding of exponentially large tables.

Theorem 1. *For the support augmentation problem (1):*

1. *On a clique path, let $\mathcal{P} = \{(i, u) : u \in M_{C_i}\}$, with*

$$(i, u) < (j, v) \iff i < j \text{ and } u_{C_i \cap C_j} = v_{C_i \cap C_j}. \quad (3)$$

This is a strict partial order, and

$$\tau_G(S) = \text{width}(\mathcal{P}) = B - \nu, \quad (4)$$

where ν is the maximum matching size in the bipartite graph having two copies of $\mathcal{P}$ and an edge $a_L b_R$ whenever $a < b$. An optimal repair is computable in polynomial time.

2. *Deciding whether $\tau_G(S) \leq k$ is NP-complete for the fixed chordal graph with maximal cliques*

$$ABC, \quad ABX, \quad ACY, \quad BCZ, \quad (5)$$

when $\mathcal{I}_A = \mathcal{I}_B = \mathcal{I}_C = [n]$ and $\mathcal{I}_X = \mathcal{I}_Y = \mathcal{I}_Z = \{0, 1\}$. Here n is part of the input; the graph has treewidth two.

The positive-MLE interpretation is recalled in Section 2. Section 4 supplies a second path algorithm, using integral flow. Section 6 gives an explicit strict gap between local and global repair numbers. On general clique trees, Section 7 gives an exact dynamic program that is polynomial when the number of configurations in each clique is bounded by a fixed constant.

Relation to earlier work

AIM's computational algebraic statistics Question 7 asks for zero-pattern conditions preventing a strictly positive MLE and for a combinatorial description of marginal-cone facets [1]. The general facial criterion and the decomposable specialization are established statistical theory [2]. The optimization problem (1) is a related augmentation problem; the results here do not classify arbitrary marginal-cone facets or address arbitrary structural zeros.

Our path formula specializes Dilworth's chain-decomposition theorem [3]; integer gluing on a clique tree is an instance of acyclic consistency for bags [5]. Minimum covers of query results provide a close database analogue. Kara and Olteanu [6] distinguish minimum from inclusion-minimal covers, give the minimum two-relation construction, and exhibit failure of local cover plans to achieve a global minimum. Thus neither the two-clique matching idea nor the existence of local-cover failures is new. In (1), only missing configurations require coverage, while all full Cartesian extensions remain available. The global compatibility order handles these requirements on a whole clique path.

Other repair objectives are different. Smallest synthetic witnesses [7] minimize the sum of tuple counts in input relations whose query result is prescribed, rather than the number of new full rows used here. Missing-answer repairs [8] concern a specified missing answer rather than simultaneous coverage of all missing clique configurations. Their complexity statements should not be transferred without identifying both the feasible repairs and the cost being minimized.

2 Positive likelihood and integer gluing

Give the cells of S arbitrary positive counts, with total $m > 0$. The graphical log-linear model on G consists of strictly positive probability distributions that factor into clique functions.

Proposition 2 (Classical decomposable criterion). *The ordinary strictly positive multinomial MLE exists if and only if $\pi_C(S) = \mathcal{I}_C$ for every maximal clique C .*

Proof. At an interior likelihood maximum, the score equations equate model and empirical clique marginals. A strictly positive distribution has positive marginals, which proves necessity.

Conversely, suppose all empirical clique marginals $\bar{n}_C$ are positive. Choose a clique tree with edge separators R_e . The consistent empirical marginals give

$$p(x) = \frac{\prod_{C \in \mathcal{C}} \bar{n}_C(x_C)}{\prod_e \bar{n}_{R_e}(x_{R_e})}. \quad (6)$$

All denominator factors are positive. Successively summing out leaf-clique variables shows that p is normalized and has the prescribed clique marginals. It is positive and factors over G . For any other positive model distribution q , $\log(q/p)$ is a sum of clique functions. Its empirical expectation therefore equals its expectation under p , giving

$$\ell(q) - \ell(p) = m \sum_x p(x) \log \frac{q(x)}{p(x)} = -mD(p||q) \leq 0.$$

Thus p is a maximum. □

This is the decomposable specialization of the standard relative-interior criterion [2, Corollary 2]. Accordingly, $\tau_G(S)$ is the minimum number of additional distinct cells whose inclusion permits a positive MLE. It depends on the support, not on the positive count magnitudes. Augmentation is a combinatorial design problem; it does not justify inventing observations in an existing statistical data set.

