SUPPLEMENT TO "Causal Structural Learning on MPHIA Individual Dataset"

Le Bao[1], Changcheng Li[2], Runze Li[1] and Songshan Yang[3]

[1]Department of Statistics, The Pennsylvania State University
University Park, PA 16802, USA
[2]School of Mathematical Sciences, Dalian University of Technology
Dalian, P.R. China
[3]Institute of Statistics and Big Data, Renmin University of China
Beijing, P.R. China

We first present proofs of Propositions 1 and 2, and then present some additional details for the six Tri90 datasets, also some additional numerical results of Section 3, a summary for the notations in the proposed algorithms, and parts of MPHIA Codebook.

S.1 Proof of Proposition 1

It is easy to get (7), (8), and (9) directly from the definition of $\mathbf{C}_X(\mathbf{S})$. Furthermore, notice that (10) can be easily derived from (9). Hence the only thing that still needs to be proved is (11).

Suppose $\mathbf{S} = \{S_1, \cdots, S_n\}$, $n \geq 1$, and $\mathbf{S}$ satisfies (2). For any $C \in \mathbf{C}_X(\mathbf{S})$, we have $C \in \mathbf{C}_X^*(\mathbf{S}) \cap \mathbf{L}_X(\mathbf{S})$, and we also have $C \not\perp X|\mathbf{S}$, and $S_i \not\perp X|(\mathbf{S}_{-i} \cup \{C\})$, $i = 1, \cdots, n$, by definition of $\mathbf{C}_X(\mathbf{S})$. Hence we have $\mathbf{C}_X(\mathbf{S}) \subseteq \{C \in \mathbf{C}_X^*(\mathbf{S}) \cap \mathbf{L}_X(\mathbf{S}), C \not\perp X|\mathbf{S}, S_i \not\perp X|(\mathbf{S}_{-i} \cup \{C\}), i = 1, \cdots, n\}$.

Furthermore, we want to prove that $\mathbf{C}_X(\mathbf{S}) = \{C \in \mathbf{C}_X^*(\mathbf{S}) \cap \mathbf{L}_X(\mathbf{S}), C \not\perp X|\mathbf{S}, S_i \not\perp X|(\mathbf{S}_{-i} \cup \{C\}), i = 1, \cdots, n\}$ by contradiction. If $\mathbf{C}_X(\mathbf{S}) \subsetneq \{C \in \mathbf{C}_X^*(\mathbf{S}) \cap \mathbf{L}_X(\mathbf{S}), C \not\perp X|\mathbf{S}, S_i \not\perp X|(\mathbf{S}_{-i} \cup \{C\}), i = 1, \cdots, n\}$, then there exists $C_0 \in \mathbf{C}_X^*(\mathbf{S}) \cap \mathbf{L}_X(\mathbf{S})$ such that $C_0 \not\perp X|\mathbf{S}$, $S_i \not\perp X|(\mathbf{S}_{-i} \cup \{C_0\})$, $i = 1, \cdots, n$, and $C_0 \notin \mathbf{C}_X(\mathbf{S})$.

[1]The authors made equal contributions to this work and are listed in the alphabetic order.

From the definition of $\mathbf{C}_X(\mathbf{S})$ and $C_0 \in \mathbf{L}_X(\mathbf{S})$, the only way for C_0 not to be in $\mathbf{C}_X(\mathbf{S})$ is for $\{C_0\} \cup \mathbf{S}$ to violate (2). So there exists $N_0 \in (\{C_0\} \cup \mathbf{S})$ and $\mathbf{S}_0 \subseteq ((\{C_0\} \cup \mathbf{S}) \backslash \{N_0\})$ such that $N_0 \perp X | \mathbf{S}_0$.

1. If $\mathbf{S}_0 \subsetneq ((\{C_0\} \cup \mathbf{S}) \backslash \{N_0\})$, then $(\{N_0\} \cup \mathbf{S}_0) \subsetneq (\{C_0\} \cup \mathbf{S})$. So there must exists i_0, $1 \leq i_0 \leq n$, such that $(\{N_0\} \cup \mathbf{S}_0) \subseteq (\{C_0\} \cup \mathbf{S}_{-i_0})$, or $(\{N_0\} \cup \mathbf{S}_0) \subseteq \mathbf{S}$. Note that from the construction of N_0 and $\mathbf{S}_0$, we know $(\{N_0\} \cup \mathbf{S}_0)$ does not satisfy (2). Furthermore, any set with $(\{N_0\} \cup \mathbf{S}_0)$ as a subset does not satisfy (2). So $\{C_0\} \cup \mathbf{S}_{-i_0}$ or $\mathbf{S}$ does not satisfy (2), which is in contradiction with $C_0 \in \mathbf{C}_X(\mathbf{S}_{-i_0})$ and $\mathbf{S}$ satisfies (2).

2. Hence we have $\mathbf{S}_0 = ((\{C_0\} \cup \mathbf{S}) \backslash \{N_0\})$.

   (a) If $N_0 = C_0$, then $\mathbf{S}_0 = \mathbf{S}$ and $C_0 \perp X | \mathbf{S}$, which is in contradiction with $C_0 \not\perp X | \mathbf{S}$.

   (b) If $N_0 \neq C_0$, then there exists i_0, $1 \leq i_0 \leq n$, such that $N_0 = S_{i_0}$. Then we have $\mathbf{S}_0 = \{C_0\} \cup \mathbf{S}_{-i_0}$, and $S_{i_0} \perp X | (\{C_0\} \cup \mathbf{S}_{-i_0})$, which is in contradiction with $S_{i_0} \not\perp X | (\{C_0\} \cup \mathbf{S}_{-i_0})$.

   In sum, we finish the proof of (11) and Proposition 1.

S.2 Proof of Proposition 2

It is easy to get (15) and (16) directly from the definition of $S_N(X, Y)$. Furthermore, notice that (17) can be easily derived from (16). Hence the only thing that still needs to be proved is (18).

From the definition of $S_N(X, Y)$, we know that $S_N(X, Y) \geq \mathrm{CI}(X, Y | \mathbf{N})$. Hence we have $S_N(X, Y) \geq \max\{S_N^*(X, Y), \mathrm{CI}(X, Y | \mathbf{N})\}$. Suppose $S_N(X, Y) = \mathrm{CI}(X, Y, \mathbf{N}_0)$, where $\mathbf{N}_0 \subseteq \mathbf{N}$.

1. If $\mathbf{N}_0 = \mathbf{N}$, then $S_N(X, Y) = \mathrm{CI}(X, Y | \mathbf{N}) \leq \max\{S_N^*(X, Y), \mathrm{CI}(X, Y | \mathbf{N})\}$.

2. If $\mathbf{N}_0 \subsetneq \mathbf{N}$, then from the construction of $S_N^*(X, Y)$, we know that $S_N(X, Y) \leq S_N^*(X, Y) \leq \max\{S_N^*(X, Y), \mathrm{CI}(X, Y | \mathbf{N})\}$.

In sum, we have

$$S_N(X, Y) \leq \max\{S_N^*(X, Y), \mathrm{CI}(X, Y | \mathbf{N})\}.$$

Furthermore, from $S_N(X, Y) \geq \max\{S_N^*(X, Y), \mathrm{CI}(X, Y | \mathbf{N})\}$, we have

$$S_N(X, Y) = \max\{S_N^*(X, Y), \mathrm{CI}(X, Y | \mathbf{N})\}.$$

Hence we finish the proof of Proposition 2.

S.3 Additional Details of Six Tri90 Datasets by Target and Gender

1. Aware: Among the 2,217 individuals included in our analysis, there are 1,720 individuals with self-reported awareness or antiretroviral (ARV) detected including 510 males and 1,210 females. So $\frac{1720}{2217} = 77.6\%$ MPHIA participants have achieved the first Tri90 goal — being aware of HIV status. We investigate important covariates and potential causal pathways for HIV awareness for males and females, respectively.

2. ART: Among the 1,720 individuals with self-reported awareness or ARV detected, there are 1,564 individuals with self-reported ART or ARV detected including 454 males and 1,110 females. So $\frac{1564}{1720} = 90.3\%$ individuals have met the second Tri90 goal — being treated. We investigate important covariates and potential causal pathways for ART coverage for males and females, respectively.

