Revisiting the Limits of MAP Inference by MWSS on Perfect Graphs

Adrian Weller
University of Cambridge

Abstract

A recent, promising approach to identifying a configuration of a discrete graphical model with highest probability (termed MAP inference) is to reduce the problem to finding a maximum weight stable set (MWSS) in a derived weighted graph, which, if perfect, allows a solution to be found in polynomial time. Weller and Jebara (2013) investigated the class of binary pairwise models where this method may be applied. However, their analysis made a seemingly innocuous assumption which simplifies analysis but led to only a subset of possible reparameterizations being considered. Here we introduce novel techniques and consider all cases, demonstrating that this greatly expands the set of tractable models. We provide a simple, exact characterization of the new, enlarged set and show how such models may be efficiently identified, thus settling the power of the approach on this class.

1 INTRODUCTION

Undirected graphical models, also called Markov random fields (MRFs), are a powerful and compact way to represent dependencies among variables, with wide use in machine learning. A fundamental problem is to identify a configuration of variables with highest probability, termed maximum a posteriori (MAP) inference. However, this is NP-hard (Shimony 1994), leading to much interest in finding methods and domains where the problem may be solved exactly in polynomial time (we call such models tractable). Previously, it was known that MAP inference is tractable with other methods (including attractive binary pairwise MRFs and weighted bipartite matching problems) where it was already known that MAP inference is tractable with other methods. Since few classes are known to be tractable, this was very encouraging and prompted a search to understand the limits of how far the approach may be applied. Weller and Jebara (2013) provided an exact characterization of the subclass of binary pairwise models which may be solved in this way provided that they made a restrictive assumption about the types of reparameterizations permitted (essentially not allowing singleton potentials to be absorbed into incident edge potentials). By relaxing this assumption and considering the full range of possible reparameterizations, we show that we may broaden the tractable subclass dramatically.

Similarly to Weller and Jebara (2013), we demonstrate that it is still possible to decompose the problem by considering each block (maximal 2-connected subgraph) of the original MRF topology. Our main result (Theorem 8) is that a binary pairwise model is tractable via a perfect NMRF iff each block is either balanced (no frustrated cycles) or almost balanced (may be rendered balanced by deleting a single vertex). Although these blocks may be handled in isolation by other methods, to our knowledge, our new NMRF approach is the first method proven to find a MAP configuration for the entire MRF, even if there are very many such blocks, in polynomial time (see §5).
2 PRELIMINARIES

We consider a model \((V, \Psi)\) with binary variables \(V = \{X_1, \ldots, X_n\} \subseteq \mathbb{B} = \{0, 1\}\), together with \((\log)\) potential functions over subsets \(c \subseteq V\), \(\Psi = \{\psi_c : c \subseteq \mathcal{P}(V)\}\), where \(\mathcal{P}(V)\) is the powerset of \(V\). Write \(x = (x_1, \ldots, x_n)\) for one particular complete configuration and \(x_c\) for a configuration just of the variables in \(c\). A potential function \(\psi_c\) maps each possible setting \(x_c\) of its variables \(c\) to a finite real number \(\psi_c(x_c)\).

Identifying a configuration of variables that is most likely, termed maximum a posteriori or MAP inference, is the combinatorial problem of identifying

\[
x^* = \arg\max_{x = (x_1, \ldots, x_n)} \sum_{c \subseteq C} \psi_c(x_c).
\]

Here we restrict attention to pairwise models, that is \(|c| \leq 2\ \forall c \subseteq C\), that are positive, i.e. each configuration \(x\) has probability \(p(x) > 0\) (when this is not the case, typically 0 may be replaced by a sufficiently small \(c\)). If \(|c| = 1\), \(\psi_c\) is a singleton potential; if \(|c| = 2\) then it is an edge potential. Binary pairwise models play a key role in computer vision, both directly and as critical subroutines in solving more complex problems \cite{boykov2001}. Note that it is possible to convert any positive discrete MRF into an equivalent binary pairwise model \cite{eaton2013}.\footnote{The same paper shows that if the MRF has configurations with 0 probability, it may still be approximated arbitrarily closely.} though this may lead to a much larger state space.

We describe the NMRF approach in \cite{eaton2013} but first introduce relevant concepts from graph theory.

2.1 Terms from Graph Theory

We follow standard definitions and omit some familiar terms, see \cite{diesel2010}.

