Quasi-Best Match Graphs

Annachiara Korchmaros¹, David Schaller¹, David Schaller¹, Marc Hellmuth², and Peter F. Stadler¹,³-⁷

¹Bioinformatics Group, Department of Computer Science & Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelestraße 16-18, D-04107 Leipzig, Germany
annachiara@bioinf.uni-leipzig.de · s david@bioinf.uni-leipzig.de · studla@bioinf.uni-leipzig.de
²Department of Mathematics, Faculty of Science, Stockholm University, SE-10691 Stockholm, Sweden
marc.hellmuth@math.su.se
³German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions Dresden-Leipzig, Leipzig Research Center for Civilization Diseases, and Centre for Biotechnology and Biomedicine at Leipzig University at Universität Leipzig
⁴Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
⁵Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Wien, Austria
⁶Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia
⁷Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501, USA

Abstract

Quasi-best match graphs (qBMGs) are a hereditary class of directed, properly vertex-colored graphs. They arise naturally in mathematical phylogenetics as a generalization of best match graphs, which formalize the notion of evolutionary closest relatedness of genes (vertices) in multiple species (vertex colors). They are explained by rooted trees whose leaves correspond to vertices. In contrast to BMGs, qBMGs represent only best matches at a restricted phylogenetic distance. We provide characterizations of qBMGs that give rise to polynomial-time recognition algorithms and identify the BMGs as the qBMGs that are color-sink-free. Furthermore, two-colored qBMGs are characterized as directed graphs satisfying three simple local conditions, two of which have appeared previously, namely bi-transitivity in the sense of Das et al. (2021) and a hierarchy-like structure of out-neighborhoods, i.e., \( N(x) \cap N(y) \in \{N(x), N(y), \emptyset\} \) for any two vertices \( x \) and \( y \). Further results characterize qBMGs that can be explained by binary phylogenetic trees.

Keywords: Colored directed graphs; hierarchies; rooted trees; phylogenetic combinatorics; best matches

1 Introduction

Best match graphs (BMGs) appear in mathematical phylogenetics to formalize the notion of evolutionary closest relatives (homologs) of a gene in a different species [1]. Phylogenetic relatedness derives from a tree \( T \) that describes the evolutionary history of the “taxa” (e.g., genes or species) at the leaves. In our setting, each leaf of \( T \) represents an extant gene, and we assume to have additional knowledge of the species \( \sigma(x) \) in which each gene \( x \) is found. Given \( T \), a leaf \( x \) is more closely related to \( y \) than to \( z \), if and only if the lowest common ancestor \( \text{lca}(x, y) \) of \( x \) and \( y \) is a proper descendant of \( \text{lca}(x, z) \) in \( T \). Moreover, \( xy \) is a best match if there is no gene \( y' \) in the same species as \( y \), i.e., \( \sigma(y') = \sigma(y) \), which is more closely related to \( x \). Best match graphs (BMGs) collect the best match information for a family of related genes from different species.

BMGs have been studied in some detail [1, 2, 3, 4, 5] because of their close connection to the practically important problem of orthology detection. Two genes \( x \) and \( y \) from different species \( \sigma(x) \neq \sigma(y) \) are orthologs if their last common ancestor coincides with the divergence of the two species \( \sigma(x) \) and \( \sigma(y) \) in which they reside [6, 7]. In the absence of horizontal gene transfer, orthologs are reciprocal best matches, i.e., \( x \) is a best match of \( y \), and \( y \) is a best match of \( x \) [1]. Identifying orthologs is an important task in several areas of computational biology. In genome annotation, orthologs are of interest because they are expected to have analogous functions in different species. In contrast, homologous genes with a less direct relationship are usually expected to have similar but distinct functions [8, 9, 10]. In phylogenomics, a large collection of unrelated
groups of orthologous genes are used because their phylogenetic tree is nearly identical to the phylogenetic tree representing the evolutionary history of the underlying species [11, 12]. In structural biology, amino acids conserved among orthologs but differing between paralogs (genes that diverged from a duplication event) are used to identify sites that determine the specificity of protein interactions [13]. A diverse class of widely used orthology detection tools is based on the concept of reciprocal best matches [13]. A key result of [1] characterizes two-colored BMGs as simpler BMGs, particularly when gene families are considered at large phylogenetic scales.

A complication in determining best matches arises when genes are so distant that evolutionary relatedness, i.e., homology, cannot be determined unambiguously. This limit on evolutionary distance will, in general, depend both on the query gene and the target genome or, equivalently, a species α, because significant sequence similarities are easier to detect for large query sequences and in smaller target genomes [17]. Thus best matches y of x are only considered if lca(x, y) is not closer to the root of T than the “detection limit” u(x, α) for target genome σ(y) = α. The resulting quasi-best match graph (qBMG) is always a subgraph of the corresponding BMG, see Figure 1 for an illustrative example. BMGs have useful properties, such as the existence of a unique “least resolved tree” that explains a given BMG [1]. Furthermore, the least resolved tree of a BMG is displayed by the true phylogenetic tree from which the BMG was derived. Here, we will be concerned with whether such properties generalize to the more general qBMGs, which appear instead of the simpler BMGs, particularly when gene families are considered at large phylogenetic scales.

An interest in qBMGs can also be motivated by studying classes of (uncolored) digraphs defined in terms of constraints on neighborhoods. A key result of [1] characterizes two-colored BMGs as bipartite graphs whose out-neighborhoods satisfy four properties, of which two have also been the subject of investigations in quite different contexts in graph theory. Sink-free graphs have been studied in the context of certain partition problems [18], the construction of orientations on graphs [19], and in conjunction with a variety of algebraic structures [20]. Bitransitive graphs were introduced in [21] in the context of bitournaments. In [22, 23] the condition that G is sink-free was lifted. As we shall see below, this also leads to qBMGs in a very natural way.
This contribution is organized as follows: After introducing basic concepts and fixing the notation in Section 2, we provide a formal definition of qBMGs in Section 3 and derive their most basic properties. In particular, we show that the qBMGs are exactly the graphs that are obtained from BMGs by deleting, for an arbitrary given list of pairs \((x, s)\) of vertices \(x\) and colors \(s\), all edges to out-neighbors of \(x\) with color \(s\). In contrast to BMGs, the qBMGs form a hereditary graph class. In Section 4, a characterization of qBMGs with an arbitrary number of colors given in terms of consistency of informative and forbidden triples. This, in particular, allows us to identify the BMGs as the qBMGs for which each gene \(x\) has some out-neighbor \(y\) for every species \(\sigma(y)\) that is distinct from \(\sigma(x)\), i.e., as the “color-sink-free qBMGs”. Section 5 and 6 are then concerned with questions related to least-resolved and binary trees to explain qBMGs. Section 7 is devoted to the 2-colored case. We establish the equivalence between 2-qBMGs and the graphs studied in [22, 23], obtain a simple characterization that exposes the hierarchy-like structure of the out-neighborhoods, and derive an explicit construction for an explaining tree and truncation map. We also characterize 2-qBMGs in a setting where only the graph but not the coloring is given.

2 Notation and Preliminaries

2.1 Graphs, Trees, and Vertex Colorings

Vertex-Colored Graphs In this paper we consider simple, directed graphs \(G = (V,E)\) with vertex set \(V = V(G)\) and edge set \(E = E(G)\) unless \(G\) is explicitly specified as undirected. A graph is simple if it has neither parallel edges nor loops. A graph \(G'\) is a subgraph of \(G\), in symbols \(G' \subseteq G\) if \(V(G') \subseteq V(G)\) and \(E(G') \subseteq E(G)\). For \(W \subseteq V\) we write \(G[W]\) for the subgraph induced by \(W\), i.e., \(V(G[W]) = W\) and \(xy \in E(G[W])\) if and only if \(x, y \in W\) and \(xy \in E(G)\).

Analogously, we write \(N_G(x) := \{y \in V(G) \mid xy \in E(G)\}\) for the set of in-neighbors of a vertex \(x \in V(G)\). If the context is clear, we omit the subscript “\(G\)”. Since \(G\) is simple, we have \(x \notin N(x)\) and \(x \notin N^-(x)\). A source is a vertex \(x\) with in-degree zero, that is, \(N^-(x) = \emptyset\), while a vertex with out-degree zero, that is, \(N(x) = \emptyset\), is called sink. A graph \(G\) is sink-free if \(N(x) \neq \emptyset\) for all \(x \in V(G)\). Two vertices \(x, y\) are independent in \(G\) if \(\{x\} \cap N(y) = \emptyset\) and \(N(x) = \emptyset\), that is, neither \(xy \in E(G)\), nor \(yx \in E(G)\). Following the usual convention, we extend the notion for out-neighbourhood of a single vertex in \(V\) to sets of vertices by setting

\[
N : 2^V \to 2^V, \quad A \mapsto N(A) := \bigcup_{x \in A} N(x). \tag{1}
\]

For simplicity we write \(N(x)\) instead of \(N(\{x\})\). By construction, \(A \subseteq B\) implies \(N(A) \subseteq N(B)\), i.e., \(N\) is an isotonic map [24].

A vertex coloring is a map \(\sigma : V(G) \to S\), where \(S\) is a finite set with \(|S| > 1\). The coloring \(\sigma\) is proper if \(\sigma(x) = \sigma(y)\) implies \(xy, yx \notin E(G)\). For every color \(s \in S \setminus \{\sigma(x)\}\) we write \(N(x, s)\) and \(N^-(x, s)\) to denote the sets of out-neighbors and in-neighbors of \(x\) with color \(s\), respectively. The tuple \((G, \sigma)\) denotes a graph \(G\) together with a vertex coloring \(\sigma : V(G) \to S\). Moreover, we write \((G, \sigma)[W]\) for the subgraph of a colored graph induced by \(W \subseteq V\) and denote the restriction of the coloring \(\sigma\) to \(W\) by \(\sigma_W\). We say that \((G, \sigma)\) is color-sink-free if \(N(x, s) \neq \emptyset\) for all \(x \in V(G)\) and all \(s \in \sigma(V) \setminus \{\sigma(x)\}\).

Leaf-Colored Rooted Trees A tree is an undirected, connected graph that does not contain cycles. We also write \(uv\) for the undirected edges in a tree \(T\). Let \(T = (V,E)\) be a tree with root \(\rho\) and leaf set \(L := L(T) \subset V\). The set of inner vertices of \(T\) is \(V \setminus L\). In case \(V \neq L\), \(\rho\) is always assumed to be an inner vertex. An edge \(e = uv \in E\) is an inner edge of \(T\) if \(u\) and \(v\) are both inner vertices. Otherwise, it is an outer edge. A vertex \(u \in V\) is an ancestor of \(v \in V\) in \(T\) if \(u\) lies on the path from \(\rho\) to \(v\). In this case, we write \(v \preceq_T u\). Whenever we write \(uv \in E\) for an edge in \(T\) we assume \(v \prec_T u\). A vertex \(v\) is a child of \(u\) if \(uv \in E\). We write \(child_T(u)\) for the set of children of \(u\) in \(T\). If \(v\) is a child of \(u\), then \(u\) is the (unique) parent of \(v\); we write \(u = parent_T(v)\). Note that \(parent_T(v)\) is defined for all vertices \(v \in V \setminus \{\rho\}\). The subtree of \(T\) rooted at \(u\) and its leaf set are denoted by \(T(u)\) and \(L(T(u))\), respectively. A tree \(T\) is phylogenetic if \(u \in V(T)\) is either a leaf or \(u\) has \(|\text{child}(u)| \geq 2\) children. All trees appearing in this contribution are assumed to be phylogenetic. The least common ancestor \(\text{lca}_T(A)\) is the unique \(\preceq_T\)-smallest vertex that is an ancestor of all vertices in \(A \subseteq V\). For brevity, we write \(\text{lca}_T(x, y)\) instead of \(\text{lca}_T(\{x, y\})\). The subtree of \(T\) rooted in a vertex \(u \in V\) is denoted by \(T(u)\).
Supposing a vertex \( v \) of degree two refers to the operation of removing \( v \) and its two incident edges \( uw \) and \( vw \) and adding the edge \( uv \). Following [25], \( T_L \) denotes the restriction of \( T \) to a subset \( L' \subseteq L(T) \), i.e., \( T_L \) is obtained from the (unique) minimal subtree of \( T \) connecting all leaves in \( L' \) by subsequently suppressing all vertices with degree two except possibly the root \( \rho_{T_{L'}} = \text{lca}_T(L') \). Note that \( \rho_{T_{L'}} \) is not necessarily the original root \( \rho_T \). The contraction of an inner edge \( e = uw \in E(T) \) is an operation that produces a new tree \( T_e \) with vertex set \( V(T_e) = (V(T) \setminus \{u, v\}) \cup \{w\} \), where the new vertex \( w \) replaces both \( u \) and \( v \). Thus we have \( \text{child}_{T_e}(w) = \text{child}_T(u) \cup \text{child}_T(v) \) and \( \text{parent}_{T_e}(w) = \text{parent}_T(u) \). We say that \( T \) displays or is a refinement of a tree \( T' \), in symbols \( T' \leq T \), if \( T' \) is obtained from a restriction \( T_L \) of \( T \) after a (possibly empty) sequence of inner edge contractions. We write \( T' < T \) for \( T' \leq T \) and \( T' \neq T \).

A leaf coloring of a tree is a map \( \sigma : L \to S \) where \( S \) is a non-empty set of colors. We consider leaf-colored trees \((T, \sigma)\) and write \( \sigma(L') \colonequals \{ \sigma(v) \mid v \in L' \} \) for subsets \( L' \subseteq L \). We say that \((T', \sigma')\) is displayed by \((T, \sigma)\) if \( T' \leq T \) and \( \sigma(v) = \sigma'(v) \) for all \( v \in L(T') \).

Hierarchies Let \( X \) be a non-empty finite set. A system \( \mathcal{H} \subseteq 2^X \) of non-empty sets is a hierarchy on \( X \) if (i) \( U, V \in \mathcal{H} \) implies \( U \cap V \subseteq \{\emptyset, U, V\} \), (ii) \( \{x\} \in \mathcal{H} \) for all \( x \in X \), and (iii) \( X \in \mathcal{H} \). There is a 1-1 correspondence between hierarchies on \( L \) and phylogenetic rooted trees with leaf set \( L \) given by \( \mathcal{H}(T) \colonequals \{L(T(v)) \mid v \in V(T)\} \), see [27, Theorem 3.5.2]. The inverse map is obtained by the Hasse diagram \( T(\mathcal{H}) \) of \( \mathcal{H} \) with respect to set inclusion. A set system \( \mathcal{H} \subseteq 2^X \) is hierarchy-like if it satisfies condition (i). In this case, \( \mathcal{H} \) still corresponds to a forest. The roots of the constituent trees correspond to the inclusions-maximal elements of \( \mathcal{H} \), and the leaves correspond to the inclusions-minimal sets, which correspond to the singletons if and only if (ii) holds. For simplicity of notation, we will think of the leaves of \( T \) as elements \( v \in X \) rather than singleton sets \( \{v\} \in 2^X \).

2.2 Best Match Graphs

Definition 2.1. Let \((T, \sigma)\) be a leaf-colored tree. A leaf \( y \in L(T) \) is a best match of the leaf \( x \in L(T) \) if \( \sigma(x) \neq \sigma(y) \) and \( \text{lca}_T(x, y) \preceq_T \text{lca}_T(x, y') \) holds for all leaves \( y' \) of color \( \sigma(y') = \sigma(y) \).

Given \((T, \sigma)\), the best match graph (BMG) associated to \((T, \sigma)\) is the digraph \( BMG(T, \sigma) = (V, E) \) with vertex set \( V = L(T) \), vertex coloring \( \sigma \), and \((xy) \in E\) if and only if \( y \) is a best match of \( x \) with respect to \((T, \sigma)\) [1].

Definition 2.2. An arbitrary vertex-colored digraph \((G, \sigma)\) is a best match graph (BMG) if there exists a leaf-colored tree \((T, \sigma)\) such that \((G, \sigma) = BMG(T, \sigma)\). In this case, we say that \((T, \sigma)\) explains \((G, \sigma)\).

A BMG \((G, \sigma)\) with vertex set \( L \) is an \( \ell \)-BMG if \( |\sigma(L)| = \ell \). Figure 1 in the introduction shows an example of a 3-BMG on 6 vertices, [22, Figure 7] gives an example of a 2-BMG on 10 vertices and [22, Figure 2(a)] gives an example of a 2-BMG on 11 vertices. Since BMGs are properly colored, every 2-BMG is bipartite. Furthermore, the subgraph of a BMG induced by the subset \( L_{rs} \colonequals \{x \in L \mid \sigma(x) \in \{r, s\} \} \) of \( L \), i.e., the vertices with two colors, is always a 2-BMG [1].

Now, we briefly review characterizations of BMGs that will be of relevance for this contribution.

Proposition 2.1 ([1, Theorem 4] and [3]). A properly two-colored graph \((G, \sigma)\) is a 2-BMG if and only if it is (i) sink-free and (ii) the out-neighborhoods satisfy the following three conditions:

\( (N1) \ x \notin N(y) \) and \( y \notin N(x) \) implies \( N(x) \cap N(N(y)) = N(y) \cap N(N(x)) = \emptyset \).

\( (N2) \ N(N(N(x))) \subseteq N(x) \).
If \( x \notin N(N(y)), y \notin N(N(x)), \) and \( N(x) \cap N(y) \neq \emptyset \) then we have \( N^{-}(x) = N^{-}(y) \) and one of the inclusions \( N(x) \subseteq N(y) \) or \( N(y) \subseteq N(x) \).

Properties (N1) and (N2) may be rephrased as in [22, 23]:

(N1) If \( x \) and \( y \) are independent vertices (or \( x = y \)), then there exist no vertices \( w \) and \( t \) such that \( xt, yw, tw \in E(G) \). (N2) If there are vertices \( x_1, x_2, y_1, y_2 \in L \) with \( x_1y_1, y_1x_2, x_2y_2 \in E(G) \), then \( x_1y_2 \in E(G) \).

These two properties imply that some directed edges must be present in a 2-BMG; see [22, Figure 6] for the edges required by (N2) and the counterpart of (N1), respectively. Property (N2) was introduced as bi-transitivity in [21] and investigated in relation to topological orderings in [22, 23]. It suffices to require (N1) for \( x \neq y \) because, for \( x = y \), \( N(x) \cap N(N(x)) = \emptyset \) is an immediate consequence of the fact that \( G \) is assumed to be bipartite. Nevertheless, we define (N1) here for all \( x, y \in V(G) \). We shall see below in Theorem 7.9 that this choice of the axiom allows us to remove the explicit requirement that \( G \) is bipartite.