3. VLS: Among the 1,564 individuals with self-reported ART or ARV detected, there are 1,428 individuals with viral load suppression (VLS) including 408 males and 1,020 females. So $\frac{1428}{1564} = 91.3\%$ individuals have met the third Tri90 goal — reaching Viral Suppression. We investigate important covariates and potential causal pathways for VLS in males and females, respectively.

S.4 Additional Tables and Figures

Table S.1: Number of important covariates for 90-90-90 goals discovered by different graphical learning methods. Aware, ART, and VLS stand for the three 90-90-90 targets of HIV awareness, ART treatment, and viral load suppression respectively. $\mathbf{d}$ is the distance from a particular 90-90-90 goal (awareness of HIV, ART, or VLS) to a covariate, and N($\mathbf{d} \leq k$), $k = 1, 2, 3$, are the number of covariates whose distances to a 90-90-90 goal are smaller than or equal to k.

		Male			Female		
Goals	Method	N($\mathbf{d} \leq 1$)	N($\mathbf{d} \leq 2$)	N($\mathbf{d} \leq 3$)	N($\mathbf{d} \leq 1$)	N($\mathbf{d} \leq 2$)	N($\mathbf{d} \leq 3$)
Aware	PC-stable	0	0	0	0	0	0
	MMPC	1	1	1	0	0	0
	IAMB	0	0	0	0	0	0
	GS	0	0	0	0	0	0
	New	6	19	49	5	18	51
ART	PC-stable	0	0	0	0	0	0
	MMPC	0	0	0	1	1	1
	IAMB	0	0	0	0	0	0
	GS	0	0	0	0	0	0
	New	4	13	38	4	11	28
VLS	PC-stable	0	0	0	0	0	0
	MMPC	0	0	0	0	0	0
	IAMB	0	0	0	0	0	0
	GS	0	0	0	0	0	0
	New	2	7	19	3	8	28

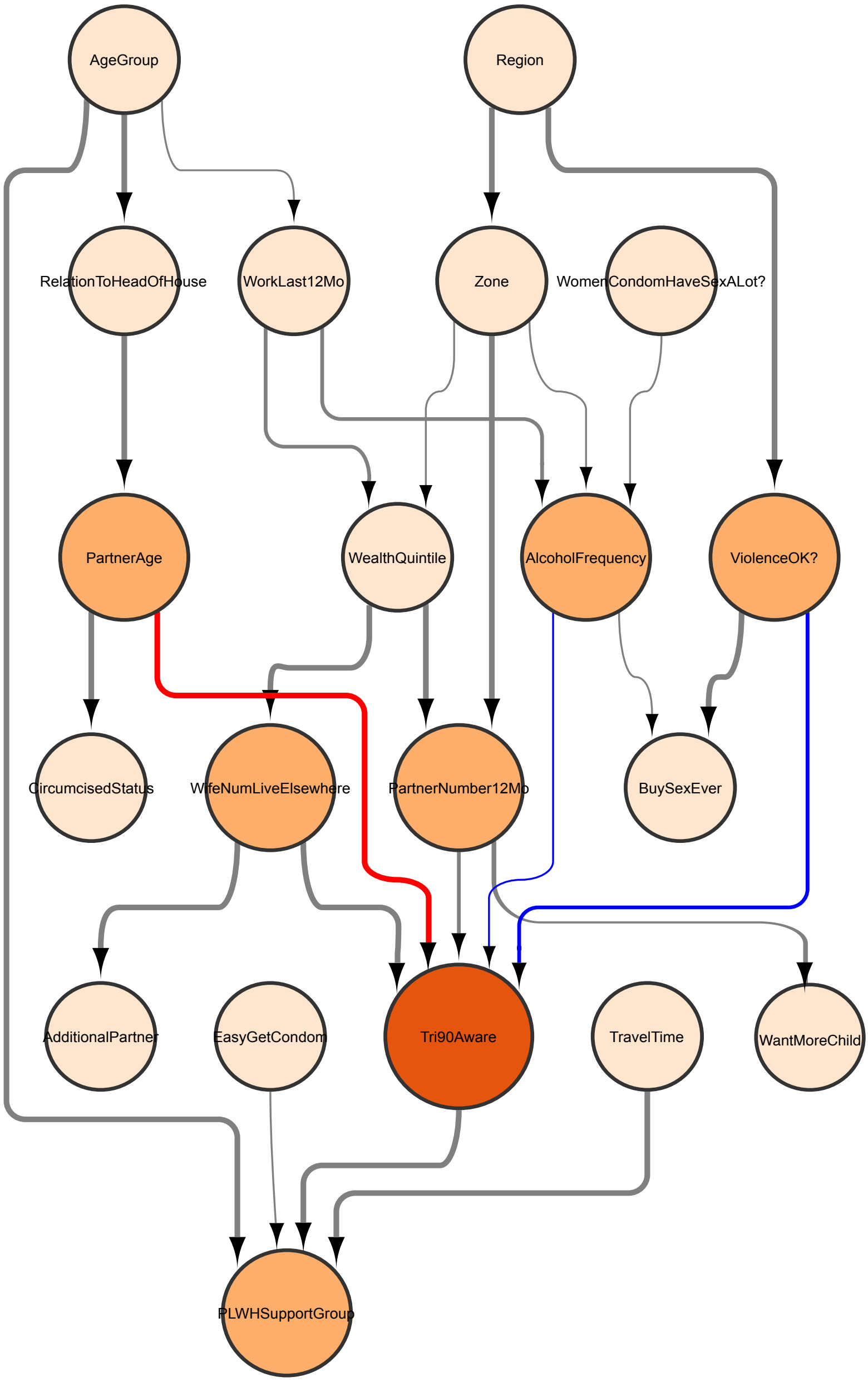


Figure S.1: 90-90-90 Awareness graph in male. Vertices representing the Tri90 goals are biggest and marked by orange; vertices closer to goals have bigger sizes and darker colors than those farther away from goals. Widths of edges reflect the significance of the non-directional connection (conditional dependence) between vertices. Red and blue edges represent positive and negative relationships with Tri90 goals, respectively. Grey edge from Tri90Aware to PLWHSupportGroup represents association of negative Tri90Aware with missingness in PLWHSupportGroup. For grey edges from WifeNumLiveElsewhere and PartnerNumber12Mo to Tri90Aware, see discussion in Section 3.2.2. Codebook can be found in Supplement S.7.3.

