



# Application of Compositional Models to Cardiology

Petr Šimeček, Radim Jiroušek, Marie Tomečková, Jana Zvárová

European Center for Medical Informatics, Statistics and Epidemiology,  
Institute of Computer Science AS CR, Prague, Czech Republic

## Introduction

The poster illustrates a new approach for the representation of multidimensional probability distributions - a suitable way of medical knowledge representation. Namely, the approach is based on the idea of composing a multidimensional distribution from a great number of low-dimensional ones. It quite naturally corresponds to the fact that global knowledge about a medical field can hardly be expressed. In textbooks and specialized monographs, it is always described a number of facts that can be called "pieces of knowledge".

This theory is implemented as a package for R which is called **MUDIM** (stands for "System for **MU**lti**DI**mensional **M**odels"). This package is free and open-source and run on a wide variety of UNIX platforms and on MS Windows 9x/2000/NT/XP.

An example concerns data from cardiology; it describes relationships among different risk factors of atherosclerosis.

## Multidimensional Compositional Models

The basic idea of this approach is that a multidimensional distribution is computed - *composed* - from a system of oligodimensional distributions by iterative application of a special operator of composition. The theoretical issues of this approach are described in [1]. This contribution uses the theory of compositional models but it does not present mathematical details.

However, we want to emphasize that compositional models can be used to build a probabilistic expert systems (Figure 1) or as a data-mining technique (Figure 2).

This poster concentrates on a description of this methodology to modeling and discovering causal relationships in cardiological knowledge.



R



R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.

MUDIM is distributed as a package for R. Thus, its strength is multiplied by all the methods and algorithms implemented in R.



C++



The code of MUDIM is written in C++. This ensures sufficient speed and object orientation. R environment is used as a front-end.

