INFERENCE ON/IN/OF NETWORKS

Research of the Statistics & Probability Unit, Groningen, NL

university of groningen
faculty of mathematics and natural sciences

P. Behrouzi, S. Mahmoudi, F. Richter, M. shafiee Kamalabad, M. Signorelli, M.Grzegorczyk, E. Wit

p.behrouzi@rug.nl, s.m.mahmoudi@rug.nl, f.richter@rug.nl, m.shafiee.kamalabad@rug.nl, m.signorelli@rug.nl,m.a.grzegorczyk@rug.nl, e.c.wit@rug.nl

Introduction

At the turn of the 21st century scientists have come to realise that a major ingredient in many modern economic, epidemiological, ecological and biological questions is to understand the **network structure** of the entities they study. Unfortunately, computational bottle-necks have meant that only the simplest analyses have been applied to these large datasets, whereas methodological bottle-necks prevented an integrative view of complex phenomena.

Rather than simplifying the methodology prior to seeing the data, modern techniques from **high-dimensional inference** allow the data to select the appropriate level of complexity. The aim of this project is to apply these techniques to the field of network analysis.

We approach networks from three different angles:

1. high-dimensional graphical models, including causal models,
2. ordinary and stochastic differential equations
3. random network models, such as stochastic blockmodels and ERGMs.

Our aim is to develop theoretically sound network inference techniques based on penalized inference. In each of these areas, the challenge is to define a sufficiently complex network models for large systems that have computationally tractable inference procedures.

A general species diversification model

Background: Sophisticated diversification models have been developed, but they perform on a case-by-case basis. We propose a general speciation in order to consider complex ecological interactions.

Challenges:

- Decay and fossilization degrade crucial evidence
- Diversification processes have many potential explanatory variables

Methods:

- We performs an EM algorithm including a monte carlo simulation in the E-step for the reconstruction of trees. (see figure)
- We embed a differential geometric path finding method (DgLars) inside the M-step of the EM algorithm. This will produce a sparse, computationally feasible and consistent model selection procedure.