Properties (N1), (N2) and (N3') were translated into a system of forbidden induced subgraphs of a 2-BMG; see [3]. In particular, the first part of (N3'), i.e., the implication \( x \notin N(N(y)) \), \( y \notin N(N(x)) \) and \( N(x) \cap N(y) \neq \emptyset \) \( \implies \) \( N^{-}(x) = N^{-}(y) \), is already covered by (N1). This makes it possible to replace (N3') by the simpler condition

(N3) If \( N(x) \cap N(y) \neq \emptyset \) then \( N(x) \subseteq N(y) \) or \( N(y) \subseteq N(x) \).

For completeness, we prove directly that (N3') may be replaced by (N3).

**Lemma 2.2.** A properly two-colored graph \((G, \sigma)\) satisfies (N1), (N2), and (N3') if and only if it satisfies (N1), (N2), and (N3).

**Proof.** Suppose \((G, \sigma)\) satisfies (N1), (N2), and (N3') and assume, for contradiction, that (N3) is violated, i.e., that there are vertices \( x, y \in V(G) \) with \( N(x) \cap N(y) \neq \emptyset \) such that neither \( N(x) \subseteq N(y) \) nor \( N(y) \subseteq N(x) \) is true. Hence, \( x \neq y \). Thus, there are distinct vertices \( u \in N(x) \setminus N(y), v \in N(y) \setminus N(x) \). By assumption, there exists \( z \in N(x) \cap N(y) \). Since \((G, \sigma)\) is properly colored, \( \sigma(x) = \sigma(y) \neq \sigma(z) = \sigma(u) = \sigma(v), \) and \( x, y, z, u, \) and \( v \) are pairwise distinct. If \( x \notin N(N(y)) \) and \( y \notin N(N(x)) \), then we immediately obtain a contradiction to (N3'). Hence, assume that there exists \( w \in V(G) \) such that \( xw, yw \in E(G) \), i.e., \( y \in N(N(x)) \). Since \( w \notin N(x), v \notin N(x) \) and \((G, \sigma)\) is properly colored, we have \( w \notin \{x, y, v\} \). However, \( xw, yw, yv \in E(G) \) together with (N2) imply that \( xv \in E(G) \) and thus, \( v \in N(x) \); a contradiction. An analogous contradiction is obtained for that case \( x \in N(N(y)) \).

Now suppose \((G, \sigma)\) satisfies (N1), (N2), and (N3) and assume, for contradiction, that \((G, \sigma)\) does not satisfy (N3'). Hence, there are two vertices \( x, y \in V(G) \) with \( x \notin N(N(y)), y \notin N(N(x)) \), and \( N(x) \cap N(y) \neq \emptyset \) such that (i) neither of the inclusions \( N(x) \subseteq N(y) \) and \( N(y) \subseteq N(x) \) holds, or (ii) \( N^{-}(x) \neq N^{-}(y) \). Case (i) immediately contradicts (N3). In case (ii), i.e., \( N^{-}(x) \neq N^{-}(y) \), we have \( x \neq y \) and we can assume that there exists \( w \in V(G) \) with \( wx \in E(G) \) and \( wy \notin E(G) \).

We also have \( yw \notin E(G) \) since otherwise \( yw, wx \in E(G) \) would contradict \( x \notin N(N(y)) \). If \( w = y \) then \( x \) and \( y \) are adjacent, and thus \( N(x) \cap N(y) = \emptyset \), contradicting our assumption. Hence, \( y \) and \( w \) are independent in \( G \). From \( N(x) \cap N(y) \neq \emptyset \), it follows that \( x \) and \( y \) have a common out-neighbor \( z \). Since \((G, \sigma)\) is properly colored and \( yz \in E(G) \) but \( yw \notin E(G) \), we can infer that that the four vertices \( x, y, z, \) and \( w \) are pairwise distinct. In summary, \((G, \sigma)\) contains the two independent vertices \( w \) and \( y \) and the two vertices \( x \) and \( z \) with \( wx, yz, xz \in E(G) \); this contradicts (N1). Hence, we conclude that \((G, \sigma)\) also satisfies (N3').

We remark that (N1) together with (N3) implies (N3'), but the converse is not true in general unless (N2) is assumed. We note that Condition (N3) may be written also as \( N(x) \cap N(y) \in \{\emptyset, N(x), N(y)\} \), i.e. the out-neighborhoods form a hierarchy-like set system on \( V(G) \). It is not a hierarchy because \( N(x) \neq V(G) \) for all \( x \in V(G) \) since \((G, \sigma)\) is bipartite.

In [3, 4, 5], BMGs have been characterized in terms of certain sets of rooted triples that easily can be constructed for every given vertex-colored digraph.

**Definition 2.3.** Let \((G, \sigma)\) be a vertex-colored digraph. Then the set of informative triples is
\[
\mathcal{R}(G, \sigma) := \{ab|b|' : \sigma(a) \neq \sigma(b) = \sigma(b'), ab \in E(G), \text{ and } ab' \notin E(G)\},
\]
and the set of forbidden triples is
\[
\mathcal{F}(G, \sigma) := \{ab|b|' : \sigma(a) \neq \sigma(b) = \sigma(b'), b \neq b', \text{ and } ab, ab' \in E(G)\}.
\]
For vertex-colored digraphs that are associated with binary trees, we will furthermore need
\[ R^B(G, \sigma) := R(G, \sigma) \cup \{ ab' | a : ab \in \mathcal{F}(G, \sigma) \} \]

Note that \( a, b, b' \) must be pairwise distinct whenever \( ab, b' \in R(G, \sigma) \), \( ab, b' \in \mathcal{F}(G, \sigma) \), or \( ab' | a \in R^B(G, \sigma) \). Moreover, the forbidden triples always come in pairs, i.e., \( ab' | b \in \mathcal{F}(G, \sigma) \) implies \( ab' | b \in \mathcal{F}(G, \sigma) \). The triples in each of the three sets are 2-colored because they refer to \((G, \sigma)[L_{\sigma(a)\sigma(b)}] \), i.e., an induced 2-BMG.

**Proposition 2.3** ([3, Lemma 3.4 and Theorem 3.5]). A properly colored graph \((G, \sigma)\) is a BMG if and only if (i) it is color-sink-free and (ii) \((R(G, \sigma), \mathcal{F}(G, \sigma))\) is consistent. In this case, every tree \((T, \sigma)\) with leaf set \(V(T)\) that agrees with \((R(G, \sigma), \mathcal{F}(G, \sigma))\) explains \((G, \sigma)\).

Binary-explainable BMGs are the vertex-colored digraphs explained by binary trees.

**Proposition 2.4** ([4, Theorem 8]). A properly colored graph \((G, \sigma)\) is a binary-explainable BMG if and only if (i) it is color-sink-free and (ii) \(R^B := R^B(G, \sigma)\) is consistent. In this case, the BMG \((G, \sigma)\) is explained by every refinement of the leaf-colored tree \((\text{BUILD}(R^B, V(G)), \sigma)\).

### 3 Basic Properties of Quasi-Best Match Graphs

We generalize the notion of best match graphs to quasi-best match graphs in order to address the limitations of methods to detect significant sequence similarities. To this end we introduce the notion of a “relative detection limit” beyond which best matches remain undetected and thus are assumed to be absent in the graphs of interest.

**Definition 3.1.** Let \((T, \sigma)\) be a leaf-colored tree with vertex set \(V\), leaf set \(L\) and \(\sigma(L) \subseteq S\). A truncation map \(u_T : L \times S \to V\) assigns to every leaf \(x \in L\) and color \(s \in S\) a vertex of \(T\) such that \(u_T(x, s)\) lies along the unique path from \(x\) to \(x\) and \(u_T(x, \sigma(x)) = x\). A leaf \(y \in L\) with color \(\sigma(y)\) is a quasi-best match for \(x \in L\) (with respect to \((T, \sigma)\) and \(u_T\)) if both conditions (i) and (ii) are satisfied:

(i) \(y\) is a best match of \(x\) in \((T, \sigma)\).
(ii) \(\lca_T(x, y) \preceq u_T(x, \sigma(y))\).

The vertex-colored digraph \(q\text{BMG}(T, \sigma, u_T)\) on the vertex set \(L\) whose edges are defined by the quasi-best matches is the quasi-best match graph \((q\text{BMG})\) of \((T, \sigma, u_T)\).

Note that, since \(xy\) is an edge only if \(\sigma(x) \neq \sigma(y)\), \(q\text{BMG}(T, \sigma, u_T)\) is always properly colored. Furthermore, if \(u_T(x, s) = x\), then \(x\) has no out-neighbors of color \(s \in S\) in \(q\text{BMG}(T, \sigma, u_T)\), but the converse is not true in general.

**Definition 3.2.** A vertex-colored digraph \((G, \sigma)\) with vertex set \(V(G) = L\) is a (colored) quasi-best match graph \((q\text{BMG})\) if there is a leaf-colored tree \((T, \sigma)\) and a truncation map \(u_T\) on \((T, \sigma)\) such that \((G, \sigma) = q\text{BMG}(T, \sigma, u_T)\).

In analogy to BMGs, a qBMG \((G, \sigma)\) with vertex set \(L\) is an \(\ell\)-qBMG if \(|\sigma(L)| = \ell\). The trivial truncation map is defined as \(u_T^\ell(x, s) := \rho_T\) for all \(x \in L(T)\) and all \(s \in S\setminus\{\sigma(x)\}\). In this case, condition (ii) becomes void so that the trivial truncation map has no effect. We state this fact as a corollary.

**Corollary 3.1.** Let \((T, \sigma)\) be a leaf-colored tree and \(u_T^\ell\) be the trivial truncation map. Then \(\text{BMG}(T, \sigma) = q\text{BMG}(T, \sigma, u_T^\ell)\), that is, \(xy\) is a quasi-best match with respect to \((T, \sigma, u_T^\ell)\) if and only if it is a best match with respect to \((T, \sigma)\). In particular, every BMG is a qBMG.

**Observation 3.1.** If \(u_1(x, r) \preceq_T u_2(x, r)\) for all \(x \in L\) and \(r \in S\), then \(q\text{BMG}(T, \sigma, u_1)\) is a subgraph of \(q\text{BMG}(T, \sigma, u_2)\).

This, together with Corollary 3.1, shows that every qBMG is a subgraph of a BMG.

**Lemma 3.2.** Let \((\hat{G}, \hat{\sigma}) := \text{BMG}(T, \sigma)\) be a BMG. Let \((G, \sigma) := q\text{BMG}(T, \sigma, u_T)\) be the qBMG arising from \((T, \sigma)\) with a truncation map \(u_T\). Then, for all \(x \in L\) and all colors \(s \in \sigma(L)\), we have either \(N_G(x, s) = N_G(x, s)\) or \(N_G(x, s) = \emptyset\).

**Proof.** If \(N_G(x, s) = \emptyset\), we are done. Hence, assume that \(N_G(x, s) \neq \emptyset\). From Definition 3.1(i), every \(y \in N_G(x, s)\) is a best match of \(x\), thus \(N_G(x, s) \subseteq N_G(x, s)\). Moreover, there is a vertex
$u \in V(T)$ such that $u = \text{lca}_T(x, y)$ for all $y \in N_G(x, s)$ and $u \preceq_T \text{lca}(x, y')$ for all $y' \notin N_G(x, s)$ with $\sigma(y) = s$. In addition, we have $u \preceq_T u_T(x, s)$ by Definition 3.1(ii). The latter properties are, in particular, valid for every $\tilde{y} \in N_G(x, s)$, which implies that $N_{\tilde{G}}(x, s) \subseteq N_G(x, s)$ and therefore, $N_G(x, s) = N_{\tilde{G}}(x, s)$. 

A leaf-colored tree $(T, \sigma)$ explaining a BMG $(G, \sigma)$ is least resolved if there is no tree $T' < T$ such that $(G, \sigma) = \text{BMG}(T', \sigma)$. By [1, Theorem 8], every BMG has a unique least resolved tree (LRT) $\hat{T}$. This result can also be used in the study of qBMGs, since, by the following lemma, every qBMG may be assumed to arise from the LRT of some BMG by means of a truncation map.

**Lemma 3.3.** Let $(T, \sigma)$ be a leaf-labeled tree with truncation map $u$ and let $(\hat{T}, \sigma)$ be the LRT of BMG$(T, \sigma)$. Then there is a truncation map $\hat{u}$ on $\hat{T}$ such that qBMG$(T, \sigma, u) = \text{qBMG}(\hat{T}, \sigma, \hat{u})$.

**Proof.** Put $\hat{G}, \sigma := \text{BMG}(T, \sigma)$ and $(G, \sigma) = \text{qBMG}(T, \sigma, u)$. Then, by definition, $(\hat{G}, \sigma) = \text{BMG}(\hat{T}, \sigma)$. Consider the truncation function $\hat{u}$ for $\hat{T}$ defined by $\hat{u}(x, s) = \rho_T$ if $N_G(x, s) \neq \emptyset$ and $\hat{u}(x, s) = x$ if $N_G(x, s) = \emptyset$. Now, let $(\hat{G}, \sigma) := \text{qBMG}(\hat{T}, \sigma, \hat{u})$. By construction, $N_{\hat{G}}(x, s) = N_G(x, s)$ if $N_G(x, s) \neq \emptyset$ and $N_G(x, s) = N_G(x, s) = \emptyset$ otherwise. Since a digraph is uniquely determined by its out-neighbors, we have $\hat{G} = G$. 

It should be noticed that Lemma 3.3 does not imply $\hat{T}$ to be a least resolved explanation for the qBMG. We will study some properties of least resolved trees of qBMGs in Section 5.

**Theorem 3.4.** A vertex-colored graph $(G, \sigma)$ with vertex set $L$ is a qBMG if and only if there is a BMG $(\hat{G}, \sigma)$ such that, for all $x \in L$ and $s \in S = \sigma(L)$, either $N_{\hat{G}}(x, s) = N_G(x, s)$ or $N_{\hat{G}}(x, s) = \emptyset$ holds.

**Proof.** Lemma 3.2 implies the only-if-direction. For the if-direction, suppose that $(\hat{G}, \sigma)$ is a BMG that is explained by $(T, \sigma)$. Let $G, \sigma$ be a vertex-colored graph such that either $N_G(x, s) = N_{\hat{G}}(x, s)$ or $N_G(x, s) = \emptyset$. Then $u$ with $u(x, s) = \rho_T$ if $N_G(x, s) = N_{\hat{G}}(x, s)$ and $u(x, s) = x$ if $N_G(x, s) = \emptyset$ is a truncation map for $T$. By construction, $(G, \sigma) = \text{qBMG}(T, \sigma, u)$ and thus, it is a qBMG.

Theorem 3.4 gives a motivation for the following definition.

**Definition 3.3.** Let $(G, \sigma)$ be a qBMG with vertex set $L$. A BMG $(\hat{G}, \sigma)$ with vertex set $L$ is associated with $(G, \sigma)$ if for all $x \in L$ and $s \in \sigma(L)$ either $N_{\hat{G}}(x, s) = N_G(x, s)$ or $N_{\hat{G}}(x, s) = \emptyset$ holds.

A necessary condition for BMGs $(\hat{G}, \sigma)$ to be associated with a qBMG $(G, \sigma)$ is that $(G, \sigma)$ is a subgraph of $(\hat{G}, \sigma)$. In this case, $(G, \sigma)$ can be obtained from $(\hat{G}, \sigma)$ by removing all out-neighbors of specific colors for specific vertices. This condition is not sufficient, however, as the example in Figure 2 shows.

By Theorem 3.4, every qBMG has at least one associated BMG. It is, therefore, possible to obtain all qBMGs from the BMGs by removing all out-edges with a subset of colors for a subset of the vertices.

In order to show that qBMGs form a hereditary graph class, we consider the deletion of a single vertex and show that both $(G, \sigma)$ and the explaining tree $(T, \sigma, u)$ are quite “well-behaved” when a single vertex is deleted.

**Lemma 3.5.** Let $(G, \sigma) = \text{qBMG}(T, \sigma, u)$ be a qBMG and $v \in L$. Then there exists a truncation map $u'$ such that $(G, \sigma)[L'] = \text{qBMG}(T_L, \sigma_L, u')$ where $L' = L \setminus \{v\}$. In particular, the digraph $(G, \sigma)[L']$ is still a qBMG.

**Proof.** Consider the phylogenetic tree $T' := T_L$ obtained by deleting the leaf $v$ and suppressing $w = \text{parent}_T(v)$ if $v$ has a single sibling $v^*$, i.e., $\text{child}_T(w) = \{v, v^*\}$, and $\sigma'(x) = \sigma(x)$ for $x \in L'$. Note that if $w$ is the root of $T$, then $v^*$ is the new root of $T'$. Depending on whether $w$ is suppressed or not, there is a 1-1 correspondence between $V(T) \setminus \{v\}$ and $V(T')$ or $V(T) \setminus \{v, w\}$ and $V(T')$, respectively, such that $L(T') = L(T) \setminus \{v\}$. In our notation, we identify corresponding vertices of $T$ and $T'$. Moreover, we will use the fact that the ancestor order $\preceq_{T'}$ is equal to $\preceq_T$ restricted to $V(T')$. Given a truncation map $u$ on $T$, we define its restriction $u'$ on $T'$ for all $x \in L'$ and $s \in S$ by

$$u'(x, s) = \begin{cases} v^* & \text{if } \text{child}_T(w) = \{v, v^*\} \text{ and thus, } u(x, s) \notin V(T') \\ u(x, s) & \text{else if } s \neq \sigma(v) \text{ or } N(x, s) \setminus \{v\} \neq \emptyset \\ x & \text{otherwise} \end{cases}$$

(2)
From now on we write \( \sigma' := \sigma|_{L'} \). The restricted leaf-colored tree \((T', \sigma')\) together with the truncation map \( u' \) explains \((G', \sigma') = qBMG(T', \sigma', u')\). We next show that \((G', \sigma')\) is precisely the induced subgraph \((G, \sigma)[L']\). Denote the out-neighborhoods of \( G \) and \( G' \) by \( N(x, s) := N_G(x, s) \) and \( N'(x,s) := N_{G'}(x, s) \), respectively. By construction, \( u'(x, \sigma(x)) = u(x, \sigma(x)) = x \) and thus \( N(x, \sigma(x)) = N'(x, \sigma(x)) = \emptyset \) for all \( x \in L' \). We proceed by comparing \( N(x, s) \) and \( N'(x, s) \) for \( x \in L' \) and \( s \in S \setminus \{ \sigma(x) \} \). To this end, note that for each \( y \in L' \), we have \( \text{lca}_{T'}(x,y) = \text{lca}_{T}(x,y) \) since otherwise, we must have \( \text{lca}_{T'}(x,y) = v \notin V(T') \), which would imply \( y = v \); a contradiction.