Figure 1: Expert System

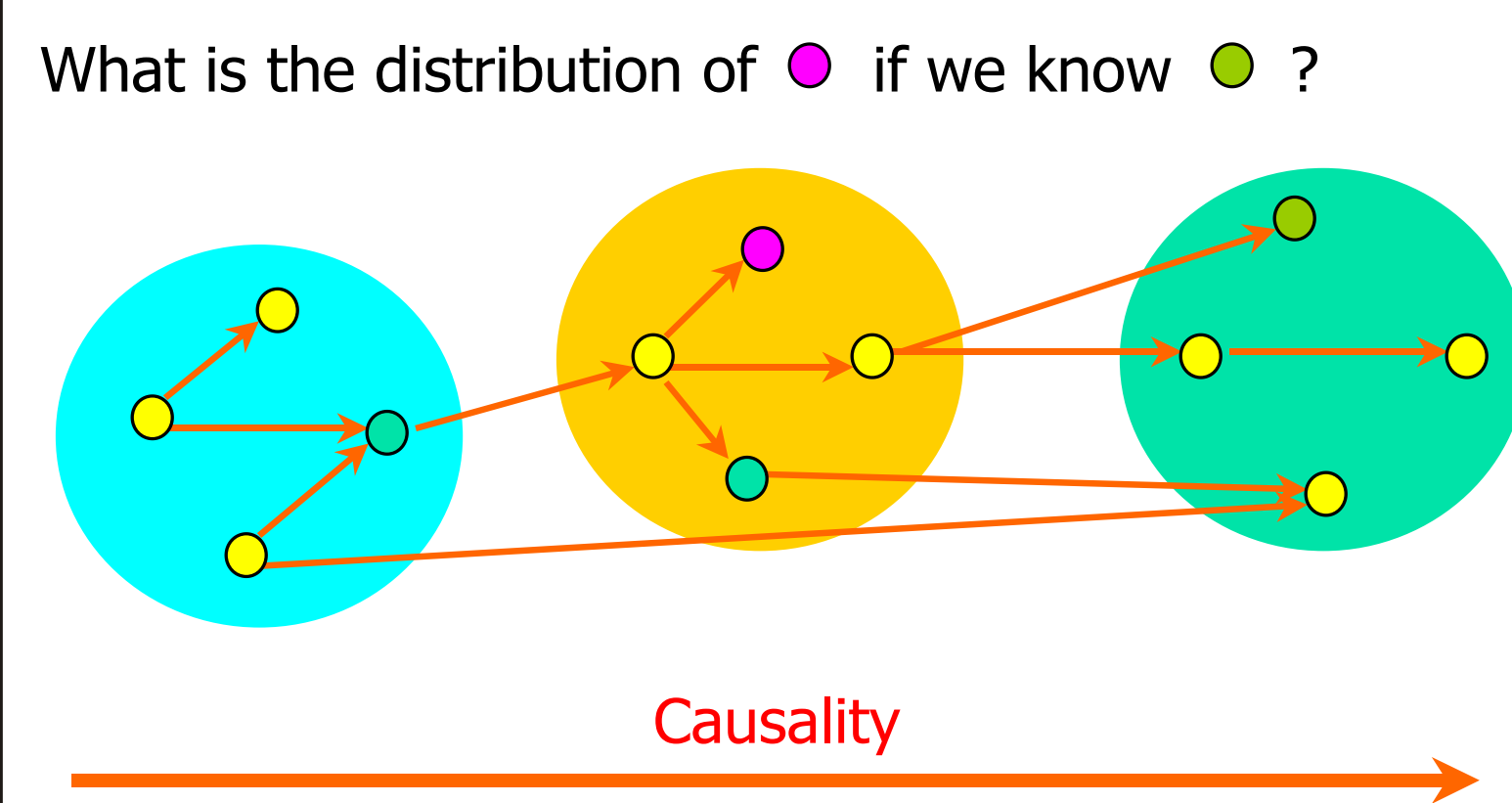


Figure 3: 464 connected pairs (p=0.05)