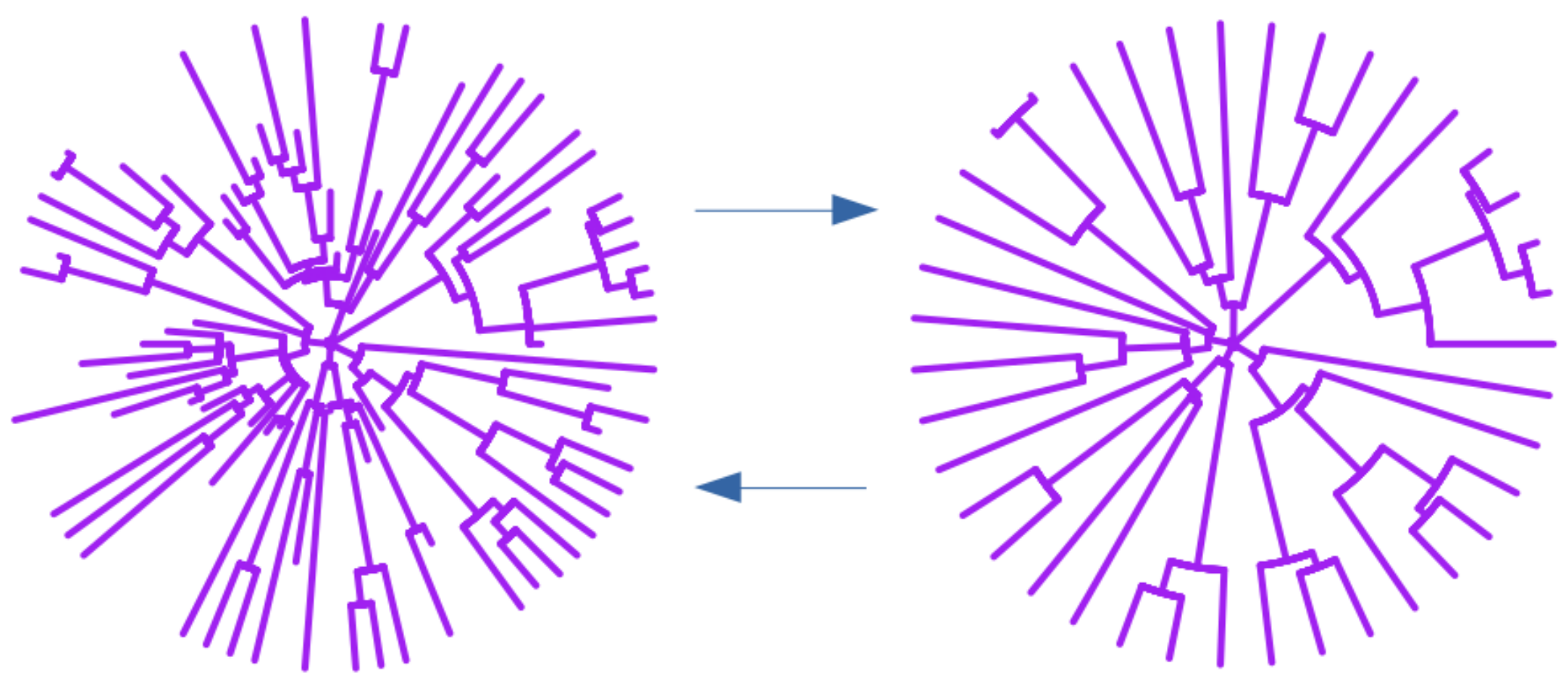


Figure 2: Loss of information on phylogenetic trees. An example of the same tree with and without extinct species.

Conclusions: With this general approach we are able to analyze and solve a wide variety of open problems in evolutionary biology including ecological interactions, geographical/spatial components, protracted and multiple speciation, migrations among many others.

Contact: Francisco Richter, f.richter@rug.nl

Causal effect in network

Motivation: Can we learn causal effects from observational data in high-dimensional systems?

Nonparanormal distribution

- $f = F^{-1} \circ \Phi$ monotone univariate function
- $f(Y) = (f_1(Y_1), ..., f_p(Y_p))^T \sim N(0, \Sigma)$ and $Y = (Y_1, ..., Y_p)^T$ has a nonparanormal distribution