Suppose first that \( v \notin N(x, s) \). If \( y \in N(x, s) \), then \( \text{lca}_{T'}(x,y) \preceq_{T} u(x, s) \) by Definition 3.1(ii), and there is no \( y' \in L \) such that \( \sigma(y) = \sigma(y') \) and \( \text{lca}_{T'}(x,y') \prec_{T'} \text{lca}_{T}(x,y) \) since \( y \) must be a best match of \( x \) by Definition 3.1(i)). The latter remains true upon restriction of \( T \) to \( L' \). Because of Equation (2) and \( y \in N(x, s) \), we have \( u'(x, s) = u(x, s) \) or \( u'(x, s) = v^* \). In the former case, we obtain \( \text{lca}_{T'}(x,y) = \text{lca}_T(x,y) \preceq_T u(x, s) = u'(x, s) \), implying \( y \in N'(x, s) \). In the latter case, i.e., if \( u'(x, s) = v^* \), we have \( u(x, s) = w \neq \text{lca}_T(x,y) \). Thus, we obtain \( \text{lca}_{T'}(x,y) \preceq_T v^* \) from \( \text{lca}_T(x,y) \preceq_T u(x, s) \) and we observe that \( \text{lca}_T(x,y) \preceq_T v \) is ruled out by \( y \in N_1(x) \). Therefore, \( y \in N'(x, s) \). If \( y \notin N(x, s) \) for some \( y \) of color \( s \), then \( u(x, s) \prec_T \text{lca}_T(x,y) \) or there is some \( y' \in L \) of color \( s' \) such that \( \text{lca}_T(x,y') \prec_T \text{lca}_T(x,y) \) and in particular \( y' \in N(x, s) \). If \( u(x, s) \prec_T \text{lca}_T(x,y) \), then by construction \( u'(x, s) \prec_T \text{lca}_T(x,y) \). In the latter case, \( y' \in N(x, s) \) implies \( y' \neq v \), and thus, \( y' \in L' \) and \( \text{lca}_{T'}(x,y') \prec_T \text{lca}_T(x,y) \). In both cases, we, therefore, have \( y \notin N'(x, s) \). In summary, \( v \notin N(x, s) \) implies \( N(x, s) = N'(x, s) \) for all \( x \in L' \).

Since \( v \in N(x, s) \) and hence \( s = \sigma(v) \), it is useful to distinguish two cases: Case (i): \( N(x, s) \setminus \{v\} \neq \emptyset \). Let \( q := \text{child}_{T}(x,v) \). If \( q \notin \text{child}_{T}(x,v) \) or \( |\text{child}_{T}(x,v)| > 2 \), then \( q \notin V(T') \) and hence \( \text{lca}_{T'}(x,y) = \text{lca}_T(x,y) \) for all \( y \in L' \) and \( u'(x, s) = u(x, s) \). This implies \( N'(x, s) = N(x, s) \setminus \{v\} \). If \( \text{child}_{T}(x,v) = \{v, v^*\} \) then \( \text{lca}_T(x,y) = v \) implies \( y = v \). Thus \( v \in N(x, s) \) implies \( N(x, s) = \{v\} \), contradicting \( N(x, s) \setminus \{v\} \neq \emptyset \). Case (ii): If \( N(x, s) = \{v\} \), then \( N'(x, s) = \emptyset \) because \( u'(x, s) = x \). In summary, \( N'(x, s) = N(x, s) \setminus \{v\} \) holds for all \( x \in L' \) and \( s \in S \). This implies that \((G', \sigma')\) is the subgraph of \((G, \sigma)\) induced by \( L' = L \setminus \{v\} \).

Since the removal of any set of vertices can be carried out step by step by deleting one vertex at a time, it turns out that the induced subgraph \((G, \sigma)[L']\) has an explanation in terms of the restriction \( T_{L'} \) of \( T \) to the leaf set \( L' \subseteq L \). Therefore, the following result is obtained.

**Corollary 3.6.** Every induced subgraph of a qBMG is a qBMG, i.e., the qBMGs form a hereditary graph class.

**Observation 3.2.** The disjoint union \((G, \sigma) = \bigcup_{i=1}^{m} (G_i, \sigma_i)\) is a qBMG if and only if each of \((G_i, \sigma_i)\) is a qBMG.

**Proof.** Since each connected component of \( G \) is an induced subgraph the only-if-direction is trivial.
Conversely, consider explanations \((T_i, \sigma_i, u_i)\) for the disjoint graphs \((G_i,\sigma_i)\). Let \(T\) be the tree whose vertices are those of \(T_i\) together with a new root \(\rho\) added as the common parent of the roots \(\rho_i\) of \(T_i\). It has the leave set \(L(T) = \bigcup L(T_i)\). Define \(\sigma(x) = \sigma_i(x)\) and \(\nu_T(x,r) = \nu_{T_i}(x,r)\) for \(x \in L_i\). There are no edges joining vertices from different components since \(u(x,r) \preceq_{T_i} \rho_{T_i} \preceq_{T} \rho\) while for each connected component, we have \(G[L(T_i)] = G_i\). Thus \((G,\sigma) = \text{qBMG}(T,\sigma,u)\). 

The corresponding result for BMGs requires the additional condition that all connected components use the same set of colors, i.e. \(\sigma(V(G_i)) = \sigma(V(G_j))\), see [1].

4 Recognition of qBMGs and Rooted Triangles

The recognition problem for BMGs, and more generally for qBMGs is of practical interest as part of workflows for orthology detection. The graph \((G,\sigma)\) recording for each query gene the most similar sequences in each target genome can be computed efficiently [28] and serves as an empirical approximation for a BMG or qBMG. The empirical estimate \((G,\sigma)\) may contain false positive and false negative edges; thus, it is, in general, neither a BMG nor a qBMG. Solving the recognition problem is the first key step toward the more difficult problem of identifying potential errors in the input data. For BMGs, the recognition problem has been solved in [3], see Proposition 2.3 in terms of two conditions: (i) the absence of color-sinks and (ii) consistency of a collection of informative and forbidden triples.

We have already seen that qBMGs are not color-sink-free in general, i.e., they may violate condition (i), see Figure 1. In contrast, triple consistency, i.e., condition (ii), remains valid as shown by the following lemma, which generalizes Lemmas 2.11 and 3.2 of [3] from BMGs to qBMGs:

**Lemma 4.1.** Let \((G,\sigma)\) be a qBMG explained by \((T,\sigma,u)\). Then \(T\) displays all triples in \(\mathcal{R}(G,\sigma)\) but none of the triples in \(\mathcal{F}(G,\sigma)\). In particular, \((\mathcal{R}(G,\sigma),\mathcal{F}(G,\sigma))\) is consistent.

**Proof.** First, suppose that \(ab'b' \in \mathcal{R}(G,\sigma)\), i.e., \(ab \in E(G)\) and \(ab' \notin E(G)\). Since \(ab \in E(G)\), there is no \(b''\) of color \(\sigma(b'') = \sigma(b')(= \sigma(b'))\) such that \(\text{lca}_T(a,b'') \prec_T \text{lca}_T(a,b)\) and that \(\text{lca}_T(a,b) \preceq_T u(a,\sigma(b))\). In particular, \(\text{lca}_T(a,b') \preceq_T \text{lca}_T(a,b)\). On the other hand, if \(\text{lca}_T(a,b') = \text{lca}_T(a,b) \preceq_T u(a,\sigma(b))\), then \(ab' \in E(G)\) since \((T,\sigma,u)\) explains the qBMG \((G,\sigma)\). Therefore, \(\text{lca}_T(a,b) \prec_T \text{lca}_T(a,b')\) is the only remaining possibility and \(T\) displays the triple \(ab'b'\).

Now suppose that \(ab'b' \in \mathcal{F}(G,\sigma)\), i.e., \(ab \in E(G)\) and \(ab' \in E(G)\). By similar arguments as above, this implies \(\text{lca}_T(a,b') \preceq_T \text{lca}_T(a,b')\) and \(\text{lca}_T(a,b') \preceq_T \text{lca}_T(a,b)\), respectively. Therefore, \(\text{lca}_T(a,b) = \text{lca}_T(a,b')\) and hence \(T\) does not display the triple \(ab'b'\). Similarly, \(T\) does not display \(ab'\) either.

**Theorem 4.2.** A properly-colored digraph \((G,\sigma)\) with vertex set \(L\) is a qBMG if and only if \((\mathcal{R}(G,\sigma),\mathcal{F}(G,\sigma))\) is consistent. In this case, for every tree \(T\) on \(L\) that agrees with \((\mathcal{R}(G,\sigma),\mathcal{F}(G,\sigma))\), there is a truncation map \(\nu\) such that \((T,\sigma,\nu)\) explains \((G,\sigma)\).

**Proof.** If \((G,\sigma)\) is a qBMG, then consistency of \((\mathcal{R}(G,\sigma),\mathcal{F}(G,\sigma))\) follows from Lemma 4.1. Now suppose that \((\mathcal{R}(G,\sigma),\mathcal{F}(G,\sigma))\) is consistent. From \(L_{\mathcal{R}(G,\sigma),\mathcal{F}(G,\sigma)} \subseteq L\), at least one tree with leaf set \(L\) displays all triples in \(\mathcal{R}(G,\sigma)\) and none of the triples in \(\mathcal{F}(G,\sigma)\). Let \((T,\sigma)\) be any tree with this property. Set \(S = \sigma(L)\) and consider the truncation map \(\nu: L \times S \to V(T)\) that is given by \(u(x,s) = x\) if \(N_G(x,s) = \emptyset\), and \(u(x,s) = \rho_T\) otherwise, for all \(s \in S\) and \(x \in L\). Note that \(u\) is a well-defined truncation map for \((T,\sigma)\), since \((G,\sigma)\) is properly colored and hence \(u(x,\sigma(x)) = x\) for all \(x \in L\).

Let \((G,\sigma) = \text{qBMG}(T,\sigma,u)\) be the qBMG explained by \((T,\sigma,u)\), \(x \in L\) and \(s \in S\). It remains to show that \((G,\sigma) = (G,\sigma)\). For \(N_G(x,s) = \emptyset\), we have set \(u(x,s) = x\) and thus \(N_G(x,s) = \emptyset\) as a consequence of condition (ii) in Definition 3.1. Now suppose \(N_G(x,s) \neq \emptyset\). This is possible only if \(\sigma(x) \neq s\). By construction, we have \(u(x,s) = \rho_T\), and therefore, \(y \in N_G(x,s)\) with \(\sigma(y) = s\) if and only if \(y\) is a best match of \(x\) in \((T,\sigma)\). Consider a vertex \(y \in N_G(x,s)\) and assume, for contradiction, that \(y\) is not a best match of \(x\) in \((T,\sigma)\), i.e., there is some \(y'\) of color \(s\) such that \(\text{lca}_T(x,y') \prec_T \text{lca}_T(x,y)\). If \(y' \in N_G(x,s)\), then \(xy'y' \in \mathcal{R}(G,\sigma)\). If \(y' \notin N_G(x,s)\), then \(xy'y' \notin \mathcal{R}(G,\sigma)\). In both cases, the agreement of \(T\) with \((G,\sigma),\mathcal{F}(G,\sigma))\) contradicts that \(\text{lca}_T(x,y') \prec_T \text{lca}_T(x,y)\).

Hence, \(y\) must be a best match of \(x\) and thus \(y \in N_G(x,s)\). Now assume \(y \notin N_G(x,s)\). In this case, \(N_G(x,s) \neq \emptyset\) implies the existence of some \(y' \in N_G(x,s)\) of color \(s\) and distinct from \(y\). Hence, \(xy'y' \in \mathcal{R}(G,\sigma)\) is displayed by \(T\) and therefore, \(\text{lca}_T(x,y') \prec_T \text{lca}_T(x,y)\). Therefore, \(y\) is not a best match of \(x\) in \((T,\sigma)\), and \(y \notin N_G(x,s)\). In summary, we have \(N_G(x,s) = N_G(x,s)\) for all \(x \in L\)
Algorithm 1 qBMG recognition

Input: A vertex-colored digraph \((G, \sigma)\) with \(\sigma: V(G) \to S\)
Output: A tree \((T, \tau, u)\) that explains \((G, \sigma)\) if it is a qBMG and, otherwise, false

1: if \((G, \sigma)\) is not properly colored then
2: return false
3: Compute \(\mathcal{R}(G, \sigma)\) and \(\mathcal{F}(G, \sigma)\) according to Definition 2.3
4: Use MTT to check if \((\mathcal{R}(G, \sigma), \mathcal{F}(G, \sigma))\) is consistent and, in the positive case, compute an agreeing tree \(T\); otherwise return false
5: Initialize \(u(x, s) = x\) for all \((x, s) \in L \times S\)
6: for all \(xy \in E(G)\) do
7: \(u(x, \sigma(y)) = \rho_T\)
8: return \((T, \tau, u)\)

and all \(s \in S \setminus \{\sigma(x)\}\). Therefore, we conclude that \((G, \sigma) = (\tilde{G}, \sigma)\) and that \((T, \tau, u)\) explains the qBMG \((\tilde{G}, \sigma)\).

As a direct consequence of Proposition 2.3 and Theorem 4.2 we obtain

**Theorem 4.3.** A properly colored graph \((G, \sigma)\) is a BMG if and only if \((G, \sigma)\) is a color-sink-free qBMG.

The proof of Theorem 4.2 is constructive and thus provides an algorithm to decide computationally whether \((G, \sigma)\) is qBMG and, if so, to compute an explanation \((T, \tau, u)\) for \((G, \sigma)\). The procedure is summarized in Algorithm 1. It relies on the polynomial-time algorithm MTT [29], named for the “mixed triplets problem restricted to trees”, which decides whether \((\mathcal{R}, \mathcal{F})\) is consistent and – in the affirmative case – constructs a corresponding tree \(T\). MTT in turn can be understood as a generalization of the well-known BUILD algorithm [26]. Given a set of rooted triples \(\mathcal{R}\) defined on a set of leaves \(L\), BUILD produces an undirected auxiliary graph, called Aho et al. graph and denoted by \([\mathcal{R}, L]\), with vertex set \(L\) and edges \(xy\) if and only if there is some \(z \in L\) such that \(xyz \in \mathcal{R}\). BUILD then recurses on the connected components of \([\mathcal{R}, L]\) with singleton vertex sets serving as base cases. The algorithm returns a tree \(T\) (on \(L\)) displaying all triples in \(\mathcal{R}\), which is determined by the recursion hierarchy and denoted by BUILD\((\mathcal{R}, L)\), or fails if no such tree exists. The latter is the case if and only if, at some recursion step with \(|L'| > 1\), the Aho et al. graph \([\mathcal{R}', L']\) is connected.

**Corollary 4.4.** Algorithm 1 with input \((G = (L, E), \sigma)\) can be implemented to run in \(O(|E||L|^2 \log |L|)\) and decides whether \((G, \sigma)\) is a qBMG and, in the affirmative case, constructs a tree \((T, \tau, u)\) that explains \((G, \sigma)\).

**Proof.** It takes \(O(|E|)\) time to verify whether \((G, \sigma)\) is a properly colored digraph. The triple sets \(\mathcal{R}(G, \sigma)\) and \(\mathcal{F}(G, \sigma)\) may be obtained in \(O(|L||E|)\) time since every triple in \(\mathcal{R}(G, \sigma) \cup \mathcal{F}(G, \sigma)\) is identifiable by an edge \(e\) and a vertex not incident with \(e\). Given a pair \((\mathcal{R}, \mathcal{F})\) of triple sets defined on \(L\), the algorithm MTT decides in \(O(|\mathcal{R}||L| + |\mathcal{F}||L| \log |L| + |L|^2 \log |L|)\) time whether \((\mathcal{R}, \mathcal{F})\) is consistent and, if so, returns a corresponding tree \(T\) with the same time complexity [29]. Since \(|\mathcal{R}| \leq |L| \cdot |E|\) and \(|\mathcal{F}| \leq |L| \cdot |E|\), we obtain an upper bound of \(O(|E||L|^2 \log |L|)\) for MTT. Finally, the truncation map \(u\) is constructed in \(O(|L||S| + |E|)\) by first initializing \(u(x, s) = x\) for all \((x, s) \in L \times S\). Then the edges are visited in arbitrary order. We set \(u(x, s) := \rho_T\) if there is an edge \(e = xy\) with \(\sigma(y) = s\). Since only colors in \(\sigma(L)\) are considered, we may assume \(|S| \leq |L|\). The total effort is therefore dominated by MTT.

The following technical result shows that the informative and forbidden triples in a subgraph \((G, \sigma)[V']\) induced by \(V' \subseteq V(G)\) are exactly the respective sets of triples of the original graph restricted to \(V'\).

**Observation 4.1** ([4], Observation 2). Let \((G, \sigma)\) be a vertex-colored digraph and \(V' \subseteq V(G)\). Then \(\mathcal{R}(G, \sigma)V' = \mathcal{R}(G[V'], \sigma_{V'})\) holds for each \(R \in \{\mathcal{R}, \mathcal{F}, \mathcal{R}^D\}\).

Theorem 4.2 and Observation 4.1 yield an alternative proof for the fact that qBMGs form a hereditary graph class.

**Alternative Proof of Corollary 3.6.** Let \((G, \sigma)\) be a qBMG. Then \((\mathcal{R} := \mathcal{R}(G, \sigma), \mathcal{F} := \mathcal{F}(G, \sigma))\) is consistent by Theorem 4.2. By Observation 4.1, we have \(\mathcal{R}' := \mathcal{R}(G[V'], \sigma_{V'}) \subseteq \mathcal{R}\) and \(\mathcal{F}' := \mathcal{F}(G[V'], \sigma_{V'}) \subseteq \mathcal{F}\).

