Resolving direct causal effects from causal networks

Gibran Hemani

2017-01-30

Contents

Code available at: https://scmv-ieugit.epi.bris.ac.uk/gh13047/graph_mr

Summary

Suppose we have M traits, each of which can be instrumented. We can calculate the causal relationships of $M \times M$ pairs of traits. This constructs a matrix of pairwise causal estimates of the total effects of each trait on each of the other traits, many of which can be null. This paper attempts to deconvolve those

'total' effects, but to deconvolve into a set of 'direct' effects, it looks like simply taking the inverse of this matrix is quite reliable, and performs favourably compared to other methods of deconvolution.

Introduction

Growth in summary data for GWASs on phenotypes is on a steep trajectory, such that we may soon be asymptoting towards a situation where we can use two-sample MR to test 'everything against everything'. This premise, however, requires methodological advancements to address several issues including:

- Multiple testing
- Decomposing many indirect or total effects into a terse set of direct effects

The latter can be summarised as follows. Suppose there are five variables of interest, 1-5, and the causal relationships are

```
1 -> 2
2 -> 3
3 -> 4
4 -> 5
```

This can be depicted in graph form as in Figure 1.

If, however, we performed MR of `1 -> 3`, `1 -> 4`, etc, we would identify associations because they exist indirectly. Hence, after testing everything against everything our graph would look like Figure 2.

The task is to decompose the complete set of associations into the direct effects only. This has the following advantages:

- Identify direct pathways through which a particular exposure influences an outcome
- Identify instances of partial mediation, which might suggest that there are unknown variables that remain to be uncovered that mediate the path from exposure to outcome

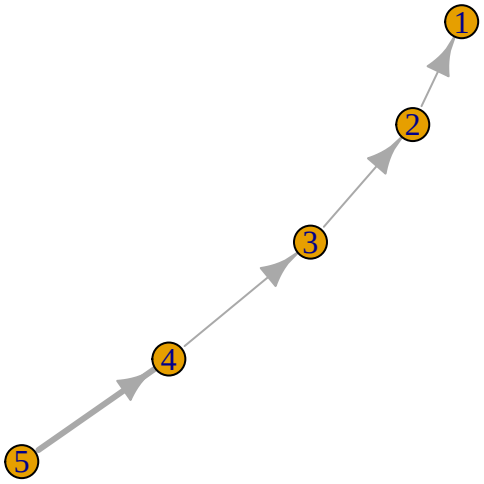


Figure 1: Simulated causal relationships

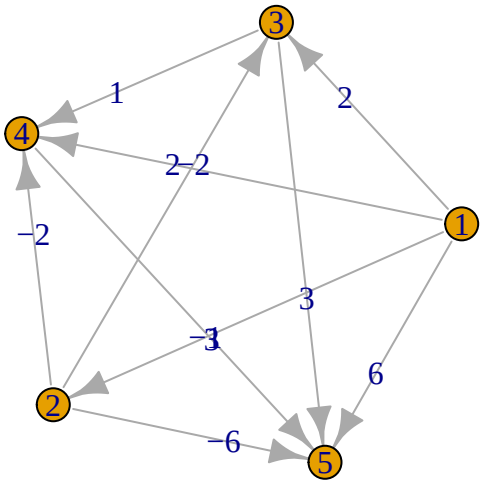


Figure 2: Empirical associations

The procedure to do this would be as follows:

1. Estimate the causal effect of every variable against every other variable. e.g. for a 5 variable graph, this means performing $5 \times 5 = 25$ MR analysies
2. Deconvoluting this set of estimates into a matrix of direct effects

Part 1 is crucial, and as MR techniques continue to improve (i.e. avoiding problems due to invalid instruments, improving power etc) this will become increasingly accurate. In these analyses I'll just assume that these total effects have been calculated reliably. Part 2 is the focus of this work.

MR for mediation (AKA network MR) has already been developed, especially for the case where there are only three phenotypic variables (Burgess et al. 2015). Here, the direct effect of trait 1 on trait 2, $\beta_{1\Rightarrow 2}$, is straightforward

$$\beta_{1\Rightarrow 2} = \beta_{1\rightarrow 2} - \beta_{1\rightarrow 3}\beta_{3\rightarrow 2}$$

With four variables it looks like:

$$\begin{aligned}\beta_{1\Rightarrow 2} = \beta_{1\rightarrow 2} &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 4}\beta_{4\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 2} \\ &- \beta_{1\rightarrow 4}\beta_{4\rightarrow 2}\end{aligned}$$

With five variables it looks like:

$$\begin{aligned}\beta_{1\Rightarrow 2} = \beta_{1\rightarrow 2} &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 4}\beta_{4\rightarrow 5}\beta_{5\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 4}\beta_{4\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 5}\beta_{5\rightarrow 2} \\ &- \beta_{1\rightarrow 4}\beta_{4\rightarrow 5}\beta_{5\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 5}\beta_{4\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 2} \\ &- \beta_{1\rightarrow 4}\beta_{4\rightarrow 2} \\ &- \beta_{1\rightarrow 5}\beta_{5\rightarrow 2}\end{aligned}$$

and this is performed for each of the 5×5 possible pairwise combinations of variables, ultimately reducing a matrix of total effect relationships, R_t e.g.

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]   -1    1    0    0    0
## [3,]    2   -2    1    0    0
## [4,]    2   -2    1    1    0
## [5,]    6   -6    3    3    1
```

into a matrix of direct relationships, R_d e.g.

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]   -1    1    0    0    0
## [3,]    0   -2    1    0    0
## [4,]    0    0    1    1    0
## [5,]    0    0    0    3    1
```

There are two problems with this approach. First, the combinatorial increase in the number of terms that are required for calculating the direct effects gets large very quickly. For example, for a graph with 10 variables

there are 52 unique paths for each of the 100 elements in the matrix, and identifying those paths itself is a computationally slow process. Second, perhaps more importantly, this method may not actually generalise beyond the three-variable graph.