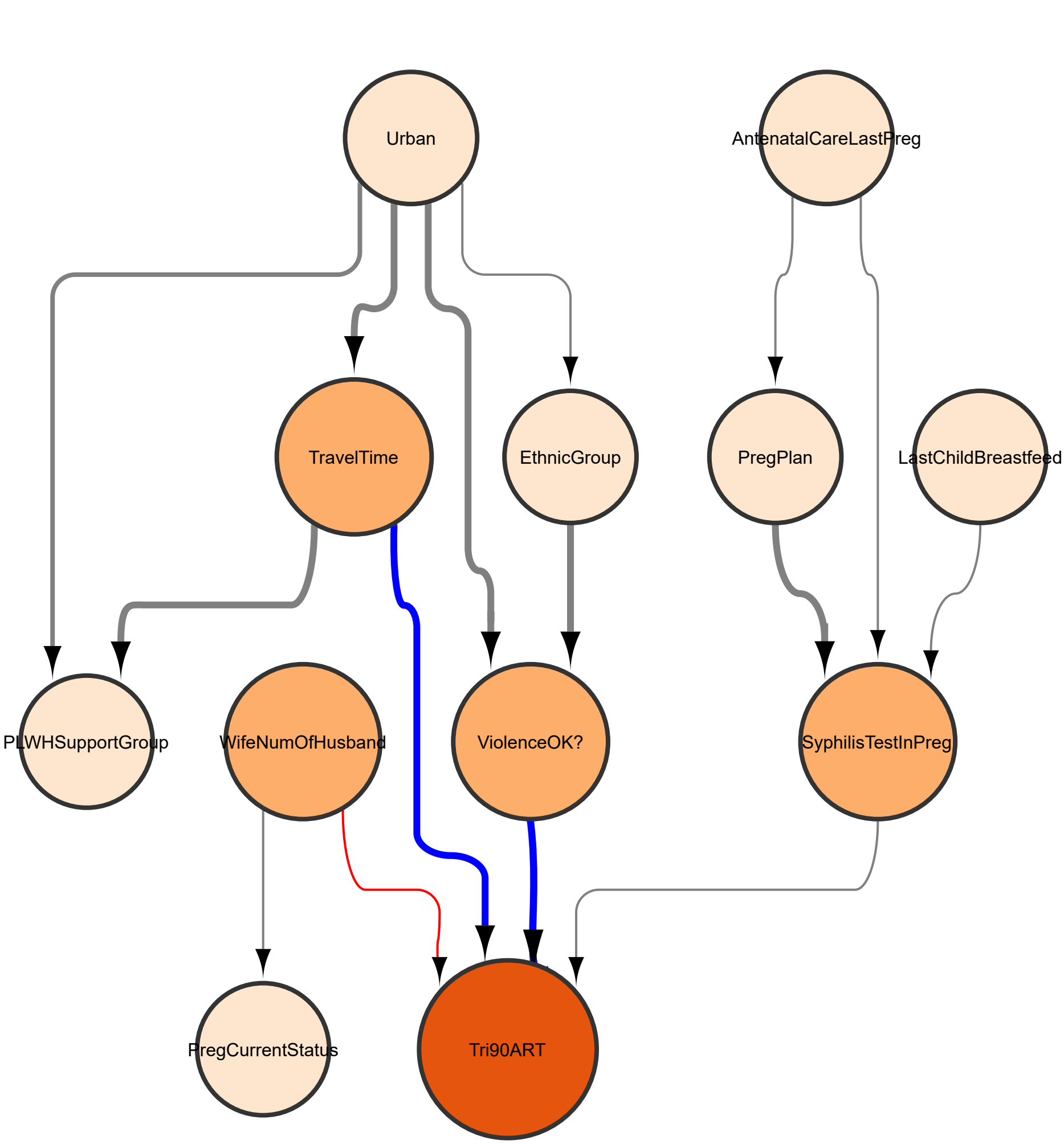


Figure S.2: 90-90-90 ART graph in female. Vertices representing the Tri90 goals are biggest and marked by orange; vertices closer to goals have bigger sizes and darker colors than those farther away from goals. Widths of edges reflect the significance of the non-directional connection (conditional dependence) between vertices. Red and blue edges represent positive and negative relationships with Tri90 goals, respectively. The grey edge from SyphilisTestInPreg to Tri90ART is discussed in Section 3.2.3. Codebook can be found in Supplement S.7.4.

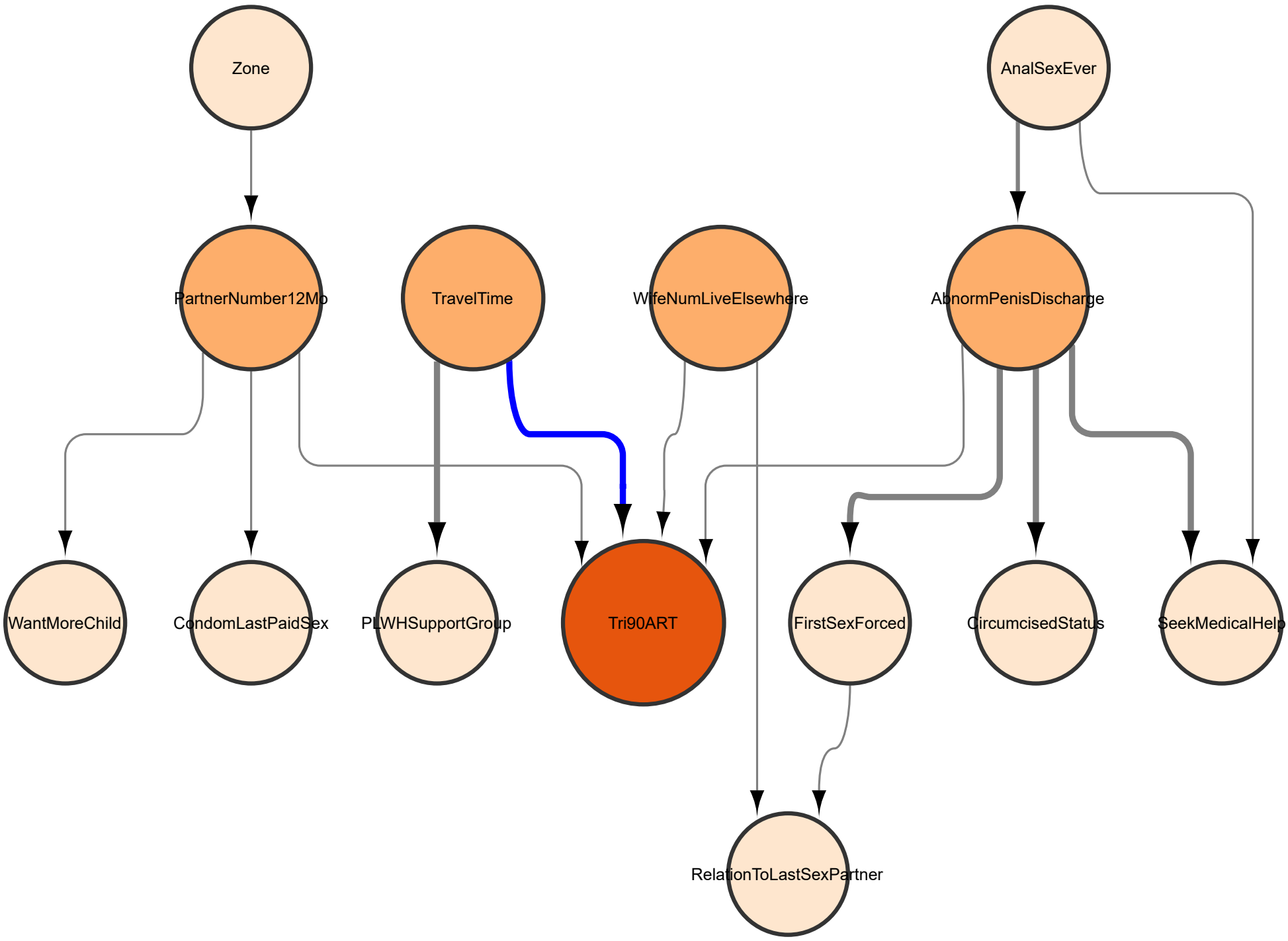


Figure S.3: 90-90-90 ART graph in male. Vertices representing the Tri90 goals are biggest and marked by orange; vertices closer to goals have bigger sizes and darker colors than those farther away from goals. Widths of edges reflect the significance of the non-directional connection (conditional dependence) between vertices. Red and blue edges represent positive and negative relationships with Tri90 goals, respectively. Grey edge from AbnormPenisDischarge to Tri90ART represents association of missingness in AbnormPenisDischarge with Tri90ART. For grey edges from WifeNumLiveElsewhere and PartnerNumber12Mo to Tri90ART, see discussion in Section 3.2.3. Codebook can be found in Supplement S.7.5.