10
\( \mathcal{F}(G[V'], \sigma_{V'}) \subseteq \mathcal{F} \) for every \( V' \subseteq V(G) \). Therefore, the pair \((R', \mathcal{F}')\) is clearly still consistent. By Theorem 4.2, \((G, \sigma)[V']\) is therefore again a qBMG.

Figure 3: Example for a qBMG \((G, \sigma)\) that is not explained by \((T_1 = \text{BUILD}(G, \sigma), V(G)), 0_1, u)\) for any truncation map \( u \). The set of informative triples is \( \mathcal{R}(G, \sigma) = \{xy|yz', yz'|z'\} \) and forbidden triples \( \mathcal{F}(G, \sigma) = \{xy|y', yx'|y', y'|y\} \). The Aho et al. graph \( H \) constructed at the top-level recursion step of \( \text{BUILD} \). The gray frames indicate the connected components of \( H \) and the green boxes indicate the sets in the auxiliary partition produced by the \( \text{MTT} \) algorithm. The tree \( T_1 \) cannot explain the edge \( xy' \) for any truncation map because \( \sigma(y) = \sigma(y') \) and \( \text{lca}_{T_1}(x, y) \prec_{T_1} \text{lca}_{T_1}(x, y') \). The tree produced by \( \text{MTT} \) with input \( \mathcal{R} \) and \( \mathcal{F} \) is \( T_2 \).

In [5], we obtained a characterization of BMGs that uses not only the informative triples and does not explicitly utilize the set of forbidden triples. More precisely, a vertex-colored digraph \((G, \sigma)\) with vertex set \( L \) is a BMG if and only if (i) the set of informative triples \( \mathcal{R} := \mathcal{R}(G, \sigma) \) is consistent, and (ii) \((G, \sigma) = \text{BMG(}\text{BUILD}(L, \sigma), \sigma)\) [5, Theorem 1]. This provides a procedure to recognize BMGs that is different from the one based on Proposition 2.3.

However, the example in Figure 3 shows that an analogous result does not hold for qBMGs. The counterexample \((G, \sigma)\) has informative triples \( \mathcal{R}(G, \sigma) = \{xy|y', yx'|y', y'|y\} \) and forbidden triples \( \mathcal{F}(G, \sigma) = \{xy|y', yx'|y', y'|y\} \). The Aho et al. graph \( H \) constructed at the top-level recursion step of \( \text{BUILD} \) has three connected components. The final output of \( \text{BUILD} \) is the tree \( T_1 \). Since \( \text{lca}_{T_1}(x, y) \prec_{T_1} \text{lca}_{T_1}(x, y') \), the edge \( xy' \) can never be contained in the qBMG explained by \((T_1, \sigma, u)\) for any truncation map \( u \). Hence, condition (ii) does not hold. The \( \text{MTT} \) algorithm also constructs the Aho et al. graph in each recursion step but merges two components \( C \) and \( C' \) whenever there is a forbidden triple \( ab|c \) such that \( a, b \in C \) and \( c \in C' \). This yields an auxiliary partition \( \mathcal{D}(L) \) of the leaf set \( L \) whose sets serve as input for the recursive calls instead of the connected components of the Aho et al. graph. In the example, \( \text{MTT} \) merges the two components \( \{x, y', z\} \) and \( \{y\} \) in the top-level recursion step in response to the forbidden triple \( xy|y \in \mathcal{F}(G, \sigma) \). The final result is the tree \( T_2 \). Also, one can easily verify that a truncation map \( u' \) can be found for \( T_2 \) such that \((T_2, \sigma, u')\) explains \((G, \sigma)\). Later we will use the following property of the trees produced by algorithm \( \text{MTT} \).

Lemma 4.5. Let \((\mathcal{R}, \mathcal{F})\) be a consistent pair of two triple sets defined on leaf set \( L \). Then \( \text{MTT} \) returns a least resolved tree \( T \) on \( L \) that agrees with \((\mathcal{R}, \mathcal{F})\), i.e., there is no tree \( T' \) on \( L \) with \( T' < T \) that still agrees with \((\mathcal{R}, \mathcal{F})\).

Proof. Since \((\mathcal{R}, \mathcal{F})\) is consistent, \( \text{MTT} \) returns a tree \( T \) on \( L \) that agrees with \((\mathcal{R}, \mathcal{F})\), see [29, Theorem 1]. To prove that \( T \) is least resolved, we show first that every inner edge in \( T \) is distinguished by some triple in \( \mathcal{R} \). Assume, for contradiction, that there is an inner edge \( vw \in E(T) \) that is not distinguished by a triple in \( \mathcal{R} \). Since \( vw \) is an inner edge, \( w \) has children \( w_1, \ldots, w_k \), \( k \geq 2 \). Consider the recursion step of \( \text{MTT} \) on \( L' := L(T(v)) \) and \((\mathcal{R}_{L'}, \mathcal{F}_{L'})\). The algorithm constructs an auxiliary partition \( \mathcal{D}(L') \) by starting with the connected components of \([\mathcal{R}_{L'}, L']\) and then merging two components \( C \) and \( C' \) stepwisely as long as there is a forbidden triple \( ab|c \) in \( \mathcal{F}_{L'} \) such that \( a, b \in C \) and \( c \in C' \). Note that \( L(T(w_i)) \cup \ldots \cup L(T(w_k)) = L(T(w)) \in \mathcal{D}(L') \). There cannot be a triple \( ab|c \in \mathcal{R}_{L'} \) with \( a \in L(T(w_j)) \) and \( b \in L(T(w_j)) \) for distinct children \( w_i \) and \( w_j \) of \( w \). To see this, consider that \( c \in L'(T(v)) \). In this case, \( \text{lca}_{T'}(a, b) = w \) and \( \text{lca}_{T'}(a, b, c) = v \), i.e., the triple \( ab|c \) would distinguish the edge \( vw \). If, on the other hand, \( c \in L(T(w)) \), then \( T \) clearly cannot display the triple \( ab|c \in \mathcal{R}_{L'} \subseteq \mathcal{R} \) also as a consequence of \( \text{lca}_{T'}(a, b) = w \). Hence, no such triple exists. Therefore, no two vertices \( a \in L(T(w_i)) \) and \( b \in L(T(w_j)) \) for distinct children \( w_i \) and \( w_j \) of \( w \) are adjacent in \([\mathcal{R}_{L'}, L']\). It follows that since \( w \) has at least two children and \( L(T(w)) \in \mathcal{D}(L') \), the set \( L(T(w)) \) must have been emerged as the disjoint union of \( \geq 2 \) connected components \( C_1, \ldots, C_l \) of \([\mathcal{R}_{L'}, L']\) in response to forbidden triples in \( \mathcal{F}_{L'} \). In particular, \( C_i \subseteq L(T(w_j)) \) for some \( 1 \leq j \leq k \) for each \( 1 \leq i \leq l \). Consider the series of merging steps that involve sets \( C_1, \ldots, C_l \) and unions of these sets. There must be a first merging step of two sets \( C \) and \( C' \) in this series that satisfy \( C \subseteq L(T(w_i)) \) and \( C' \subseteq L(T(w_j)) \) for distinct \( w_i, w_j \in \text{child}_{T}(w) \) in response to some triple.
Let $\sigma$ be a tree obtained from $T$ by a non-empty series of edge contractions, say of edges $e_1, \ldots, e_m$. By the arguments above, $e_1$ is distinguished by a triple $ab|c \in \mathcal{R}$. Therefore, we have $lca_{T'}(a, b) = lca_{T'}(\{a, b, c\})$ after contraction of $e_1$, and thus, $ab|c$ is not displayed by $T_{e_1}$. Since further contraction of edges does not introduce newly displayed triples, see [25, Theorem 1]. It follows that $T'$ does not display $\mathcal{R}$ and hence $T$ is least resolved.

Finally, we point out that increasing the number of colors by splitting a color class preserves the qBMG property.

**Proposition 4.6.** If $(G, \sigma)$ is a $\ell$-qBMG with $\sigma(L) = \ell < |L|$ colors, then there is a proper coloring $\sigma'$ such that $(G, \sigma')$ is an $(\ell + 1)$-qBMG.

**Proof.** Write $L_s := \{x \in L \mid \sigma(x) = s\}$. If $\ell < |L_s|$ there is a color $s$ with $|L_s| \geq 2$. Let $\sigma'$ be a coloring of $G$ obtained by arbitrarily partitioning $L_s = L_s' \cup L_s''$ into two non-empty subsets with new colors $s'$ and $s''$. Let $(T, \sigma, u)$ be an explanation of $(G, \sigma)$. For the leaf-colored tree $(T, \sigma')$, we construct the truncation map $u'$ as follows: For $x \in L_s'$, we set $u'(x, s') = u(x, s') = x$. For $x \notin L_s'$, we set $u'(x, s') = x$ if $N_G(x) \cap L_s' = \emptyset$ and $u'(x, s'') = x$ if $N_G(x) \cap L_s'' = \emptyset$; otherwise $u'(x, s') = u(x, s)$ and $u'(x, s'') = u(x, s)$, respectively. Finally, for $x \in V$ and $t \notin \{s', s''\}$, we set $u'(x, t) = u(x, t)$. It is not difficult to check that, by construction, $N_G(x, s') = N_G(x, s'') \cup N_G(x, s''')$ for all $x \in V$, while the out-neighborhoods of all vertices and all other colors remain unchanged. Thus $(G, \sigma') = qBMG(T, \sigma', u')$. Since $\sigma'$ is a proper $(\ell + 1)$-coloring, $(G, \sigma')$ is an $(\ell + 1) - qBMG$.

## 5 Least Resolved Trees for qBMGs

In general, many trees explain a given qBMG $(G, \sigma)$. Among these trees, the least resolved ones are of particular interest since they describe the phylogenetic information implicit in $(G, \sigma)$ without adding internal vertices and, thus, evolutionary events that are not implied by the available best matches. Least-resolved trees (LRTs) thus are the most parsimonious explanations. It is of particular interest whether there is a unique LRT and thus unambiguous phylogenetic information or whether there are conflicting explanations in the form of mutually inconsistent trees. BMGs are explained by a unique LRT $\hat{T}$ [1]. As we shall see below, this is no longer true for qBMGs. Although we lose uniqueness, the LRTs of qBMGs still have some convenient properties. In the following paragraphs, we briefly summarize the situation.

**Definition 5.1.** Let $(T, \sigma, u)$ be a leaf-colored tree with truncation map $u$. An edge $e \in E(T)$ is redundant with respect to the explained qBMGs if there is a truncation map $u'$ such that $qBMG(T_e, \sigma, u') = qBMG(T, \sigma, u)$, i.e., $(T, \sigma, u)$ and $(T_e, \sigma, u')$ explain the same qBMG. An edge $e \in E(T)$ that is not redundant is essential with respect to the explained qBMGs. Moreover, $(T', \sigma, u)$ is least resolved if there is no tree $T' < T$ and truncation map $u'$ such that $(T', \sigma, u')$ explains $qBMG(T, \sigma, u)$.

**Lemma 5.1.** Let $(G, \sigma)$ be a qBMG explained by $(T, \sigma, u)$ and let $e = vw \in E(T)$ with $w \prec_T v$. Then $e$ is essential if and only if $w \in L(T)$ or there are $x, y, y' \in L(T)$ such that $w = lca_T(x, y)$, $xy \in E(G)$, $v = lca_T(x, y')$ and $\sigma(y') = \sigma(y)$.
Proof. We start with the if-direction. Clearly, if \( w \in L(T) \), then \( L(T) \neq L(T_e) \) and thus, there is no truncation map \( u \) such that \( (T_e, \sigma, u') \) still explains \( (G, \sigma) \). Now suppose there are \( x, y \in L(T) \) such that \( w = \text{lca}_T(x, y) \) and \( xy \in E(G) \), and \( y' \in L(T) \) with \( \text{lca}_T(x, y') = v \) and \( \sigma(y') = \sigma(y) \). Since \( \text{lca}_T(x, y) = w \prec_T v = \text{lca}_T(x, y') \), we have \( xy' \notin E(G) \). After contraction of \( e \), we have \( \text{lca}_T(x, y) = \text{lca}_T(x, y') \). Hence, we have \( xy \in E(\text{qBMG}(T_e, \sigma, u')) \) if and only if \( xy' \in E(\text{qBMG}(T_e, \sigma, u')) \) for any truncation map \( u' \). Therefore, \((G, \sigma)\) cannot be explained by any tree of the form \((T_e, \sigma, u')\) and thus \( e \) is essential.

We continue with the only-if-direction. To this end suppose, for contraposition, that \( e = vw \) is an inner edge and that there are no \( x, y \in L(T) \) such that \( w = \text{lca}_T(x, y) \), \( xy \in E(G) \), \( v = \text{lca}_T(x, y') \) and \( \sigma(y') = \sigma(y) \). By Lemma 4.1, \( T \) agrees with \( (R(G, \sigma), I(G, \sigma)) \). Since \( T_e \prec_T T \), every triple that is displayed by \( T_e \) is also displayed by \( T \), see [27, Theorem 6.4.1]. Therefore, \( T_e \) does not display any of the forbidden triples in \( I(G, \sigma) \). Now suppose that \( T_e \) does not display some triple \( ab'b' \in R(G, \sigma) \). By definition, we have \( ab \in E(G) \), \( ab' \notin E(G) \) and \( \sigma(a) \neq \sigma(b) = \sigma(b') \). By Lemma 4.1, \( T \) displays \( ab'b' \), i.e., \( \text{lca}_T(a, b) \prec_T \text{lca}_T(a, b') = \text{lca}_T(b, b') \). Since \( T \) and \( T_e \) differ only by contraction of \( e \) and \( T_e \) does not display \( ab'b' \), it must hold that \( \text{lca}_T(a, b) = w = \text{lca}_T(a, b') = v \). Together with \( \sigma(b) = \sigma(b') \) and \( ab \in E(G) \), this contradicts the assumption. Hence, \( T_e \) does not display all triples in \( R(G, \sigma) \). In summary, \( T \) agrees with \( (R(G, \sigma), I(G, \sigma)) \). By Theorem 4.2, there is a truncation map \( u' \) such that \((T_e, \sigma, u')\) explains \( \text{qBMG} (G, \sigma) \). Hence, \( e \) is redundant and thus not essential.

The conditions \( w = \text{lca}_T(x, y) \), \( v = \text{lca}_T(x, y') \), and \( \sigma(y') = \sigma(y) \) in Lemma 5.1 imply that \( xy' \notin E(G) \). Together with \( xy \in E(G) \), we obtain

**Corollary 5.2.** Let \((G, \sigma)\) be a \( \text{qBMG} \) explained by \((T, \sigma, u)\). An inner edge \( vw \in E(T) \) is essential if and only if it is distinguished by an informative triple \( xy'y \in I(G, \sigma) \).

The following result shows that it is possible to characterize least resolved trees for \( \text{qBMGs} \) by the absence of redundant edges:

**Lemma 5.3.** A tree \((T, \sigma, u)\) is least resolved if and only if it does not contain a redundant edge.

Proof. Let \((G, \sigma)\) be the \( \text{qBMG} \) explained by \((T, \sigma, u)\). If \( e \in E(T) \) is redundant, there is a truncation map \( u' \) such that \((T_e, \sigma, u')\) explains \( (G, \sigma) \). Since \( T_e \prec_T T \), \((T, \sigma, u)\) is not least resolved.

For the converse, suppose \((T, \sigma, u)\) does not contain redundant edges and, for contradiction, assume that there is a tree \( T' \prec_T T \) and a truncation map \( u' \) such that \((T', \sigma, u')\) explains \((G, \sigma)\). Therefore, we have \( L(T) = L(T') \), and thus \( T' \) is obtained from \( T \) by a series of at inner-edge contractions \( e_1, \ldots, e_k \) with \( k \geq 2 \) since otherwise \( e_1 \) would be redundant by definition. Since \( e_1 = vw \) with \( w \prec_T v \) is essential, we have, by Lemma 5.1, \( x, y \in L(T) \) such that \( w = \text{lca}_T(x, y) \) and \( y \in N_G(x, \sigma(y)) \) and \( y' \in L(T) \) with \( \text{lca}_T(x, y') = v \) and \( \sigma(y') = \sigma(y) \). In particular, therefore, \( y' \notin N_G(x, \sigma(y)) \). After contraction of \( e_1 \) in \( T \), we have \( \text{lca}_{T_{e_1}}(x, y) = \text{lca}_{T_{e_1}}(x, y') \). Hence, there is no set \( A \in \mathcal{H}(T_{e_1}) \) such that \( (a, x, y) \in A \) and \( y' \notin A \) or \( (b, x, y') \in A \) and \( y \notin A \). Since \( T' \prec_T T_{e_1} \) and \( L(T) = L(T_{e_1}) \), we have \( \mathcal{H}(T_{e_1}) \subset \mathcal{H}(T_{e_1}) \). Therefore, there is also no set \( A \in \mathcal{H}(T') \) such that \( (a, x, y) \in A \) or \( (b, x, y') \in A \). It follows that \( \text{lca}_T(x, y) = \text{lca}_T(x, y') \). Since \((T', \sigma, u')\) explains \((G, \sigma)\), this immediately implies that either \( y, y' \in N_G(x, \sigma(y)) \) or \( y, y' \notin N_G(x, \sigma(y)) \); a contradiction.

Even though Lemma 5.3 generalizes the situation in \( \text{BMGs} \), we no longer have uniqueness of the least resolved tree. Figure 4 gives a simple counterexample with only two colors. Nevertheless, as a consequence of Theorem 4.2 and Lemma 4.5, we obtain

**Proposition 5.4.** Let \((G, \sigma)\) be a \( \text{qBMG} \) with vertex set \( L \). Then \((T, \sigma, u)\), where \( T \) is the tree produced by algorithm \( \text{MTT} \) with input \((R(G, \sigma), I(G, \sigma)) \) and \( L \) and \( u \) is a suitable truncation map is least resolved and explains \((G, \sigma)\).