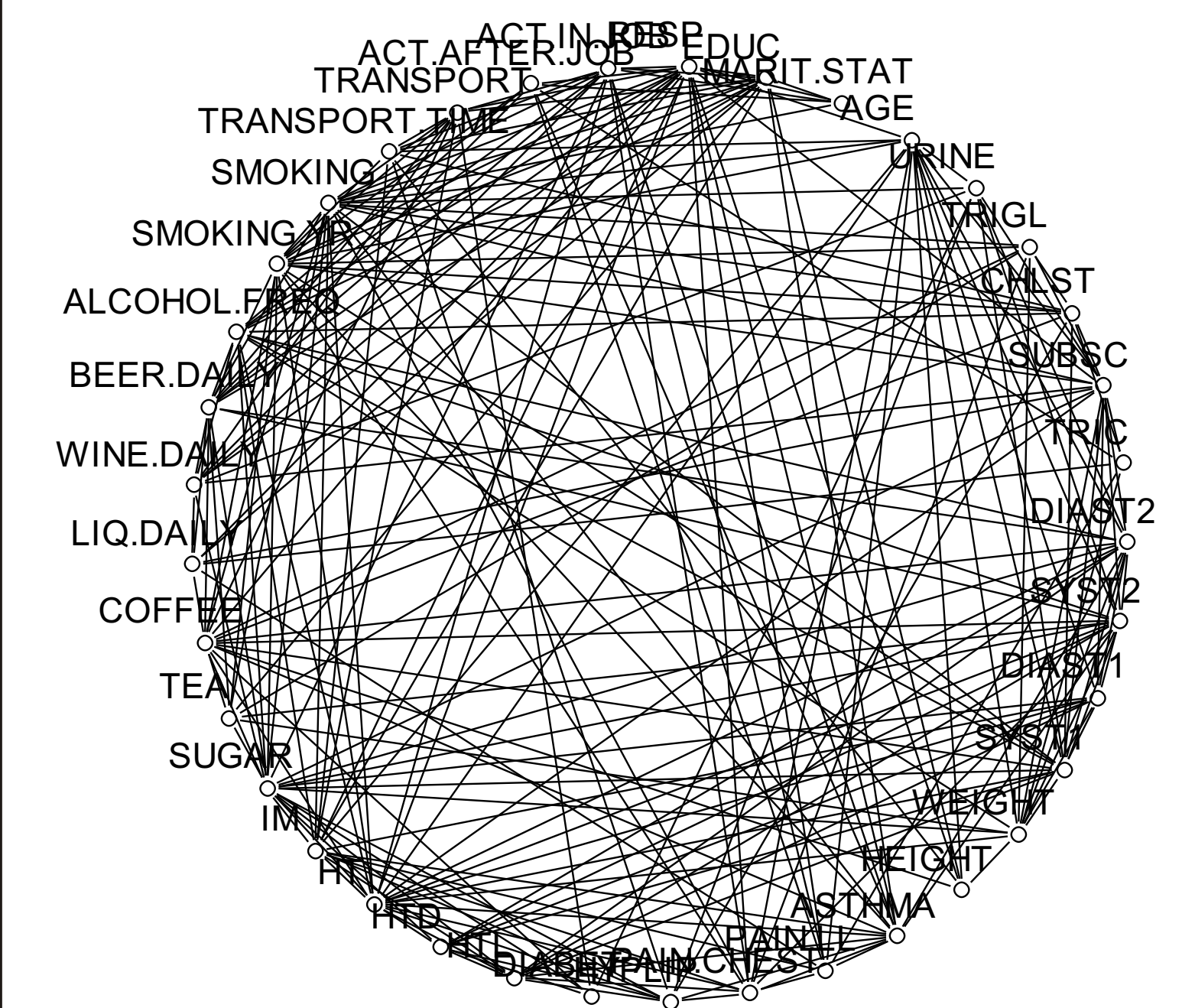


Figure 4: 160 connected pairs (p=0.05/666)