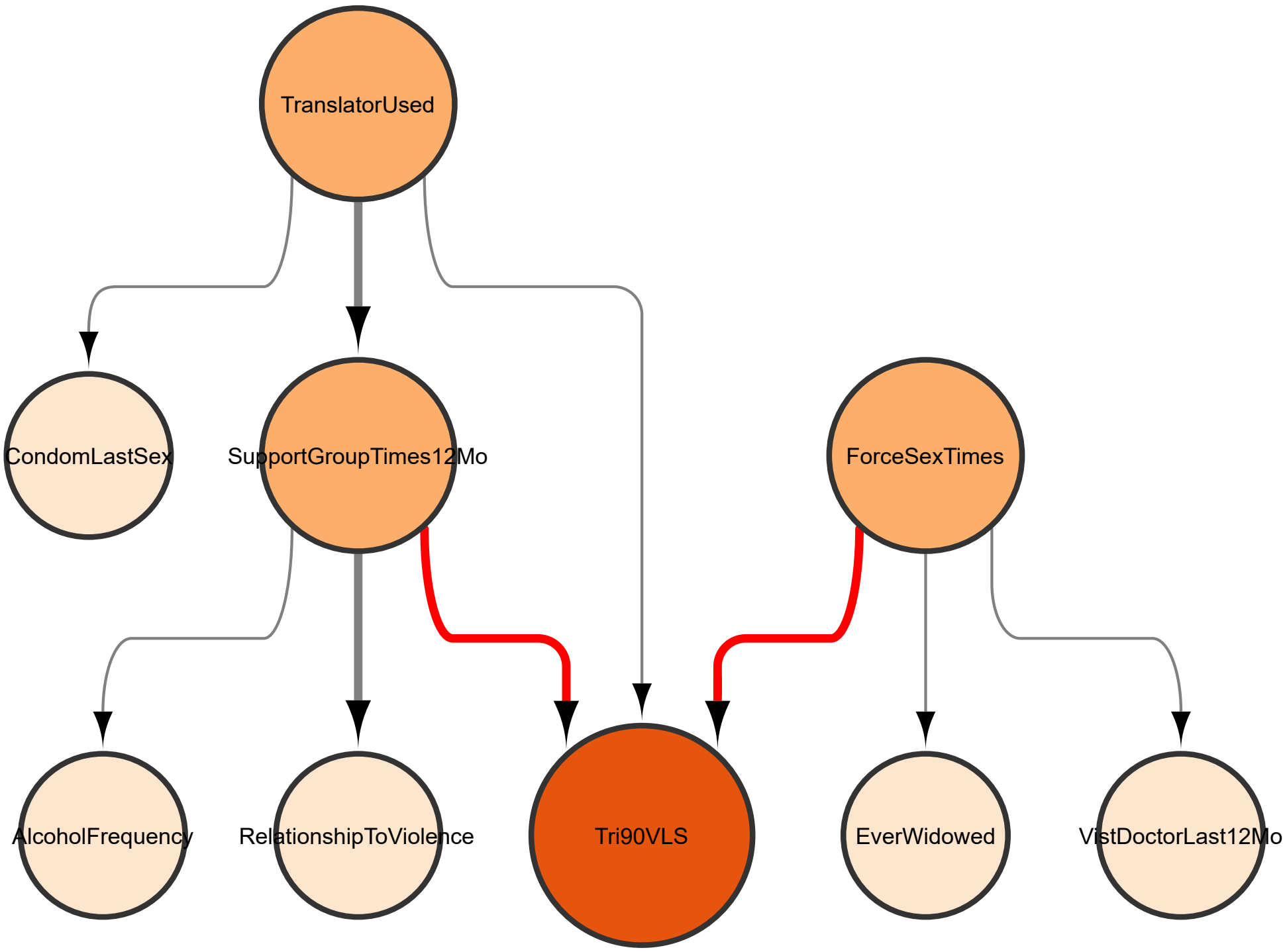


Figure S.4: 90-90-90 VLS graph in female. Vertices representing the Tri90 goals are biggest and marked by orange; vertices closer to goals have bigger sizes and darker colors than those farther away from goals. Widths of edges reflect the significance of the non-directional connection (conditional dependence) between vertices. Red and blue edges represent positive and negative relationships with Tri90 goals, respectively. The grey edge from TranslatorUsed to Tri90VLS represents neither positive nor negative relationship as TranslatorUsed has multiple levels. Codebook can be found in Supplement S.7.6.

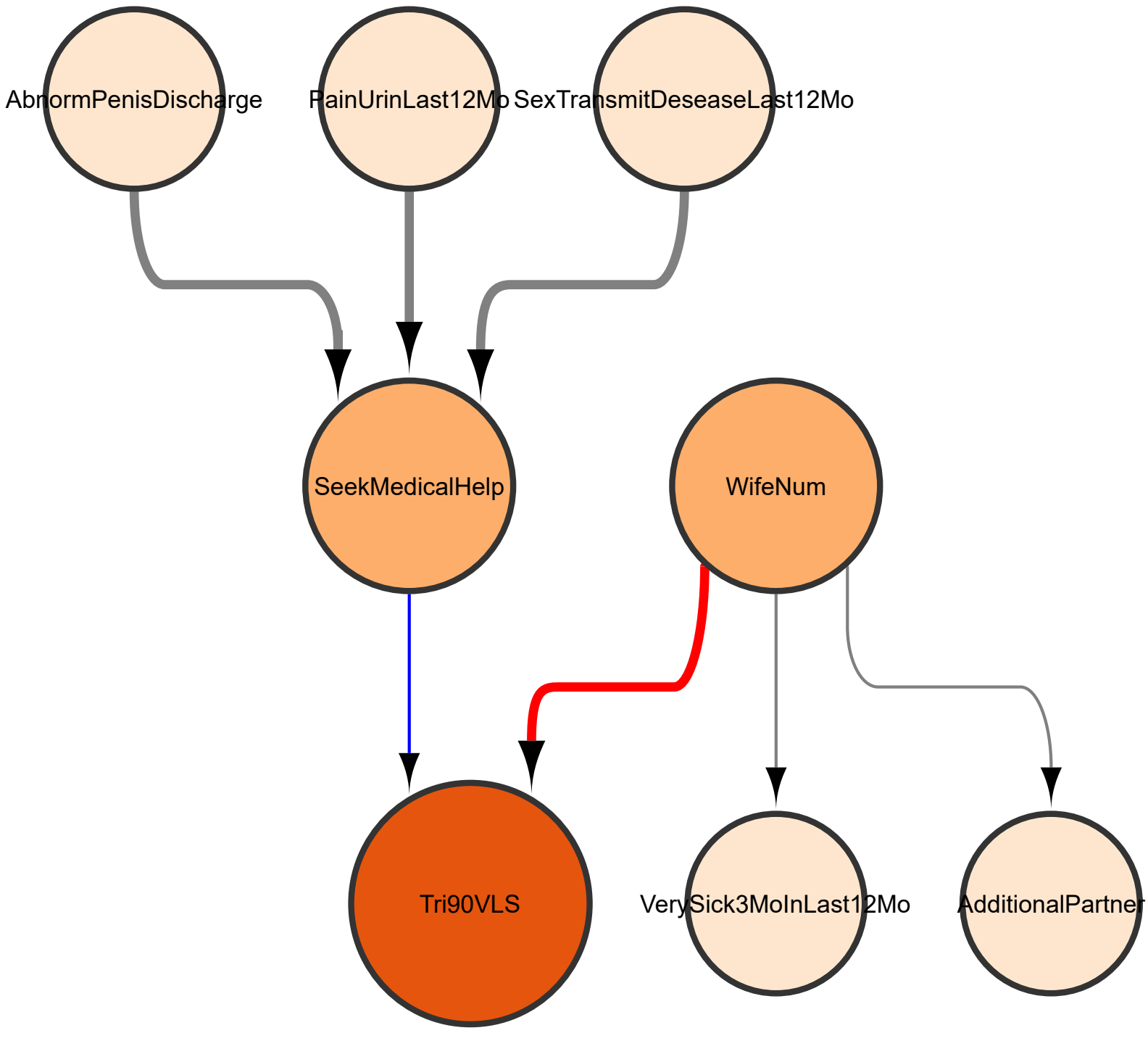


Figure S.5: 90-90-90 VLS graph in male. Vertices representing the Tri90 goals are biggest and marked by orange; vertices closer to goals have bigger sizes and darker colors than those farther away from goals. Widths of edges reflect the significance of the non-directional connection (conditional dependence) between vertices. Red and blue edges represent positive and negative relationships with Tri90 goals, respectively. Codebook can be found in Supplement S.7.7.

Table S.2: Empirical true positive rates and true negative rates of the proposed algorithm with different M_{CI} (in percentage). Aware, ART, and VLS stand for the three 90-90-90 targets of HIV awareness, ART treatment, and viral load suppression respectively.

