

# Pseudocode for MRPC Update

Jarred M. Kvamme  
University of Idaho  
Department of Statistical Science

January 18, 2022

## 1.1 - The Trio Specific Case:

In a network consisting of three nodes:  $G(V_1, T_1, T_2)$  we can identify the 5 possible topologies laid out under MRPC using results of the coefficient tests from the pair of regressions on each non-instrumental variable (Please refer to the **Appendix** at the end for Tables, Figures, and mathematical definitions).

**Step 1.** Calculate the frequency of the minor variant  $f_{\text{minor}}$ . If  $V_1$  is an eQTL, then  $f_{\text{minor}}$  is the frequency of the minor allele. If  $V_1$  is not an eQTL, then calculate  $f_{\text{minor}}$  as the frequency of the least common count or class.

**Step 2.** Perform two regressions treating each non-instrumental variable as the response once:

$$T_1 = \beta_0 + \beta_{11}V_1 + \beta_{21}T_2 + \mathbf{\Gamma U} + \epsilon \quad (1)$$

$$T_2 = \beta_0 + \beta_{12}V_1 + \beta_{22}T_1 + \mathbf{\Gamma U} + \epsilon \quad (2)$$

**Step 2.2** - If the minor variant frequency  $f_{\text{minor}}$  (allele frequency or copy number variation) from **Step 1.** is less than the predetermined threshold  $\gamma$ , perform the permuted regressions described in **Section 1.2**.

**Step 3.** calculate the correlation matrix for  $[V_1, T_1, T_2]$  and perform hypothesis testing on  $r_{V_1, T_2}$  and  $r_{V_1, T_1}$  i.e the marginal relationships between  $V_1, T_2$  and  $V_1, T_1$

**Step 4.** following **Steps 1 - 3** obtain the vector of  $p$ -values for hypothesis tests on  $\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}, r_{V_1, T_2}$ , and  $r_{V_1, T_1}$  (in this order) as the vector  $\mathbf{p}$

**Step 4.1** Convert the vector of  $p$ -values  $\mathbf{p}$  into the indicator vector  $\mathbf{I}_p$  where 1 denotes a significant  $p$ -value at threshold  $\alpha$  and 0 denotes a nonsignificant  $p$ -value

**Step 5.** - Compare  $\mathbf{I}_p$  with the expected results for each model topology given in **Tables 1 and 2** in the **Appendix** and allocate the trio to model type for which it matches (note that models M0, M1, and M2 have two cases each depending on the directions of the edges). If no match is available allocate the trio the class "other".

**Step 6.** - Return the correct adjacency matrix from the inferred model type in **Step 5**

## 1.2 - Permuted Regression for Rare Variants

- We will apply the permuted regression described by Yang et al., 2017 whenever the instrumental variable contains a rare count at frequency  $< \gamma$ . The permuted regression

is preformed to obtain a robust estimate of the mediation effect between the nodes  $T_i$  and  $T_j$  in  $G(V_k, T_i, T_j)$  which may be masked in the standard regression when  $V_k$  contains few observations for the minor variant. The algorithm for preforming the permuted regression is as follows:

**Step 1.** - Let  $f_{\text{minor}}$  be the frequency of the minor variant of  $V_1$ . If  $f_{\text{minor}} < \gamma$  proceed to **Step 2**.

**Step 2.** - Preform the standard regressions given in (1) and (2) and retain the observed  $t$ -statistics for the tests on  $\beta_{21}$  and  $\beta_{22}$  (which we will denote as  $t_{\text{obs}_{21}}$  and  $t_{\text{obs}_{22}}$ , respectively)

**Step 2.2** - Similarly, retain the observed  $p$ -values for the tests on  $\beta_{11}$  and  $\beta_{12}$  (which we will denote  $p_{\beta_{11}}$  and  $p_{\beta_{12}}$  respectively)

**Step 3.** - permute  $T_2$  in (1) within the levels of  $V_1$  denoted  $T_2^*$ . Similarly, permute  $T_1$  in (2) within the levels of  $V_1$  denoted  $T_1^*$ .