A graph \(G(V, E)\) is a set of vertices \(V\), and edges \(E \subseteq V \times V\). Throughout this paper, all graphs are finite and simple, that is a vertex may not be adjacent to itself (no loops) and each pair of vertices may have at most one edge (no multiple edges). The complete graph on \(n\) vertices, written \(K_n\), has all \(\binom{n}{2}\) edges.

A signed graph \cite{heider1946, harary1953} is a graph \((V, E)\) together with one of two possible signs for each edge. This is a natural structure when considering binary pairwise models, where we characterize edges as either attractive or repulsive, see \cite[3.1.1]{harary1959}. A frustrated cycle in a signed graph is a cycle with an odd number of repulsive edges. A signed graph is balanced if it contains no frustrated cycles. We say a signed graph is almost balanced if it contains a vertex such that deleting it renders the remaining graph balanced. Observe that, with our definition, a balanced graph is also almost balanced.

A graph is connected if there is a path connecting any two vertices. A cut vertex of a connected graph \(G\) is a vertex \(v \in V\) such that deleting \(v\) disconnects \(G\). A graph is 2-connected (or biconnected) if it is connected and contains no cut vertex. A block is a maximal 2-connected subgraph. Every block is either \(K_2\) (two vertices joined by an edge) or contains a cycle. Different blocks of \(G\) overlap on at most one vertex, which must be a cut vertex. Hence \(G\) can be written as the union of its blocks with every edge in exactly one block. These blocks are connected without cycles in the block tree for each connected component of \(G\).

A stable set in a graph is a set of vertices, no two of which are adjacent. A weighted graph \((V, E, w)\) is a graph with a nonnegative real value for each vertex, called its weight \(w(v)\). Of all stable sets in a weighted graph, a maximum weight stable set (MWSS) is one with maximum weight. A maximal maximum weight stable set (MMWSS) is a MWSS of maximum cardinality (this is useful in our context since, after reparameterization, we may have many nodes with 0 weight, see \cite[3.3]{eaton2013}).

A clique in a graph is a set of vertices, of which every pair is adjacent. The clique number of a graph \(G\), written \(\chi(G)\), is the maximum size of a clique in \(G\).

The complement of a graph \(G(V, E)\) is the graph \(G(V, F)\) on the same vertices with an edge in \(F\) iff it is not in \(E\). For example, a clique is the complement of a stable set and vice versa.

A coloring of a graph is a map from its vertices to the integers (considered the colors of the vertices) such that no two adjacent vertices share the same color. The chromatic number of a graph \(G\), written \(\chi(G)\), is the minimum number of colors required to color it. Observe that clearly \(\chi(G) \geq \omega(G)\) for any graph \(G\).

An induced subgraph \(H(U, F)\) of a graph \(G(V, E)\) is a graph on a subset of the vertices \(U \subseteq V\), inheriting all edges with both ends in \(U\), so \(F = \{(v, w) \in E : v, w \in U\}\).

A graph \(G\) is perfect iff \(\chi(H) = \omega(H)\) for all induced subgraphs \(H\) of \(G\). Examples include any bipartite or chordal graph. Related concepts (see Theorem 1): a hole in a graph \(G\) is an induced subgraph which is a cycle of length \(\geq 4\) (note this means the cycle must be chordless); an antihole is an induced subgraph whose complement is a hole. A hole or antihole is odd if it has an odd number of vertices. Note that, as a special case, a hole with 5 vertices is isomorphic to an antihole of the same size. It is easily shown that a graph containing an odd hole or antihole is not perfect. Remarkably, the converse holds, see Theorem 1.

\cite{harary1959} described such signed graphs as having point index \(\leq 1\).
2.2 Properties of Perfect Graphs

There is a rich literature on perfect graphs. We highlight a key result used in this paper, which confirmed a conjecture by Claude Berge that had been open for several decades.

**Theorem 1** (Strong Perfect Graph Theorem, Chudnovsky et al., 2006). A graph is perfect iff it contains no odd hole or antihole (equivalently, iff neither the graph nor its complement contains an odd hole).

The NMRF approach to MAP inference is to reduce the problem to finding a maximum weight stable set in a derived weighted NMRF graph, as described in [3]. By using Theorem 1, we can check if the NMRF contains odd holes or antiholes to see if it is perfect, and if it is, then a key property is that a MWSS may be found in time polynomial in \( n \), the number of nodes in the NMRF, via the ellipsoid method (Grötschel et al., 1984).

