7 SUPPLEMENTARY MATERIALS

7.1 Definitions

In this subsection presents the definitions [11] used in the paper.

Definition 7.1 (Blocked Paths [11]). *Given a Bayesian network, a path between two nodes X and Y is blocked by a set Z of nodes if*

1. *Either that path contains a node Z that is in Z and the connection at Z is either serial or diverging.*
2. *Or that the path contains a node W such that W and its descendants are not in Z and the connection at W is a converging connection.*

Definition 7.2 (d-separation [11]). *Two nodes X and Y are d-separated by a set Z if all paths between X and Y are blocked by Z.*

Definition 7.3 (I-map [11]). *A DAG $\mathcal{G}$ is an I-map of a distribution $\mathcal{P}$ over a set of random variables V, if for any three disjoint subsets of variables X, Y, and Z, Z d-separates X and Y in $\mathcal{G}$ implies $X \perp\!\!\!\perp_P Y|Z$.*

Definition 7.4 (D-map [11]). *A DAG $\mathcal{G}$ is a D-map of a distribution $\mathcal{P}$ over a set of random variables V, if for any three disjoint subsets of variables X, Y, and Z, $X \perp\!\!\!\perp_P Y|Z$ implies Z d-separates X and Y in $\mathcal{G}$.*

Definition 7.5 (perfect map [11]). *A DAG $\mathcal{G}$ is a perfect map of a distribution $\mathcal{P}$ if it is both an I-map and a D-map.*

7.2 Proofs

In this subsection we present proofs of lemmas and theorem in Section 3.

Lemma 3.3. *Let $\mathrm{MB}_{\mathcal{B}'}(\Phi_t)$ be a Markov blanket of Φ_t in the surrogate network $\mathcal{B}'$ that does not contain any contribution variables, denote $\mathcal{M}$ as the subset of $\mathrm{MB}_{\mathcal{B}'}(\Phi_t)$, which includes all feature variables only; mathematically,*

$$\mathcal{M} = \left\{F_{v_i}^{(j)} : \exists j \in [0..d_{\zeta(v_i)}] \text{ s.t. } F_{v_i}^{(j)} \in \mathrm{MB}_{\mathcal{B}'}(\Phi_t),\ \text{for all } v_i \text{ in } G_t\right\},$$

and let $\mathcal{V}$ be the set of vertexes in the input graph that are associated with variables in $\mathrm{MB}_{\mathcal{B}'}(\Phi_t)$, i.e.

$$\mathcal{V} = \left\{v_i : \exists V_i \in \mathrm{MB}_{\mathcal{B}'}(\Phi_t)\right\}. \tag{9}$$

If for all $v_i \in \mathcal{V}$, there exists a subset $\mathcal{M}' \subseteq \mathcal{M}\backslash\mathcal{F}_{\phi,t}(v_i)$ such that

$$\Phi_t \perp\!\!\!\perp_{\mathcal{B}'} F_{v_i}^{(j)} \Big| \mathcal{M}',\ \forall F_{v_i}^{(j)} \in \mathcal{K}_{\Phi_t}(v_i)\backslash\mathcal{F}_{\phi,t}(v_i), \tag{10}$$

where $\mathcal{F}_{\phi,t}(v_i) = \left\{F_{v'}^{(j)} \in \mathrm{MB}_{\mathcal{B}'}(\Phi_t) : v' = v_i, j \in [0..d_{\zeta(v_i)}]\right\}$ and $\mathcal{K}_{\Phi_t}(v_i)$ is defined in Equation (5), then $\mathcal{M}$ is a Markov blanket of Φ_t in $\mathcal{B}^*$.