Causal Effect for Nonparanormal Graphical Models

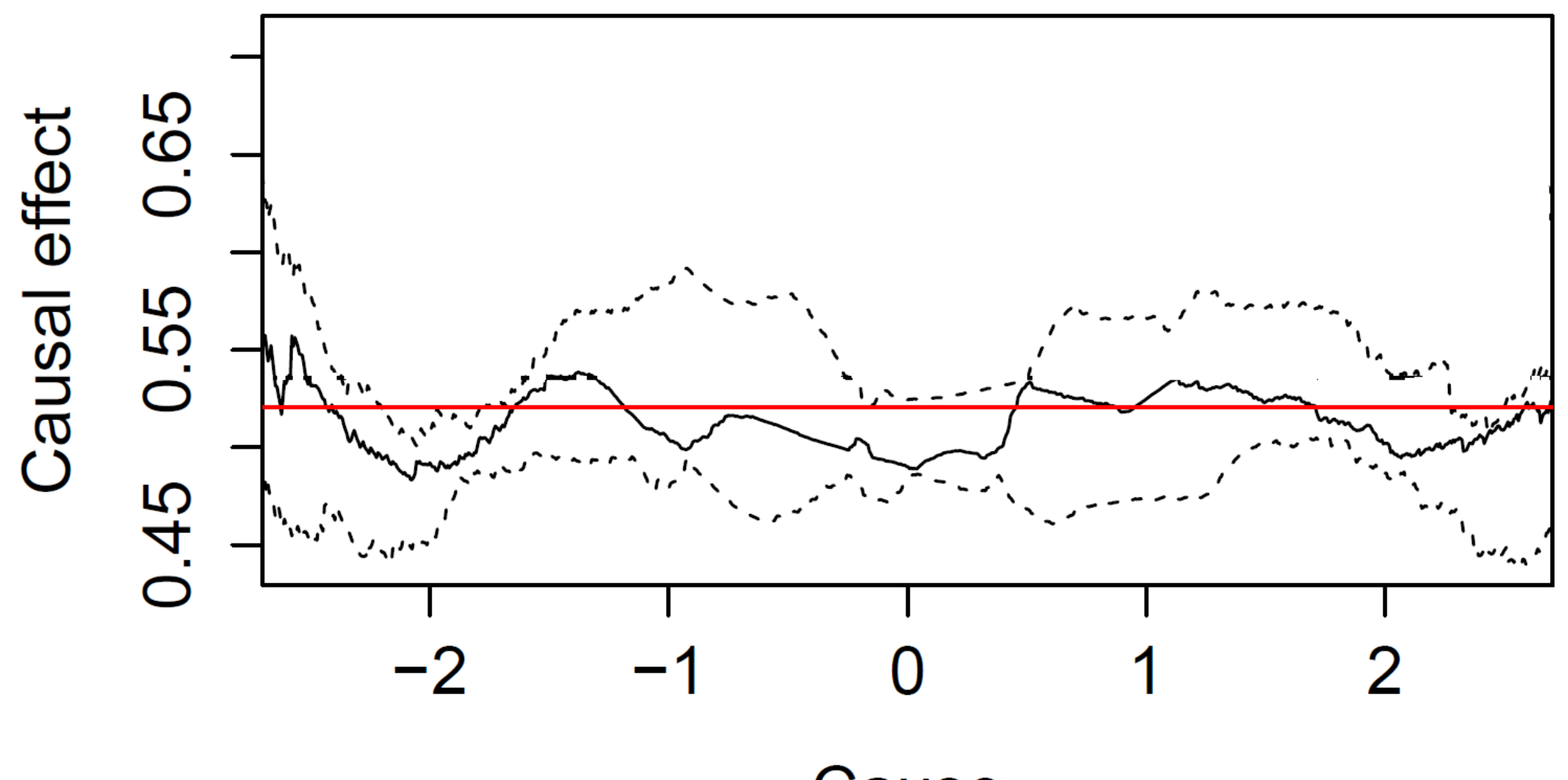
We are interested in the causal effect of Y_i on Y_p for $i \in (1, ..., p-1)$

- Gaussian distribution: $E(Y_p|Y_i = y_i, pa_i) = \beta_0 + \beta_i y_i + \beta_{pa_i}^T pa_i$

$$\frac{\partial}{\partial y_i} E[Y_p|do(Y_i = y_i)] \equiv \beta_i$$

- Non-Gaussian distribution

$$\begin{aligned}\frac{\partial}{\partial y_i} E[Y_p|do(Y_i = y_i)] &\cong f_p^{'}(z_{0j})\delta_i(f_i^{-1})^{'}(y_i) \\ &= f_p^{'}(z_{0i})\beta_i(f_j^{-1})^{'}(y_i)\end{aligned}$$



Simplification of complex networks

RESEARCH QUESTION: large networks challenge our capacities to visualize and interpret them. Often, one can exploit information on communities to derive a reduced graph summarizing relations between them.

EXAMPLES:

1. use ontologies/pathways to summarize gene regulatory networks;
2. groups of individuals in social networks (e.g., parties in a Parliament).

METHODS:

1. significance test based on number of links between communities: implemented in `R` package `neat` (→ *arXiv:1604.01210*);
2. penalized stochastic blockmodels (→ *arXiv:1607.08743*).