**Step 3.1** - Next preform the regressions in **Section 1.1** using the permuted variables:

$$T_1 = \beta_0 + \beta_{11}V_1 + \beta_{21}^*T_2^* + \mathbf{\Gamma U} + \epsilon \quad (3)$$

$$T_2 = \beta_0 + \beta_{12}V_1 + \beta_{22}^*T_1^* + \mathbf{\Gamma U} + \epsilon \quad (4)$$

**Step 3.2** - store the observed  $t$ -statistic for  $\beta_{21}^*$  in the vector  $\mathbf{\Theta}_{21}$  and the  $t$ -statistic for  $\beta_{22}^*$  in the vector  $\mathbf{\Theta}_{22}$

**Step 4.** repeat **Steps 3 - 3.2**  $m$  times to obtain the  $m \times 1$  vectors of observed  $t$ -statistics  $\mathbf{\Theta}_{21}$  and  $\mathbf{\Theta}_{22}$ .

**Step 4.1** - Using  $\mathbf{\Theta}_{21}$ ,  $\mathbf{\Theta}_{22}$ ,  $t_{\text{obs}_{21}}$ , and  $t_{\text{obs}_{22}}$ , calculate the permutation nominal  $p$ -values for  $\beta_{21}$  and  $\beta_{22}$  (which we will denote  $p_{\beta_{21}}^*$  and  $p_{\beta_{22}}^*$  respectively)

**Step 5.** calculate the correlation matrix for  $[V_1, T_1, T_2]$  and preform hypothesis testing on  $r_{V_1, T_2}$  and  $r_{V_1, T_1}$  i.e the marginal relationships between  $V_1, T_2$  and  $V_1, T_1$

**Step 6.** - allocate the vector of  $p$ -values for hypothesis tests on  $\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}, r_{V_1, T_2}$ , and  $r_{V_1, T_1}$  (in this order) as the vector  $\mathbf{p}$  (using the nominal  $p$ -values for  $\beta_{21}$  and  $\beta_{22}$ ) and proceed to **Step 4.1** in **Section 1.1**

### 1.3 - Inferring Trios Without Variants

We can infer the graph skeleton for any 3-node network  $G(T_i, T_j, T_k)$  and uniquely infer the graph for M2. The algorithm for this is as follows:

**Step 1.** - Perform the regressions in **Section 1.1, Step 2** replacing the instrumental variable  $V_1$  in the regressions with  $T_3$  (skipping **Step 2.2**).

**Step 2.** - Perform **Section 1.1 Steps 3-4.1**

**Step 3.** - Compare  $\mathbf{I}_p$  with the expected results for M2 given in **Tables 1 and 2** in the **Appendix**.

**Step 4.** - If  $\mathbf{I}_p$  matches the expectation for M2, return the correct adjacency matrix for M2.

## 2. General Algorithm

**Step 1.** - Given a data matrix  $\mathbf{X}$  of  $q$  instrumental variables,  $p$  non-instrumental variables, and  $g$  confounders. Calculate the partial correlation matrix  $\mathbf{H}$  and extract the first  $\lambda = \{1 : p + q\}$  rows and columns of  $\mathbf{H}$  which represents the partial correlations between all nodes in the graph  $G(V_1, V_2, \dots, V_q, T_1, T_2, \dots, T_p)$ .

**Step 1.1** - perform a partial correlation test on all non-diagonal entries in  $\mathbf{H}[\lambda, \lambda]$  to obtain the  $(p + q \times p + q)$  matrix of  $p$ -values  $\mathbf{P}$  corresponding to the nodes in  $G(V_1, V_2, \dots, V_q, T_1, T_2, \dots, T_p)$ .

**Step 1.2** -For each non-diagonal entry in  $\mathbf{P}$  replace significant  $p$ -values at threshold  $\alpha$  with 1 and nonsignificant  $p$ -values with 0 to obtain the  $(p + q \times p + q)$  adjacency matrix  $\mathbf{A}$  for the skeleton of  $G(V_1, V_2, \dots, V_q, T_1, T_2, \dots, T_p)$

**Step 2.** - pre-allocate the data matrices for all  $q \times \binom{p}{2}$  possible 3-node networks involving the instrumental variable(s) into a list. Each entry in the list is the  $(n \times g + 3)$  data matrix  $[V_k, T_i, T_j, \mathbf{U}]$

**Step 2.1** - Similarly, construct the list of all  $\binom{p}{3}$  possible 3-node networks involving only the non-instrumental variable nodes. Each entry in the list is the  $(n \times g + 3)$  data matrix  $[T_k, T_i, T_j, \mathbf{U}]$

**Step 3.** - Determine the directed structure of all trios with an instrumental variable from **Step 2** using the regressions and tests outlined in **Sections 1.1-1.2**.