Simulations

I will assume that there are M variables measured in N samples represented in a $N \times M$ matrix P. Further, each variable has a valid instrument, hence there are M instruments also, represented in a $N \times M$ matrix G. As stated above, in this analysis I am assuming that every causal estimate made by MR is reliable.

DAGs are simulated such that the M variables are related to each other by random causal effects. Cycles are avoided. Following on, two-stage least squares is used to calculate all pairwise causal relationships e.g.

$$R_t(1 \to 2) = \frac{cov(P_{,2}, G_{,1} cov(P_{,1}, G_{,1}/var(G_{,1})))}{var(G_{,1} cov(P_{,1}, G_{,1}/var(G_{,1})))}$$

Three methods are then used to try to deconvolve the graph R_t into R_d.

Method 1 - mediation by MR

This is as described in the Introduction

Method 2 - inversion

Simply a method to orthogonalise the matrix by

$$R_d = R_t^{-1}$$

method 3 - Feizi deconvolution

In (Feizi et al. 2013) a method is outlined for network deconvolution that is primarily aimed at correlation matrices (i.e. symmetric, non-causal versions of R_t). The method is:

$$R_d = R_t(I + R_t)^{-1}$$

Standard errors

To obtain the standard errors of the direct effects for the inversion method, we can use bootstrapping. Here, each element in the matrix of total effects is resampled with $R_t(i,j)* \sim N(R_t(i,j), se(R_t(i,j)))$. The inversion method is then applied to the resampled matrix R_t* and the results R_d* are stored. This is performed 1000 times to obtain a distribution of effects for each element of the R_d matrix. The standard deviation of the distribution from each element is taken to be the standard error of that direct effect estimate.

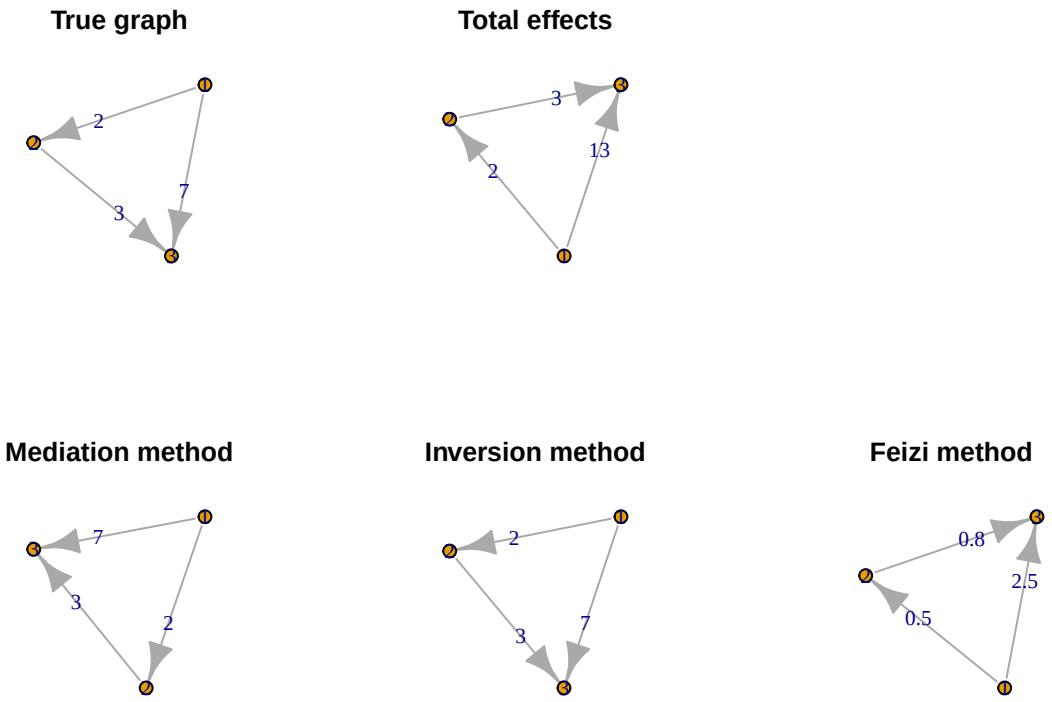


Figure 3: Three variables, N=500000

Results

The analysis is divided into two sections. First, a demonstration that in most cases the inversion method is rapid and returns the true direct effects; and second, an evaluation of statistical power, particularly in comparing the ability to detect a total effect vs detecting a chain of .

Simulation 1 - three-variable networks

Simulate 500000 samples with three variables, with the following real causal structure

The three methods agree on the causal structures as the true graph. The total effects graph has much larger effects for $1 \to 3$ as expected, because this is the sum of the direct and indirect effects. The Feizi method shrinks the effect sizes somewhat.

Simulation 2 - four-variable networks

Simulate 500000 samples with three variables, with the following real causal structure

Similar story to the three variable network, but of course there aresome paths in the R_t which should not be present in the estimates of R_d. Another way to evaluate agreement is to plot the values of the matrix elements against the true matrix elements, e.g.

Here we can see that the mediation and inversion methods are reasonably accurate, but the Feizi method seems to be very close to the total effects.