		True Positive Rate					True Negative Rate				
Goals	Gender	$M_{\text{CI}}=2$	3	4	5	∞	$M_{\text{CI}}=2$	3	4	5	∞
Aware	Male	41.4	43.1	43.2	43.3	43.3	98.1	98.0	98.0	98.0	98.0
	Female	38.4	38.2	38.3	38.4	38.4	98.6	98.4	98.4	98.4	98.4
ART	Male	45.2	46.4	46.4	46.4	46.4	98.2	98.0	98.0	97.9	97.9
	Female	37.7	37.5	37.6	37.8	38.2	98.3	98.1	98.1	98.0	98.0
VLS	Male	41.6	44.7	44.7	44.7	44.7	97.9	97.9	97.9	97.9	97.9
	Female	40.4	40.0	40.0	40.1	40.1	98.5	98.4	98.4	98.4	98.3

S.5 Additional Simulation Studies

S.5.1 Chosen of M_{CI}

In this simulation study, we use a simulation setting similar to Section 4 to check the performance of the proposed algorithm with different values of M_{CI}. More specifically, we use the DAGs learned by the proposed algorithm as the truth to generate the simulation data. That is to say, let $\mathcal{G}_k$ be the DAG learned by the proposed algorithm on the 90-90-90 MPHIA data set $\mathbf{D}_k$ for $k = 1, 2, \cdots, 6$. Then we fit the data distribution $\mathcal{P}_k$ based on $\mathcal{G}_k$ on the data $\mathbf{D}_k$. We further randomly generate M simulated data sets $\mathcal{D}_k = (\mathbf{D}_{k,1}, \mathbf{D}_{k,2}, \cdots, \mathbf{D}_{k,M})$ based on the distribution $\mathcal{P}_k$ with the sample size n. Here we set $n = 500$ for a sample size similar to the MPHIA datasets. And we further apply the proposed algorithm with $M_{\text{CI}} = 2, 3, 4, 5, \infty$ on the generated dataset. The whole simulation is repeated 500 times, and we summarize the empirical averages of true positive rates and true negative rates of edges disregarding the orientation in Table S.2.

The left and right panels of Table S.2 summarize the empirical true positive and negative rates of the proposed algorithm with different M_{CI}, respectively. From Table S.2, we can see that there is no significant difference among the true positive rates and true negative rates for the proposed algorithm with $M_{\text{CI}} = 2, 3, 4, 5, \infty$. It shows that the proposed algorithm is quite robust to the choice of M_{CI}. As discussed by other causal structural learning literature such as Yan and Zhou (2020), we recommend to use $M_{\text{CI}} = 3$ for sparse or moderate sparse graphs. Notice that the choice of value of M_{CI} depends on the sample size, the types of covariates, the property of the true graph (the

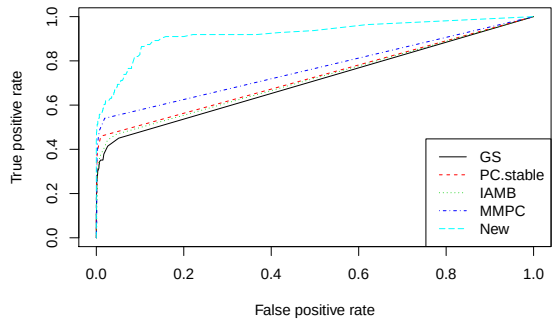
degree of the graph), etc as discussed in the remark for Algorithm 1, it is also possible to try different values of M_{CI} and to cross-validation methods, BIC scores, or simulation studies to have a more sophisticated chosen of M_{CI}.

S.5.2 Simulation Study to Check Receiver Operating Characteristic (ROC) Curve

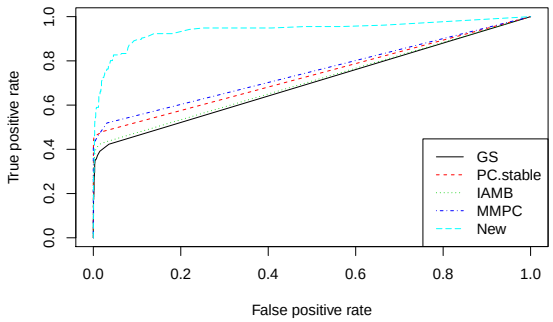
In this simulation study, we use a simulation setting similar to Section 4. We use the receiver operating characteristic (ROC) curve and the area under the (ROC) curve (AUC) to have a closer look at the performance of the different structural learning algorithms. More specifically, we use the DAGs learned by the proposed algorithm as the truth to generate the simulation data since the BIC criterion shows that the graphs learned by the proposed algorithm are better fits for the MPHIA data than those learned by the other algorithms. That is to say, let $\mathcal{G}_k$ be the DAG learned by the proposed algorithm on the 90-90-90 MPHIA data set $\mathbf{D}_k$ for $k = 1, 2, \cdots, 6$. Then we fit the data distribution $\mathcal{P}_k$ based on $\mathcal{G}_k$ on the data $\mathbf{D}_k$. We further randomly generate M simulated data sets $\mathcal{D}_k = (\mathbf{D}_{k,1}, \mathbf{D}_{k,2}, \cdots, \mathbf{D}_{k,M})$ based on the distribution $\mathcal{P}_k$ with the same sample size as the original data set $\mathbf{D}_k$. Applying graphical learning algorithm A_i, for $i = 1, \cdots, 5$, on the simulated data sets $\mathcal{D}_k$, we have M DAGs $(\mathcal{G}_{k,i,1}, \mathcal{G}_{k,i,2}, \cdots, \mathcal{G}_{k,i,M})$. We set $M = 500$, so the whole simulation is repeated 500 times.

Here we use the receiver operating characteristic (ROC) curve and the area under the (ROC) curve (AUC) to measure the edge discovery in 90-90-90 graphs to have a better understanding of the Type I and Type II error and their trade-off for different structural learning algorithms. To get the ROC curve, from the M DAGs learned in the Monte Carlo simulations $(\mathcal{G}_{k,i,1}, \mathcal{G}_{k,i,2}, \cdots, \mathcal{G}_{k,i,M})$, we first calculate an average graph $\bar{\mathcal{G}}_{k,i}$. $\bar{\mathcal{G}}_{k,i}$ is an undirected graph with weighted edges where the weight of an edge $X - Y$ is the empirical frequency of the existence of edge $X - Y$ in $(\mathcal{G}_{k,i,1}, \mathcal{G}_{k,i,2}, \cdots, \mathcal{G}_{k,i,M})$ disregarding the direction of edges. Hence $\bar{\mathcal{G}}_{k,i}$ reflects the "confidence" in edges for algorithm A_i. Then for each cut-off value λ, we can get an undirected graph $\mathcal{G}_{k,i,\lambda}$ by keeping all the edges in $\bar{\mathcal{G}}_{k,i}$ with weights greater than or equal to λ, calculate the true positive rate (TPR) and false positive rate (FPR) of edges, and obtain the AUC score and the ROC curve. The AUC score and the ROC curve for edge discovery in 90-90-90 graphs of different structural learning algorithms calculated from 500 Monte Carlo simulations are summarized in Table S.3 and Figures S.6, S.7, respectively.

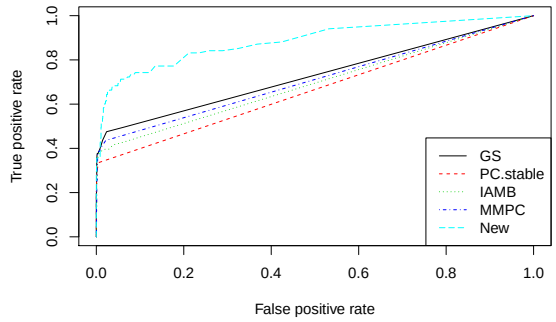
In Table S.3, we can see that the proposed algorithm achieves better (larger) AUC compared to other structural learning algorithms across all the three 90-90-90 goals and



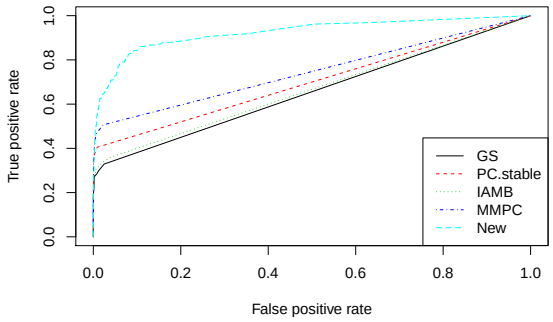
(a) ROC for male awareness.