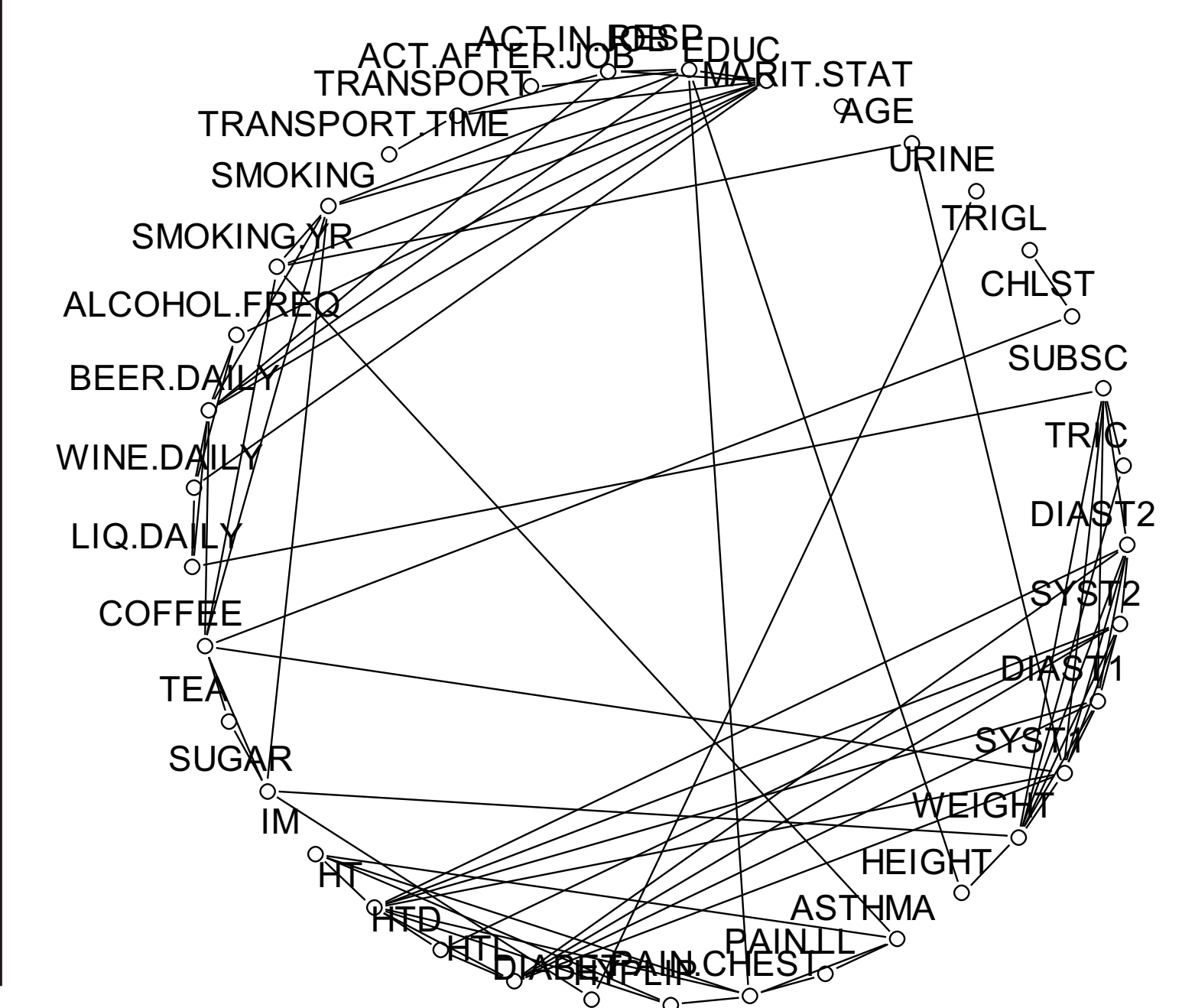


Figure 2: Data mining

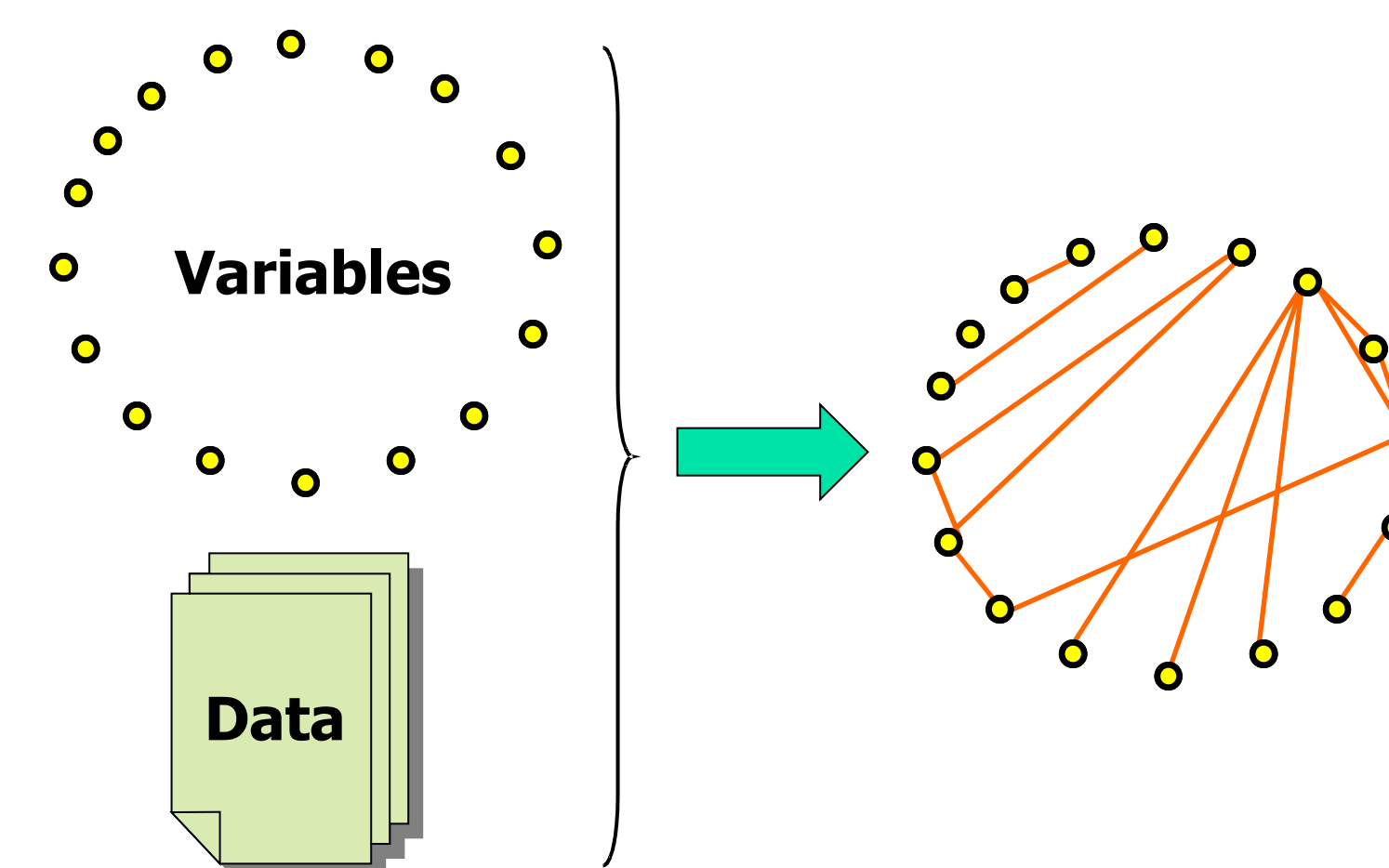


Figure 5: 59 direct causes found by MUDIM

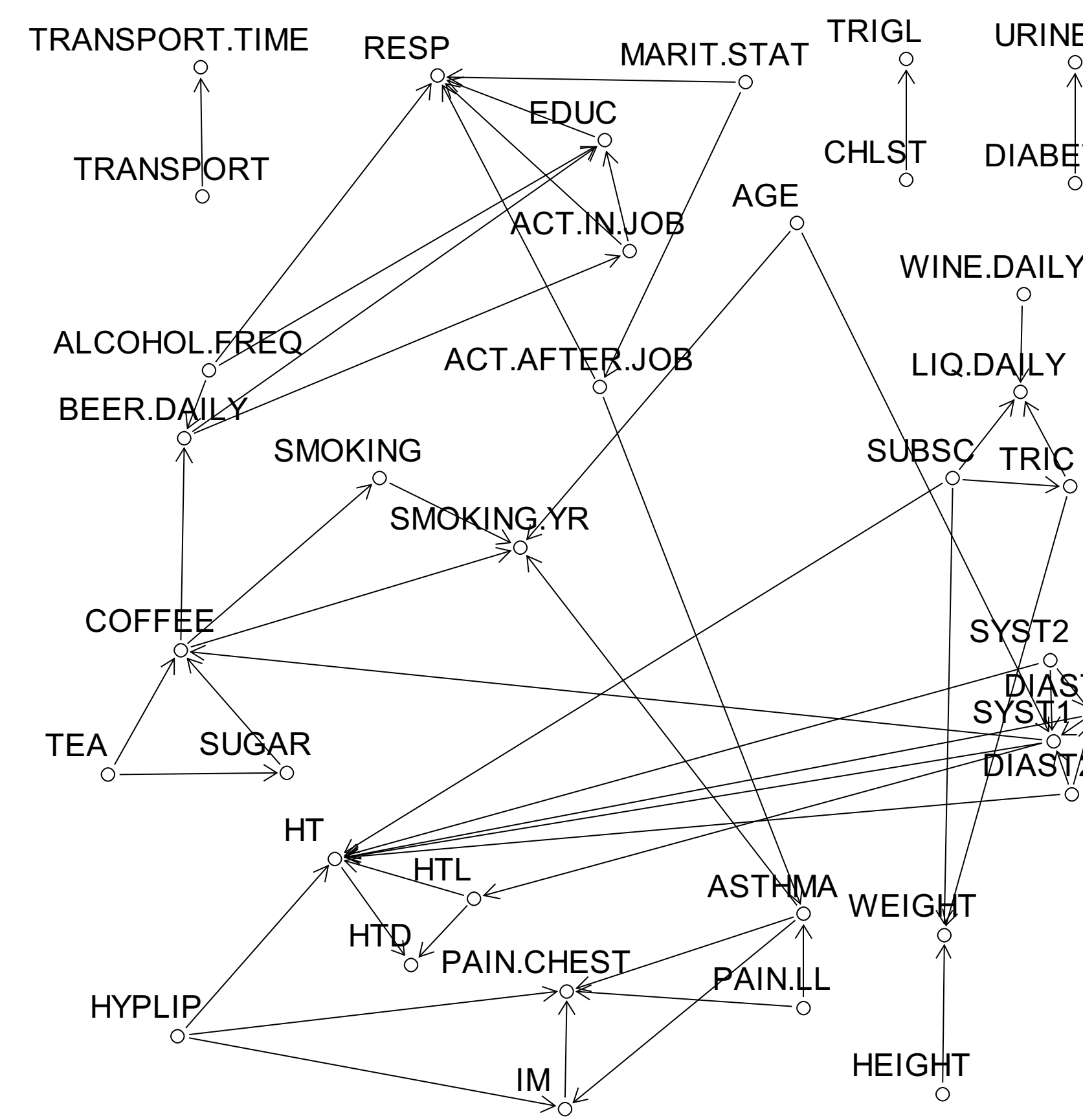


Figure 6: Logistic regression

```
>summary(glm(HT~HYPLIP+IM+AGE+SUBSC,data=C,family="binomial"))

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.322730   1.274252  -3.392 0.000693 ***
              IM         1.246937   0.513342   2.429 0.015138 **
              HYPLIP     1.126383   0.333971   3.373 0.000744 ***
              SUBSC      0.009521   0.003978   2.393 0.016699 *
              AGE        0.245182   0.136678   1.794 0.072835 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
                0.1 ' ' 1
```

## STULONG Dataset

In the early seventies of the twentieth century, a project of extensive epidemiological study of atherosclerosis primary prevention was developed in the Czechoslovakia. It was entitled National Preventive Multifactor Study of Hard Attacks and