Sensitivity of Inference in Bayesian Networks to Assumptions about Founders

Julia Mortera

Università Roma Tre

mortera@uniroma3.it

Peter Green

University of Bristol

P.J.Green@bristol.ac.uk

Durham – July 2008

Bayesian networks, with inferences computed by probability propagation methods ("junction tree algorithms"), offer an appealing practical modelling framework for structured systems involving discrete variables in numerous domains, including forensic genetics.

However, when allowing for uncertainty in some of the probability distributions specifying the model, exact calculation of conditional probabilities by propagation methods is not so straightforward.

In forensic genetics there is uncertainty about the gene frequency distribution.

The algorithms cannot be applied in systems where the discrete variables have continuous parents. This rules out having continuously distributed unknown parameters in the distributions of the discrete variables.

Overview

Forensic Identification

Example 1: Criminal Identification

Object-Oriented Bayesian Networks (OOBN)

Variations in Standard Assumptions Uncertain Gene Frequency UGF; Identity by Descent IBD; Subpopulations

Example 2: DNA Mixtures

Results

Forensic Identification

The following hypotheses (queries) are typical of forensic identification:

Criminal case Did individual A leave the DNA trace found at the scene of the crime?

Criminal case- mixed trace: Did A and B both contribute to a stain found at the scene of the crime? Who contributed to the stain?

Disputed paternity: Is individual A the father of individual B?

Immigration: Is A the mother of B? How is A related to B?

Computation of LR

The weight of the evidence is reported as a likelihood ratio

$$LR = \frac{P(E|H = \text{true})}{P(E|H = \text{false})}.$$

This can be computed in a Bayesian network using uniform prior probabilities $\Pr(H = \text{false})/\Pr(H = \text{true})$ from:

$$LR = \frac{\Pr(E \,|\, H = \text{true})}{\Pr(E \,|\, H = \text{false})} = \frac{\Pr(H = \text{true} \,|\, E)}{\Pr(H = \text{false} \,|\, E)} \frac{\Pr(H = \text{false})}{\Pr(H = \text{true})}.$$

Forensic Genetics: Criminal Identification

A simple case of criminal identification we have a DNA profile found at the scene of the crime and the DNA profile of a suspect which matches the crime profile. We denote this evidence by E.

The query or hypothesis H to be investigated: Did the suspect leave the trace at the crime scene? (suspect is guilty?)

Genetic Background

An identified area (locus)on a chromosome is a *gene* and the DNA composition on that area is an *allele*.

A gene thus corresponds to a (random) variable and an allele to its realised state.

A DNA *marker* is a known locus where the allele can be identified in the laboratory.

Short Tandem Repeats (STR) are markers with alleles given by integers. If an STR allele is 5, a certain word (e.g. CAGGTG) is repeated exactly 5 times at that locus:

. . . CAGGTGCAGGTGCAGGTGCAGGTGCAGGTG. . .

Standard Assumptions

A genotype of an individual at a locus is an unordered pair of genes.

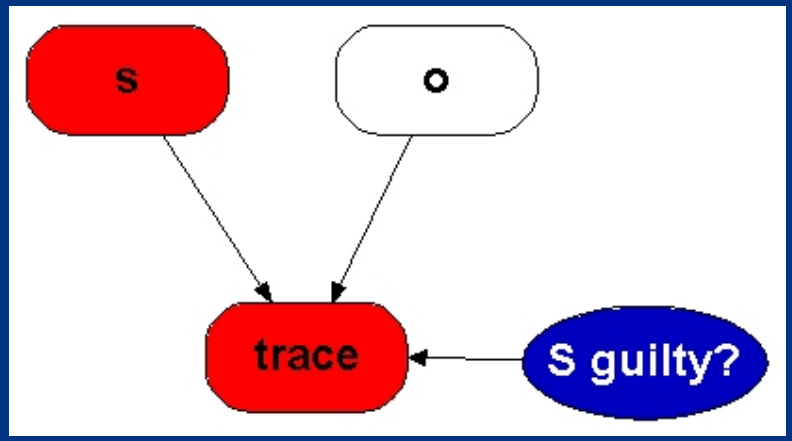
Marker	Genotype	Frequency f_0
D13	$\{9, 14\}$	{0.08,0.05}
FGA	$\{21, 22\}$	{0.19, 0.22}

It's customary to assume that all individuals are drawn from a *homogeneous population* in *Hardy-Weinberg equilibrium*, with *known* gene frequencies f_0.