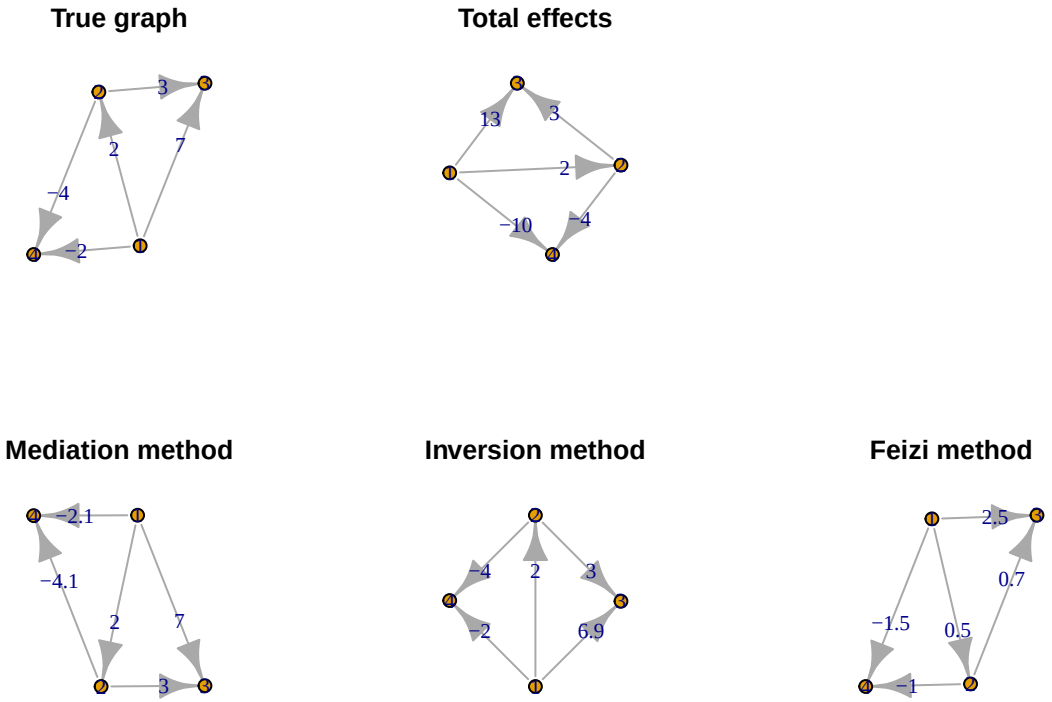


Figure 4: Graphs of four variables, N=500000

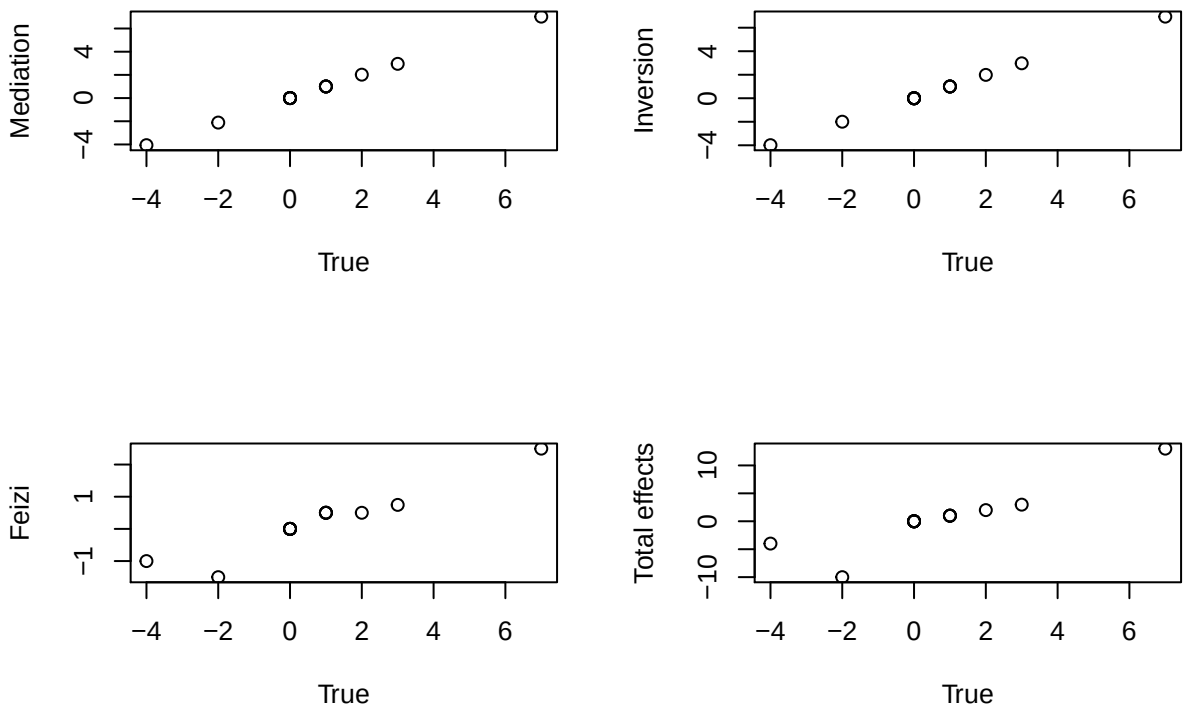


Figure 5: Matrix elements of four variables, N=500000

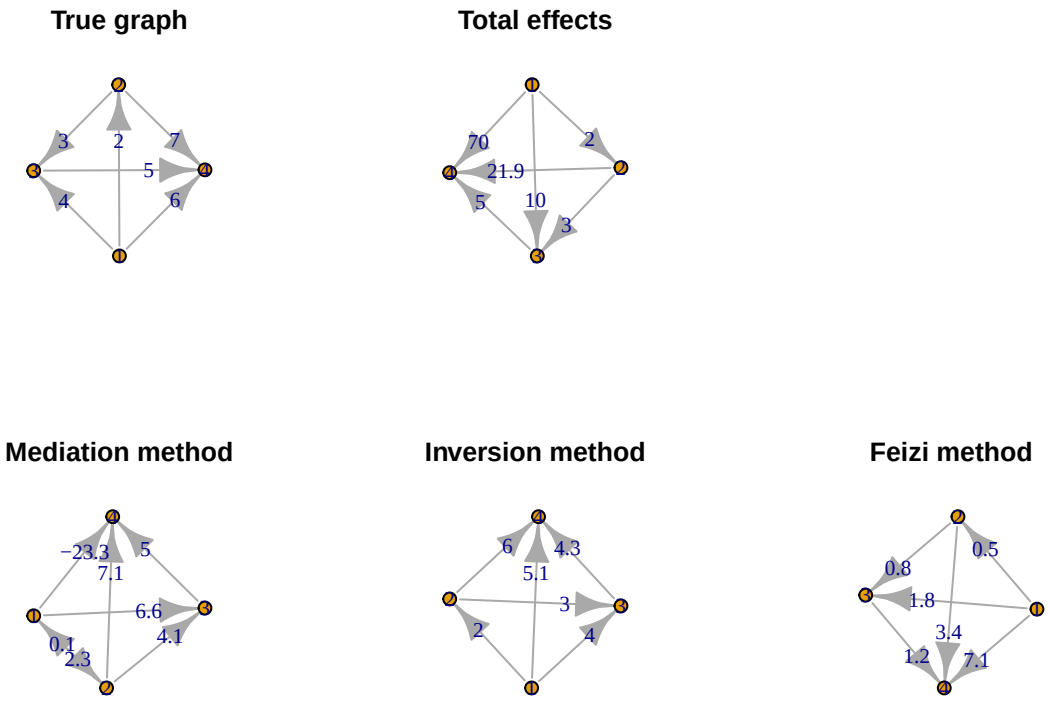


Figure 6: Graphs of four variables, N=500000

Simulation 3 - four-variable networks, more complex

Here the graph is slightly more complex

And it can be seen that the mediation method does not resolve the direct effects accurately for one particular edge

---

Simulation 4 - five variables, causal chain

Increasing to 5 variables, and there is a clear problem with the mediation and Feizi methods

The inversion method seems to be fairly reliable

Simulation 5 - 15 variables

After around 7 variables the mediation method becomes too slow to run, so that will be ignored for the subsequent simulations. Here, with 15 variables there is still reasonably good performance from the inversion method.

Simulation 6 - 20 variables, more complex

Here the inversion method has some issues, there are some indirect effects that should be set to 0 in the direct graph but have failed to have been removed.