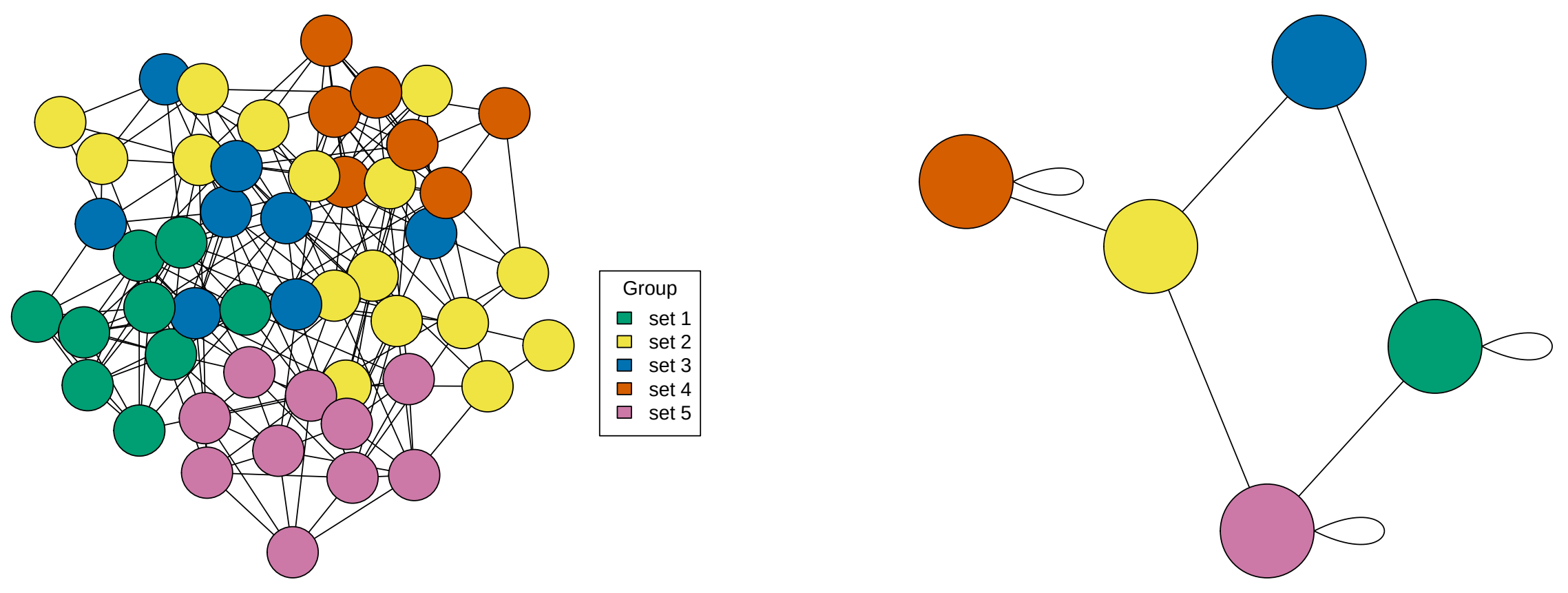


Figure 1: We develop statistical methods that allow to summarize relations between communities in large graphs (left) with a reduced graph (right).

CONCLUSION: reduced graphs displaying relations between communities can provide a synthetic and meaningful insight on complex networks.

CONTACT: Mirko Signorelli, m.signorelli@rug.nl.

Inference of complex networks

Background: In crossing two parents in plants/animals, certain combinations of genomic regions of parents do not function well when brought together in the genome of the offspring, thus resulting in sterility, low fertility, or even complete lethality.

Aim: Detecting genomic regions that result disorders in offspring.

Challenges:

- Dealing with high-dimensional setting ($p >> n$);
- Reconstructing conditional independence graph for ordinal data.
- Traditional approach, multiple testing, is under-powered.

Method: The network estimation relies on penalized Gaussian copula graphical models which is implemented in `R` package `EpiMultiLoc`.