Strokes. Institute of Clinical and Experimental Medicine (IKEM) in Prague was the co-ordinating center. The study was planned over a long period of time including six sites in the Czech and Slovak Republics.

Most of the data was transferred into electronic form by EuroMISE (European Center for Medical Informatics, Statistics and Epidemiology) with the support of European project (see [2]).

Thirty-seven attributes of entry examination has been selected to this study describing age, height, weight, blood pressure, marital status, education, occupation, physical activity, way of transport to job, smoking, drinking tea, coffee, beer, wine and liquor, cholesterol, triglycerides, myocardial infarction, hypertension, ictus, hyperlipidemia, diabetes, chest and lower limbs pain, asthma and others.

These factors were observed for 1417 Czech middle-aged men.

## Results

Pairwise testing of statistical connection returns a large number of significantly connected pairs even with using Bonferroni correction as you can see on Figure 3 and Figure 4. But which of them are direct causality connections?

The heuristic implemented in MUDIM identifies 59 connections as direct causality connections. They are visualized on Figure 5 with using Bayesian Network notation.

Finally, standard R functions can be used for measuring the influence of direct causes. For example, on Figure 6 you can how occurrence of myocardial infarction depends on other factors.

## Acknowledgments:

This work was supported by the project LN00B107 of the Ministry of Education of the Czech Republic and by grants n. 201/04/0393 of the Grant Agency of the Czech Republic and n. IAA2075302 of the Grant Agency of the Academy of Sciences of the Czech Republic.

## References

- [1] Jiroušek R., *On approximating multidimensional probability distributions by compositional models*, ISIPTA (2003)
- [2] Stulong, <http://euromise.vse.cz/stulong-en/index.php>
- [3] MUDIM - web page, <http://www.euromise.cz/mudim>