**Step 4.** - Infer the network of all  $\binom{p}{3}$  possible 3-node networks involving only the non-instrumental variable nodes (referred to as T-nodes) using the regressions and tests

outlined in **Section 1.3**. This step is to identify v-structures among T-nodes and direct additional edges in the network.

**Step 5.** - Use the results from **Steps 3 and 4** to update the network graph/adjacency matrix

### 3. Simulation Strategy to Validate Section 1.1

To verify that the indicator vectors  $\mathbf{I}_p$  derived from the hypothesis tests in **Section 1.1** are sufficient in identifying the proposed model structures from MRPC given in **Tables 1 and 2**, we propose the following simulation methodology:

- (1.) - Using the linear models for the 5 basic topologies described by Badsha and Fu 2019, we propose to simulate each trio topology under varying signal strengths (and possibly sample size) using the simulation functions in the *R* package *MRPC*.
- (2.) - we then apply the algorithm proposed in **Section 1.1** to determine if the expected results given in **Tables 1 and 2** are identified. we compare the indicator vector  $\mathbf{I}_p$  to the expected indicator vectors for each topology and allocate the trio to one of the 5 topologies as laid out in **Section 1.1 Step 5**. Our goal is to determine if the inference for each trio is adequate for identifying the generating model or if unexpected indicator vectors arise.
- (3.) - Investigate, if any, the indicator vectors allocated to the "Other" class.

# Appendix

## Definitions

$V_i$  - The  $i^{th}$  instrumental variable when  $q > 1$

$T_i$  - a non-instrumental variable/node

$p$  - the number of non-instrumental variables/nodes in a network

$q$  - the number of instrumental variables

$g$  - the number of confounding variables selected for a network

$m$  - the number of permutations to preform in a permuted regression (mediation test)

$n$  - the sample size of the data

$\mathbf{U}$  - the  $(n \times g)$  matrix whose columns represent confounding variables

$\mathbf{X}$  - the  $(n \times p + q + g)$  data matrix of all variables/nodes and all confounders

$\mathbf{H}$  - the  $(p + q + g \times p + q + g)$  precision matrix

$A$  - a  $(p + q \times p + q)$  adjacency matrix for the network

$G(A, B, C)$  - a graph with nodes A, B, and C

$F(\cdot)$  - the Fisher transformation function

$f_{\text{minor}}$  - The frequency of the minor variant of  $V_i$  (when  $V$  represents a type of genetic variation)

$\gamma$  - the threshold frequency of the minor variant for which we determine if a permuted regression is needed

$\rho_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}}$  - the partial correlation between the  $i^{th}$  and  $j^{th}$  columns/variables of  $\mathbf{X}$

Table 1: - Expected results for the tests on the regression coefficients under each model scenario (trios with variants only).

Model	$\beta_{11}$	$\beta_{21}$	$\beta_{12}$	$\beta_{22}$	$V_1 \perp\!\!\!\perp T_2$	$V_1 \perp\!\!\!\perp T_1$
<b>M0</b>	$\neq 0$	$= 0$	$= 0$	$= 0$	Yes	
	$= 0$	$= 0$	$\neq 0$	$= 0$		Yes
<b>M1</b>	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	No	
	$= 0$	$\neq 0$	$\neq 0$	$\neq 0$		No
<b>M2</b>	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	Yes	
	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$		Yes
<b>M3</b>	$\neq 0$	$= 0$	$\neq 0$	$= 0$	No	
<b>M4</b>	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	No	
<b>Conditionally: <math>Y \sim</math></b>	$V_i T_j, \mathbf{U}$	$T_j V_i, \mathbf{U}$	$V_i T_i, \mathbf{U}$	$T_j V_i, \mathbf{U}$		