For a nonnegative integer table y_C and $R \subseteq C$, define the marginal by summing multiplicities:

$$(\pi_R y_C)(s) = \sum_{u \in \mathcal{I}_C: u_R = s} y_C(u).$$

Lemma 3 (Integer gluing). *Fix a clique tree. The number $\tau_G(S)$ is the minimum N for which there are integer tables $y_C \geq b_C$ with common total N such that, on every tree edge CD ,*

$$\pi_{C \cap D} y_C = \pi_{C \cap D} y_D. \quad (7)$$

Proof. A repair provides such tables by counting its projections. Conversely, list the root table as N rows with multiplicities. Process child cliques in rooted order. A new clique intersects the previous union exactly in its parent separator: a variable on both sides must occur along their connecting tree path. Equation (7) permits the rows to be matched within each separator stratum. Continuing in this way produces N full rows with the prescribed tables. This is the usual acyclic bag-gluing argument; see [5, Theorem 3].

The resulting multiset covers every missing configuration. If an optimal multiset contained a row of S , deleting it would lose no missing configuration. If it contained a duplicate, one copy could be deleted without losing coverage. Either deletion would contradict optimality. Thus a minimum multiset consists of distinct new rows, proving equality with (1). □

For $q > 1$, the equalities on separators also force equal totals, including when a separator is empty. For $q = 1$, the common-total condition simply names the total of the single table. Integer gluing does not imply integrality of an arbitrary continuous lower-bound relaxation on a branching clique tree.

3 The clique path formula

Proof of Theorem 1(1). The only nontrivial order property in (3) is transitivity. If $i < j < k$, running intersection implies $C_i \cap C_k \subseteq C_j$. Thus agreement of u with v and of v with w forces agreement of u with w on $C_i \cap C_k$. Hence $(i, u) < (j, v) < (k, w)$ implies $(i, u) < (k, w)$.

The missing configurations covered by a single full row form a chain in $\mathcal{P}$. Assign every missing configuration to one of the rows covering it. A repair therefore gives a chain partition with at most as many parts as rows. Conversely, the partial assignments in a chain agree on every overlap. Their union extends to a full cell by choosing arbitrary values on the remaining variables. Every nonempty chain contains a missing configuration, so its extension is outside S . A partition into k chains gives a repair with at most k rows; coincident extensions can be merged. Consequently $\tau_G(S)$ equals the minimum number of chains in a partition of $\mathcal{P}$.

Dilworth's theorem [3, Theorem 1.1] identifies this minimum with $\text{width}(\mathcal{P})$. To obtain the matching formula, orient a matched edge $a_L b_R$ from a to b . Every vertex has in-degree and out-degree at most one, and strict order forbids cycles. A matching of size ν therefore gives vertex-disjoint directed paths partitioning $\mathcal{P}$ into $B - \nu$ chains. Conversely, linking consecutive vertices in each of k chains gives a matching of size $B - k$. This proves (4).

There are B vertices in each bipartite part and at most B^2 edges. Each compatibility test uses only variable labels and local assignments. A maximum matching and the chain extensions can therefore be found in polynomial time in the stated input size. At optimality the extensions cannot coincide, since that would give a repair smaller than the proved lower bound. $\square$

Any antichain gives a lower bound because one row covers at most one of its elements. Thus an antichain and a repair of equal cardinality certify optimality. If $B = 0$, the empty poset has width zero and the empty repair is optimal.