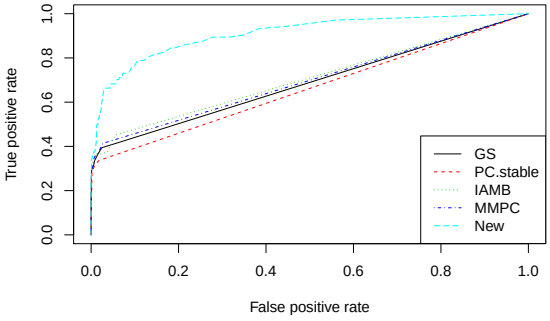
(b) ROC for female awareness.



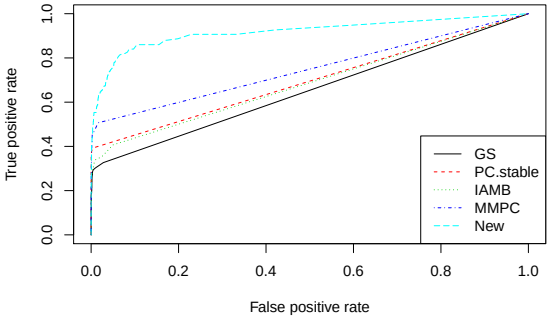
(c) ROC for male ART.



(d) ROC for female ART.



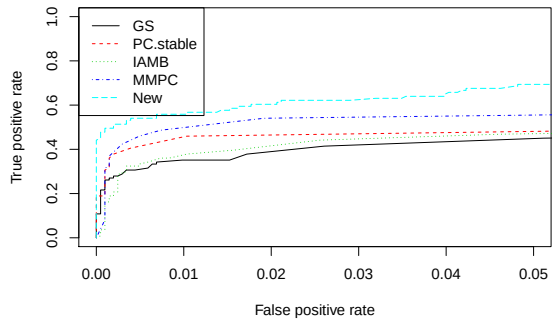
(e) ROC for male VLS.



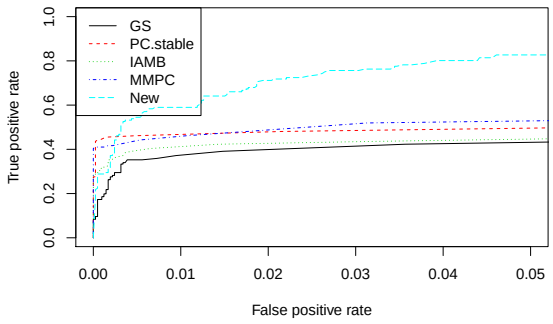
(f) ROC for female VLS.

Figure S.6: ROC curves for edge discovery of different structural learning algorithms for three 90-90-90 targets of both genders calculated from 500 Monte Carlo simulations.

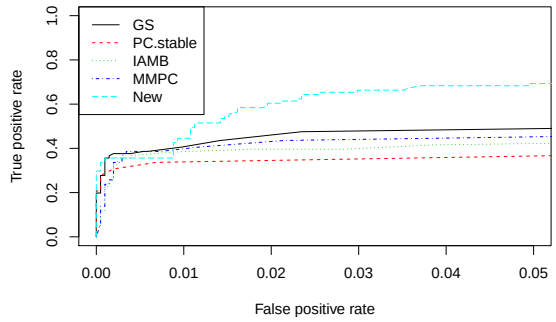
both genders. To understand why the proposed method achieves a better AUC, let us look at Figures S.6 and S.7. In Figure S.6, we can see that when the false positive rate (FPR) is extremely small, the proposed algorithm has a similar true positive rate



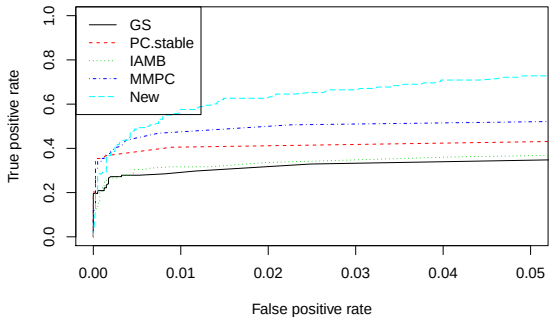
(a) Part of ROC for male awareness.



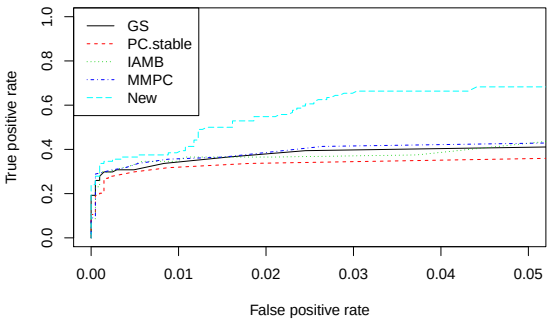
(b) Part of ROC for female awareness.



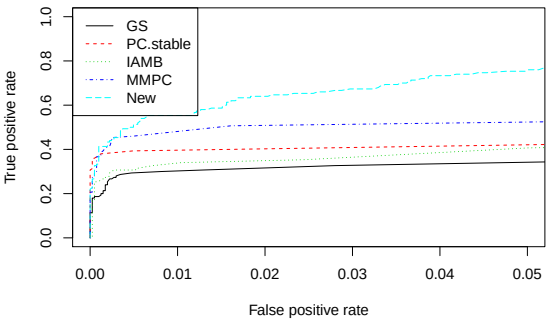
(c) Part of ROC for male ART.



(d) Part of ROC for female ART.



(e) Part of ROC for male VLS.



(f) Part of ROC for female VLS.

Figure S.7: Part of ROC curves with FPR ≤ 0.05 for edge discovery of different structural learning algorithms for three 90-90-90 targets of both genders calculated from 500 Monte Carlo simulations.

(TPR) as the other algorithms; while with larger FPR, the proposed algorithm has better TPR than the other algorithms. We can see the details for the ROCs with small

Table S.3: AUC of Different 90-90-90 Targets, Genders, and Causal Structural Learning Algorithms. Aware, ART, and VLS stand for the three 90-90-90 targets of HIV awareness, ART treatment, and viral load suppression respectively.

	Aware		ART		VLS	
Method	Male	Female	Male	Female	Male	Female
PC-stable	0.726	0.734	0.666	0.700	0.662	0.695
GS	0.708	0.700	0.730	0.655	0.688	0.653
IAMB	0.719	0.708	0.694	0.666	0.706	0.686
MMPC	0.764	0.751	0.711	0.747	0.697	0.749
New	0.922	0.940	0.885	0.925	0.910	0.918

FPR more clearly in Figure S.7 and find that the proposed algorithm achieves a better or comparable TPR with FPR ≥ 0.01 and a much better TPR with FPR ≥ 0.02 across all 90-90-90 goals and genders. The similar TPR of all algorithms for extremely small FPR illustrates that all algorithms have similar performance in the discovery of the most important relationship from the data. Furthermore, the better TPR of the proposed algorithm for larger FPR shows that while the existing structural learning algorithms cannot discover weaker signals beyond a cut-point; the proposed algorithm has a better ability in picking up relatively weak signals, which is the reason for the better AUC of the proposed algorithm in Table S.3.

S.5.3 Simulation Study with Continuous Variables and Different Graphical Densities and Signal Strengths

In this simulation study, we check the empirical performance of the proposed algorithm on synthetic data sets with continuous variables and different levels of "sparsity" and signal strengths of edges. Let K be the number of vertices and $\rho \in (0, 1)$ be the parameter that controls the level of "sparsity" of edges, we generate the simulation data randomly using the following procedure:

1. We first generate a DAG $\mathcal{G}^*$. Generate $K(K-1)/2$ random variables $E^*_{i,j}$ i.i.d from Bernoulli(ρ), $1 \leq i < j \leq K$. For vertices i and j, the edge $i \to j$ exists in $\mathcal{G}^*$ if and only if $E^*_{i,j} = 1$. Further generate $K(K-1)/2$ random variables $S^*_{i,j}$ i.i.d from Normal$(0, 1)$, $1 \leq i < j \leq K$.