Yildirim and Fan-Orzechowski (2006) derived a faster exact approach in \( O(n^6) \) time based on semidefinite programming. This may be improved using primal-dual methods (Chan et al., 2009). Alternatively, linear programming may be used for the MWSS problem but this requires \( O(n^3\sqrt{n_K}) \) time, where \( n_K \) is the number of maximal cliques in the graph (Jebara 2009, 2014). When \( n_K \) is small, this can be more efficient than semidefinite programming. However, in the worst case, \( n_K \) may be exponentially large in \( n \), thus linear programming is useful only in some settings. Another possibility is to use message-passing methods (Poulds et al., 2011; Jebara, 2014), though these also become inefficient for graphs with many cliques.

Where other methods exist for solving exact MAP inference, the reduction to MWSS is typically not the fastest approach in \( O(n^3) \) time, where \( n \) is small, this can be more efficient than semidefinite programming. However, in the worst case, \( n_K \) may be exponentially large in \( n \), thus linear programming is useful only in some settings. Another possibility is to use message-passing methods (Poulds et al., 2011; Jebara, 2014), though these also become inefficient for graphs with many cliques.

Where other methods exist for solving exact MAP inference, the reduction to MWSS is typically not the fastest method, yet there is hope for improvement since the field is advancing rapidly, with significant breakthroughs in recent years (Chudnovsky et al., 2006; Faenza et al., 2011).

3 THE NMRF APPROACH TO MAP INFERENCE

We describe the reduction of MAP inference to MWSS on an MRF. Given an MRF model, construct a **nond Markov random field** (NMRF) \( N \), defined as follows:

- A weighted graph \( N(V_N, E_N, w) \) with vertices \( V_N \), edges \( E_N \) and a weight function \( w : V_N \rightarrow \mathbb{R}_+ \).
- Each \( c \in C \) of the original model maps to a clique in \( N \), which we call a clique group. This contains one node for each possible configuration \( x_c \), with all these nodes pairwise adjacent in \( N \).
- Nodes in \( N \) are adjacent iff they have inconsistent settings for any variable \( X_i \).
- Nonnegative weights of each node in \( N \) are set as \( \psi_c(x_c) = \min_{x_c} \psi_c(x_c) \), see 3.1 for an explanation of the subtraction.

We differentiate between two types of nodes in an NMRF:

- **An snode** relates to a setting of a single variable from its MRF. Equivalently, it is a node from a clique group deriving from a singleton potential with \( c = \{ X_i \} \) for some \( i \).
- **An enode** is a node from a clique group deriving from some \( c \in C \) with \( |c| = 2 \), that is an enode derives from an edge potential of the MRF. See Figure 1 for an example of a MRF mapping to an NMRF containing only enodes.

![Figure 1: An example of mapping a binary MRF with edge potentials (shown as a factor graph) to its NMRF. Subscripts denote the edge variables, superscripts denote the configuration number.](image)

Jebara (2014) proved that a maximal cardinality set of consistent nodes in \( N \) with greatest total weight, i.e. a MMWSS of \( N \) (see 2.1), will identify a globally consistent configuration of all variables of the original MRF that solves the MAP inference problem (1).

**Sketch proof:** (Slightly different to Jebara (2014), this allows the result to be extended after discussing pruning in 3.1). A MMWSS \( S \) is consistent by construction and clearly contains at most one node from each clique group. It remains to show it has at least one node from each clique group. Suppose a clique group has no representative. Identify a member of this group which could be added to \( S \), establishing a contradiction since \( S \) is maximal, as follows: the group overlaps with some variables of \( S \), copy the settings of these; for all other variables in the group, pick any setting. Note that if we do not insist on a maximal MWSS, it is possible that we do not get a representative for some clique groups and hence do not obtain a complete MAP configuration for the initial MRF.

3.1 Reparameterizations, Pruning and Efficiency

A **reparameterization** is a transformation

\( \{ \psi_c \} \rightarrow \{ \psi'_c \} \) s.t. \( \forall x_c \sum_{c \in C} \psi_c(x_c) = \sum_{c \in C} \psi'_c(x_c) + \text{constant} \).

This clearly does not modify a MAP solution (1) but can be helpful to make the problem easier.
One simple reparameterization is just to add a constant to any \( \psi_e \) function, since any consistent configuration has exactly one setting for each group of variables \( c \). Hence we may subtract the minimum \( \psi_e(x_c) \) and assume that in each clique group of \( N \), the minimum weight of a node is exactly zero. The reduction result above holds provided we insist on a maximal MWSS (MMWSS). This is helpful since to find a MMWSS, it is sufficient first to remove or prune the zero weight nodes, find a MWSS on the remaining, smaller graph, then reintroduce a maximal number of the zero weight nodes while maintaining stability of the set.