Table 2: - Expected indicator table for the tests on the regression coefficients under each model scenario (trios with variants only). Note that 1 indicates a rejection of  $H_0$  and 0 indicates a failure to reject

Model	$H_0 : \beta_{11} = 0$	$H_0 : \beta_{21} = 0$	$H_0 : \beta_{12} = 0$	$H_0 : \beta_{22} = 0$	$H_0 : V_1 \perp\!\!\!\perp T_2$	$H_0 : V_1 \perp\!\!\!\perp T_1$
<b>M0</b>	1	0	0	0	0	
	0	0	1	0		0
<b>M1</b>	1	1	0	1	1	
	0	1	1	1		1
<b>M2</b>	1	1	1	1	0	
	1	1	1	1		0
<b>M3</b>	1	0	1	0	1	
<b>M4</b>	1	1	1	1	1	

Table 3: - The set up for the adjacency matrix for each 3-node network denoted by the graph  $G(T_i, T_j, T_k)$   $i, j, k \in \{1 : p\}; \forall i \neq j \neq k$  found using the regressions outlined in **Section 1.3**. Each entry in the table shows the null hypothesis used for testing the edge between the node in the row and node in the column. An entry with 1 means we keep the edge between the nodes (i.e we reject  $H_0$ ) and a 0 means we remove the edge between the nodes (i.e we fail to reject  $H_0$ )

Response	$T_i$	$T_j$	$T_k$
$T_i$	0	$H_0 : T_i \perp\!\!\!\perp T_j   T_k, \mathbf{U}$	$T_i \perp\!\!\!\perp T_k   T_j, \mathbf{U}$
$T_j$	$H_0 : T_j \perp\!\!\!\perp T_i   T_k, \mathbf{U}$	0	$T_j \perp\!\!\!\perp T_k   T_i, \mathbf{U}$
$T_k$	$H_0 : T_k \perp\!\!\!\perp T_i   T_j, \mathbf{U}$	$H_0 : T_k \perp\!\!\!\perp T_j   T_i, \mathbf{U}$	0

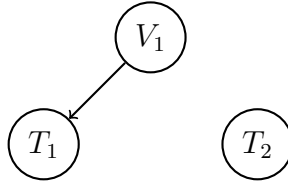


Figure 1: M0 -  $V_1 \not\perp\!\!\!\perp T_1; V_1 \perp\!\!\!\perp T_2; T_1 \perp\!\!\!\perp T_2$

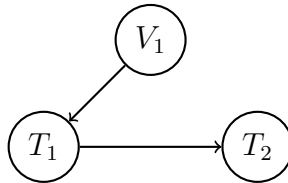


Figure 2: M1 -  $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; V_1 \perp\!\!\!\perp T_2 | T_1$

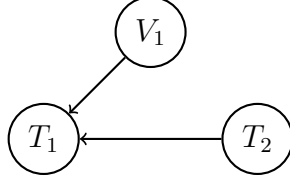


Figure 3: M2 -  $V_1 \not\perp\!\!\!\perp T_1; V_1 \perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; V_1 \not\perp\!\!\!\perp T_2|T_1$

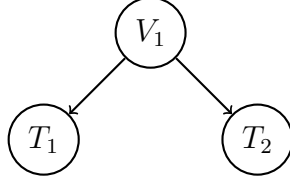


Figure 4: fig: M3 -  $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; T_1 \perp\!\!\!\perp T_2|V_1$

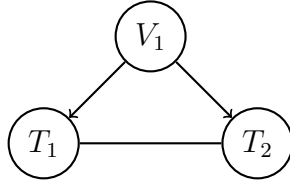


Figure 5: fig: M4 -  $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2|V_1$