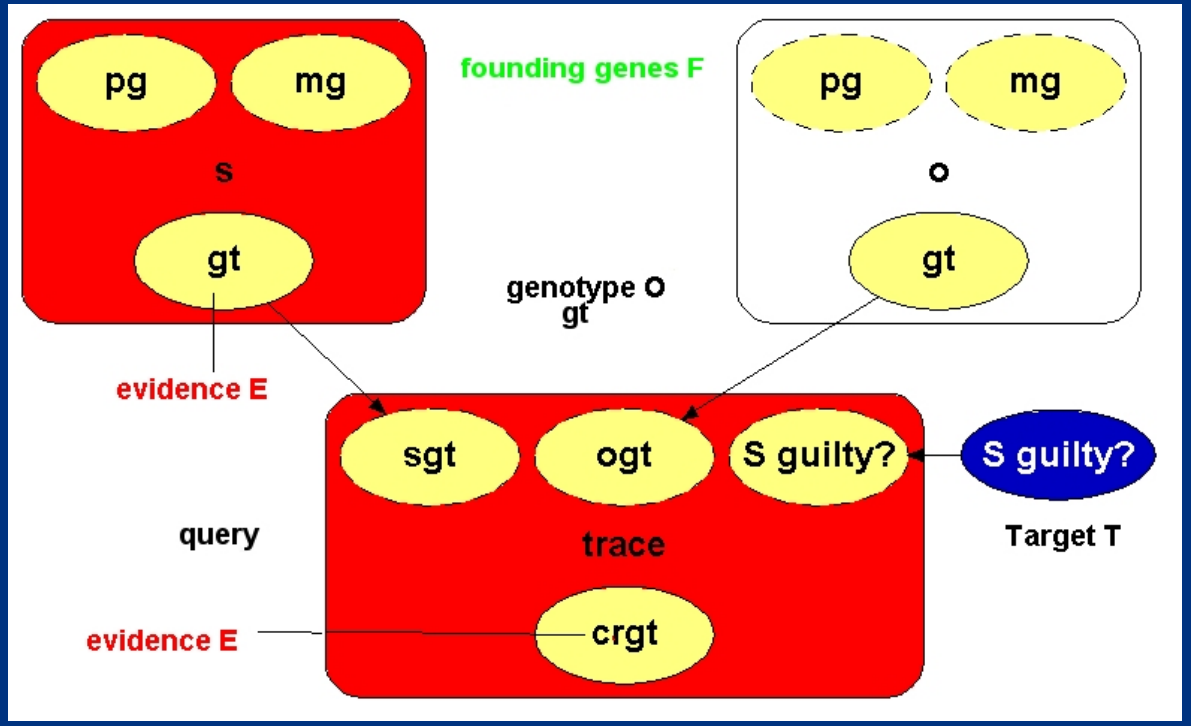
Forensic Genetics: Criminal Identification

Table 1: **Crime and suspect's DNA profile (excerpt)**

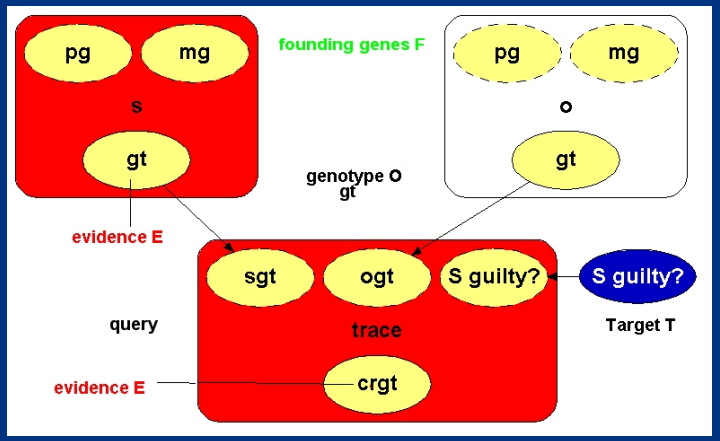
Marker	D13	D3	D5	D7	FGA
Evidence E_m	9 14	11 17	9 11	10	21 22
Frequency f_0	.08 .05	.002 .125	.05 .38	.24	.19 .22



OOBN for Criminal Identification



Joint distribution of all Variables



$$p(\texttt{S guilty?}) \prod_m [p(\texttt{spg}_m)p(\texttt{smg}_m)p(\texttt{opg}_m)p(\texttt{omg}_m)]$$

$$\times \prod_m [p(\texttt{sgt}_m|\texttt{spg}_m, \texttt{smg}_m)p(\texttt{ogt}_m|\texttt{opg}_m, \texttt{omg}_m)$$

$$\times\, p(\texttt{trace}_m|\texttt{sgt}_m, \texttt{ogt}_m, \texttt{S guilty?})]$$

Marginal posteriors in a Bayesian network

The set of nodes in a BN for forensic genetics can be partitioned disjointly as

$$X = F \cup T \cup O \cup E,$$

F Founding genes, T Targets ($T = 0, 1$ corresponding to the hypotheses $H =$ true and $H =$ false), O Others and E Evidence. Interest is in

$$h(f) = \log LR = \log \frac{P\{T = 1|E\}}{P\{T = 0|E\}} = \log \frac{p_1^t f}{p_0^t f},$$

as a function of the distribution f of F with $P\{F = \boldsymbol{i}\} = f_{\boldsymbol{i}}$. We wish to evaluate variations in $h(f)$ as f varies from the baseline f_0.