Different reparameterizations will yield different pruned NMRFs. By the earlier argument: MWSS will find one member from each of some of the clique groups, then we can always find one of the zero weight nodes to add from each of the remaining groups using any greedy method. Hence the following result holds (Weller and Jebara 2013, Lemma 6), where efficient means in polynomial time.

**Lemma 2.** MAP inference on an MRF is tractable provided \( \exists \) an efficiently identifiable efficient reparameterization such that the MRF maps to a perfect pruned NMRF.

### 3.1.1 Singleton Transformations and Associativity

Another useful form of reparameterization is a singleton transformation, which is a change in one or more \( \psi \) functions for a single variable, with corresponding changes to a higher order term which brings it to a convenient form.

For binary pairwise models, it is easily shown that a reparameterization of an edge via singleton transformations, \( \begin{pmatrix} \psi_{00} & \psi_{01} \\ \psi_{10} & \psi_{11} \end{pmatrix} \rightarrow \begin{pmatrix} \psi'_{00} & \psi'_{01} \\ \psi'_{10} & \psi'_{11} \end{pmatrix} \), is valid iff \( \psi_{00} + \psi_{11} - \psi_{01} - \psi_{10} = \psi'_{00} + \psi'_{11} - \psi'_{01} - \psi'_{10} \).

This quantity, the associativity of the edge, is well-defined and invariant with respect to any singleton transformation.

An edge is either attractive\(^1\) in which case it tends to pull its two end vertices toward the same value, or repulsive, in which case it tends to push its two end vertices apart to different values, according to whether its associativity is \( \geq 0 \) or \( < 0 \). A binary pairwise model is attractive iff every one of its edges is attractive.

An attractive edge may be reparameterized such that three of its entries are 0, and therefore may be pruned, leaving only either \( \psi_{00} \) or \( \psi_{11} \) enodes, with form \( (x = 0, y = 0) \) or \( (x = 1, y = 1) \), with a positive value. Similarly, we may reparameterize a repulsive edge to leave just one enode with form \( (x = 0, y = 1) \) or \( (x = 1, y = 0) \). For our purpose of mapping to a perfect NMRF, this ability to prune away many enodes can be helpful, though see \(^1\)

\( \psi \)

### 3.1.2 Connecting nodes

Whenever a singleton transformation is used to reparameterize an edge so as to leave one enode, it is easily checked that the result on the singleton potentials is always to raise the score of the individual connected nodes, i.e. s nodes with opposite settings to the enode, thus making it more likely for the enode to connect in the NMRF. As an example, we show how a symmetric attractive edge with associativity \( W > 0 \) is transformed to leave only \( \psi'_{00} \):

\[
\begin{pmatrix} \frac{W}{2} & 0 \\ 0 & \frac{W}{2} \end{pmatrix} \rightarrow \begin{pmatrix} 0 & W \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & \frac{W}{2} \\ -\frac{W}{2} & 0 \end{pmatrix}
\]

Here \( \oplus \) indicates combining the potentials appropriately, with a constant of \( -\frac{W}{2} \). Observe that each singleton potential had its setting of 1 raised by \( \frac{W}{2} > 0 \).

If singleton potentials are very high relative to edge weights then connecting enodes may be avoided, preventing holes and antiholes, and making inference easier. However, 0 singleton potentials and symmetric edge potentials are valid settings, which will lead to connecting enodes when edges are reparameterized so as to obtain one enode per edge.

### 3.2 Decomposition

We shall use the following decomposition theorem in the proof of our main result in §4.2.1.

**Theorem 3.** (MRF Decomposition, Weller and Jebara 2013, Theorem 7). If \( MRF_A(V_A, \Psi_A) \) and \( MRF_B(V_B, \Psi_B) \) both map to perfect NMRFs \( N_A \) and \( N_B \), and have exactly one variable \( s \) in common, i.e. \( V_A \cap V_B = \{s\} \), with the same snode for variable \( s \) (there must be at least one), then the combined MRF \( (V_A \cup V_B, \Psi_A \cup \Psi_B) \) maps to an NMRF \( N' \) which is also perfect. The converse is true by the definition of perfect graphs.

By repeatedly breaking apart an MRF at cut vertices, this allows the question of which MRFs are tractable with the NMRF approach to be focused only on sub-models whose topologies form blocks of the original graph (see §2.1 for definitions). Note the requirement that \( N_A \) and \( N_B \) contain the same snodes for the overlapping variable \( s \), which we shall examine in §4.2.1.