## 6 Binary-explainable \( \text{qBMGs} \)

Binary Trees play a prominent role in phylogenetics because evolutionary events are usually assumed to give rise to only two descendant genes or species. Polyomtoms, i.e., vertices in \( T \) with three or more children, are, therefore, in most cases interpreted as the consequence of insufficient phylogenetic information (soft polyomtoms) rather than as true multifurcations (hard polyomtoms). The latter appear occasionally but are considered very rare by most authors [30, 31]. It is of interest, therefore, to determine whether a \( \text{qBMG} \) can be explained by a binary tree \( T \). BMGs that can be explained by binary trees have been studied in [4]. In this section, we derive analogous results for \( \text{qBMGs} \).
**Definition 6.1.** A vertex-colored digraph \((G, \sigma)\) is a binary-explainable qBMG if there is a binary tree \(T\) and a truncation map \(u\) such that \((T, \sigma, u)\) explains \((G, \sigma)\).

**Theorem 6.3.** A properly vertex-colored digraph \((G, \sigma)\) with vertex set \(L\) is a binary-explainable qBMG if and only if \(\mathcal{R}^B = \mathcal{R}^B(G, \sigma)\) is consistent. In this case, for every refinement \(T\) of the tree \(\text{BUILD}(\mathcal{R}^B, L)\), there is a truncation map \(u\) such that \((T, \sigma, u)\) explains \((G, \sigma)\).

**Proof.** For the only-if direction, suppose \((G, \sigma)\) is a binary-explainable qBMG, i.e., it is explained by a binary leaf-colored tree with truncation map \((T, \sigma, u)\). By Theorem 4.2, \(T\) agrees with \((\mathcal{R}(G, \sigma), \mathcal{F}(G, \sigma))\). In particular, \(T\) displays all triples in \(\mathcal{R}(G, \sigma) \subseteq \mathcal{R}^B\). Now assume that there is a triple \(bb'/a \in \mathcal{R}(G, \sigma) = \{bb'|a:\ abb' \in \mathcal{F}(G, \sigma), \sigma(b) = \sigma(b')\}\). By definition, \(abb' \in \mathcal{F}(G, \sigma)\) implies \(a, b\) \(\in \mathcal{F}(G, \sigma)\). Therefore, of the three possible triples \(ab'b, ab'h, and bb'|a\) on \(\{a, b, b'\}\), only \(bb'|a\) may be displayed by \(T\). This, together with the hypothesis that \(T\) is binary, implies that \(T\) indeed displays \(bb'|a\). Since this is true for any \(bb'|a \in \mathcal{R}^B \setminus \mathcal{R}(G, \sigma)\), \(T\) displays all triples in \(\mathcal{R}^B\), and thus, \(\mathcal{R}^B\) is consistent.

For the if direction, suppose that \(\mathcal{R}^B\) is consistent. Hence, the tree \(\text{BUILD}(\mathcal{R}^B, L)\) exists and displays all triples in \(\mathcal{R}^B\). Consider an arbitrary refinement \(T\) of \(\text{BUILD}(\mathcal{R}^B, L)\) (note that \(T = \text{BUILD}(\mathcal{R}^B, L)\) is possible). By [25, Theorem 1], \(T\) also displays all triples in \(\mathcal{R}^B\). Since \(\mathcal{R}(G, \sigma) \subseteq \mathcal{R}^B\), \(T\) displays all triples in \(\mathcal{R}(G, \sigma)\). Now assume there is a triple \(ab'b \in \mathcal{F}(G, \sigma)\) where \(\sigma(b) = \sigma(b')\). By definition, we have \(bb'|a \in \mathcal{R}^B\), and thus, \(bb'|a\) is displayed by \(T\). Therefore, \(T\) does not display the triple \(ab'b\). Since this is true for any \(ab'b \in \mathcal{F}(G, \sigma)\), \(T\) displays none of the triples in \(\mathcal{F}(G, \sigma)\).

**Corollary 6.2.** Every induced subgraph of a binary-explainable qBMG is again a binary-explainable qBMG.

As an immediate consequence of Theorem 6.1, Observation 4.1, and the fact that subsets of consistent triple sets are again consistent, we obtain that binary-explainable qBMGs form a hereditary class of colored digraphs.

**Definition 6.2.** An hourglass in a properly vertex-colored digraph \((G, \sigma)\), denoted by \([xy, x'y']\), is a subgraph \((G(Q), \sigma_Q)\) induced by a set of four pairwise distinct vertices \(Q = \{x, x', y, y'\} \subseteq V(G)\) such that (i) \(\sigma(x) = \sigma(x') \neq \sigma(y) = \sigma(y')\), (ii) \(xy, yx', x'y', y'x \in E(G)\), (iii) \(x'y, x'y' \notin E(G)\). A properly vertex-colored digraph is hourglass-free if it does not have an hourglass as an induced subgraph.

The definition of hourglasses is illustrated in Figure 5 (leftmost digraph). We will use the following technical result to link hourglasses to the inconsistency of the triple set \(\mathcal{R}^B(G, \sigma)\).

**Proposition 6.3** ([25, Theorem 2]). A set of triples \(\mathcal{R}\) defined on a leaf set \(L\) is consistent if and only if \(\mathcal{R}_{L'} \setminus L'\) is disconnected for every subset \(L' \subseteq L\) with \(|L'| \geq 3\).

**Lemma 6.4.** Every binary-explainable qBMG \((G, \sigma)\) is hourglass-free.
Proof. Suppose that \((G, \sigma)\) has an hourglass \([xy \xrightarrow{x} x'y']\) as a subgraph induced by \(L' := \{x, x', y, y'\}\). By the definition of hourglasses, informative triples, and forbidden triples, we have \(x'y'|y,y'x'\in \mathcal{R}(G,\sigma)\) and \(xy'y'|y,yx'|x'x'|x\in \mathcal{F}(G,\sigma)\). This in turn yields \(x'y'|y,y'x'|x,yy'|x,x'x'|x\in \mathcal{F}(G,\sigma)\). Therefore, \([xy,y',xx']\) is not binary-explainable since \(\mathcal{R}(G',\sigma') = \mathcal{R}_\mathcal{B}\). However, \([xy,y',xx']\) is not binary-explainable since \(\mathcal{R}(G',\sigma') = \mathcal{R}_\mathcal{B}\). Hence, \([x'y'|y,y'x'|x,yy'|x,x'x'|x]\) is connected and thus \(\mathcal{B}(G',\sigma')\) is consistent, see Theorem 6.1. By construction, for all \(\sigma(1)\) with at most \(k\) colors such that \(\sigma(x) = \sigma(x') = \sigma(y) = \sigma(y')\). By Theorem 6.1, \(\mathcal{B}(G',\sigma') = \mathcal{R}(G',\sigma')\). Hence, \([xy,y',xx']\) is not binary-explainable since \(\mathcal{R}(G',\sigma') = \mathcal{R}_\mathcal{B}\). In contrast to BMGs, see [32, Proposition 8], the converse of Lemma 6.4 is not true for qBMGs. Figure 5 shows an example of a hourglass-free qBMG \((G',\sigma')\) that is not binary-explainable. In general, no finite set of forbidden induced subgraphs characterizes binary-explainable qBMGs among qBMGs. Figure 6A, arising from a geometric configuration, provides a counterexample. For an integer \(k \geq 2\), consider a regular \(4k\)-gon \(x_1, z_1, \ldots, x_{2k-2}, x_{2k}\) in the Euclidean plane. Its vertices are colored with \(2k\) colors such that \(x_i, z_i\) receive the same color for all \(1 \leq i \leq 2k\). Then, we insert directed edges \(x_1, x_{2k+1}, x_{k+1}, z_1, z_2, x_2\) and the symmetric edge \(x_1, x_{k+1}\). Corresponding arcs are inserted for each of the \(2k - 1\) rotation of the polygon by an angle of \(\pi i/k\) for \(1 \leq i < k - 1\) around its center; see Figure 6A. A formal presentation of this graph is given in the following lemma, in which we use brackets to denote indices that are taken modulo \(2k\).

**Lemma 6.5.** For an integer \(k \geq 2\), let \((G,\sigma)\) be a vertex-colored digraph with vertex set \(L\) and edge set \(E\) where

(i) \(L = \{x_1, z_1, x_2, z_2, \ldots, x_{2k-2}, z_{2k}\}\),

(ii) \(\sigma(L) := \{c_1, c_2, \ldots, c_{2k}\}\) comprises \(2k\) pairwise distinct colors such that \(\sigma(x_i) = \sigma(z_i) = c_i\) for all \(1 \leq i \leq 2k\), and

(iii) \(E = \bigcup_{n=1}^{2k} \{x_i, x_{i+k}, x_{i+k}, z_i, x_{i+1}\}\) where \([n] := n \mod 2k\).

Then \((G,\sigma)\) is a qBMG that is not binary-explainable. Moreover, every induced subgraph of \((G,\sigma)\) with at most \(|L| - 2\) vertices is a binary-explainable qBMG.

**Proof.** By construction, for all \(i \in \{1, \ldots, 2k\}\), we have \(z_iz_{i+1} \in E\) but \(z_iz_{i+1} \notin E\). This and \(\sigma(z_i) \neq \sigma(x_{i+1}) = \sigma(z_{i+1})\) implies \(z_iz_{i+1}\in \mathcal{R}(G,\sigma)\) for all \(i \in \{1, \ldots, 2k\}\). Therefore, \(\mathcal{R} := \{z_iz_{i+1} | z_{i+1} : 1 \leq i \leq 2k\} \subseteq \mathcal{R}(G,\sigma)\).

Assume, for contradiction, that \(\mathcal{R}(G,\sigma) \setminus \mathcal{R} \neq \emptyset\), i.e., \(\mathcal{R}(G,\sigma)\) contains further informative triples \(ab|b' \) in which case \(ab \in E(G), ab' \notin E(G)\) and \(\sigma(a) \neq \sigma(b) = \sigma(b')\). By construction, \(ab \in E(G)\)
implies that only \( a = x_i \) and either \( b = x_{i+k} \) or \( b = z_{i+k} \) is possible for some \( i \in \{1, \ldots, 2k\} \). In either case, \( \sigma(b) = \sigma(b') \) implies that \( b' \in \{x_{i+k}, z_{i+k}\} \setminus \{b\} \) and thus \( ab' \in E(G) \); a contradiction. Hence, there are no other informative triples, i.e., \( \mathcal{R} = \mathcal{R}(G, \sigma) \). Similar arguments imply

\[
\mathcal{F} := \mathcal{F}(G, \sigma) = \bigcup_{i=1}^{2k} \{x_i x_{i+k} \mid |z_{i+k}|, x_i z_{i+k} \mid |x_{i+k}|\}
\]

and thus

\[
\mathcal{R}^B := \mathcal{R}^B(G, \sigma) = \bigcup_{i=1}^{2k} \{z_i x_{i+1} \mid |z_{i+1}|, x_{i+k} z_{i+k} \mid |x_i|\}.
\]

Using the tree \((T, \sigma)\) in Figure 6B, one easily verifies that this tree displays all triples in \( \mathcal{R} \) and none of the triples in \( \mathcal{F} \). Therefore, \((\mathcal{R}, \mathcal{F})\) is consistent, and by Theorem 4.2, \((G, \sigma)\) is a qBMG. However, the graph \( H := [\mathcal{R}^B, L] \) is connected (see also Figure 6A), and thus Proposition 6.3 implies that \( \mathcal{R}^B \) is not consistent. By Theorem 6.1, this implies that \( (G, \sigma) \) is not binary-explainable.

We next show that \((G, \sigma)[L \setminus \{x_i\}]\) is a binary-explainable qBMG for every \( 1 \leq i \leq 2k \). By the symmetric roles of \( x_i \) in \((G, \sigma)\), it suffices to show this claim for \( x_1 \). Thus consider \( L' = L \setminus \{x_1\} \) and \((G', \sigma') := (G, \sigma)[L']\). First note that the induced subgraph \((G', \sigma')\) is again a qBMG by Corollary 3.6. By Observation 4.1, we have

\[
\mathcal{R}^B(G', \sigma') = \mathcal{R}^B(G, \sigma)L' = \left( \bigcup_{i=1}^{2k} \{z_i x_{i+1} \mid |z_{i+1}|, x_{i+k} z_{i+k} \mid |x_i|\} \right) \setminus \{z_2 k x_1 \mid z_1, x_{k+1} z_{k+1} \mid x_1\}.
\]

Since no second triple of the form \( x_{k+1} z_{k+1} \mid y \) exists, the edge \( x_{k+1} z_{k+1} \) is not present in the Aho et al. graph \( H'' = [\mathcal{R}^B(G', \sigma'), L'] \). In particular, \( H'' \) is disconnected (see Figure 6C) and has two connected components \( C = \{z_1, x_2, z_2, \ldots, x_k, z_k, x_{k+1}\} \) and \( C' = \{z_{k+1}, x_{k+2}, z_{k+2}, x_{2k}, z_{2k}\} \). One easily verifies that there is no triple of the form \( x_{i+k} |z_{i+k}| x_i \) in \( \mathcal{R}^B(G', \sigma') \) that is contained in \( \mathcal{R}^B(G', \sigma')_C \) or \( \mathcal{R}^B(G', \sigma')_{C'} \). In particular, this implies that \( \mathcal{R}^B(G', \sigma')_C \) and \( \mathcal{R}^B(G', \sigma')_{C'} \) contain only informative triples of \((G', \sigma')\). Since \((G', \sigma')\) is a qBMG, \( \mathcal{R}(G', \sigma') \) is consistent by Theorem 4.2. The latter arguments together imply that \textsc{build} with input \( \mathcal{R}^B(G', \sigma') \) and \( L' \) never encounters a connected Aho et al. graph with more than two vertices, i.e., \( \mathcal{R}^B(G', \sigma') \) is consistent. By Theorem 6.1, this implies that \((G', \sigma')\) is binary-explainable.

Now consider an induced subgraph \((G'', \sigma'') := (G, \sigma)[L'']\) where \( L'' = L \setminus \{y, \tilde{y}\} \) for two distinct vertices \( y, \tilde{y} \in L \). If there is some \( 1 \leq i \leq 2k \) such that \( y = x_i \) or \( \tilde{y} = x_i \), then \((G'', \sigma'')\) is the induced subgraph of a binary-explainable qBMG \((G, \sigma)[L \setminus \{x_i\}]\) and thus also a binary-explainable qBMG by Corollary 6.2. Now suppose that this is not the case, i.e., \( y = z_i \) or \( \tilde{y} = z_i \) and \( 1 \leq i < j \leq 2k \). By construction of the Aho et al. graph and since \( \mathcal{R}^B(G'', \sigma'') = \mathcal{R}^B(G', \sigma')L'' \) by Observation 4.1, \([\mathcal{R}^B(G'', \sigma''), L'']\) is a subgraph of \( H[L''] \). Using Figure 6A, we therefore observe that \( H[L''] \) and thus also \([\mathcal{R}^B(G'', \sigma''), L'']\) has at least two connected components. In particular, \textsc{build} with input \( \mathcal{R}^B(G'', \sigma'') \) and \( L'' \) recurses on these connected components, and none of them can contain both \( x_i \) and \( x_j \). Let \( C \) be such a connected component and suppose that \( x_i \notin C \). Recall that \textsc{build} recurses on \( \mathcal{R}^B(G', \sigma')_C \) and \( C \). By Observation 4.1, we have \( \mathcal{R}^B(G', \sigma')_C = \mathcal{R}^B(G'[C], \sigma''_C) \). Since \((G'[C], \sigma''_C)\) is an induced subgraph of a binary-explainable qBMG \((G, \sigma)[L \setminus \{x_i\}]\), it is also a binary-explainable qBMG by Corollary 6.2. It follows that \( \mathcal{R}^B(G', \sigma')_C \) must be consistent. Since this is true for all connected components in the top-level recursion step, \textsc{build} never encounters a connected Aho et al. graph with more than one vertex, and thus \( \mathcal{R}^B(G', \sigma'') \) is consistent. By Theorem 6.1, this implies that \((G'', \sigma'')\) is binary-explainable.

Finally an induced subgraph \((G^*, \sigma^*)\) of \((G, \sigma)\) with at most \(|L| - 2\) vertices is in particular, also an induced subgraph of a graph \((G, \sigma)[L'']\) with \(|L''| = |L| - 2\). By the arguments above, \((G, \sigma)[L'']\) is a binary-explainable qBMG. By Corollary 6.2, \((G^*, \sigma^*)\) is also a binary-explainable qBMG.

Since \( k \geq 2 \) in Lemma 6.5 can be chosen arbitrarily large, we conclude

**Corollary 6.6.** There is no finite set of forbidden induced vertex-colored subgraphs that characterize the subclass of binary-explainable qBMGs among the class of qBMGs.

In particular, there are minimal forbidden induced subgraphs with an arbitrarily large number of colors.
7 Two-Colored qBMGs

The restrictions of qBMGs to the vertices with two colors in a sense form “building blocks” for the more general case. These bipartite graphs deserve a more detailed investigation. With only two colors, say \( \sigma(L) = \{r, s\} \), we have \( N(x) = N(x, s) \) if \( \sigma(x) = r \) and \( N(x) = N(x, r) \) if \( \sigma(x) = s \). We can, therefore, largely omit explicit references to the vertex colors \( r \) and \( s \) and focus entirely on the sets \( N(x) \) and \( N^-(x) \). In particular, properly 2-colored digraphs are color-sink-free precisely if they are sink-free. To avoid the explicit treatment of trivial cases, we will consider the monochromatic edge-less graph, and in particular singleton graphs \( K_1 \), also as 2-qBMGs.

In [5] 2-BMGs are characterized as sink-free graphs satisfying (N1), (N2), and (N3), and in Theorem 4.3 we have identified BMGs as the sink-free qBMGs. This suggests but does not imply, that 2-qBMGs are characterized by the neighborhood conditions (N1), (N2), and (N3). In the following, we show that this conjecture is indeed true.

Lemma 7.1. Every 2-qBMG \((G, \sigma)\) satisfies (N1), (N2), and (N3).

Proof. Let \((\tilde{G}, \sigma)\) be a 2-BMG associated with the qBMG \((G, \sigma)\) and \(L\) the common vertex set of \((G, \sigma)\) and \((\tilde{G}, \sigma)\). By Proposition 2.1 and Lemma 2.2, \((G, \sigma)\) satisfies (N1), (N2), and (N3).