Proof. The lemma can be proved by contradiction. Assume $\mathcal{M}$ is *not* a Markov blanket of Φ_t in $\mathcal{B}^*$, then $\mathcal{M}$ cannot d-separate Φ_t from all other variables in $\mathcal{B}^*$ that are not in $\mathcal{M}$, which means there exists at least one active path p connecting to Φ_t that is not blocked by $\mathcal{M}$. Since Φ_t is a leaf node in $\mathcal{B}^*$, p must contain a direct parent of Φ_t, denoted by $\bar{F}$. Note that $\bar{F}$ is not in $\mathcal{M}$ as assumed. Since $\bar{F}$ is a direct parent of Φ_t in $\mathcal{B}^*$, it is connected to Φ_t in $\mathcal{B}'$ by a contribution variable $\bar{\xi}$ and a synthetic variable $\bar{V}$ according to the rules of constructing $\mathcal{B}'$, then $\bar{V}$ is a direct parent of Φ_t in $\mathcal{B}'$ and $\bar{V}$ must be included in $\mathrm{MB}_{\mathcal{B}'}(\Phi_t)$ because direct parents are always included in the Markov blanket [27]. Hence the vertex $\bar{v}$ associated with $\bar{V}$ is included in $\mathcal{V}$ and there exists a subset $\mathcal{M}' \subseteq \mathcal{M}\backslash\mathcal{F}_{\phi,t}(\bar{v})$ such that $\Phi_t \perp\!\!\!\perp_{\mathcal{B}'} \bar{F} \big| \mathcal{M}'$ since $\bar{F} \in \mathcal{K}_{\Phi_t}(\bar{v})\backslash\mathcal{F}_{\phi,t}(\bar{v})$, which means $\mathcal{M}'$ d-separates $\bar{F}$ and Φ_t in $\mathcal{B}'$, namely, all paths from $\bar{F}$ to Φ_t are blocked by $\mathcal{M}'$. However, there is a path from $\bar{F}$ to Φ_t : $\bar{F} \to \bar{\xi} \to \bar{V} \to \Phi_t$, while $\mathcal{M}$ does not contain any contribution variable ξ_i^j or synthetic variable V_i, hence the above path can never be blocked by any subset of $\mathcal{M}$, which means there does not exist a subset $\mathcal{M}' \subseteq \mathcal{M}$ that can d-separate $\bar{F}$ from Φ_t. At this point, we reach a contradiction. In conclusion, there does not exists such an active path p in $\mathcal{B}^*$ that is not blocked by $\mathcal{M}$, and $\mathcal{M}$ is a Markov blanket of Φ_t in $\mathcal{B}^*$. □

Lemma 3.4. *In the surrogate network $\mathcal{B}'$, for any synthetic variable V_i, let Z be a set of variables that does not contain any of its grandparent feature variable $F_{v_i}^{(\cdot)}$'s or contribution variables, if there exists some $F_{v_i}^{(j)}$ such that $F_{v_i}^{(j)} \not\perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$, then $V_i \not\perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$.*

Proof. Since $F_{v_i}^{(j)} \not\perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$, then Z cannot d-separate $F_{v_i}^{(j)}$ from Φ_t, there exists at least one active path p in $\mathcal{B}'$ connecting $F_{v_i}^{(j)}$ and Φ_t that is not blocked by Z, which is in either of the below cases:

(1) p connects $F_{v_i}^{(j)}$ to Φ_t through $F_{v_i}^{(j)} \to \xi_i^j \to V_i$; or

(2) p connects $F_{v_i}^{(j)}$ to Φ_t through some parent node of $F_{v_i}^{(j)}$.

In case (1), V_i is obviously not blocked from Φ_t as it is contained in p. In case (2), V_i is also connected to Φ_t through a path p' formed by concatenating p with $F_{v_i}^{(j)} \to \xi_i^j \to V_i$. As p is connect to $F_{v_i}^{(j)}$ though its parent, $F_{v_i}^{(j)} \to \xi_i^j \to V_i$ and p have a junction at $F_{v_i}^{(j)}$ with *chain* pattern, *i.e.* it is a serial path and it has no node contained in Z, hence p' extends p to connect V_i while remaining an active path that is not blocked by Z [28]. In conclusion, if $F_{v_i}^{(j)} \not\perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$, then there exists an active path in $\mathcal{B}'$ connecting V_i and Φ_t that is not blocked by Z, thus $V_i \not\perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$. □

To illustrate the two cases in the proof, take V_2 in Figure 2(d) as an example. For case (1), p can be **P2**-"*DB2*"$\to \xi_2^2 \to V_2 \to$ **A1**-"*DB*" $\to \xi_4^1 \to V_4 \to \Phi_t$, then V_2 is connect to Φ_t. For case (2), p' can be $V_2 \leftarrow \xi_2^2 \leftarrow$ **P2**-"*DB2*" $\leftarrow V_1 \to$ **P3**-"*DB2*" $\to \xi_3^4 \to V_3 \to \Phi_t$; note that at V_1, the junction pattern is *fork*, where p' is not blocked as long as V_1 is not in Z [28]. Lemma 3.5. *In the surrogate network $\mathcal{B}'$, for any synthetic variable V_i, let Z be a set of variables that does not contain any of its grandparent feature variable $F_{v_i}^{(\cdot)}$'s or contribution variables, if $V_i \perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$, then all its grandparent feature variable and parent contribution variables are independent from the prediction variable conditioned on Z, namely, $F_{v_i}^{(j)} \perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$ and $\xi_i^j \perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$ for all $j \in [0..d_{\zeta(v_i)}]$.*