### 4 NEW RESULTS

We clarify the objective. Our aim is to classify which binary pairwise MRFs, identified only by their signed graph topology (see §2.1), are tractable via a perfect NMRF for any valid potentials, i.e. any finite singleton and edge potentials consistent with the signed topology. If a topology is intractable for some set of valid potentials, then it does not meet our criterion for tractability. Note that without this requirement, any MRF with finite edge potentials would be tractable with sufficiently strong singleton potentials, since they would directly force each variable’s assignment.

\(^1\)Other equivalent terms used are associative, ferromagnetic or regular. This is equivalent to \( \psi \) for the edge being supermodular, or having submodular cost function.
We highlight important properties of perfect graphs, then clarify the difference between our analysis and that of Weller and Jebara [2013], henceforth ‘WJ’.

From the definition of a perfect graph (s.1.1), if nodes are removed from a graph, it can only make it easier to show that it is perfect. By Theorem 1, perfection may be checked by looking for possible odd holes or antiholes. In an NMRF $N$, intuitively edge enodes are typically more likely than singleton snodes to form part of an odd hole or antihole (and hence lead to the graph not being perfect). This is because enodes each contain settings for two variables, leading to more adjacencies in $N$ (see s.3.1.1). Hence, it seems reasonable always to reparameterize edges using singleton transformations so that each edge results in just one enode (see s.3.1.1). Since this could cause arbitrary reparameterized singleton potentials, WJ made the assumption that all possible snodes could be present. We call this the snode assumption.

In contrast, here we shall consider the full range of possible reparameterizations, including where snodes are sometimes absent, typically effected through being ‘absorbed’ by an edge potential, see s.4.1.1. Accordingly, for such absorbing edges, we must consider that all enodes from its clique group could be present in the pruned NMRF. Perhaps surprisingly, we shall show that the net effect of adding enodes and removing snodes can be very helpful.

A central tool in WJ’s method was the following result.

**Lemma 4** (Weller and Jebara [2013], Lemma 16). **Subject to the snode assumption**: For some valid potentials, any cycle $C$ in a binary pairwise MRF $M$ generates an induced (chordless) cycle $H$ in its NMRF $N$ with size at least as great, and with the same parity (odd/even number of vertices) as the number of repulsive edges (odd/even) in the MRF’s cycle. In particular, if $M$ has any frustrated cycle with $\geq 4$ edges, or with $3$ edges requiring any snode to link the enodes in $N$, then this maps to an odd hole in $N$.

Essentially, each repulsive edge, which will have an enode of the form $(s = 0, t = 1)$ or $(s = 1, t = 0)$, flips the parity of the end variable as we move around the cycle. In s.4.1.1 we show that such holes may be ‘broken’ by using absorbing edges to remove connecting snodes.

### 4.1 Absorbing Singleton Potentials, Breaks, Phantom Edges and Surrogate snodes

Here we consider edges that absorb incident singleton potentials. For example, consider the reparameterization $\psi'(x_i, x_j) = \psi_i(x_i) + \psi_j(x_i, x_j) + \psi_j(x_j), \psi'(x_i) = 0, \psi'(x_j) = 0$. This removes all snodes at $X_i$ and $X_j$ from the pruned NMRF, though now any of the 4 possible enodes for such an absorbing edge could be present.

**Definition 5.** Given a particular reparameterization of a sub-MRF, a **break** at a variable vertex is a missing snode in the pruned NMRF (typically because it has been absorbed by an incident edge) such that it is not possible for enodes from some incident edges to connect through the vertex. A particular reparameterization of a sub-MRF signed topology is **unbroken** if it contains no breaks.