In order to show that \((G, \sigma)\) satisfies (N1), we first consider a pair of independent vertices \(x, y \in G\). Since \(G \subseteq \tilde{G}\), \(x, y\) are independent in \(G\). Using \(E(G) \subseteq E(\tilde{G})\), the conclusion of property (N1) remains true for \(G\) whenever \(x, y\) are independent in \(G\). Now consider two vertices \(x, y\) independent in \(G\) but not in \(\tilde{G}\) and assume that \(xy \in E(\tilde{G})\). Thus, \(N_\tilde{G}(x) \neq N_\tilde{G}(y)\) and Lemma 3.2 implies that \(N_\tilde{G}(x) = \emptyset\). Hence, \((G, \sigma)\) trivially satisfies (N1) since there cannot be a vertex \(t\) with \(xt \in E(G)\).

To see that \((G, \sigma)\) satisfies (N2) assume that \(xy, yz, xz \in E(G)\). Recall that \(E(G) \subseteq E(\tilde{G})\) and \((\tilde{G}, \sigma)\) satisfies (N2) and thus, \(xy \notin E(\tilde{G})\). Hence, if \(xy \notin E(G)\), we have \(N_\tilde{G}(x) \neq N_\tilde{G}(y)\) and Lemma 3.2 implies that \(N_\tilde{G}(x) = \emptyset\), contradicting the assumption \(xy \in E(\tilde{G})\). Thus we have \(xy \notin E(G)\), and hence \((G, \sigma)\) satisfies (N2).

It remains to show that \((G, \sigma)\) satisfies (N3). Consider \(x, y \in L\) with a common out-neighbor \(z \in G\). By Lemma 3.2, this implies that \(N_\tilde{G}(x) = N_\tilde{G}(y)\) and \(N_\tilde{G}(x, y) = N_\tilde{G}(y, x)\). In particular, \(z\) is also a common out-neighbor of \(x, y\) in \(G\) since \(G \subseteq \tilde{G}\). Now assume, for contradiction, that neither \(N_G(x) \subseteq N_\tilde{G}(x)\) nor \(N_G(y) \subseteq N_\tilde{G}(y)\). Hence, there are vertices \(v\) such that \(xv \in E(G)\) and \(xv \notin E(\tilde{G})\), and \(w\) such that \(yw \in E(G)\) and \(yw \notin E(\tilde{G})\). Since \(E(G) \subseteq E(\tilde{G})\), we have \(v \in N_\tilde{G}(x)\) and \(w \in N_\tilde{G}(y)\). Since \((G, \sigma)\) satisfies (N3) and \(x, y\) have a common out-neighbor \(z\) in \(\tilde{G}\), it follows that \(N_\tilde{G}(x) \subseteq N_\tilde{G}(y)\) or \(N_\tilde{G}(y) \subseteq N_\tilde{G}(x)\). If \(N_G(x) \subseteq N_\tilde{G}(x)\), we have \(v \in N_G(x)\), and \(yv \in E(G)\); a contradiction. If \(N_G(y) \subseteq N_\tilde{G}(y)\), we similarly obtain the contradiction that \(xw \in E(G)\). Therefore, we conclude that \((G, \sigma)\) satisfies (N3).

The hierarchy-like structure of the out-neighborhood, i.e., property (N3), suggests that the out-neighborhoods contain information on the structure on the tree(s) explaining a 2-colored qBMG. This connection, however, is less straightforward than one might expect. Following [1], we consider the reachable sets
\[
R(x) := N(x) \cup N(N(x)) \cup N(N(N(x))) \cup \cdots
\]
and the corresponding isotonic map \(R : 2^V \to 2^V\), \(A \mapsto R(A) := \bigcup_{x \in A} R(x)\). As a direct consequence of (N2), we have \(R(x) = N(x)\cup N(N(x)) \subseteq N(N(N(A))) \subseteq N(N(A))\), and thus also \(N(N(N(A))) \subseteq N(N(A)) \subseteq R(A)\), which implies \(R(R(A)) \subseteq R(A)\).

Lemma 7.2. Let \((G, \sigma)\) be a properly 2-colored graph satisfying (N1) and (N2). If \(N(x) \cap N(y) = \emptyset\), then \(N(N(x)) \cap N(N(y)) = \emptyset\).

Proof. Assume, for contradiction, that \(N(x) \cap N(y) = \emptyset\) and there is a vertex \(w \in N(N(x)) \cap N(N(y))\). Then neither \(x\) nor \(y\) is a sink, and \(x, y\) are distinct, and, since \((G, \sigma)\) is properly colored, we have \(\sigma(x) = \sigma(y) = \sigma(w)\). Moreover, there are vertices \(u \in N(x)\) and \(v \in N(y)\) such that \(xu, uw, yv, vw \in E(G)\). In particular, \(u \neq v\) and \(xv, yu \notin E(G)\) because \(N(x) \cap N(y) = \emptyset\). Therefore, \(x, y, u, v\) must all be pairwise distinct. If \(x = w\), then \(yw, vw = vx, xu \in E(G)\) and \(yu \notin E(G)\) contradicts (N2). The case \(y = w\) yields an analogous contradiction. Since \((G, \sigma)\) is properly colored, it remains to consider the case when all five vertices \(x, y, u, v, w\) are all pairwise distinct. Suppose that \(x\) and \(v\) are independent. Together with \(xu, vu, uw\) this contradicts (N1). Hence, we must have \(vx \in E(G)\). But then \(yu, xv, xu \in E(G)\) and (N2) imply \(yu \notin E(G)\); a contradiction. Therefore, we conclude that a vertex \(w \in N(N(x)) \cap N(N(y))\) cannot exist. \(\square\)
We are now in the position to generalize [1, Lemma 9], which is equivalent to the next statement if one assumes \((G, \sigma)\) to be connected.

**Lemma 7.3.** Let \((G, \sigma)\) be a properly 2-colored graph satisfying (N1), (N2), and (N3). Then the set of reachable sets \(R := \{R(x) \mid x \in V(G)\}\) forms a hierarchy-like set system.

**Proof.** Let \(x, y \in V(G)\). If \(x = y\), we trivially have \(R(x) \cap R(y) = R(x) = R(y)\). If \(x\) or \(y\) is a sink, we have \(R(x) \cap R(y) = \emptyset\). Now assume that \(R(x) \cap R(y) \neq \emptyset\). By Lemma 7.2, this implies either (i) \(N(x) \cap N(y) \neq \emptyset\), or (ii) \(N(x) \cap N(y) = \emptyset\) and \(N(x) \cap N(N(y)) \neq \emptyset\) or \(N(N(x)) \cap N(y) \neq \emptyset\).

Since \((G, \sigma)\) is properly colored, Case (i) implies \(\sigma(x) = \sigma(y)\) while Case (ii) can only occur if \(\sigma(x) \neq \sigma(y)\). In Case (i), (N3) implies \(N(x) \subseteq N(y)\) or \(N(y) \subseteq N(x)\), which by isotony implies \(R(x) \subseteq R(y)\) or \(R(y) \subseteq R(x)\), respectively.

For Case (ii), assume \(\sigma(x) \neq \sigma(y)\) and \(N(x) \cap N(N(y)) \neq \emptyset\) and let \(v \in N(x) \cap N(N(y))\). From \(\sigma(x) \neq \sigma(y)\), we infer \(\sigma(v) = \sigma(y)\). If \(x \in R(y)\), then isotony of the map \(R\) and the arguments above imply \(R(x) \subseteq R(R(y)) \subseteq R(y)\). Similarly, \(y \in R(x)\) implies \(R(y) \subseteq R(x)\). Now assume \(x \notin R(y)\) and \(y \notin R(x)\) and thus in particular \(x \notin N(y)\) and \(y \notin N(x)\). Together with \(v \in N(x) \cap N(N(y))\), this implies that \(x, y, v\) are pairwise distinct. Moreover, there must be a vertex \(w \in N(y)\) with \(yw, we \in E(G)\) and \(w \notin \{x, y, v\}\). Thus, we have \(xv, yu, uv \in E(G)\) and \(x\) and \(y\) are independent vertices \(x\) and \(y\); a contradiction to (N1). Analogous arguments show that only \(R(x) \subseteq R(y)\) or \(R(y) \subseteq R(x)\) are possible if \(N(N(x)) \cap N(y) \neq \emptyset\). Therefore, we have \(R(x) \cap R(y) \in \{\emptyset, R(x), R(y)\}\) for all \(x, y \in V\).

We note that the proof of Lemma 7.3 in [1] relies on the assumption that \((G, \sigma)\) is sink-free [5].

The proof above shows that this additional assumption is not necessary.

As in the special case of BMGs, however, the reachable sets \(R\) do not coincide with the hierarchy \(\mathcal{H}(T)\) of a tree explaining \((G, \sigma)\). This is easily seen by considering

\[
\bigcup_{x \in V(G)} R(x) = \{y \in V(G) \mid N^-(y) \neq \emptyset\} = V \setminus V_{\text{source}} \tag{4}
\]

since a source vertex \(y\) with \(N^-(y) = \emptyset\) is never contained any set \(N(x)\) and thus also not in any reachable set \(R(x)\). As discussed in [1] for the special case of BMGs, it is not sufficient to simply add \([x]\) to \(R(x)\), however. Instead, larger sets are required to handle source vertices as well as vertices that are indistinguishable in terms of their in-neighborhood. Here, we consider a slightly modified construction that also accommodates sinks. To this end, we define for any digraph \(G\) and every vertex \(x \in V(G)\) the set

\[
Q(x) := \{y \in V(G) \mid N^-(y) = N^-(x) \text{ and } \emptyset \neq N(y) \subseteq N(x)\}. \tag{5}
\]

This definition differs from the specification in [1] in two aspects: (i) it adds the conditions "\(\emptyset \neq N(y)\)" , which is always satisfied in sink-free graphs and thus in BMGs. (ii) Each vertex is considered separately here instead of being aggregated into so-called thinness classes. We collect several simple properties of \(Q(x)\) in Lemma 7.4 below. The straightforward arguments are essentially the same as in [1, Lemma 10].

**Lemma 7.4.** Let \((G, \sigma)\) be a properly 2-colored digraph.

(a) If \(x\) is a sink, i.e., \(N(x) = \emptyset\), then \(Q(x) = \emptyset\); otherwise \(N(x) \neq \emptyset\) implies \(x \in Q(x)\).

(i) \(y \in Q(x)\) implies \(\sigma(x) = \sigma(y)\).

(ii) \(y \in Q(x)\) implies \(Q(y) \subseteq Q(x)\).

(iii) \(y \in Q(x)\) implies \(N(y) \subseteq N(x)\).

(iv) If \(x \notin N(y)\), then \(Q(x) \cap N(y) = \emptyset\).

(v) If \(x \notin N(N(y))\), then \(Q(x) \cap N(N(y)) = \emptyset\).

(vi) \(N(x) \cap N(y) = \emptyset\) implies \(Q(x) \cap Q(y) = \emptyset\).

**Proof.** (o) If \(x\) is a sink, the condition becomes \(\emptyset \neq N(y) \subseteq \emptyset\) and thus no such \(y\) exists. Otherwise, both conditions are trivially true for \(y = x\).

(i) Since \(y \in Q(x)\) implies that \(x\) and \(y\) share at least one out-neighbor in the bipartite graph \(G\), \(x\) and \(y\) are in the same color class.

(ii) Let \(y \in Q(x)\) and \(z \in Q(y)\). Then by definition of \(Q\), \(N^-(z) = N^-(y) = N^-(x)\) and \(\emptyset \neq N(z) \subseteq N(y) \subseteq N(x)\), which implies \(z \in Q(x)\) for all \(z \in Q(y)\) and thus \(Q(y) \subseteq Q(x)\).

(iii) follows immediately from the definition of \(Q\).
(iv) Suppose $x \notin N(y)$ but there is $z \in Q(x) \cap N(y)$. Then $y \in N^-(z) = N^-(x)$ and thus $x \in N(y)$; a contradiction.

(v) Suppose $x \notin N(N(y))$ but there is $z \in Q(x) \cap N(N(y))$. Thus there is $w \in N(y)$ such that $w \in N^-(z) = N^-(x)$ and therefore $x \in N(N(y))$; a contradiction.

(vi) Let $N(x) \cap N(y) = \emptyset$ and assume, for contradiction, that $Q(x) \cap Q(y) \neq \emptyset$. Thus, there is a $z \in Q(x) \cap Q(y)$. By definition of $Q$, $\emptyset \neq N(z) \subseteq N(x), N(y)$ and thus $N(x) \cap N(y) \neq \emptyset$; a contradiction. \hfill \square

With the help of $Q$, we are now in the position to define the extended reachable set for any digraph $G$ and every vertex $x \in V(G)$ as follows:

$$R'(x) := R(x) \cup Q(x)$$

by analogy to the construction of 2-BMGs. For BMGs, the $R'(x)$ reduces to the corresponding sets in [1] since BMGs have no sinks. Note, if $x$ is a sink, i.e., $N(x) = \emptyset$, then also $N(N(x)) = \emptyset$ and thus $R'(x) = \emptyset$. Conversely, $N(x) \subseteq R'(x) = \emptyset$ implies that $x$ is a sink. Therefore, we have

Observation 7.1. $R'(x) = \emptyset$ if and only if $x$ is a sink.

Observation 7.2. For a properly 2-colored digraph $(G, \sigma)$, we have $y \in N(x)$ if and only if $y \in R'(x)$ and $\sigma(x) \neq \sigma(y)$.

Proof. While the only-if-direction is trivial, the if-direction follows from the fact that all vertices in $N(N(x))$ and $Q(x)$ must have the same color as $x$. \hfill \square

The following result mirrors [1, Lemma 11].

Lemma 7.5. If $(G, \sigma)$ is properly two-colored graph satisfying (N1), (N2), and (N3), then $R' := \{R'(x) \mid x \in V(G)\} \setminus \{\emptyset\}$ forms a hierarchy-like set system.

Proof. Let $R'(x), R'(y) \in R'$ for two distinct vertices $x, y \in V(G)$. By definition of $R'$, neither of $R'(x)$ and $R'(y)$ is empty. This and Observation 7.1 implies that neither $x$ nor $y$ is a sink.

We first show that $y \in R'(x)$ implies $R'(y) \subseteq R'(x)$. Let $y \in R'(x)$. If $y \in Q(x)$, then Lemma 7.4 yields $Q(y) \subseteq Q(x)$ and $N(y) \subseteq N(x)$, and thus also $R(y) \subseteq R(x)$ and finally $R'(y) \subseteq R'(x)$. If $y \in R(x)$ then $R(y) \subseteq R(x) \subseteq R'(x)$ since $y \in R(x)$ implies that everything that is reachable from $y$ is also reachable from $x$. By definition of $Q$, $N^-(z) = N^-(y)$ for all $z \in Q(y)$. Thus $y \in R(x)$ implies $N(x) \cap N(y) = y \in N(x) \cap N(N(x)) = R(x) \subseteq R'(x)$. Hence, we have $Q(x) \cap Q(y)$ and, in summary, $R'(y) = R(y) \cup Q(y) \subseteq R'(x)$. By analogous arguments, $x \in R'(y)$ implies $R'(x) \subseteq R'(y)$.

Now suppose $y \notin R'(x)$ and $x \notin R'(y)$. We will show that this implies $R'(x) \cap R'(y) = \emptyset$. Since $x \notin R'(y)$, we have $y \notin N(y)$ and $x \notin N(N(y))$. This together with Lemma 7.4(iv)-(v) implies $Q(x) \cap N(y) = \emptyset$ and $Q(x) \cap N(N(y)) = \emptyset$, respectively. Thus $Q(x) \cap R(y) = \emptyset$. By similar arguments, we obtain $R(x) \cap Q(y) = \emptyset$ from $y \notin R'(x)$. From $y \notin R'(x)$ and $x \notin R'(y)$, we have $y \notin N(x)$ and $x \notin N(y)$, which together with (N1) implies $N(N(x)) \cap N(y) = N(x) \cap N(N(y)) = \emptyset$. Taken together, the latter arguments imply $R'(x) \cap R'(y) = (N(x) \cap N(y)) \cup (N(N(x)) \cap N(N(y))) = (Q(x) \cap Q(y))$. If $N(x) \cap N(y) = \emptyset$, Lemma 7.2 implies $N(N(x)) \cap N(N(y)) = \emptyset$ and Lemma 7.4(iv) implies $Q(x) \cap Q(y) = \emptyset$. Therefore, $R'(x) \cap R'(y) = \emptyset$. Now assume $N(x) \cap N(y) \neq \emptyset$. From (N3) the two cases $N(y) \subseteq N(x)$ or $N(x) \subseteq N(y)$ arise. In the former case, (N3'), which is satisfied as a consequence of Lemma 2.2, together with $x \notin N(y)$, $y \notin N(x)$ and $N(x) \cap N(y) \neq \emptyset$ implies $N^-(x) = N^-(y)$ and thus $Q(y) \subseteq Q(x)$. Furthermore, we have $N(N(y)) \subseteq N(N(x))$ by isometry of $N$, and thus $R'(y) \subseteq R'(x)$. Since $y \in R'(y)$, this contradicts $y \notin R'(x)$. In the latter case, analogous arguments yield $R'(x) \subseteq R'(y)$, contradicting $x \notin R'(x)$. Therefore, the case $N(x) \cap N(y) \neq \emptyset$ cannot occur and we indeed have $R'(x) \cap R'(y) = \emptyset$.

In summary, $R'$ forms a hierarchy-like set system. \hfill \square

It should be stressed that we have made no assumptions about the connectedness of $(G, \sigma)$, which has not been the case in the discussion in [1]. The Hasse diagram $T(R')$ therefore will in general be a forest rather than a tree. Still, the hierarchy-like set system $R'$ can easily be extended to a hierarchy by adding all singletons as minimal elements and $V(G)$ as maximal element:

Proposition 7.6. Let $(G, \sigma)$ be a properly 2-colored digraph satisfying (N1), (N2), and (N3). Then

$$\mathcal{H}(G, \sigma) := R' \cup \{x \mid x \in V(G)\} \cup \{V(G)\}$$

is a hierarchy on $V(G)$. 

19
The following result parallels part of the proof of [1, Theorem 4].

**Lemma 7.7.** Let \((G, \sigma)\) be a properly 2-colored digraph satisfying (N1), (N2), and (N3), let \(T\) be the tree with hierarchy \(\mathcal{H}(T) = \mathcal{H}(G, \sigma)\) and define for all \(x \in V(G)\) and \(r \neq \sigma(x)\) the truncation map by \(u(x, r) = pr\) if \(N(x) \neq \emptyset\) and \(u(x, r) = x\) if \(N(x) = \emptyset\). Then \((G, \sigma) = q\text{BMG}(T, \sigma, u)\).