Corollary 4 (Two cliques). *For two maximal cliques with separator R , let $m_i(s) = |\{u \in M_{C_i} : u_R = s\}|$. Then*

$$\tau_G(S) = \sum_{s \in \mathcal{I}_R} \max\{m_1(s), m_2(s)\}.$$

Proof. The matching graph is a disjoint union, over s , of complete bipartite graphs of part sizes $m_1(s)$ and $m_2(s)$. Apply (4). $\square$

The same separator-stratum matching is the minimum two-relation cover construction in [6]. The path theorem chooses all compatibilities together, rather than committing to successive locally optimal pairings.

4 A lower-bound circulation algorithm

Put $R_i = C_i \cap C_{i+1}$ for $1 \leq i < q$, and $R_0 = R_q = \emptyset$. Layer i , for $0 \leq i \leq q$, contains one vertex for each configuration of R_i . Every $u \in \mathcal{I}_{C_i}$ supplies a forward arc

$$u_{R_{i-1}} \longrightarrow u_{R_i}$$

from layer $i - 1$ to layer i , with lower bound $b_{C_i}(u)$, upper bound B , and cost zero. Keep parallel arcs: different clique configurations can have identical separator restrictions. Add an arc from the unique layer- q vertex back to the unique layer-0 vertex, with capacity B and unit cost.

Proposition 5. *If $B > 0$, the minimum integral circulation cost in this network is $\tau_G(S)$. Decomposing an optimal circulation yields an optimal repair.*

Proof. Arc flows at clique C_i form the table y_{C_i} . Conservation at a separator vertex is precisely the corresponding equality in (7). The return flow is the common table total N . Lemma 3 therefore identifies the minimum cost with $\tau_G(S)$; the bound (2) ensures that capacity B is sufficient.

The forward network is layered and acyclic. Decompose an integral circulation into forward paths and return arcs. Along a forward path, adjacent clique assignments agree on separators; running intersection makes all repeated variable assignments agree. Since the maximal cliques cover V , a path defines a full cell. The deletion argument of Lemma 3 shows that at optimum these cells are distinct and outside S . $\square$

Integer minimum-cost circulation gives a polynomial algorithm. Alternatively, fix the return flow to N , reduce lower-bound feasibility to an integral maximum-flow problem, and binary-search $0 \leq N \leq B$. Feasibility is monotone: an arbitrary full row can be added to a feasible multiset. The value B is feasible by extending the missing configurations individually. When $B = 0$, return the empty repair without constructing the network.

There are $\sum_{i=0}^q |\mathcal{I}_{R_i}|$ vertices and $1 + \sum_{i=1}^q |\mathcal{I}_{C_i}|$ arcs, with capacities at most $B \leq \sum_i |\mathcal{I}_{C_i}|$. This explains both polynomial input size and the need to retain parallel arcs and empty separators.

5 Hardness on a fixed branching graph

Proof of Theorem 1(2). The graph (5) is chordal: remove the simplicial vertices X, Y, Z and then the triangle ABC . Its maximum clique size is three, hence its treewidth is two. Its cliques have the star-shaped clique tree centered at ABC .

Reduce from completion of a valid partial Latin square P of order n , an NP-complete problem [4]. Define three 0/1 tables:

$$\begin{aligned} d_{AB}(a, b) &= 1 && \text{if position } (a, b) \text{ is blank,} \\ d_{AC}(a, c) &= 1 && \text{if symbol } c \text{ is absent from row } a, \\ d_{BC}(b, c) &= 1 && \text{if symbol } c \text{ is absent from column } b. \end{aligned}$$

If P has N blanks, each table has exactly N ones, since filled entries do not repeat a symbol within a row or column. Let S consist of all full cells satisfying

$$\begin{aligned} d_{AB}(a, b) = 1 &\implies x = 0, \\ d_{AC}(a, c) = 1 &\implies y = 0, \\ d_{BC}(b, c) = 1 &\implies z = 0. \end{aligned} \tag{8}$$

This support is explicitly enumerable in at most $8n^3$ rows. Every (a, b, c) occurs with $x = y = z = 0$, so the root projection is full. The missing ABX configurations are exactly $(a, b, 1)$ with $d_{AB}(a, b) = 1$. To see that there are no others, fix any allowed x and set $y = z = 0$. The analogous assertion holds for ACY and BCZ . Thus the root has no missing configurations and each leaf has exactly N .