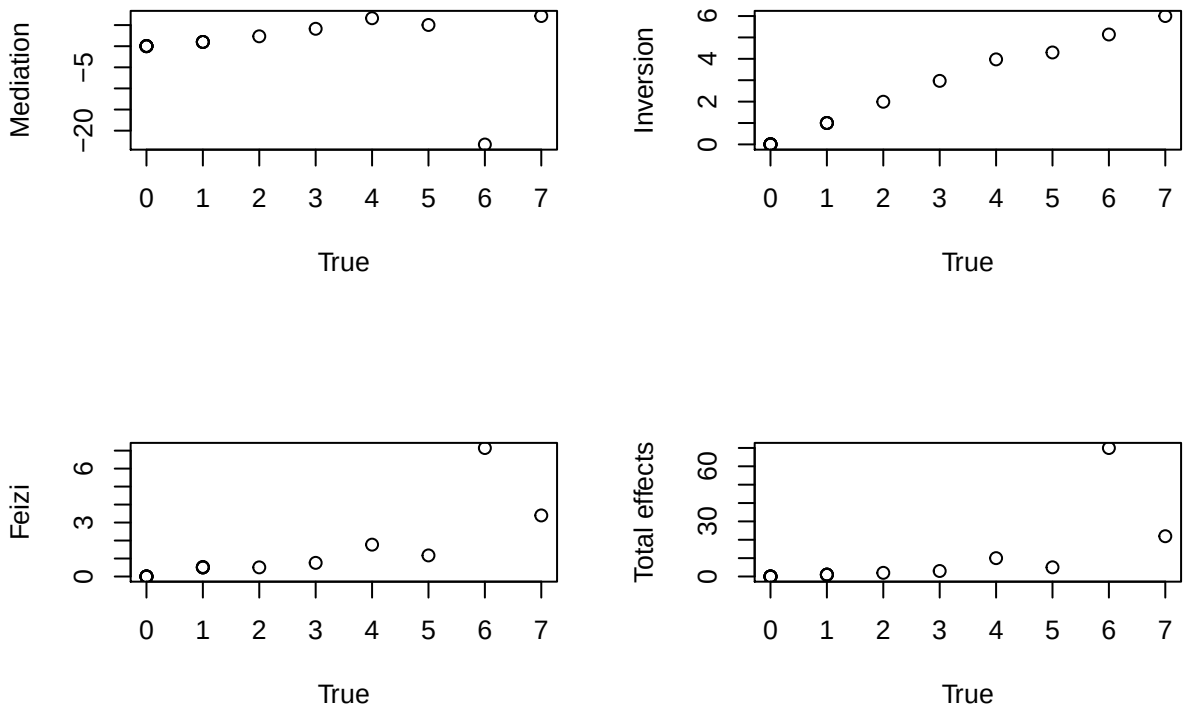


Figure 7: Matrix elements of four variables, N=500000

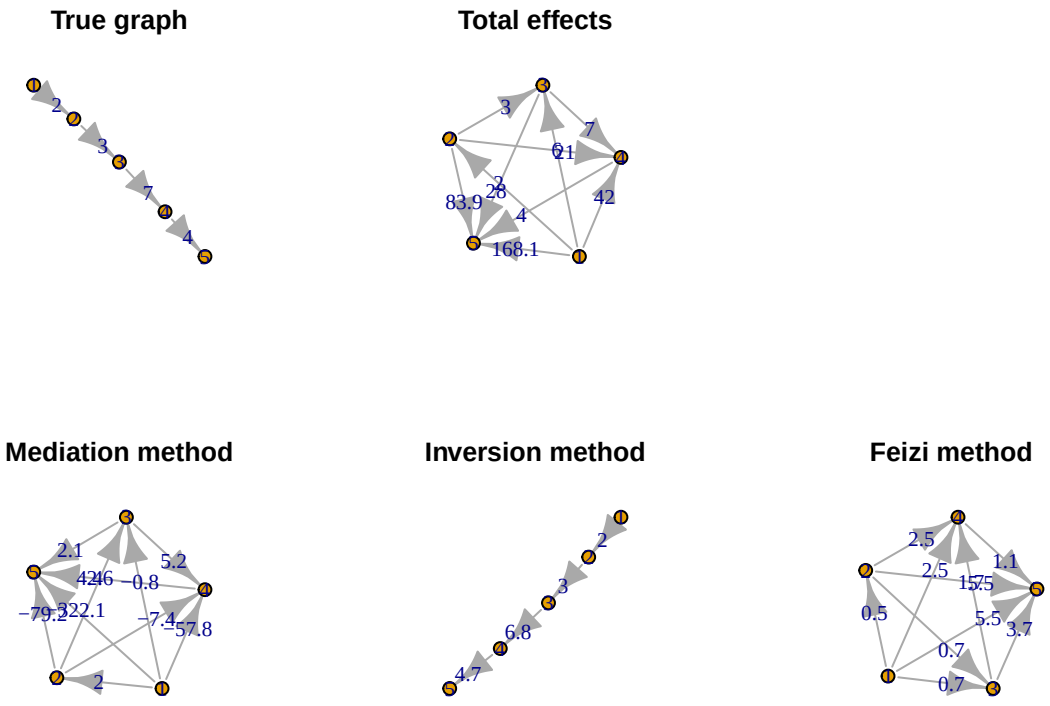


Figure 8: Graphs of five variables, N=500000

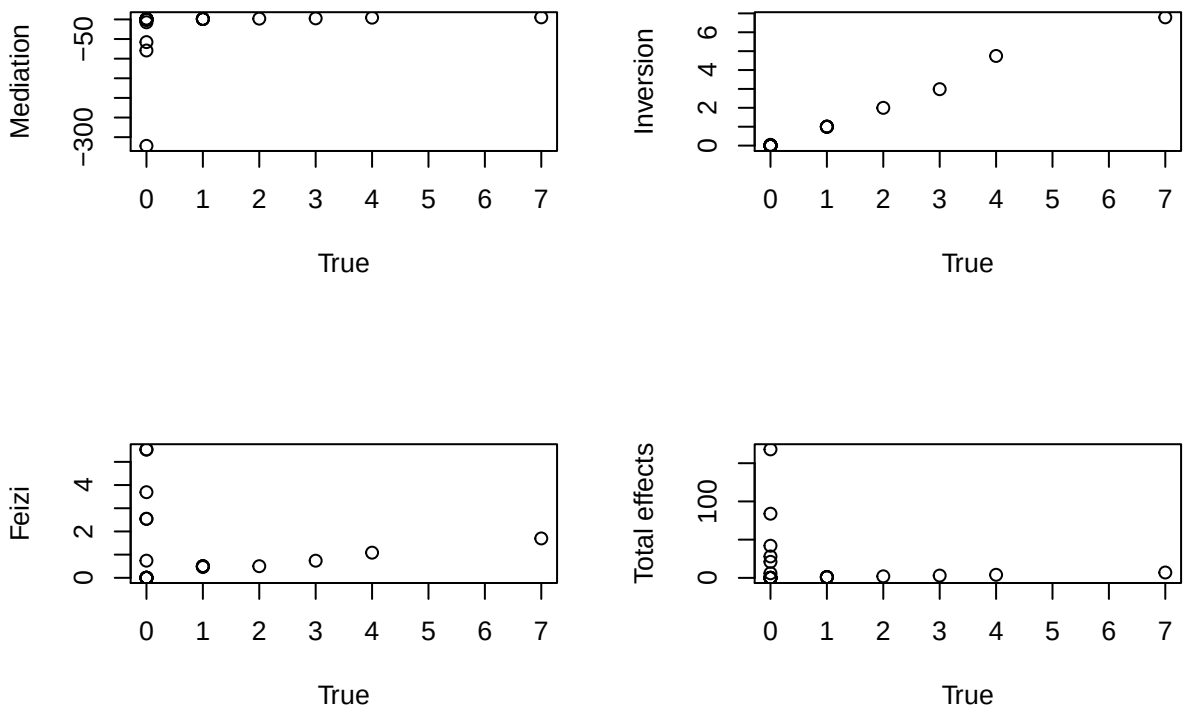


Figure 9: Matrix elements of five variables, N=500000

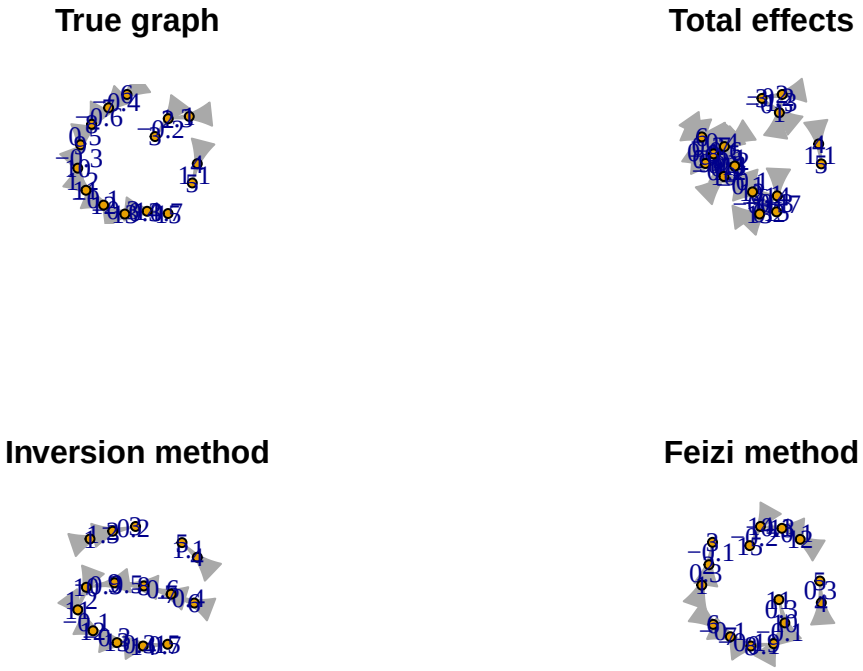


Figure 10: Graphs of 15 variables, N=300000

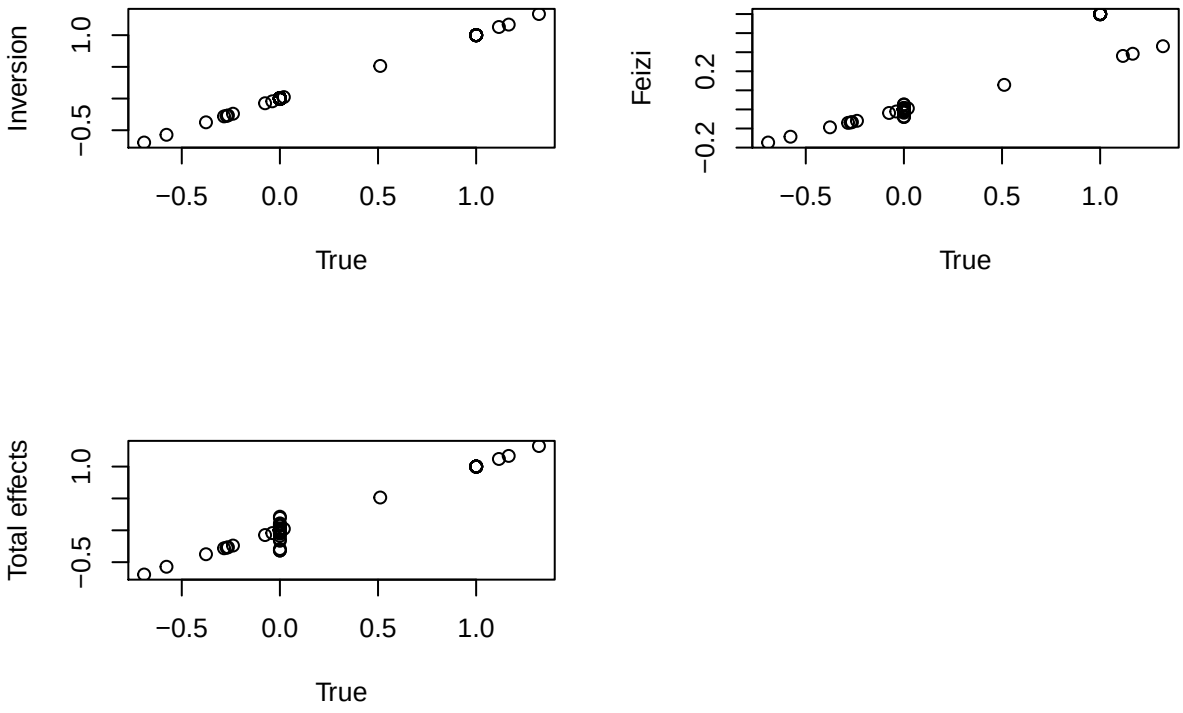


Figure 11: Matrix elements of 15 variables, N=300000