A break can be very helpful but often raises new difficulties. We illustrate the idea and introduce related new terms with the following example. Consider a frustrated 5-cycle $v_1, \ldots, v_5$ that has been reparameterized as in s.3.1.1 to have one enode per edge. For a range of singleton potentials, this will be unbroken and hence will form an odd hole in the pruned NMRF, as described in Lemma 4. However, if we arrange that a particular vertex, say $v_3$, is broken with respect to the incident edges $v_2 - v_3$ and $v_3 - v_4$, then the odd hole could be avoided. This might be achieved as follows: (i) first reparameterize as in s.3.1.1 to get one nonzero enode per edge, choosing a reparameterization such that both the $v_2 - v_3$ enode and the $v_3 - v_4$ enode have setting $v_3 = 0$, hence they only connect at $v_3$ via the snode $(v_3 = 1)$; (ii) now add a **phantom edge** $v_1 - v_3$, which did not exist in the original MRF; this initially has $\psi_{13}(x_1, x_3) = 0 \forall x_1, x_3$, but then is reparameterized to absorb the singleton potentials $\psi_1(x_1)$ and $\psi_3(x_3)$, i.e. $\psi_{13}'(x_1, x_3) = \psi_1(x_1) + \psi_3(x_3), \psi_1' = 0, \psi_3' = 0$. This now breaks the original odd hole at $v_3$, preventing it from connecting and apparently solving the problem. However, at least one new odd hole has been introduced, formed by NMRF nodes from the $v_3 - v_4 - v_5 - v_1$ section of the original MRF together with either one or two enodes from the new phantom $v_1 - v_3$ edge. To see this, recall that we have chosen the $v_3 = 0$ setting and suppose that the $v_5 - v_1$ enode has setting $v_1 = a \in \{0, 1\}$. Let $\bar{a} = 1 - a$. We assume that the phantom $v_1 - v_3$ edge could have all 4 enodes, so in particular this includes $(v_1 = \bar{a}, v_3 = 1)$ which would connect the ends (i.e. the enode for $v_5 - v_4$ and that for $v_5 - v_1$) with one enode, and $\{(v_1 = v_3 = 0), (v_1 = a, v_3 = 1)\}$ which would connect the ends with two enodes. Thus, there is a new odd hole - we solved one problem but introduced another, for no net benefit.

If the 5-cycle were part of a larger MRF, including say $v_6$ that is not adjacent to any variable of the 5-cycle, one might think that we could form the break at $v_3$ yet avoid the problem above by introducing a phantom edge $v_3 - v_6$. However, this does not work: either the $(v_3 = 1, v_6 = 0)$ or $(v_3 = 1, v_6 = 1)$ enode could play the role of what we call a **surrogate snode**, i.e. it would play the same role as the initial $(v_3 = 1)$ snode did in connecting the original odd hole formed by $v_1, \ldots, v_5$.

The reasoning above yields the following results.

**Lemma 6.** An unbroken cycle in an MRF with at least 4 edges, containing an edge reparameterized as an absorbing edge, will lead to an odd hole in its NMRF for some valid potentials.
A frustrated cycle in an MRF can be broken at a vertex $v$ only if its enodes have the same setting for $v$, and by using an absorbing edge from $v$ to another vertex on the same cycle, which absorbs the connecting snode.

Despite these difficulties, we shall show that the break idea will allow us to extend significantly the range of signed MRF structures that may be handled efficiently with the NMRF method, for any valid potentials.

4.2 Main Result

In this Section, we present and prove our main result, see §2.1 for definitions.

Theorem 8. A binary pairwise MRF maps efficiently to a perfect pruned NMRF for any valid reparameterization iff each block of the MRF is almost balanced.

For comparison, the class of tractable models identified by Weller and Jebara (2013) is much smaller. WJ's class consists of blocks only of the following 3 types: (i) $B_R$ balanced subgraphs; (ii) $T_{m,n}$ frustrated multi-triangles on a common repulsive base $s-t$, where each additional variable is connected using either 2 repulsive or 2 attractive edges; and (iii) $U_n$ frustrated multi-triangles on a common attractive base, where each additional variable is connected by 1 attractive and 1 repulsive edge. Note that this admits frustrated cycles only of size 3, and even then only in very restricted configurations. See Figure 2 for examples.

Each of these 3 types are clearly almost balanced (the first is balanced; the multi-triangles may be rendered acyclic by deleting $s$ or $t$) but our result adds much richer models to the tractable range of the NMRF approach. For example, we add frustrated cycles of any size and indeed include frustrated blocks of any treewidth (as a specific example, consider a balanced $K_n$ structure together with one additional variable connected to all the others using any edges that lead to a frustrated cycle).

We shall first prove sufficiency then necessity of the condition. In order to test whether a derived NMRF is perfect, we use Theorem 1 hence must check for possible odd holes or antiholes. Since an odd antihole of size 5 is isomorphic to a hole of the same size, we need only check for odd holes, together with odd antiholes of size $\geq 7$ (see §2.1).

4.2.1 Sufficiency of the Condition

We provide a constructive proof that if a sub-MRF $M_1$ on $(V_1, E_1)$ of the MRF is almost balanced, with any valid potentials, then there is an efficient reparameterization that leads to a perfect pruned sub-NMRF $N_1$. Next we show that sub-NMRFs for each block may be pasted together to yield a perfect NMRF for the whole MRF.

