A A SUPPLEMENTARY MATERIAL TO GENERAL IDENTIFIABILITY WITH ARBITRARY SURROGATE EXPERIMENTS

A.1 DERIVATION

We derive an expression for Fig. 1a as follows

$$\begin{aligned}
P_{x_1,x_2}(y) &= \sum_w P_{x_1,x_2}(y,w) \\
&= \sum_w P_{w,x_1,x_2}(y) P_{y,x_1,x_2}(w) \\
&= \sum_w P_{x_2,w,x_1}(y) P_{x_1}(w) \\
&= \sum_w P_{x_2,w}(y) P_{x_1}(w) \\
&= \sum_w P_{x_2}(y|w) P_{x_1}(w)
\end{aligned}$$

The query $P_{x_1,x_2}(y)$ is rewritten as $\sum_w P_{x_1,x_2}(w,y)$ and factorized $\sum_w P_{w,x_1,x_2}(y)P_{y,x_1,x_2}(w)$ based on c-component form. For the first term, by Rule 3 and 2 of do-calculus, $P_{x_2,w,x_1}(y) = P_{x_2,w}(y) = P_{x_2}(y|w)$. For the second term, $P_{y,x_1,x_2}(w) = P_{x_1}(w)$ by Rule 3 of do-calculus. Hence, $P_{x_1,x_2}(y) = \sum_w P_{x_2}(y|w)P_{x_1}(w)$.

For Fig. 2a, it only requires a single application of Rule 3 of do-calculus. Simply put, intervened variables outside the ancestors of an outcome variable have no effect on the outcome variable. Hence, $P_{x_1,x_2}(y_1) = P_{x_1}(y_1)$ and $P_{x_1,x_2}(y_2) = P_{x_2}(y_2)$.

A.2 NON-IDENTIFIABILITY MAPPING

Lemma 9. *Let* $\mathbf{X}$, $\mathbf{Y}$ *be disjoint sets of variables in* $\mathcal{G}$. *Let* $\mathcal{J}$ *be a nonempty subgraph of* $\mathcal{G}$ *with root set* $\mathbf{R}$, *where* $\mathbf{R} \subseteq An(\mathbf{Y})_{\mathcal{G}_{\underline{\mathbf{X}}}}$. *Let* $\mathcal{M}_1$ *and* $\mathcal{M}_2$, *which are compatible with* $\mathcal{J}$, *satisfy*

$$\sum_{\mathbf{r}|\bigoplus\mathbf{r}=1} P^1_{\mathbf{x}\cap\mathcal{J}}(\mathbf{r}) \neq \sum_{\mathbf{r}|\bigoplus\mathbf{r}=1} P^2_{\mathbf{x}\cap\mathcal{J}}(\mathbf{r})$$

for some $\mathbf{x}$ *where all variables in* $\mathbf{R}$ *are binary. Then, there are two models* $\mathcal{M}'_1$ *and* $\mathcal{M}'_2$ *compatible with* $\mathcal{G}$ *such that* $P'^1_{\mathbf{x}}(\mathbf{y}) \neq P'^2_{\mathbf{x}}(\mathbf{y})$ *for some* $\mathbf{y}$.

Proof. Similar results appear in identifiability literature, e.g., [Shpitser and Pearl, 2006, Thm. 4]. We first employ their strategies in the proof, and discuss about some theoretical oversight. By the condition $An(\mathbf{Y})_{\mathcal{G}_{\underline{\mathbf{X}}}}$, there exist directed downward paths from $\mathbf{R}$ to $\mathbf{Y}$ where no $\mathbf{X}$ appear in-between and each node has at most one child. That is, one can parametrize each node (which is binary) in the paths as an exclusive-or of its observable parents. Then, the discrepancy in bit-parity for $\mathbf{R}$ in $\mathcal{M}_1$ and $\mathcal{M}_2$ will also be happened at $\mathbf{Y}$ in $\mathcal{M}'_1$ and $\mathcal{M}'_2$ under $do(\mathbf{x})$ (n.b. values of $\mathbf{x}$ outside $\mathcal{J}$ are irrelevant to $\mathbf{Y}$).

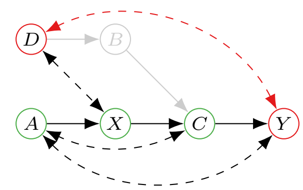

Figure 6: A causal graph $\mathcal{G}$ with a hedge $\langle\mathcal{F},\mathcal{F}'\rangle$ for $P_x(y)$ where $\mathcal{F} = \mathcal{G} \setminus \{B\}$ with $\mathcal{F}'$ shown in red and variables in $\mathcal{F}''$ shown in green. Bit-parity of D and Y should be mapped to Y through B and C where C is in the top of the hedge.

A possible oversight is that the downward paths might cross $\mathcal{J}$ without passing $\mathbf{X}$ (see Fig. 6 for an example). The remedy is simple. For nodes appearing in the directed downward paths from $\mathbf{R}$ to $\mathbf{Y}$, we can assign an additional bit to pass bit parity information from $\mathbf{R}$ to $\mathbf{Y}$. Further, given a probability distribution $P_{\mathbf{w}}(\mathbf{z})$ on which $\mathcal{M}_1$ and $\mathcal{M}_2$ agree ($\mathbf{W},\mathbf{Z} \subseteq \mathbf{V}(\mathcal{J})$), $\mathcal{M}'_1$ and $\mathcal{M}'_2$ will also agree on $P_{\mathbf{w}\cup\mathbf{b}}(\mathbf{z})$ for any $\mathbf{b} \in \mathfrak{X}_{\mathbf{B}}$ where $\mathbf{B} \subseteq \mathbf{V}(\mathcal{G}) \setminus \mathbf{V}(\mathcal{J})$ for two reasons: Variables outside the paths from $\mathbf{R}$ to $\mathbf{Y}$ and $\mathcal{J}$ are ignored. Both models $\mathcal{M}'_1$ and $\mathcal{M}'_2$ behave exactly the same for nodes between $\mathbf{R}$ to $\mathbf{Y}$. $\square$