2. We then generate a dataset $\mathbf{D}^*$ of size n according to DAG $\mathcal{G}^*$. For $j = 1, \cdots, K$,

generate x_j^* recursively from the following linear regression models:

$$x_j^* = \theta \sum_{i=1}^{j-1} x_i^* E_{i,j}^* S_{i,j}^* + \epsilon_j^*, \tag{S.1}$$

where θ controls the strengths of signals and ϵ_j^* i.i.d. follows the standard normal distribution, $j = 1, \cdots, K$. And we repeat this step n times to generate an n by K data set $\mathbf{D}^*$.

After generation of the dataset $\mathbf{D}^*$, we permute the order of variables and use the permutation to obtain an n by K data set $\mathbf{D}$ and the corresponding DAG $\mathcal{G}$. We then carry out the proposed algorithm together with the aforementioned PC-stable, GS, MMPC, and IAMB algorithms on the n by K data set $\mathbf{D}$. Furthermore, we calculate true positive rates and true negative rates of edges disregarding the orientation for each algorithm.

Table S.4: Empirical true positive rates and true negative rates of different causal structural learning algorithms (in percentage).

		True Positive Rate					True Negative Rate				
ρ	θ	GS	PC-stable	IAMB	MMPC	New	GS	PC-stable	IAMB	MMPC	New
0.01	0.125	32.30	35.6	35.4	35.5	37.8	99.3	99.3	99.3	99.3	99.1
	0.25	50.60	61.5	60.4	60.8	63.7	99.4	99.4	99.5	99.4	99.1
	0.5	55.08	77.3	74.6	76.0	79.9	99.6	99.6	99.6	99.6	99.2
	0.75	54.14	82.7	78.8	81.3	85.3	99.6	99.7	99.7	99.7	99.3
0.02	0.125	29.29	34.8	34.0	34.1	37.5	99.4	99.4	99.4	99.4	99.1
	0.25	39.84	59.3	55.0	55.8	62.6	99.6	99.6	99.7	99.6	99.2
	0.5	33.44	74.2	63.4	67.9	78.4	99.7	99.8	99.8	99.8	99.4
	0.75	26.94	76.9	63.0	70.1	81.9	99.7	99.9	99.9	99.9	99.4
0.04	0.125	23.25	33.2	30.0	30.3	36.4	99.6	99.5	99.6	99.6	99.2
	0.25	22.26	56.2	42.5	44.1	60.8	99.7	99.8	99.9	99.9	99.4
	0.5	9.70	66.4	41.6	48.3	73.2	99.8	99.9	100.0	100.0	99.5
	0.75	5.46	65.0	38.2	48.3	73.9	99.8	99.9	100.0	100.0	99.4

Here we set $K = 100$ and $n = 500$ for a similar number of covariates and sample size with our real data. We set $\rho = (0.01, 0.02, 0.04)$ for different levels of "sparsity" of the true graph and $\theta = (0.125, 0.25, 0.5, 0.75)$ for different strengths of signals. Note that in this simulation, we set the upper bound of sizes of conditional sets $M_{\text{CI}} = 2$ and the size of the conditional independence test $\alpha = 0.01$ for all the causal structural learning algorithms to reduce the computation time. We repeat the Monte Carlo simulation 1,000 times for each setting and summarize the results in Table S.4. The left and right panels of Table S.4 summarize the empirical true positive and negative rates of the proposed

algorithm as well as those of existing algorithms, respectively. From the right panel of Table S.4, we can see that the proposed algorithm has similar true negative rates with existing algorithms. Furthermore, from the left panel of Table S.4, we can see that the proposed algorithm has better true positive rates than existing algorithms. In sum, we have similar conclusions to those of Section 4.

Comparing the simulation results in Table S.4 with those in Table 5, notice that the simulation settings in Section 4 are more challenging than those in this section in terms of the true positive rate. This is because there are many categorical variables in the simulation in Section 4, while there are only continuous ones in the simulation in this section. Since the conditional set of categorical variables takes more degrees of freedom away from the conditional independence tests than the continuous ones, categorical variables in the simulation in Section 4 can lead to more Type II errors and more contradictory/inconsistent statistical testing results than the simulation in this section. Hence, we can see that the improvement in the true positive rates of the proposed algorithm over the existing ones in Table 5 is larger than the improvement in Table S.4 in this section. In sum, we can see that the proposed algorithm is more beneficial in the true positive rate in the case of categorical variables.

S.6 Summary for Algorithm Notations

In Algorithms 1, 2, and 3, notations are used for numbers, vertices, vertex sets, and sets of vertex sets. A summary of the notations is provided in Table S.5. It is followed by a short explanation for some of the important notations.

Notation for Numbers. α refers to the significance level for the conditional independence test. M_{CI} refers to the upper-bound of sizes of conditional sets used in the algorithms. $\text{CI}(X, Y|\mathbf{S})$ is the p-value of the chosen conditional independence test for vertices X and Y given some vertex set $\mathbf{S}$. $S_N(X, Y)$ and $Q_X(\mathbf{N})$ are defined as the following.

$$\begin{aligned} S_N(X, Y) &:= \max_{S \subseteq N} \text{CI}(X, Y|\mathbf{S}), \\ Q_X(\mathbf{N}) &:= \min_{M_i \in \mathbf{V} \setminus (\mathbf{N} \cup \{X\})} S_N(M_i, X). \end{aligned} \tag{S.2}$$

$S_N(X, Y)$ measures whether any subset of $\mathbf{N}$ makes the vertices X and Y conditional independent, and $Q_X(\mathbf{N})$ measures how well it is for subsets of $\mathbf{N}$ to "separate" the vertex X from any vertices not in $\mathbf{N} \cup \{X\}$ in Algorithm 2. More information and discussion on $S_N(X, Y)$ and $Q_X(\mathbf{N})$ can be found in Section 2.5.

Table S.5: A summary for notations used in Algorithms 1, 2, and 3.

	Algorithm 1	Algorithm 2	Algorithm 3
Number	α M_{CI}	r $Q_X(\mathbf{N})$ $S_N(X, Y)$ $\text{CI}(X, Y\|\mathbf{S})$	α
Vertex	X C	X M_i	X Y Z
Vertex Set	$\mathbf{V}$ $\mathbf{N}$ $\mathbf{S}$ $\mathbf{C}_X(\mathbf{S})$ $\mathbf{C}_X^*(\mathbf{S})$ $\mathbf{L}_X(\mathbf{S})$	$\mathbf{V}$ $\mathbf{N}$ $\mathbf{N}_X$ $\mathbf{M}$	$\mathbf{V}$ $\mathbf{N}_X$ $\mathbf{N}_Y$ $\mathbf{S}(X, Y)$
Set of Vertex Sets	$\mathcal{N}_X$ $\mathcal{S}$ $\mathcal{R}$	$\mathcal{N}_X$	

Notation for Vertex. X refers to the vertex we want to find the neighborhood in Algorithms 1 and 2.

Notation for Vertex Set. $\mathbf{V}$ always refers to the set of all the vertices. $\mathbf{L}_X(\mathbf{S})$ refers to the set of vertices to be considered to be added into a vertex set $\mathbf{S}$ in Algorithm 1. $\mathbf{C}_X(\mathbf{S})$ refers to the set of the vertices in $\mathbf{L}_X(\mathbf{S})$ that can be added into $\mathbf{S}$ while still satisfying equation (2). More information on $\mathbf{L}_X(\mathbf{S})$, $\mathbf{C}_X(\mathbf{S})$, and $\mathbf{C}_X^*(\mathbf{S})$ can be found in Section 2.4. $\mathbf{S}(X, Y)$ refers the d-separation set between vertices X and Y in Algorithm 3.