Reparameterization Let $s$ be a vertex whose deletion renders the remainder of the sub-MRF balanced. Let $V'_1 = V_1 \backslash \{s\}$. Let $E'_1 \subseteq E_1$ be all edges not incident to $s$. Pick any vertex $v \in V'_1$ and a setting $a \in \{0, 1\}$. Let $\bar{a} = 1 - a$. First we shall reparameterize each edge in $E'_1$ to have one enode, as in [3, 1.1]4 Start with edges incident to $v$ and always choose the enode with setting $v = a$: if $(v, u) \in E'_1$ is attractive then use $(v = a, u = a)$, if repulsive then use $(v = a, u = \bar{a})$. Hence, the parity of the setting of an end vertex flips iff an edge is repulsive. Now iteratively extend outward to the rest of $V'_1$, as in breadth-first search. Whenever a vertex $w \in V'_1$ is first reached by an edge, mark it as either $a$ or $\bar{a}$ according to the setting of $w$ in the enode for the edge. Then for all other edges in $E'_1$ incident to $w$, use this same setting for $w$. Since the signed graph on $V'_1$ is balanced, this is guaranteed to yield an efficient consistent reparameterization across all edges in $E'_1$. After this is complete, the singleton potentials will have various values.

Now use absorbing edges between $s$ and all variables in $V'_1$, creating phantom edges where necessary. These will absorb all snodes: $\forall w \in V'_1$, the edge $s - w$ absorbs the singleton potential of $w$; the singleton potential of $s$ can be shared among all the absorbing edges in any way (for example, all could be absorbed into $s - v$).

See Figure 3 for an illustration of the construction for a frustrated cycle on 6 variables.

---

3 This algorithm yields a consistent marking for all variables in $V'_1$ unless two paths from $v$ to some $w \in V'_1$ contain a different number of repulsive edges, which happens iff joining the paths would yield a frustrated cycle, contradicting the assumption that the topology on $V'_1$ is balanced. It can be extended to check all edges in $E'_1$, with consistency iff the topology on $V'_1$ is balanced.
**Lemma 9.** With the reparameterization above, a perfect pruned sub-NMRF $N_1$ is obtained.

**Proof.** Let $u-v$ be an edge of $E'_1$ (i.e. an edge not incident to $s$). By construction, in $N_1$ there is one $u-v$ enode, say $(u = b, v = c)$ with $b, c \in \{0, 1\}$ and its neighbors are all either $s-u$ or $s-v$ enodes.

(i) **No odd holes.** Suppose an odd hole $H$ exists and consider candidate members. Consider first the enode from edge $u-v$. To form an odd hole, we must have $(s = a, u = b)$ and $(v = c, s = a)$ for some $a \in \{0, 1\}$ else we have a triangle. To continue past $(v = c, s = a)$, we cannot have anything with setting $s = \bar{a}$ else it would form a chord, so we must have $(v = c, s = a)$; but then there is no way to continue without forming a chord. Hence there is no such $u-v$ enode in $H$. The only remaining candidates are the enodes from absorbing edges. More than one absorbing edge must be involved to have sufficient nodes. To connect across different edges, there must be enodes with different settings for $s$ but there is no way to do this without at least one enode being adjacent to $\geq 3$ others, contradiction.

(ii) **No odd antiholes of size $\geq 7$.** Suppose an antihole $A$ of size $\geq 7$ exists. Consider if the enode from edge $u-v$ could be in $A$. It has 4 neighbors given by $(s = 0, u = b), (s = 1, u = \bar{b}), (s = 0, v = c), (s = 1, v = \bar{c})$, each of which is adjacent to 2 of the others. But to be in $A$, there must be 2 that are adjacent to 2 of the others, and 2 that are adjacent to 1 of the others, contradiction. The only remaining possible nodes of $A$ are enodes from absorbing edge clique groups. If one of them, say $p$, is in $A$, then there must be a set of 4 neighbors of $p$, say $q_1, q_2, q_3$ and $q_4$, all in order going around $A$, with $p$ and $q_1$ both adjacent to $q_3$ and $q_4$, and $q_3$ not adjacent to $q_4$. For $q_3$ and $q_4$ not to be adjacent, they must be in different clique groups with the same setting for $s$, say $s = a$. To be adjacent to both, $p$ and $q_1$ must each have setting $s = \bar{a}$. Now $q_2$ is not adjacent to $q_1$, so has setting $s = \bar{a}$ but then $q_2$ is adjacent to $q_3$, contradiction. \(\square\)

**Pasting Blocks Together** It remains to show that perfect sub-NMRFs for each block of the original MRF may be pasted together to yield a perfect NMRF for the whole MRF. We shall use the decomposition result of Theorem 3 but must take care since there it was required that when sub-NMRFs are pasted together on a variable, they must each have the same snodes for that overlapping variable, with at least one snode present, yet in the construction above we explicitly removed all snodes.