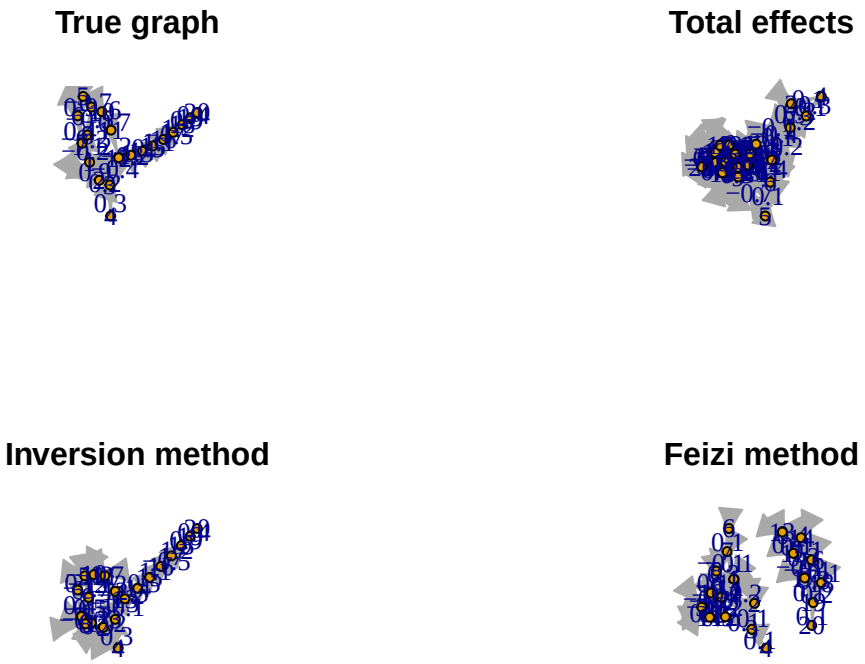


Figure 12: Graphs of 20 variables, N=300000

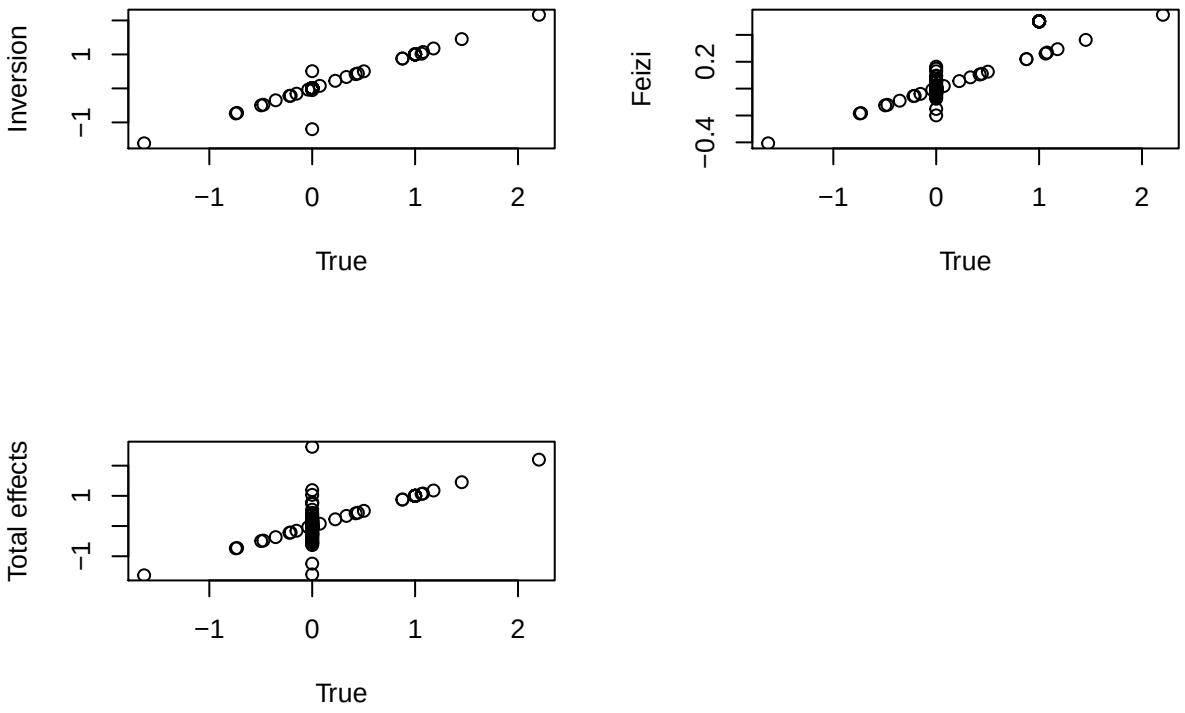


Figure 13: Matrix elements of 20 variables, N=300000

Simulation 7 - 30 variables, non-gaussian direct effects

Here the effect sizes are sampled from an exponential distribution rather than a normal distribution. The inversion method seems relatively robust to this.

Statistical power

As an illustration, in the following analyses a 5-trait network is created as in figure 1. Is it statistically more efficient to

1. estimate the total effect of trait 1 on trait 5 using standard MR, or
2. obtain the direct effects of the 5-trait network, and estimate the causal chain from trait 1 -> 2 -> 3 -> 4 -> 5.

Strategy (2) entails finding a 'significant' p-value at each of the 4 direct relationships that map trait 1 to trait 5.

The simulation was performed using 5000 samples, and the simulated direct effect sizes for each chain were sampled from $\beta\ N(0,1)$.