Notation for Set of Vertex Sets. $\mathcal{N}_X$ always refers to the set of all the candidate neighbor set for X that (approximately) satisfies equation (2), which is the output for Algorithm 1 and the input for Algorithm 2.

S.7 Parts of MPHIA Codebook

S.7.1 Codebook for Covariates in Table 4

1. AbnormPenisDischarge: During the last 12 months, have you had an abnormal discharge from your penis?
2. AgeGroup: Age groups for population pyramid
3. AlcoholFrequency: How often do you have a drink containing alcohol?
4. EasyGetCondom: If you wanted a condom, would it be easy for you to get one?
5. Education: Level of school respondent ever attended
6. ForceSexTimes: How many times in your life have you been physically forced to have sex?
7. PartnerAge: How old is your partner? Please give your best guess.
8. PartnerNumber12Mo: Number of people they had sex with in the last 12 months
9. PLWHSupportGroup: Have you ever attended a support group for people living with HIV?
10. PregNum: How many times have you been pregnant including a current pregnancy?
11. SeekMedicalHelp: Did you see a doctor, clinical officer or nurse because of these problems?
12. SupportGroupTimes12Mo: In the last 12 months, how many times did you attend a support group?
13. SyphilisTestInPreg: When you were pregnant, were you offered a test for syphilis?
14. TranslatorUsed: whether or translator is used or not.
15. TravelTime: At your last HIV care visit, approximately how long did it take you to travel from your home (or workplace) one way?
16. ViolenceOK?: Do you believe it is right for a man to hit or beat his wife/partner?
17. WifeNum: Altogether, how many wives or partners do you have?

18. WifeNumLiveElsewhere: How many wives/partners do you have who live elsewhere?

19. WifeNumOfHusband: Including yourself, in total, how many wives or live-in partners does your husband or partner have?

S.7.2 Codebook for Covariates in Figure 1

1. AgeGroup: Age groups for population pyramid

2. ChildNumSince2012: How many children have you given birth to since 2012?

3. CircumcisedHIVRisk: Relationship of circumcision and risk of HIV?

4. EasyGetCondom: If you wanted a condom, would it be easy for you to get one?

5. Education: Level of school respondent ever attended

6. EthnicGroup: What is your ethnic group?

7. MarrigeStatus: What is your marital status now: are you married, living together with someone as if married, widowed, divorced, or separated?

8. PartnerNumber12Mo: Number of people they had sex with in the last 12 months

9. PLWHSupportGroup: Have you ever attended a support group for people living with HIV?

10. PregNum: How many times have you been pregnant including a current pregnancy?

11. RelationToHeadOfHouse: What is your relationship to the head of the household?

12. SellSexEver: Have you ever sold sex for money?

13. SyphilisTestInPreg: When you were pregnant, were you offered a test for syphilis?

14. TravelTime: At your last HIV care visit, approximately how long did it take you to travel from your home (or workplace) one way?

15. Urban: Urban Area Indicator

16. ViolenceOK?: Do you believe it is right for a man to hit or beat his wife/partner?

17. WorkLast12Mo: Have you done any work in the last 12 months for which you received a paycheck, cash or goods as payment?

18. Zone: Zone name

S.7.3 Codebook for Covariates in Figure S.1

1. AdditionalPartner: Do you have additional spouse(s)/partner(s) that live with you?

2. AgeGroup: Age groups for population pyramid

3. AlcoholFrequency: How often do you have a drink containing alcohol?

4. BuySexEver: Have you ever paid money for sex?

5. CircumcisedStatus: Are you circumcised or planning to get circumcised?

6. EasyGetCondom: If you wanted a condom, would it be easy for you to get one?

7. PartnerAge: How old is your partner? Please give your best guess.

8. PartnerNumber12Mo: Number of people they had sex with in the last 12 months

9. PLWHSupportGroup: Have you ever attended a support group for people living with HIV?

10. Region: Region Name

11. RelationToHeadOfHouse: What is your relationship to the head of the household?

12. TravelTime: At your last HIV care visit, approximately how long did it take you to travel from your home (or workplace) one way?

13. ViolenceOK?: Do you believe it is right for a man to hit or beat his wife/partner?

14. WantMoreChild: Would you like to have a/another child?

15. WealthQuintile: Wealth quintile

16. WifeNumLiveElsewhere: How many wives/partners do you have who live elsewhere?

17. WomenCondomHaveSexALot?: Do you believe women who carry condoms have sex with a lot of men?

18. WorkLast12Mo: Have you done any work in the last 12 months for which you received a paycheck, cash or goods as payment?

19. Zone: Zone name

S.7.4 Codebook for Covariates in Figure S.2

1. AntenatalCareLastPreg: Flag if mother who gave birth 3 years preceding survey received antenatal care during last pregnancy

2. EthnicGroup: What is your ethnic group?

3. LastChildBreastfeed: Mother's current and past breast feeding status

4. PLWHSupportGroup: Have you ever attended a support group for people living with HIV?

5. PregCurrentStatus: Are you pregnant now?

6. PregPlan: When you were pregnant, did you plan to get pregnant at that time?

7. SyphilisTestInPreg: When you were pregnant, were you offered a test for syphilis?

8. TravelTime: At your last HIV care visit, approximately how long did it take you to travel from your home (or workplace) one way?

9. Urban: Urban Area Indicator

10. ViolenceOK?: Do you believe it is right for a man to hit or beat his wife/partner?

11. WifeNumOfHusband: Including yourself, in total, how many wives or live-in partners does your husband or partner have?

S.7.5 Codebook for Covariates in Figure S.3

1. AbnormPenisDischarge: During the last 12 months, have you had an abnormal discharge from your penis?

2. AnalSexEver: Have you ever had anal sex?

3. CircumcisedStatus: Are you circumcised or planning to get circumcised?

4. CondomLastPaidSex: Flag if condom was used at last paid sexual intercourse

5. FirstSexForced: The first time you had vaginal or anal sex, was it because you wanted to or because you were forced to?
6. PartnerNumber12Mo: Number of people they had sex with in the last 12 months
7. PLWHSupportGroup: Have you ever attended a support group for people living with HIV?
8. RelationToLastSexPartner: Relationship status with their last sex partner in the past 12 months
9. SeekMedicalHelp: Did you see a doctor, clinical officer or nurse because of these problems?
10. TravelTime: At your last HIV care visit, approximately how long did it take you to travel from your home (or workplace) one way?
11. WantMoreChild: Would you like to have a/another child?
12. WifeNumLiveElsewhere: How many wives/partners do you have who live elsewhere?
13. Zone: Zone name

S.7.6 Codebook for Covariates in Figure S.4

1. AlcoholFrequency: How often do you have a drink containing alcohol?
2. CondomLastSex: Indicator for used condom at last sexual encounter in the past 12 months
3. EverWidowed: Have you ever been widowed? That is, did a spouse ever die while you were still married or living with them?
4. ForceSexTimes: How many times in your life have you been physically forced to have sex?
5. RelationshipToViolence: Relationship between you and the person who give physical violence to you.
6. SupportGroupTimes12Mo: In the last 12 months, how many times did you attend a support group?

7. TranslatorUsed: whether translator is used or not.

8. VistDoctorLast12Mo: Have you seen a doctor, clinical officer or nurse in a health facility in last 12 months?

S.7.7 Codebook for Covariates in Figure S.5

1. AbnormPenisDischarge: During the last 12 months, have you had an abnormal discharge from your penis?

2. AdditionalPartner: Do you have additional spouse(s)/partner(s) that live with you?

3. PainUrinLast12Mo: During the last 12 months, have you had pain on urination?

4. SeekMedicalHelp: Did you see a doctor, clinical officer or nurse because of these problems?

5. SexTransmitDeseaseLast12Mo: In the last 12 months, did a doctor, clinical officer or nurse tell you that you had a sexually transmitted disease?

6. VerySick3MoInLast12Mo: Has name been very sick for at least 3 months during the past 12 months, that is name was too sick to work or do normal activities?

7. WifeNum: Altogether, how many wives or partners do you have?

References

Chuanxu Yan and Shuigeng Zhou. Effective and scalable causal partitioning based on low-order conditional independent tests. *Neurocomputing*, 389:146–154, 2020.