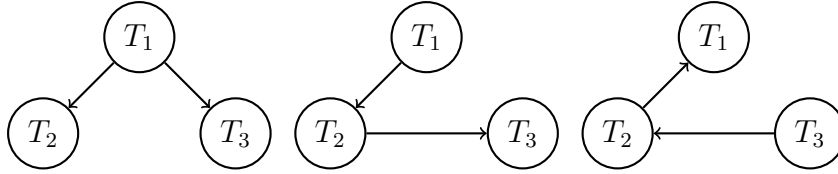


Figure 6: fig: The three graphs above are Markov equivalent meaning they share the same conditional and marginal independence relations Markove:  $T_1 \perp\!\!\!\perp T_3|T_2$  Minimality:  $T_1 \not\perp\!\!\!\perp T_2; T_2 \not\perp\!\!\!\perp T_3$  faithfulness:  $T_1 \not\perp\!\!\!\perp T_3$  and are therefore indistinguishable from each other



# Details On Statistical Methods

## Calculating Precision

Given a data matrix  $\mathbf{X}$  of  $q$  instrumental variables,  $p$  non-instrumental variables, and  $g$  confounders:

Assuming  $\mathbf{X}$  is centered:

$$X \sim N_k(\mathbf{0}, \Sigma) \quad \text{for } k = p + q + g$$

Then the precision matrix of  $\mathbf{X}$  is defined as

$$\mathbf{H} = \Sigma^{-1}$$

$\mathbf{H}$  can be scaled to the partial correlation matrix for the entries in  $\mathbf{X}$ . Given the entry in the  $i^{th}$  row and  $j^{th}$  column of  $\mathbf{H}$ :

$$\mathbf{x}_i, \mathbf{x}_j | \mathbf{x}_{-(i,j)} = -\frac{h_{ij}}{\sqrt{h_{ii}}\sqrt{h_{jj}}} = \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}}$$

which is a measure of the association between the  $i^{th}$  and  $j^{th}$  columns/variables in  $\mathbf{X}$  conditioned on all other variables.

The Fisher transformation can be used to formulate a test for each partial correlation coefficient of interest:

$$\frac{\sqrt{n - |\mathbf{x}_{-i,j}|} - 3}{2} \ln \left( \frac{1 + \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}}}{1 - \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}}} \right) \approx N(0, 1)$$

where null and alternative hypotheses are

$$H_0 : \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}} = 0 \quad H_A : \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}} \neq 0$$

$$\text{reject } H_0 \text{ if } |Z_{\text{obs}}| > Z_{1-\alpha/2}$$

by applying the cases:

$$a_{i,j} = \begin{cases} 1 & \text{if } 2 \times P(Z > |Z_{\text{obs}}|) < \alpha \\ 0 & \text{else} \end{cases} \quad \forall i, j \in \{1 : p + q\}$$

we can obtain the  $(p + q \times p + q)$  adjacency matrix  $\mathbf{A}$  for the network skeleton

## Permutated Regression mediation test

repeat  $m$  times: permute  $T_j$  in (1) within the levels of  $V_i$  denoted  $T_j^*$ . Similarly, permute  $T_i$  in (2) within the levels of  $V_k$  denoted  $T_i^*$ . Next perform the regressions using the permuted variables:

$$T_i = \beta_0 + \beta_{1i}V_k + \beta_{2i}^*T_j^* + \Gamma\mathbf{U} + \epsilon \quad (5)$$

$$T_j = \beta_0 + \beta_{1j}V_k + \beta_{2j}^*T_i^* + \Gamma\mathbf{U} + \epsilon$$

Let  $\Theta_{2i}$  and  $\Theta_{2j}$  denote the  $(m \times 1)$  vectors representing the collection of  $t$  statistics from the wald tests on  $\beta_{2i}^*$  and  $\beta_{2j}^*$  coefficients (respectively) from the permuted regressions in **Step 2.** such that:

$$\Theta_{2i} = [T_{2i}^{*(1)}, T_{2i}^{*(2)}, T_{2i}^{*(3)}, \dots]$$

$$\Theta_{2j} = [T_{2j}^{*(1)}, T_{2j}^{*(2)}, T_{2j}^{*(3)}, \dots]$$

We next test the conditional association between  $T_i$  and  $T_j$  using the nominal test defined by Yang et. al., 2017. Let  $T_{\text{obs}_{2i}}$  be the observed wald statistic from (1) and  $T_{\text{obs}_{2j}}$  be the observed wald statistic from (2). We formulate the testable hypotheses:

$$H_0 : T_{\text{obs}_{2i}} = \mu_{2i}^*, \quad H_A : T_{\text{obs}_{2i}} \neq \mu_{2i}^*$$