Proof. The *converse-negative proposition* of Lemma 3.4 shows that under the stated condition, *if* $V_i \perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$, *then* $F_{v_i}^{(j)} \perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z, \forall F_{v_i}^{(j)} \in \mathcal{K}_{\Phi_t}(v_i)$. Hence, to prove Lemma 3.5, it suffices to prove that all parent contribution variables of V_i are independent from the prediction variable conditioned on Z, *i.e.* $\xi_i^j \perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$. According to the building rules of the surrogate network, for any synthetic variable in $\mathcal{B}'$, all its parent contribution ξ_i^j are contained

in one serial path *only*, that is $V_{i'} \to F^{(j)}_{v_i} \to \xi^j_i \to V_i \to F^{(j')}_{v_{i''}}$. If V_i is d-separated from from Φ_t by Z, then the above path is blocked by Z, which means ξ^j_i is d-separated from Φ_t by Z, hence $\xi^j_i \perp\!\!\!\perp_{\mathcal{B}'} \Phi_t \big| Z$. □

Lemma 3.6. *Given a pretrained DGN ϕ and a target node t, assumes there exists a perfect map $\mathcal{B}^*$ of the model behavior distribution $\mathcal{P}_\phi(t)$ and every necessary vertex in G_t is connected to the target node through other necessary vertex(es). Let $\mathcal{E}'$ be a set of vertexes, whose induced subgraph of G_t is a connected component that contains the target t, denote its neighborhood as $\boldsymbol{Ne}(\mathcal{E}') = \bigcup_{v\in\mathcal{E}'} Adj[v] \big\backslash \mathcal{E}'$, where $Adj[v]$ is the set of all adjacent neighbor(s) of the vertex v; and denote the set of synthetic variables and feature variables induced by $\mathcal{E}'$ as*

$$\begin{aligned}\mathcal{Z}(\mathcal{E}') = &\left\{V_i : v_i \in \mathcal{E}' \cup \boldsymbol{Ne}(\mathcal{E}')\right\} \\ &\bigcup \left\{F^{(j)}_{v'} \in \mathcal{K}_{\Phi_t}(v') : v' \in \mathcal{E}' \cup \boldsymbol{Ne}(\mathcal{E}')\right\}.\end{aligned} \tag{11}$$

If for all $V' \in \{V' : v' \in \boldsymbol{Ne}(\mathcal{E}')\}$, there exists a subset $\mathcal{M}' \subseteq \mathcal{Z}(\mathcal{E}')$ such that $\Phi_t \perp\!\!\!\perp_{\mathcal{B}'} V' \big| \mathcal{M}'$, or $\Phi_t \perp\!\!\!\perp_{\mathcal{B}'} F^{(j)}_{v_i} \big| \mathcal{M}'$ for all grandparent feature variables of V', then a Markov blanket of Φ_t can be found as a subset of $\mathcal{Z}(\mathcal{E}')$.

Proof. Under the stated condition of this lemma, $\mathcal{E}'$ separates all random variables in $\mathcal{P}_{\Phi_t}(t)$ into two groups based on the topology of the input graph. To prove $\mathcal{Z}$ contains a complete Markov blanket, one needs to firstly prove all contribution variables related to vertexes in $\mathcal{E}'$ are conditionally independent with Φ_t; secondly, all random variables that are not related to vertexes in $\mathcal{E}'$ are also conditionally independent with Φ_t. The former is obvious as, for any contribution variable ξ^j_i, there is only one serial path that contains it (see proof of Lemma 3.5), thus the path can always be blocked by any set that contains its adjacent neighbors. The second statement can be proved by contradiction. Assume $\mathcal{Z}$ does not contain a complete Markov blanket of Φ_t, then there exists at least one necessary variable $\bar{V}$ outside $\mathcal{Z}$. It can be either

(1) associated with a vertex $\bar{v}$ in $\boldsymbol{Ne}(\mathcal{E}')$; or
(2) associated with a vertex $\bar{v}$ outside $\boldsymbol{Ne}(\mathcal{E}') \bigcup \mathcal{E}'$.