Suppose first that $N > 0$ and there is a repair with at most N rows. By (2) there are exactly N rows. Each must cover one missing configuration in each of the three leaves, and each missing configuration must occur exactly once. Hence every row has $x = y = z = 1$. Its triple (a, b, c) places symbol c in a blank position (a, b) . Every blank is filled once; every missing row–symbol and column–symbol pair is used once. The resulting assignments complete P to a Latin square.

Conversely, a Latin completion supplies N rows $(a, b, c, 1, 1, 1)$ at the blank positions. These rows are distinct, are outside S , and cover all missing leaf configurations. They form a repair

because the root projection was already full. For $N = 0$, P is complete and (8) gives $S = \mathcal{I}$, so the same equivalence is immediate. We have proved

$$\tau_G(S) \leq N \iff P \text{ has a Latin completion.}$$

The construction is polynomial in the explicit partial square. Finally, a yes instance admits at most $\min\{k, B\}$ new rows, by (2). Membership outside S , distinctness and all clique projections can be verified in polynomial time in the explicit input. The decision problem therefore belongs to NP. $\square$

The construction places restrictions on the *observed* support, not on which future rows are allowed. All $8n^3$ full cells remain available. Growing alphabets are essential to this fixed-graph statement; six fixed binary variables would give only a fixed finite problem. Nor does the reduction prove that every chordal graph lacking a clique path has a hard augmentation problem.

6 A local five versus global six example

Use zero-based labels and the partial Latin square

$$P = \begin{pmatrix} 0 & 1 & * \\ 1 & 0 & * \\ * & * & * \end{pmatrix},$$

where a star marks a blank. Both first rows would need symbol 2 in the last column, so P has no completion. The support (8) has 95 rows; its three leaf missing sets each have size five. The reduction proves $\tau_G(S) \geq 6$.

The following six new rows give equality. Columns are in the order A, B, C, X, Y, Z :

$$\begin{array}{ccc|ccc} 0 & 2 & 2 & 1 & 1 & 1 \\ 1 & 2 & 2 & 1 & 1 & 0 \\ 2 & 0 & 2 & 1 & 1 & 1 \\ 2 & 1 & 2 & 1 & 0 & 1 \\ 2 & 2 & 0 & 1 & 1 & 1 \\ 2 & 2 & 1 & 0 & 1 & 1 \end{array}$$

For each leaf the five missing pair values are $(0, 2), (1, 2), (2, 0), (2, 1), (2, 2)$, with its auxiliary bit equal to one. The displayed rows cover them for ABX , ACY and BCZ respectively. Each row contains a missing leaf configuration, so it is outside S .

Each individual root–leaf pair has optimum five: the root misses nothing, the leaf misses five configurations, and one extension per missing configuration suffices. Thus all three local optimum cardinalities are five while the joint optimum is six. The point is not the already-known possibility of local cover failure [6], but an explicit observed support realizing this discrepancy within (1).

7 Exact dynamic programming on a clique tree

Proposition 6. *On any clique tree, $\tau_G(S)$ can be computed by a finite dynamic program with integer separator-count messages. If $|\mathcal{I}_C| \leq K$ for every clique and K is a fixed constant, the algorithm is polynomial in the explicit input size.*

Proof. Fix $N \in \{0, \dots, B\}$ and root the clique tree. At a clique C , enumerate integer tables $y_C \geq b_C$ of total N . Retain a table precisely when its projection to each child separator occurs among the accepted messages from that child. For a nonroot clique, send the parent-separator projection of every retained table. At a leaf all locally feasible tables are retained. At the root, declare feasibility if at least one table remains.