In fact, however, using methods similar to those of the proof of Lemma 9, we shall show that there are certain *phantom snodes*, i.e. snodes with 0 weight, which may be added to the construction above for any sub-NMRF without introducing any odd holes or antiholes. In particular, we may always add the following snodes: for the special vertex $s$, we may add $(s = 0)$ or $(s = 1)$; for all other $w \in V'_1$, we may add the snode with the mark for $w$, i.e. the same setting for $w$ as for all its incident enodes.

Checking for odd holes is straightforward. To check for antiholes, suppose antihole $A$ exists of size $\geq 7$, containing an added snode $(u = a)$. There must be 4 distinct nodes, say $q_1, q_2, q_3$ and $q_4$ consecutively in $A$, all adjacent to the snode. These 4 must be enodes with setting $u = \bar{a}$. $q_1$ and $q_4$ are adjacent so have different settings for some other variable, say $q_1 = (t = 0, u = \bar{a})$ and $q_4 = (t = 1, u = \bar{a})$. But now $q_2$ must be adjacent to $q_4$, which is not possible.

The ability to add these phantom snodes is sufficient to allow the pasting we need. We show this by induction on $k$, the number of blocks in the original MRF. The basis case $k = 1$ is trivial. Now suppose the result holds for $k$ blocks. Recall from §2.1 that the blocks form a block tree. Hence when considering $k+1$ blocks, we may examine the NMRF for a leaf block $N_{k+1}$, which is connected to the rest of the NMRF $N_R = \bigcup_{i=1}^k N_i$ at only one variable, say $x$. If $N_R$ has no $x$ snode already, then add one phantom snode, with setting as above (if $x$ is a special $s$ variable, then either $(x = 0)$ or $(x = 1)$ may be picked). Now in $N_{k+1}$, if the same snode may be added according to the above rules, then we are done. If not, then we are free to flip the reparameterization for $N_{k+1}$ by picking the opposite value for $a$ in the construction above, i.e. we flip the marks of all vertices $w \in V'_{k+1}$, thus allowing the appropriate snode to be added. This completes the proof of sufficiency.

4.2.2 Necessity of the Condition

In this Section, we assume that the condition is violated, i.e. that we have a 2-connected block $(V_1, E_1)$ which is not almost balanced, and show that this block will map to
Figure 4: A minimal example of a block that is not \textit{almost balanced}, hence will lead to an odd hole in its NMRF for some valid potentials. Solid blue (dashed red) edges are attractive (repulsive). All triangles are frustrated with an odd number of repulsive edges.

This completes the proof of our main result, Theorem 8.

5 \hspace{1em} \textbf{DISCUSSION}

We have provided an exact characterization of which binary pairwise MRFs can map to perfect pruned NMRFs for any valid potentials, allowing the full range of possible reparameterizations. This is a significant theoretical contribution, defining the power of the method. It extends the work of Weller and Jebara (2013), and greatly expands the range of models that are tractable with this approach.

Detecting if a given model satisfies our condition may be performed efficiently in time $O(|E||V|)$. First identify the block structure, which is a standard application of depth-first search (Hopcroft and Tarjan [1973]) and runs in $O(|E|)$. Next for each block, delete one vertex at a time, each time testing the remainder of the block to see if it is balanced. Testing a subgraph $(V_1, E_1)$ for balance takes time $O(|E_1|)$, see [Harary and Kabell 1980], the algorithm is similar to that described in footnote 4, checking all edges.

Note that each almost balanced block is tractable by other methods in isolation: once a variable has been identified such that deleting it renders the remainder of the block balanced, simply solve for a MAP configuration conditioned on each of the 2 settings of that variable then take the combination with higher score. The MAP problem on the balanced portion of the block may be solved efficiently with a variety of methods, including the standard linear programming relaxation [Johnson 2008, p. 119]. However, as far as we are aware, our approach is the first that is guaranteed to handle an MRF containing $\Omega(|V|)$ such blocks, including high treewidth frustrated structures, automatically in polynomial time, thereby extending the family of models that are tractable with any method.

\textbf{Acknowledgements}

The author thanks Tony Jebara and Maria Chudnovsky for many helpful discussions, and David Sontag for encouraging this research direction.
References