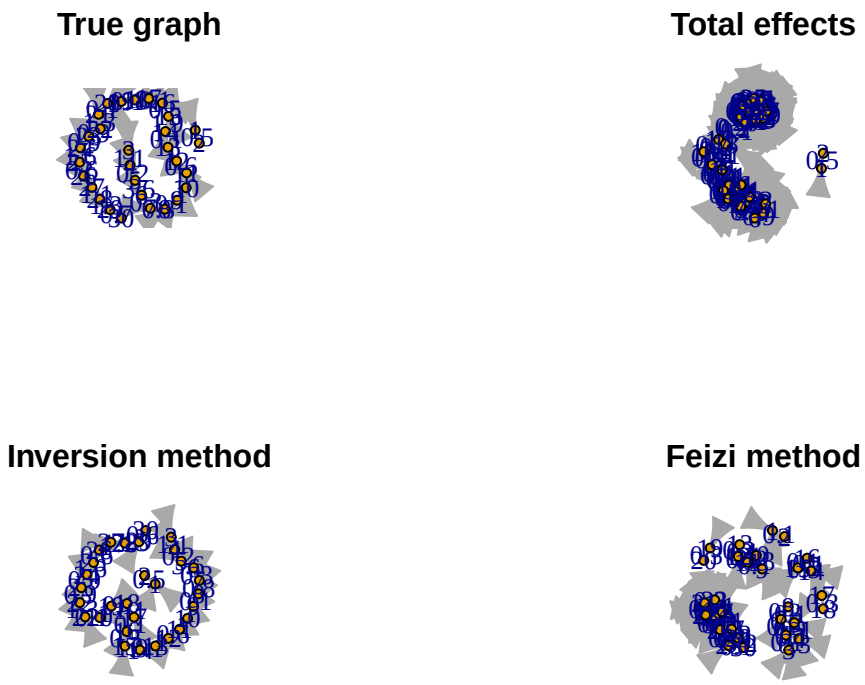


Figure 14: Graphs of 15 variables, N=300000

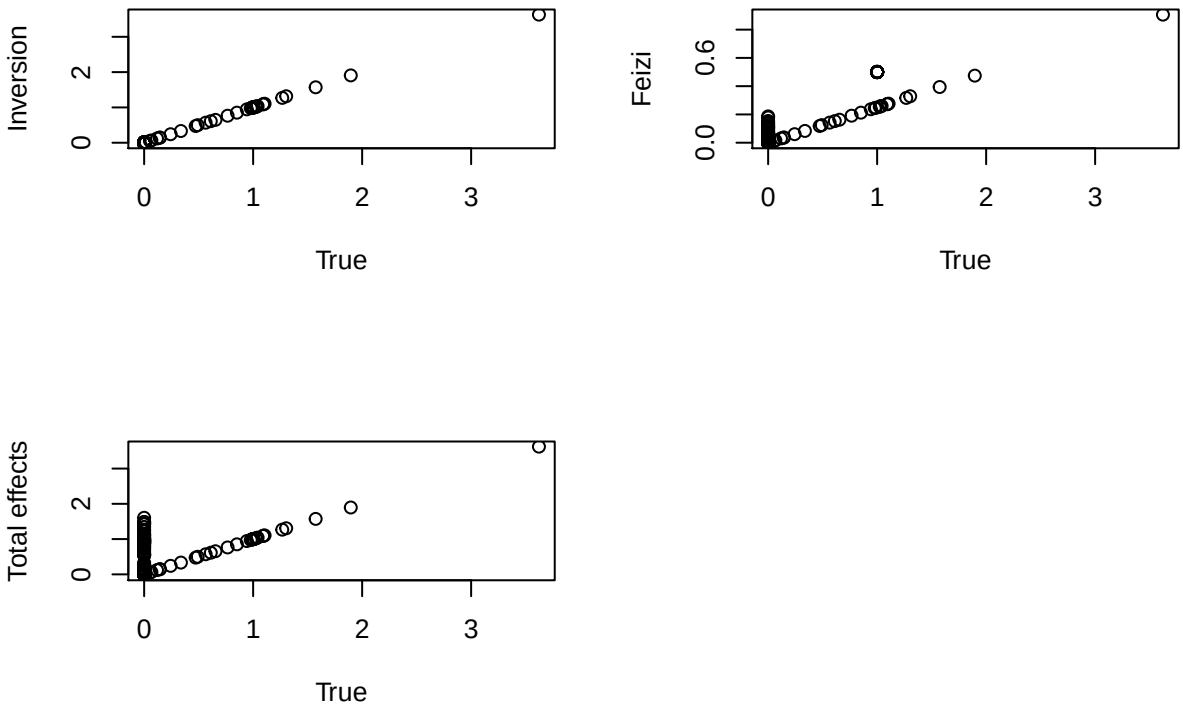
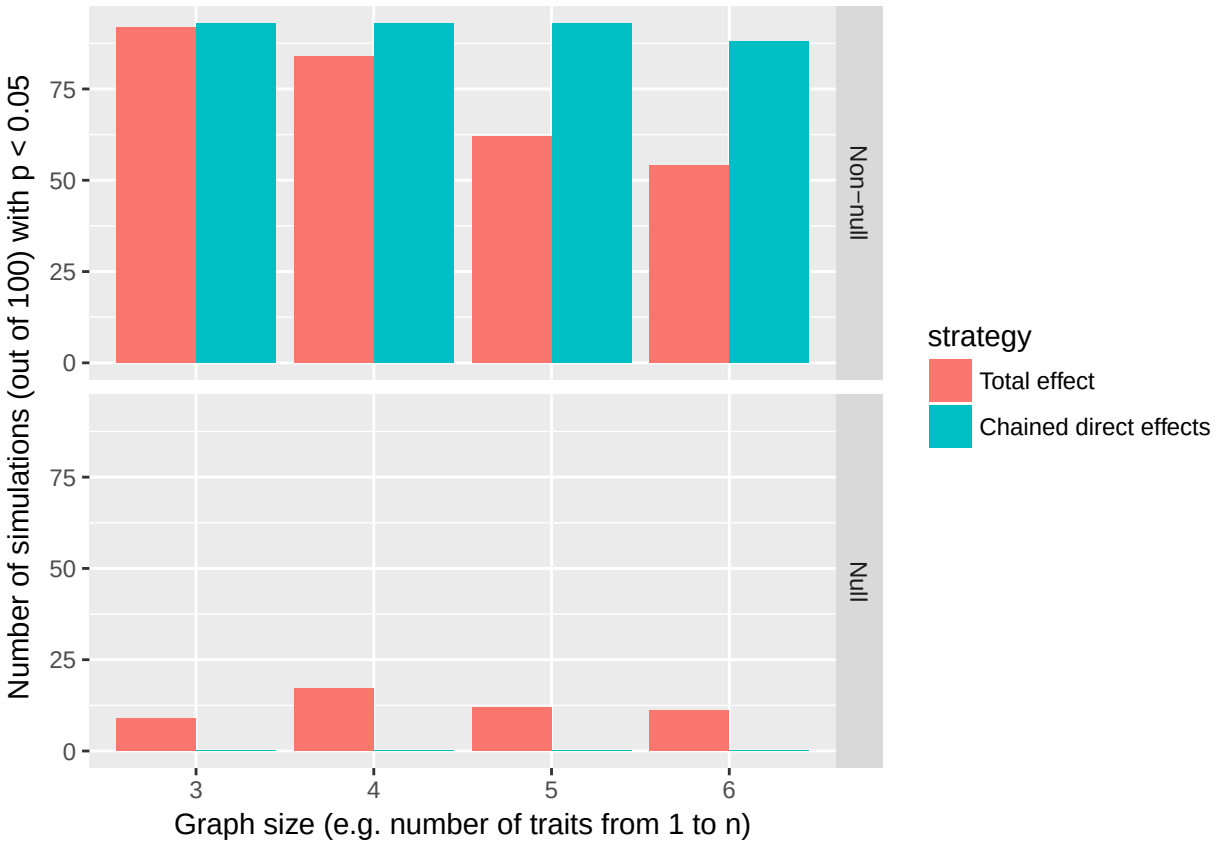


Figure 15: Matrix elements of 15 variables, N=300000



Here it is apparent that as the causal chain grows larger the statistical power of the second strategy, which uses the direct effects estimated from the graph, becomes substantially more powerful than the standard approach of identifying the total causal effect. The false discovery rate of the second strategy appears to compare favourably also.

Discussion

The inversion method appears to be reasonably accurate in a range of scenarios, though not perfect when graphs become larger and relationships more complex.

Issues:

- The mediation method might be interpreted too simplisticly here. I think there needs to be recursion beyond the 3 variable example - i.e. the direct effects are a function of indirect effects at the moment, but these indirect effects probably need to be reduced to direct effects also. This would require identifying a path through which to traverse the graph and recursively estimate the direct effects
- These simulations only looked at relatively sparse graphs, and with no loops.
- Why isn't the Feizi method working? Need to make sure it works for symmetric correlation graphs
- Need to estimate standard errors of direct effects. This could be obtained by bootstrapping
- Haven't evaluated the influence of cycles in the graph - this is hard to simulate though
- How does this improve power? e.g. If A influences F through B, C, D, and E, is it easier to identify the intermediate path than the direct relationship?
- Graphical lasso may be useful to make the matrix sparse

References

Burgess, Stephen, Rhian M Daniel, Adam S Butterworth, and Simon G Thompson. 2015. "Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways." *International Journal of Epidemiology* 44 (2): 484–95. doi:10.1093/ije/dyu176.

Feizi, Soheil, Daniel Marbach, Muriel Médard, and Manolis Kellis. 2013. "Network deconvolution as a general method to distinguish direct dependencies in networks." *Nature Biotechnology* 31 (8). Nature Research: 726–33. doi:10.1038/nbt.2635.