In case (1), there exists a set $\mathcal{M}' \subseteq \mathcal{Z}$ such that (*a*) $\Phi_t \perp\!\!\!\perp_{\mathcal{B}'} V' \big| \mathcal{M}'$, then $\bar{V}$ can be excluded from a Markov blanket, which contradicts with the assumption of $\bar{V}$ being a necessary variable; or (*b*) $\Phi_t \perp\!\!\!\perp_{\mathcal{B}'} F^{(j)}_{\bar{v}} \big| \mathcal{M}'$ for all grandparent feature variables of $\bar{V}$, then every $F^{(j)}_{\bar{v}}$ is d-separated by $\mathcal{M}'$ from Φ_t, thus $\bar{V}$ must be d-separated by $\mathcal{M}'$ from Φ_t as well, otherwise $F^{(j)}_{\bar{v}}$'s can be connected to Φ_t through $\bar{V}$ (see Proof. of Lemma 3.4), hence $\bar{V}$ can be excluded from a Markov blanket and contradicts with the assumption again. In case (2), because all necessary vertex is connected to the target by other necessary ones, then there must exist a path in the input graph composed of necessary vertexes that connects $\bar{v}$ to t, hence there must exist some necessary variable in $\boldsymbol{Ne}(\mathcal{E}')$, which is exactly case (1). Combining both cases, we prove that there does not exist a necessary variable outside $\mathcal{Z}$, then all synthetic variables outside $\mathcal{Z}$ are independent with Φ_t conditioned on some subset of $\mathcal{Z}$. According to Lemma 3.5, all their parent contribution variables and grandparent feature variables are independent from Φ_t conditioned on the same set. In conclusion, any variable outside $\mathcal{Z}$ is conditionally independent with Φ_t. To summarize, we have proved all contribution variables related to vertexes in $\mathcal{E}'$ and all other variables that are *not* related to vertexes in $\mathcal{E}'$ are both conditionally independent with Φ_t, and one can find a Markov blanket of Φ_t in $\mathcal{Z}$ under the stated condition. □

Theorem 3.8. *Given a pretrained DGN ϕ and a target node t, assumes there exists a perfect map $\mathcal{B}^*$ of the model behavior distribution $\mathcal{P}_\phi(t)$ and every necessary vertex in G_t is connected to the target node through other necessary vertex(es), running Algorithm 1 to solve the heterogeneity-agnostic multi-level explanation generation problem defined in Definition 2.3 outputs the topological-level explanation $\mathcal{E}_{topo}$. Let $\mathcal{R} = \{V_i : \exists v_i \in \mathcal{E}_{topo}\}$, then $\mathcal{R}$ is the set of all direct parents of Φ_t in $\mathcal{B}'$, i.e. $\mathcal{R} = \boldsymbol{Pa}(\Phi_t)$.*

Proof. We first prove (i) $\boldsymbol{Pa}(\Phi_t) \subseteq \mathcal{R}$. It has been proven that $\{\hat{V}\text{'s}\} \bigcup \mathcal{M}$ is a Markov blanket of Φ_t in $\mathcal{B}'$ in Theorem 3.7, then it must contain all direct parents of Φ_t by definition (direct parents are always included in Markov blanket). $\mathcal{R}$ differ from $\{\hat{V}\text{'s}\}$ in that it does not include those synthetic variables whose grandparent feature variables are conditionally independent with Φ_t, which means these feature variables are not direct parents of Φ_t in the original perfect map $\mathcal{B}^*$, then the corresponding synthetic variables are not direct parents of Φ_t in $\mathcal{B}'$ either according to the building rule of surrogate network $\mathcal{B}'$. Next, we prove (ii) $\mathcal{R} \subseteq \boldsymbol{Pa}(\Phi_t)$ by contradiction, if there exist a variable in $\mathcal{R}$ that is not in $\boldsymbol{Pa}(\Phi_t)$, denote the variable $\bar{V}$, then $\bar{V}$ is not a direct parent of Φ_t in $\mathcal{B}'$, then all its grandparent feature variables are not direct parent of Φ_t in $\mathcal{B}^*$. Hence, $\bar{V}$ can be d-separated from Φ_t by some subset of $\mathcal{M}$, which contradict with *line 21* in the algorithm. Combining (i) and (ii), we prove that $\mathcal{R} = \boldsymbol{Pa}(\Phi_t)$. □