**Proof.** By Proposition 7.6, \(\mathcal{H}(G, \sigma)\) is a hierarchy on \(V(G)\) and thus \(L(T) = V(G)\). In particular, \((G, \sigma)\) and \(q\text{BMG}(T, \sigma, u)\) have the same vertex set. Denote by \(N(x)\) the out-neighbors of \(x \in V(G)\) in \(q\text{BMG}(T, \sigma, u)\) and write \(N(x)\) for the out-neighbors in \((G, \sigma)\). We prove that \(y \in N(x)\) if and only if \(y \in N(x)\) for all \(x, y \in V(G)\). By Observation 7.2, \(y \in N(x)\) is equivalent to \(y \in R'(x)\) and \(\sigma(y) \neq \sigma(x)\).

First assume \(y \in \tilde{N}(x)\). By construction, we have \(\sigma(y) \neq \sigma(x)\) and \(N(x) \neq \emptyset\) since otherwise \(u(x, \sigma(y)) = x\) contradicts \(y \in \tilde{N}(x)\). Assume, for contradiction, that \(y \notin R'(x)\). Since \(N(x) \neq \emptyset\), there is a vertex \(y' \in N(x) \subseteq R'(x)\) such that \(y' \neq y\) and \(\sigma(y') = \sigma(y)\). Since \(\mathcal{H}(G, \sigma)\) is a hierarchy, see Proposition 7.6, there is a unique inclusion-minimal set \(R^* \in \mathcal{H}(G, \sigma)\) with \(x, y \in R^*\). However, \(x \in R'(x) \cap R^*\) together with \(y \notin R'(x)\) and the fact that \(R'(x)\) and \(R^*\) are both sets in the hierarchy \(\mathcal{H}(G, \sigma)\) implies that \(R'(x) \subseteq R^*\). Denoting by \(v\) and \(v^*\) the vertices of \(T\) satisfying \(L(T(v)) = R'(x)\) and \(L(T(v^*)) = R^*\), respectively, we have \(v \prec_T v^*\). Since \(R^*\) is inclusion-minimal with respect to the property \(x, y \in R^*\), we have \(lca_T(x, y) = v^*\). Similarly, \(x, y' \in R'(x)\) implies that \(lca_T(x, y') \preceq_T v\). In summary, we obtain \(lca_T(x, y') \preceq_T v \prec_T v^* = lca_T(x, y)\) and \(\sigma(y) = \sigma(y')\). Thus \(y\) cannot be a best match of \(x\), which contradicts \(y \in \tilde{N}(x)\). Therefore, we have \(y \in R'(x)\) and thus \(y \in N(x)\).

For the converse, assume \(y \in N(x)\), i.e., \(y \in R'(x)\) and \(\sigma(y) \neq \sigma(x)\). In particular, therefore, the truncation map \(u(x, \sigma(y)) = pr\) imposes no further constraint and \(y \in N(x)\) if and only if \(y\) is a best match of \(x\). Now assume, for contradiction, that \(y\) is not a best match of \(x\), i.e., there is some vertex \(y' \in V(G)\) of color \(\sigma(y') = \sigma(y)\) such that \(v := lca_T(x, y') \prec_T lca_T(x, y)\). Thus, we have \(x, y' \in L(T(v))\) and \(y \notin L(T(v))\). In particular, by construction of \(R^*\), there is a vertex \(z \in V(G)\) such that \(L(T(v)) = R'(z)\). Since \(y \in R'(x)\) and \(y \notin R'(x)\), we have \(x \neq z\). We have to consider three cases according to the constituents of \(R'(z) = N(z) \cup N(N(z)) \cup Q(z)\): If \(x \in N(z)\) then \(y \in N(N(z)) \subseteq R'(z)\); a contradiction. If \(x \in N(N(z))\), then there is \(w \notin \{x, y, z\}\) such that \(zw, wx, xy \in E(G)\) which, together with \(N(z)\), implies \(y \in N(z) \subseteq R'(z)\); a contradiction. Finally, if \(x \in Q(z)\), then by definition \(y \in N(x) \subseteq N(z) \subseteq R'(z)\); a contradiction. Since none of these cases is possible, we conclude that no vertex \(y'\) as specified can exists and thus \(y\) is a best match of \(x\), i.e., \(y \in \tilde{N}(x)\).

Now, \(y \in \tilde{N}(x)\) if and only if \(y \in N(x)\) for all \(x, y \in V(G)\) immediately implies \(\tilde{N}(x) = N(x)\) for all \(x \in V(G)\), and thus the two 2-colored graphs coincide. \(\square\)

Combining Lemma 7.1 and Lemma 7.7 we obtain the desired characterization of qBMGs:

**Theorem 7.8.** A properly 2-colored graph is a qBMG if and only if it satisfies (N1), (N2), and (N3).

The axioms (N1), (N2), and (N3) are independent of the coloring \(\sigma\). It is of interest, therefore, to consider digraphs satisfying these three conditions without considering a coloring. If \(G\) is bipartite, then a proper 2-coloring \(\sigma\) exists that turns \((G, \sigma)\) into a qBMG. This begs the question of whether the axioms already imply that \(G\) is bipartite. We give an affirmative answer in the following theorem.

**Theorem 7.9.** Let \(G\) be a digraph satisfying (N1) and (N2). Then \(G\) is bipartite.

**Proof.** A directed graph \(G\) is a minimal counterexample to the assertion of the theorem if it is minimal among the induced subgraphs of non-bipartite digraph satisfying (N1) and (N2). \(G\) is not bipartite if and only if its underlying undirected graph contains an odd cycle. In the setting of directed graphs, an odd cycle is a subgraph of \(G\) with vertex set \(C = \{x_1, x_2, \ldots, x_m\} \subseteq V(G)\) of odd cardinality \(m = |C|\) such that \(x_{i-1}x_i \in E(G)\) or \(x_ix_{i-1} \in E(G)\) for \(1 \leq i \leq \ell\), where \(x_0 = x_m\). Two vertices \(x_i\) and \(x_j\) are consecutive if \(j = i \pm 1\). Since (N1) and (N2) are hereditary, in particular \(G[C]\) satisfies (N1) and (N2). A minimal counterexample \(G\), therefore, contains only the vertices of the odd cycle but possibly additional edges.

If \(G\) contains a triangle \(\{x, y, z\}\) then, w.l.o.g., there is a directed path \(xyz\). Then \(xz \in E(G)\), since otherwise \(xz \in E(G)\) must complete the triangle in which case \(z \in N(N(N(x)))\), contradicting (N2). However, \(xz \in E(G)\) implies \(z \in N(x) \cap N(N(x))\), which contradicts (N1). Thus \(G\) is triangle-free and \(|C| \geq 5\). In particular, therefore, there are two distinct, non-consecutive vertex \(x\) and \(y\).
Denote by $C_{xy} \subseteq C$ and $C_{yx} \subseteq C$ the vertex sets of the distinct “underlying undirected” paths connecting $x$ and $y$ in $C$. A pair of vertices $\{x_i, x_j\}$ is a chord in $C$ if $x_i x_j \in E(G)$ and $x_i$ and $x_j$ are non-consecutive along $C$, i.e., $|i - j| \neq 1$ for $1 \leq i \leq m$ with $x_i = x_0$. If $\{x, y\}$ is a chord in $C$, then both $G[C_{xy}]$ and $G[C_{yx}]$ are cycles and $|C_{xy}|, |C_{yx}| \geq 3$ since $x$ and $y$ are non-consecutive along $C$. Since $C_{xy} \cap C_{yx} = \{x, y\}$ we have $|C_{xy}| + |C_{yx}| = |C| + 2$ and thus $|C_{xy}|, |C_{yx}| \leq |C| - 1$. Therefore, $G[C]$ is composed of two shorter cycles, of which one is even, and the other is odd; say $C_{xy}$ is odd. This odd cycle is shorter than $C$. Again, since (N1) and (N2) are hereditary, in particular $G[C_{xy}]$ satisfies (N1) and (N2), and thus $G[C_{xy}]$ is a counterexample contradicting minimality of $G[C]$. Hence, a minimal counterexample $G$ is isomorphic to a chordless odd cycle $C$ that is not a triangle and thus comprises $|C| \geq 5$ vertices.

Assume that the minimal counterexample $G$ contains a path with three consecutive edges with the same orientation, $x \rightarrow a \rightarrow b \rightarrow y$. Then $y \in N(N(N(x)))$ and (N2) implies $y \in N(x)$. If $x \neq y$, this implies that $G$ is either an even cycle or $G$ has a chord $\{x, y\}$; a contradiction. If, on the other hand, $x = y$, we obtain a contradiction to $G$ being a simple graph. Thus $G$ may contain at most two consecutive edges with the same orientation. If two such edges exist, then the next edge must have the opposite orientation. More precisely, there exists a (not necessarily induced) subgraph $x \rightarrow a \rightarrow z \leftarrow y$ in $G$. Hence, $z \in N(N(x)) \cap N(y)$, and thus (N1) implies $x \in N(y)$ or $y \in N(x)$. Therefore, $\{x, y\}$ is a chord whenever $|V(G)| \geq 5$, a contradiction to the observation above that a minimum counterexample is isomorphic to a chordless cycle $C$ with $|C| \geq 5$. Therefore, a minimal counterexample $G[C]$ cannot contain a pair of consecutive edges with the same orientation. Hence, any two consecutive edges must have alternating orientations. This implies that $|C|$ is even. Therefore, no minimal counterexample exists.

**Corollary 7.10.** Let $G$ be a graph satisfying (N1), (N2), and (N3). Then $(G, \sigma)$ is a qBMG for every proper 2-coloring $\sigma$ of $G$.

**Proof.** Consider the decomposition $G = \bigcup_i G_i$ of $G$ into its connected components. Theorem 7.9 implies that $G$ and thus each of its connected components $G_i$ are bipartite. Let $\sigma_i$ be a 2-coloring of $G_i$. Furthermore, by heredity, the $G_i$ satisfies (N1), (N2), and (N3). This together with Theorem 7.8 implies that $(G_i, \sigma_i)$ is a 2-qBMG and thus, has an explanation $G_i(T_i, \sigma_i, u_i)$. Now exchange color $r$ and $s$ in $G_i$, and set $u'_i(x, r) := u_i(x, s)$ and $u'_i(x, s) := u_i(x, r)$ for all $x \in V(G)$. Clearly, $G_i(T_i, \sigma'_i, u'_i)$ is an explanation for $(G_i, \sigma'_i)$. As a consequence of Observation 3.2 there is an explaining tree $(T, \sigma, u)$ for every proper 2-coloring of $G$.

Recall that color-sink-freeness of properly 2-colored digraphs is equivalent to claiming that they are sink-free. Another way of expressing that 2-BMGs are the (color-)sink-free 2-qBMGs is to say that 2-BMGs are the 2-qBMGs that have an explanation $(T, \sigma, u)$ with $u(x, s) = \rho_T$ for all $x \in L$ and $s \neq \sigma(x)$, see Theorem 4.3. This begs the question of whether there are interesting subclasses of 2-qBMGs that are more general than 2-BMGs. In practical applications, it is plausible to assume that sequence similarities can reliably identify homologs, at least within some range of evolutionary divergence. This suggests restricting the truncation map $u$ to be bounded away from the leaves. In the following, we will investigate in which cases a 2-qBMG can be explained by a leaf-colored tree $(T, \sigma, u)$ with truncation map $u$ that satisfies

$(M) \quad z \prec_T u(z, r)$ for $r \neq \sigma(z)$ and all $z \in V(G)$.

In other words, a truncation map $u$ satisfies $(M)$ if and only if the statement “$u(x, s) = x$ if and only if $s = \sigma(x)$” is satisfied. We start with the following simple technical result.

**Lemma 7.11.** Let $(G, \sigma)$ be a qBMG explained by $(T, \sigma, u)$ and $v \in V(T)$ such that $|\sigma(L(T(v)))| = 1$. Then $N^-(x) = N^-(y)$ holds for all $x, y \in L(T(v))$.

**Proof.** Let $x, y \in L(T(v))$. Since $|\sigma(L(T(v)))| = 1$, we have $\sigma(x) = \sigma(y)$. Assume there is a vertex $q \in N^-(x)$. Then, we have $\sigma(x) \neq \sigma(q)$ and thus $q \notin L(T(v))$. Hence, $v \prec_T \text{lca}_T(x, q) = \text{lca}_T(y, q)$. Set $w = \text{lca}_T(x, q)$. Now $q \in N^-(x)$ implies $w \preceq_T u(q, \sigma(x))$. Since $\sigma(x) = \sigma(y)$ we have $u(q, \sigma(x)) = u(q, \sigma(y))$ and thus, $q \in N^-(y)$. By the same arguments, $q \in N^-(y)$ implies $q \in N^-(x)$. Therefore, we obtain $N^-(x) = N^-(y)$.

**Definition 7.1.** A graph $(G, \sigma)$ satisfies condition (N4) if, for all $x \in V(G)$ with $N(x) = \emptyset$, there is $y \in V(G) \setminus \{x\}$ such that $\sigma(y) = \sigma(x)$, and $N^-(x) = N^-(y)$.

**Proposition 7.12.** Let $(G, \sigma)$ be a 2-qBMG. Then the following two statements are equivalent.

(i) $(G, \sigma)$ is explained by a tree $T$ and a truncation map $u$ on $T$ satisfying $(M)$. 


(ii) $(G,\sigma)$ satisfies condition (N4).

Proof. Set $L = V(G)$ and let $\sigma : L \to S$ be a proper 2-coloring of $G$. Suppose first that $(G,\sigma)$ can be explained by $(T,\sigma,u)$ satisfying (M). Define $x_u := u(x,s)$ to be the truncation vertex of $x$ for the color $s \neq \sigma(x)$ in $T$, and consider the subtree $T(x_u)$. By the definition of the truncation maps, we have $x \preceq_T x_u$ and thus, $x \in L(T(x_u))$. Suppose $|\sigma(L(T(x_u)))| = 2$. Hence, there is at least one $z \in L(T(x_u))$ of color $\sigma(z) \neq \sigma(x)$ such that $z$ is a best match of $x$ and $\text{lca}(x,z) \preceq_T x_u$. Hence, $z \in N(x)$; contradicting the assumption that $N(x) = \emptyset$. Hence, $|\sigma(L(T(x_u)))| = 1$. Since $x_u$ is not a leaf (because $(T,\sigma,u)$ satisfies (M)) and since $T$ is phylogenetic, there is a leaf $y \in L(T(x_u))$ with $y \neq x$ and $|\sigma(L(T(x_u)))| = 1$ implies $\sigma(x) = \sigma(y)$. Together with Lemma 7.11 this yields $N^{-}(x) = N^{-}(y)$. We emphasize that $N^{-}(x) = N^{-}(y) = \emptyset$ is still possible.

For the converse, assume that the 2-qBMG $(G,\sigma)$ satisfies (N4). By assumption, $|\sigma(L)| = 2$ and thus $(G,\sigma)$ contains at least one vertex for each of two distinct colors and at least two vertices. As a consequence of the latter, every leaf of an explaining tree has a parent. We will prove the statement by constructing a finite sequence of trees $(T_i,\sigma,u_i)$ with $1 \leq i \leq k$, each of which explains $(G,\sigma)$ and such that $(T_i,\sigma,u_i)$ differs from its predecessor and successor, and $(T_k,\sigma,u_k)$ satisfies the desired property (M).

Since $(G,\sigma)$ is a qBMG, there is always a tree explaining $(G,\sigma)$. Denote this tree by $(T_1,\sigma,u_1)$. If $(T_1,\sigma,u_1)$ satisfies (M), $k = 1$ and the proposition follows. Otherwise, we construct $(T_2,\sigma,u_2)$ in the following way: Since $(T_1,\sigma,u_1)$ does not satisfy (M), there is at least one vertex $x \in L$ such that $x_u = x$ for the color $r \neq \sigma(x)$ and thus $N(x) = \emptyset$. Therefore, (N4) ensures that there is a vertex $y \in L' := L \setminus \{x\}$ such that $\sigma(y) = \sigma(x)$ and $N^{-}(x) = N^{-}(y)$. Set $w = \text{parent}_T(x)$. We distinguish two cases: (a) $\sigma(L(T(w))) = \{\sigma(x)\}$ and (b) $\{\sigma(x)\} \not\subseteq L(T(w))$.

In Case (a) set $T_2 = T_1$, $x_{u_2} = x_u$, and $z_{u_2} = z_u$ for all $z \in V(G) \setminus \{x\}$, and put $(G_2,\sigma) = \text{qBMG}(T_2,\sigma,u_2)$. We observe that $x$ is still a sink in $G_2$. By construction, the out-neighbors of all $x \in V(G) \setminus \{x\}$ also have not changed. Therefore, qBMG$(T_2,\sigma,u_2) = (G,\sigma)$.

In Case (b) $|L| \geq 3$ and thus $|L'| \geq 2$ because $G$ contains vertices $x$ and $y$ and at least one vertex of the opposite color. In order to construct $(T_2,\sigma,u_2)$, we first restrict the tree $T_1$ to $L'$ say $T'$, and then obtain $T_2$ from $T'$ by splitting the edge parent$_T(y)p$, i.e., we replace the edge by a newly-created vertex $p$ together with the edges parent$_T(y)p$ and $py$, and attaching $x$ as the second child of $p$. Recall from the proof of Lemma 3.5 that if $w' \in V(T_1) \setminus V(T')$, then $(V(T_1) \setminus V(T_2) \setminus \{x\})$ is only possible if either $w' = x$ or $w' = w$ and child$_T(w) = \{x,v^*\}$. Now take $z_{u_2} = z_u$, if $z_u \in V(T_2)$ (that is if either $z_u \neq w$ or $w$ has more than two children in $T_1$) and $z_{u_2} = v^*$ otherwise. Finally, set $x_u = p$. Our notation identifies corresponding vertices of $T_1$ and $T_2$. Moreover, observe that the ancestor order $\preceq_T$ is preserved with respect to $\preceq_T$ restricted to $T_2 \setminus \{p\}$. Now we show that $(T_2,\sigma,u_2)$ also explains $(G,\sigma)$. Denote by $(G',\sigma')$ the digraph explained by $(T_2,\sigma,u_2)$. Since $\sigma(x) = \sigma(y)$, we have by construction $|\sigma(L(T_2(p)))| = 1$. Together with Lemma 7.11, this implies $N_G'(x) = N_G'(y)$. Moreover, $N_G'(x) = \emptyset = N_G(x)$ follows from $x_u = p$. Therefore, we may restrict ourselves to vertices $x',y' \in L'$ with $\sigma(x') \neq \sigma(y')$. Observe that lca$_T(x',y') \not\in V(T_2)$ implies lca$_T(x',y') \preceq_T w$ and that $w$ has only two children $\{x,v^*\}$, which yields $x \in \{x',y'\}$; a contradiction. Thus, lca$_T(x',y') \in V(T_2)$.