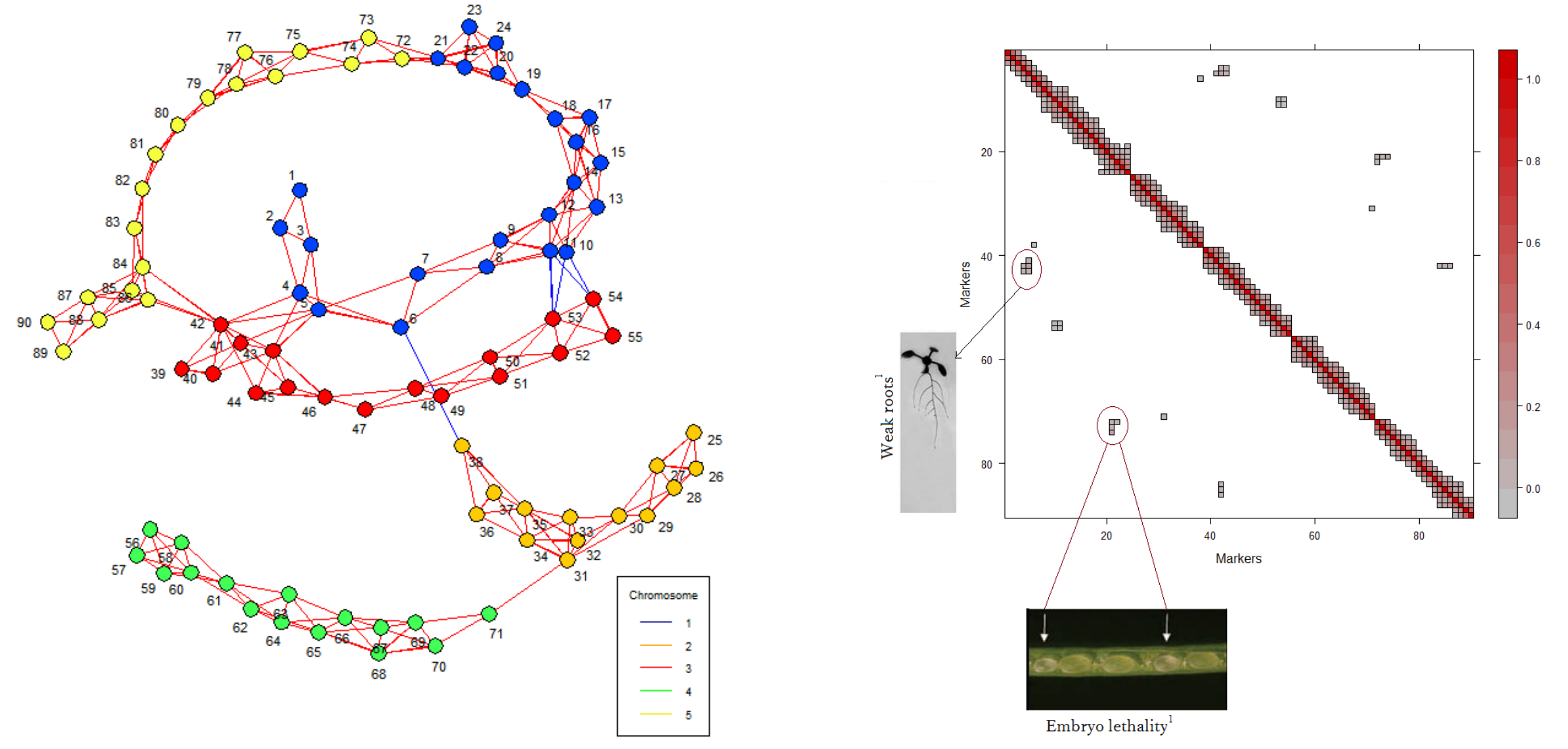


Figure 4: We discovered two regions in Arabidopsis that interact epistatically, which had prior been shown to cause arrested embryo development and root growth impairments. We discovered a few other trans-chromosomal interactions.

Conclusion: We developed a novel statistical method to explore aberrant marker-marker interactions in any diploid and polyploid genome.

Contact: Pariya Behrouzi, p.behrouzi@rug.nl.