and

$$H_0 : T_{\text{obs}_{2j}} = \mu_{2j}^*, \quad H_A : T_{\text{obs}_{2j}} \neq \mu_{2j}^*$$

where  $\mu_{2i}^*$  and  $\mu_{2j}^*$  denote the centers of the non-central  $t$ -distributions of  $\Theta_{2i}$  and  $\Theta_{2j}$  respectively. Therefore the mediation test statistic is:

$$Z_{\text{obs}_{ij}} = \frac{T_{\text{obs}_{ij}} - \frac{\sum \Theta_{ij}}{m}}{SE(\Theta_{ij})}$$

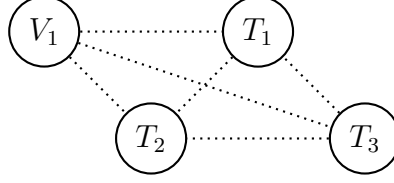
where we

$$\text{reject } H_0 \text{ if } 2 \times P(Z > |Z_{\text{obs}_{ij}}|) < \alpha$$

## Inferring the Network Among Non-Instrumental Variables

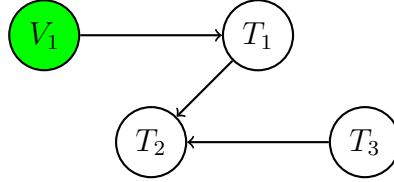
Because we infer the network among all 3-node networks involving an instrumental variable first, we have additional information for each 3-node network among T-nodes. When we lack an instrumental variable, we can only uniquely infer M2. As a result we cannot distinguish M1 and M3 due to Markov equivalence, but we are able to determine the undirected graphs of M0 and M4. (**Figure 6**). Exploiting the information provided by

the associations between the T-nodes and the instrumental variable(s) we obtain in **Step 3** of the **General Algorithm** we can actually view the problem as a partially inferred network among 4 nodes i.e the graph  $G(V_1, T_1, T_2, T_3)$ .



The above graph demonstrates how we may view the problem of classifying trios of T-nodes, by exploiting whatever relationships exist in the subnetwork of the graphs  $G(V_1, T_1, T_2)$ ,  $G(V_1, T_1, T_3)$  and  $G(V_1, T_2, T_3)$ .

**Example 1: v-structure in  $G(T_1, T_2, T_3)$**

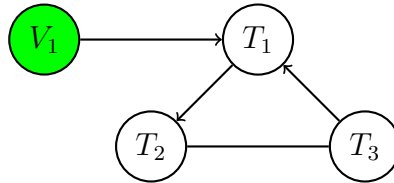


Take the example above as one such network that may exist among the nodes  $V_1, T_1, T_2, T_3$  defined by the individual graph types:

$$G(V_1, T_1, T_2, T_3) = \begin{cases} G(V_1, T_1, T_2) : M1 \\ G(V_1, T_1, T_3) : M0 \\ G(V_1, T_2, T_3) : Other \\ G(T_1, T_2, T_3) : M2 \end{cases}$$

Therefore, we require the step of inferring  $G(T_1, T_2, T_3)$  to identify v-structures in the non-instrumental variable network of T-nodes. If we did not preform a separate trio analysis for the nodes  $T_1, T_2, T_3$  we could still infer all the edges in the above graph, however the edge between  $T_2$  and  $T_3$  would be left undirected. Therefore, by including this step in the algorithm we are able to direct an additional edge whenever the relationship between  $T_i, T_j, T_k$  is of type M2:

**Example 2: No v-structure in  $G(T_1, T_2, T_3)$**



$$G(V_1, T_1, T_2, T_3) = \begin{cases} G(V_1, T_1, T_2) : M1 \\ G(V_1, T_1, T_3) : M2 \\ G(V_1, T_2, T_3) : Other \\ G(T_1, T_2, T_3) : M4 \end{cases}$$

This is a counter example where the inference for the subnetwork for  $T_1, T_2, T_3$  would be returned as M4 which is undirectable in the absence of the PMR assumption. As a result the subnetworks for the combinations of nodes involving the variant fully specify the graph  $G((V_1, T_1, T_2, T_3))$