Now suppose that $y' \in N_G'(x')$. Therefore, lca$_T(x',y') \preceq_T x'_u$. If $x'_u \in V(T_2)$, we have $x'_u = x_u'$. Otherwise, $x'_u = w$ and child$_T(w) = \{x,v^*\}$. This implies lca$_T(x',y') \preceq_T v^* \preceq_T w$ because lca$_T(x',y') = w$ occurs only if $x \in \{x',y'\}$. Therefore, lca$_T(x',y') \preceq_T x'_u$ in both cases. Now assume, for contradiction, that $y' \not\in N_G'(x')$. Since lca$_T(x',y') \preceq_T x'_u$, this implies that there is $y'' \in L \cap N_G(x')$ of color $\sigma(y'') = \sigma(y')$ such that lca$_T(x',y'') \preceq_T \text{lca}_T(x',y')$, contradicting $y' \in N_G(x')$. Now $x = y'' \in N_G(x')$ and $N_{G'}(x) = N_G'(y)$ imply $y \in N_G(x')$ and thus lca$_T(x',y') = \text{lca}_T(x',x) \preceq_T \text{lca}_T(x',y')$. Since $x',y' \in V(T_1)$ and lca$_T(x',y') \in V(T_1)$ implies lca$_T(x',y') \preceq_T \text{lca}_T(x',x)$, which together with $\sigma(y) = \sigma(x) = \sigma(y')$ contradicts $y'' \in N_G(x')$. Hence, $y' \not\in N_G(x')$.

Now assume $y' \in N_G(x')$. Thus either $x'_u \preceq_T \text{lca}_T(x',y')$ or there is a $y'' \in L$ with $\sigma(y'') = \sigma(y')$ such that lca$_T(x',y'') \preceq_T \text{lca}_T(x',y')$. Suppose $x'_u = \text{lca}_T(x',y')$. Then either $x'_u \in V(T_2)$ or $x'_u \not\in V(T_2)$. In the first case, $x'_u = x_u' \preceq_T \text{lca}_T(x',y')$. Since $v^* \in V(T_2)$, we obtain $x'_u \preceq_T \text{lca}_T(x',y')$, and thus, $y' \not\in N_G(x')$. Finally, assume there is $y'' \in N_G(x')$ with $y'' \neq y'$. If $y'' \neq x$, then lca$_T(x',y'')$, lca$_T(x',y') \in V(T_2)$ implies lca$_T(x',y'') \preceq_T \text{lca}_T(x',y')$, and thus $y' \not\in N_G(x')$. On the other hand, if $y'' = x$, then $N_G(x) = N_G(y)$ implies $y \in N_G(x')$, and thus lca$_T(x',y') = \text{lca}_T(x',x')$. Since lca$_T(x',y') \in V(T_2)$, we obtain lca$_T(x',y') = \text{lca}_T(x',y'') \preceq_T \text{lca}_T(x',y')$ which, together with $\sigma(y) = \sigma(x)$, yields $y' \not\in N_G(x')$. 22
On the other hand, there is some \( z \prec v \) thus \( N^-(y) = N^-(y') \). Hence, Property (N4) is not hereditary.

In summary, we have shown that \((G, \sigma) = (G', \sigma)\), and hence \((T_2, \sigma, u_2)\) also explains \((G, \sigma)\). We show that, for every \( z \in L \), \( z_{u_2} \) cannot be a leaf whenever \( z_{u_1} \) is not a leaf. Observe that \( z \prec_{T_1} z_{u_1} \) implies \( z \neq x \). If \( z_{u_1} \in V(T_2) \), then \( z_{u_2} = z_{u_1} \) and \( z_{u_2} \) is clearly still an inner vertex. On the other hand, \( z_{u_1} \notin V(T_2) \) ensures \( z_{u_1} = w \) and \( \text{child}_{T_1}(w) = \{ x, v^* \} \). Since \( z \neq x \) we have \( z \geq_{T_1} v^* \). If \( \sigma(z) = \sigma(x) \), then \( \{ \sigma(x) \} \neq \sigma(L(T_1(w))) \) implies that there exists \( z' \in L(T_1(v^*)) \) such that \( \sigma(z') = \sigma(z) \) and thus \( z' \neq z \). Hence, \( v^* = z_{u_2} \) is not a leaf. On the other hand, if \( \sigma(z) \neq \sigma(x) \), then there is some \( v^{**} \geq_{T_1} v^* \) with \( \sigma(v^{**}) = \sigma(x) \) implying that \( v^{**} \) is an inner vertex or, if not, \( x \notin N_G(z) \) as \( \text{lca}_{T_1}(z, x) = w = x_{u_1} \). In the latter case, \( N_G^-(x) = N_G^-(y) \) yields \( y \notin N_G(z) \), and thus \( w = \text{lca}_{T_1}(z, x) = \text{lca}_{T_1}(z, y) \). However, \( x \neq y \) and \( w = \text{parent}_{T_2}(x) \) imply that there is \( v^{**} \in \text{child}_{T_2}(w) \setminus \{ x, v^* \} \); a contradiction. Hence, the case \( z_{u_1} \notin V(T_2) \) and \( \sigma(z) \neq \sigma(x) \) cannot occur.

In both Cases (a) and (b), therefore, \((T_2, \sigma, u_2)\) explains \((G, \sigma)\) and \( z_{u_2} \) can only be a leaf if \( z_{u_1} \) is a leaf. Furthermore, \( z_{u_1} \) is not a leaf, while \( z_{u_1} = x \), i.e., a leaf. Therefore we have decreased the number of truncation vertices that are leaves. Since the number of leaves is finite, we can repeat this procedure and eventually obtain a tree \((T_k, \sigma, u_k)\) for which \( z_{u_k} \notin L \) for all leaves \( z \). Thus the tree \((T_k, \sigma, u_k)\) satisfies property (M).

As a simple example, consider the digraph of two vertices with a single edge. It satisfies (N1), (N2), and (N3), and thus is a qBMG, but it violates condition (N4) and thus has no explanation without a truncation at a leaf. Proposition 7.12 implies that, for a 2-qBMG \((G, \sigma)\) that satisfies (N4), we always find an explaining leaf-colored tree with truncation map \((T, \sigma, u)\) that satisfies (M). However, a 2-qBMG \((G, \sigma)\) satisfying (N4) may also have explanations that do not satisfy (N4), as shown by the example in Figure 7.

Moreover, like sink-freeness, Property (N4) is not hereditary. To see this, let \( x \) be a sink and suppose that \( \hat{x} \in L' \) is the only vertex with \( \sigma(x) = \sigma(\hat{x}) \) and \( N^-(\hat{x}) = N^-(x) \). Then \((G, \sigma)[L']\) is still a qBMG Lemma 3.6. We have \( \hat{x} \notin N^-(x') \) for any \( x' \in L' \) of color \( \sigma(x) \) because of \( \sigma(x) = \sigma(\hat{x}) \), and thus \( N^-(x') \) remains unchanged by deleting \( \hat{x} \). Hence, \((G, \sigma)[L']\) does not satisfy (N4).

8 Concluding Remarks and Open Questions

In this contribution, we investigated a generalization of best match graphs of [1], which we termed qBMGs. In the two-colored case, which in particular occurs in the form of the subgraph induced by two vertex colors, qBMGs are characterized by three simple conditions, (N1), (N2), and (N3). The first two conditions are already sufficient to ensure that the graph is bipartite. In the general case, qBMGs are characterized by their induced subgraphs on three vertices with two colors, which translate to sets of “informative” and “forbidden” triples. It is, therefore, possible to recognize \( \ell \)-colored qBMGs in polynomial time. In the positive case, an explaining tree and a corresponding truncation map can also be constructed in polynomial time.

In contrast to BMGs, being a qBMG is a hereditary property. On the other hand, BMGs have unique least resolved trees (LRTs), a property that is no longer true for qBMGs in general. It will be interesting to ask how different alternative LRTs can become. Since they must display the informative triples, one would expect that there is some common “core”. The uniqueness of LRTs on BMGs suggests considering (maximal) induced subgraphs that are BMGs. Are their LRTs displayed by all LRTs of the qBMG? Is there an efficient algorithm to find all maximal induced BMGs in a qBMG?
From an application point of view, the truncation map $u(x, r)$ describes the phylogenetic scope within which a homolog $y$ of the query gene $x$ can be found. If the target genomes have similar size and organization, it becomes a reasonable approximation to assume that phylogenetic scope depends only on the dissimilarity between query and target gene but not on the identity of the target genome. In this case, the truncation map becomes independent of the color, i.e., we have $u(x, r) = u(x)$ for all $r \neq \sigma(x)$. Assuming further that all genes have similar size and internal structure further restricts $u$ to “cutting” the tree at a certain height. It remains an open question whether these subclasses of qBMGs also have interesting mathematical properties. In the special case of two colors, a simple characterization was obtained for qBMGs that can be explained by truncation maps that exclude the leaves. Quasi-best match graphs that can be explained by a tree $(T, \sigma, u)$ satisfying (M) are also of interest for more than two colors.

In Corollary 7.10 we characterized the digraphs $G$ that admit a 2-coloring $\sigma$ such that $(G, \sigma)$ is a 2-qBMG. Naturally, one might want to ask which directed graphs admit an $\ell$-coloring such that $(G, \sigma)$ is an $\ell$-qBMG. On the other extreme, setting $S = V(G)$ and using coloring $\sigma(x) = x$, we can use the rooted star tree with truncation map $u(x, \sigma(y)) = r$ if $xy \in E(G)$ and $u(x, \sigma(y)) = x$ if $xy \notin E(G)$ as explanation for $(G, \sigma)$. Thus every directed graph $G$ can be colored to be a $|V(G)|$-qBMG.

This suggests to consider the qBMG-coloring problem: Given a digraph $G$ and an integer $1 \leq \ell \leq |V(G)|$, is there a coloring $\sigma$ with $\sigma(V) = \ell$ such that $(G, \sigma)$ is an $\ell$-BMG? Proposition 4.6, furthermore, shows that every coloring $\sigma$ of $G$ for which $(G, \sigma)$ is an $\ell$-qBMG with $\ell < |V(G)|$ can be transformed into an $(\ell + 1)$ coloring $\sigma'$ by arbitrarily splitting a color class resulting in an $(\ell + 1)$-qBMG. Thus, for every digraph, there is a minimum integer $\ell_qBMG$, the “qBMG-chromatic number”, such that an $\ell_qBMG$-coloring $\sigma$ exists for which $(G, \sigma)$ is a qBMG. Figure 8 shows an example of a bipartite digraph $G$ that can be colored to be a 3-qBMG, while there is no 2-coloring $\sigma$ such $(G, \sigma)$ is a 2-qBMG.

Acknowledgments. This work was support in part by the German Research Foundation (DFG, STA 850/49-1).

References

[1] M. Geiß, E. Chávez, M. González Laffitte, A. López Sánchez, B. M. R. Stadler, D. I. Valdivia, M. Hellmuth, M. Hernández Rosales, and P. F. Stadler, “Best match graphs,” J. Math. Biol., vol. 78, pp. 2015–2057, 2019.

[2] M. Geiß, M. E. González Laffitte, A. López Sánchez, D. I. Valdivia, M. Hellmuth, M. Hernández Rosales, and P. F. Stadler, “Best match graphs and reconciliation of gene trees with species trees,” J. Math. Biol., vol. 80, pp. 1459–1495, 2020.

[3] D. Schaller, P. F. Stadler, and M. Hellmuth, “Complexity of modification problems for best match graphs,” Theor. Comp. Sci., vol. 865, pp. 63–84, 2021.

[4] D. Schaller, M. Geiß, M. Hellmuth, and P. F. Stadler, “Best match graphs with binary trees,” in Algorithms for Computational Biology, 8th AlCoB (C. Martín-Vide, M. A. Vega-Rodríguez, and T. Wheeler, eds.), vol. 12715 of Lect. Notes Comp. Sci., pp. 82–93, 2021.

Figure 8: Example of a bipartite digraph $G$ with a qBMG-chromatic number of 3. The only proper 2-coloring (up to exchanging the colors) is $\sigma_2$. However, $(G, \sigma_2)$ violates both (N1) and (N2) and thus, it is not a qBMG. For $\sigma_3$, we can easily find a truncation map $u$ such that $(G, \sigma_3) = qBMG(T, \sigma_3, u)$ by setting $u(x, \sigma_3(z)) = u(y, \sigma_3(v)) = u(x, \sigma_3(y)) = u(z, \sigma_3(x)) = \rho_T$, $u(x, \sigma_3(z)) = x$, $u(v, \sigma_3(y)) = v$, $u(z, \sigma_3(y)) = z$ and $u(y, \sigma_3(z)) = y$. 

2
[5] D. Schaller, M. Geiß, E. Chávez, M. González Laffitte, A. López Sánchez, B. M. R. Stadler, D. I. Valdivia, M. Hellmuth, M. Hernández Rosales, and P. F. Stadler, “Corrigendum to “Best Match Graphs”,” *J. Math. Biol.*, vol. 82, p. 47, 2021.

[6] W. M. Fitch, “Distinguishing homologous from analogous proteins,” *Syst Zool.*, vol. 19, pp. 99–113, 1970.

[7] W. M. Fitch, “Homology: a personal view on some of the problems,” *Trends Genet.*, vol. 16, pp. 227–231, 2000.

[8] E. Koonin, “Orthologs, paralogs, and evolutionary genomics,” *Ann. Rev. Genetics*, vol. 39, pp. 309–338, 2005.

[9] A. M. Altenhoff, R. A. Studer, M. Robinson-Rechavi, and C. Dessimoz, “Resolving the orthology conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs,” *PLoS Comp. Biol.*, vol. 8, no. e1002514, 2012.

[10] T. Gabaldón and E. V. Koonin, “Functional and evolutionary implications of gene orthology,” *Nat. Rev. Genet.*, vol. 14, pp. 360–366, 2013.

[11] D. Frédéric, H. Brinkmann, and H. Philippe, “Phylogenomics and the reconstruction of the tree of life,” *Nature Rev. Gen.*, vol. 6, pp. 361–375, 2005.

[12] A. D. Young and J. P. Gillung, “Phylogenomics – principles, opportunities and pitfalls of big-data phylogenetics,” *Syst. Entomology*, vol. 45, pp. 225–247, 2020.

[13] L. A. Mirny and M. S. Gelfand, “Using orthologous and paralogous proteins to identify specificity determining residues,” *Genome Biol.*, vol. 3, p. preprint0002.1, 2002.

[14] B. T. L. Nichio, J. N. Marchaukoski, and R. T. Raittz, “New tools in orthology analysis: A brief review of promising perspectives,” *Front. Genet.*, vol. 8, p. 165, 2017.

[15] L. Y. Rusin, E. Lyubetskaya, K. Y. Gorbunov, and V. Lyubetsky, “Reconciliation of gene and species trees,” *BioMed Res. Int.*, vol. 2014, p. 642089, 2014.

[16] J. C. Setubal and P. F. Stadler, “Gene phylogenies and orthologous groups,” in *Comparative Genomics* (J. C. Setubal, P. F. Stadler, and J. Stoye, eds.), vol. 1704, pp. 1–28, Heidelberg: Springer, 2018.

[17] S. Karlin and S. F. Altschul, “Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes,” *Proc. Natl. Acad. Sci. USA*, vol. 87, pp. 2264–2268, 1990.

[18] A. Ódám, “On certain partitions of finite directed graphs and of finite automata,” *Acta Cybernetica*, vol. 6, pp. 331–346, 1984.

[19] H. Cohn, R. Pemantle, and J. G. Propp, “Generating a random sink-free orientation in quadratic time,” *Electr. J. Comb.*, vol. 9, p. R10, 2002.

[20] G. Abrams and J. K. Sklar, “The graph menagerie: Abstract algebra and the mad veterinarian,” *Math. Mag.*, vol. 83, pp. 168–179, 2010.

[21] S. Das, P. Ghosh, S. Ghosh, and S. Sen, “Oriented bipartite graphs and the goldbach graph,” *Discrete Math.*, vol. 344, p. 112497, 2021.

[22] A. Korchmaros, “The structure of 2-colored best match graphs,” *Discr. Appl. Math.*, vol. 304, pp. 397–416, 2021.

[23] A. Korchmaros, “Circles and paths in 2-colored best match graphs,” Tech. Rep. math.CO/2006.04100v1, arXiv, 2020.

[24] P. C. Hammer, “Extended topology: Structure of isotonic functions,” *J. reine angew. Math.*, vol. 213, pp. 174–186, 1964.
[25] D. Bryant and M. Steel, “Extension operations on sets of leaf-labeled trees,” *Adv. Appl. Math.*, vol. 16, pp. 425–453, 1995.

[26] A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman, “Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions,” *SIAM J Comput*, vol. 10, pp. 405–421, 1981.

[27] C. Semple and M. Steel, *Phylogenetics*. Oxford UK: Oxford University Press, 2003.

[28] J. E. Hernández-Salmerón and G. Moreno-Hagelsieb, “Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and mmseqs2,” *BMC Genomics*, vol. 21, p. 741, 2020.

[29] Y.-J. He, T. N. D. Huynh, J. Jansson, and W.-K. Sung, “Inferring phylogenetic relationships avoiding forbidden rooted triplets,” *J. Bioinf. Comp. Biol.*, vol. 4, pp. 59–74, 2006.

[30] P. Górecki and O. Eulenstein, “Refining discordant gene trees,” *BMC Bioinformatics*, vol. 15, p. S3, 2014.

[31] E. Sayyari and S. Mirarab, “Testing for polytomies in phylogenetic species trees using quartet frequencies,” *Genes*, vol. 9, p. 132, 2018.

[32] D. Schaller, M. Geiß, P. F. Stadler, and M. Hellmuth, “Complete characterization of incorrect orthology assignments in best match graphs,” *J. Math. Biol.*, vol. 82, p. 20, 2021.