Inductively, a message is accepted exactly when its table extends to compatible integer tables throughout the descendant subtree. The induction step chooses a witnessing table in each child subtree; all relevant compatibility equations are then satisfied. Lemma 3 identifies root feasibility with existence of a covering multiset of N full rows. Minimize N and retain back-pointers to recover the tables and then the repair.

There are at most $(N + 1)^K$ candidate tables per clique. Since $B \leq qK$, enumeration over N , projection comparison and message storage take polynomial time for fixed K . $\square$

The state must retain multiplicities, not just the sets of separator values occurring. The bound on K controls the number of clique configurations; fixed graph treewidth alone does not do so when alphabet sizes grow. This is consistent with Theorem 1(2) and does not conflict with bounded-treewidth algorithms for deciding MLE existence [2], which solve a different problem.

8 Exact checks and remaining questions

The source package includes two standard-library Python checkers. One compares integer lower-bound circulation with memoized exact set cover on 735 path instances, checks 85 partial-Latin instances, and verifies the six-row witness. The other compares matching, circulation and exact set cover on all 65,535 nonempty supports of a binary four-variable path and on 81 additional boundary cases. Normal and optimized runs, including isolated replays, agree. All optimization arithmetic is integral; explicit exceptions remain active under optimized Python. These are finite regressions, not substitutes for the proofs. The general clique-tree dynamic program is proved above but is not separately benchmarked by these checkers.

The classification of tractable branching structures is left open. Other questions include weighted costs for added rows, restricted allowed cell sets, and repair criteria for nonchordal models. None is answered by treating separator supports as counts or by applying the path order outside its running-intersection hypothesis.

This manuscript was developed with AI assistance in mathematical derivation, computational checking and preparation. It is an unrefereed preprint; no independent review, proof-assistant certification or absolute priority is claimed.

References

- [1] American Institute of Mathematics, *Computational algebraic statistics*, workshop problem list, 27 March 2004, Question 7, pp. 4–5.
<https://aimath.org/WWN/compalgstat/compalgstat.pdf>.
- [2] N. Eriksson, S. E. Fienberg, A. Rinaldo and S. Sullivant, Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models, *J. Symbolic Comput.* **41** (2006), 222–233. doi:10.1016/j.jsc.2005.04.003. Preprint: [arXiv:math/0405044](https://arxiv.org/abs/math/0405044).
- [3] R. P. Dilworth, A decomposition theorem for partially ordered sets, *Ann. of Math.* **51** (1950), 161–166. doi:10.2307/1969503.

- [4] C. J. Colbourn, The complexity of completing partial Latin squares, *Discrete Appl. Math.* **8** (1984), 25–30. [doi:10.1016/0166-218X\(84\)90075-1](https://doi.org/10.1016/0166-218X(84)90075-1).
- [5] A. Atserias and P. G. Kolaitis, Consistency of relations over monoids, *J. ACM* (2025). [doi:10.1145/3721855](https://doi.org/10.1145/3721855). Theorem numbering refers to [arXiv:2312.02023v1](https://arxiv.org/abs/2312.02023v1) (2023).
- [6] A. Kara and D. Olteanu, Covers of query results, *ICDT 2018*, LIPIcs **98**, article 16. [doi:10.4230/LIPIcs.ICDT.2018.16](https://doi.org/10.4230/LIPIcs.ICDT.2018.16). Preprint: [arXiv:1709.01600](https://arxiv.org/abs/1709.01600).
- [7] A. Esmailpour, B. Glavic, X. Hu and S. Sintos, Smallest synthetic witnesses for conjunctive queries, *Proc. ACM Manag. Data* **3**(2) (2025), article 113. [doi:10.1145/3725250](https://doi.org/10.1145/3725250).
- [8] J. Comer and V. Tannen, The complexity of finding missing answer repairs, *ICDT 2026*, LIPIcs **365**, article 12. [doi:10.4230/LIPIcs.ICDT.2026.12](https://doi.org/10.4230/LIPIcs.ICDT.2026.12).