Bayesian Network: BN

We wish to assess sensitivity by devising a BN whose structure implies a variety of alternative settings for f:

- **unknown** allele frequencies (UGF)
- **identity by descent** (IBD) among founders
- **heterogeneity** (HET), i.e. the existence of *subpopulations*

These variations in standard assumptions generate dependence between founding genes. This can be studied by considering the effect of perturbing the joint distribution of the founding genes on the posterior inferences of interest.

Marker data may not be CI

Usually, the likelihood ratio LR for $E = \{E_m\}$ on $m = 1, 2, \ldots, M$ markers is given by the *product rule*:

$$LR = \frac{P\{E|T=1\}}{P\{E|T=0\}} = \prod_{m=1}^{M} \left\{ \frac{P\{E_m|T=1\}}{P\{E_m|T=0\}} \right\}.$$

For IBD and HET the product rule (PR) fails to apply (they have latent variables common to all markers).

Uncertain Allele Frequencies

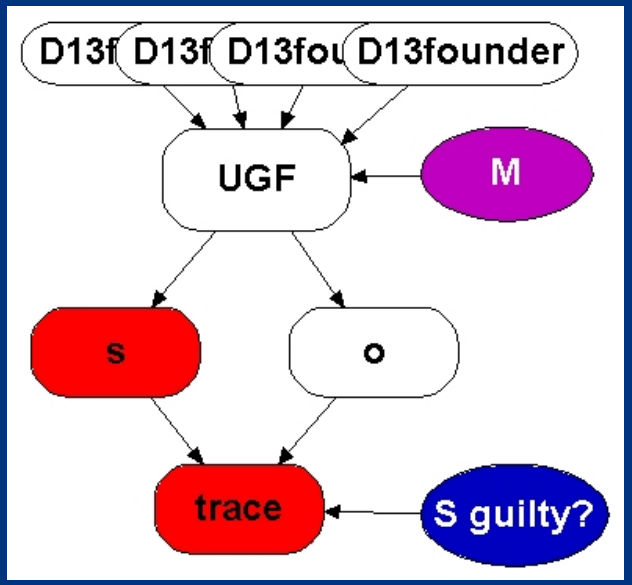
Allele frequencies are *not* fixed probabilities, but empirical frequencies in a database.

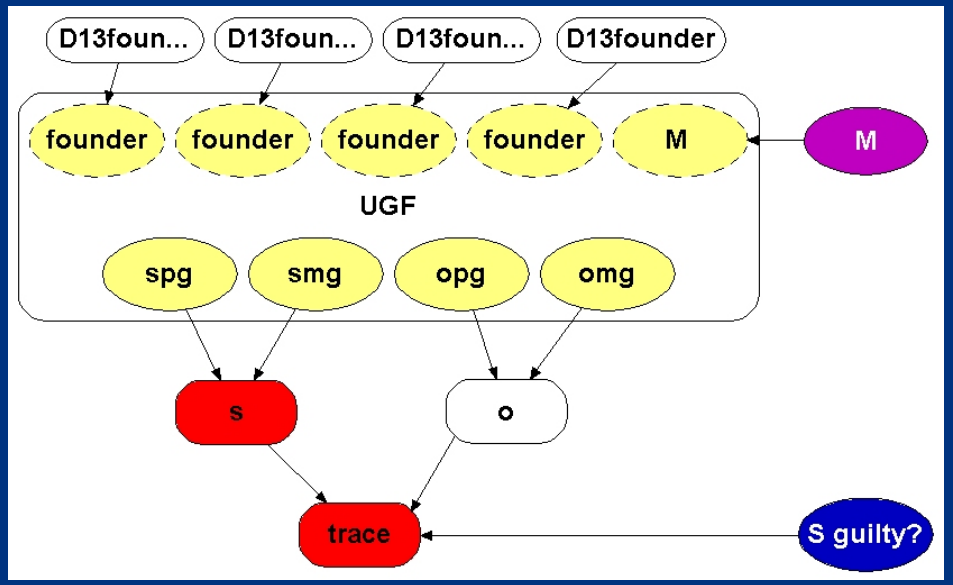
Assuming a Dirichlet prior and multinomial sampling the posterior distribution of a set of probabilities $\mathbf{r}$ is Dirichlet $(M\rho(1), M\rho(2), \ldots, M\rho(k))$.

The founding genes (spg, smg, opg, omg) are drawn i.i.d. from the distribution $\mathbf{r}$ across alleles, which has the above Dirichlet distribution where M is the sample size and ρ are the database allele frequencies.

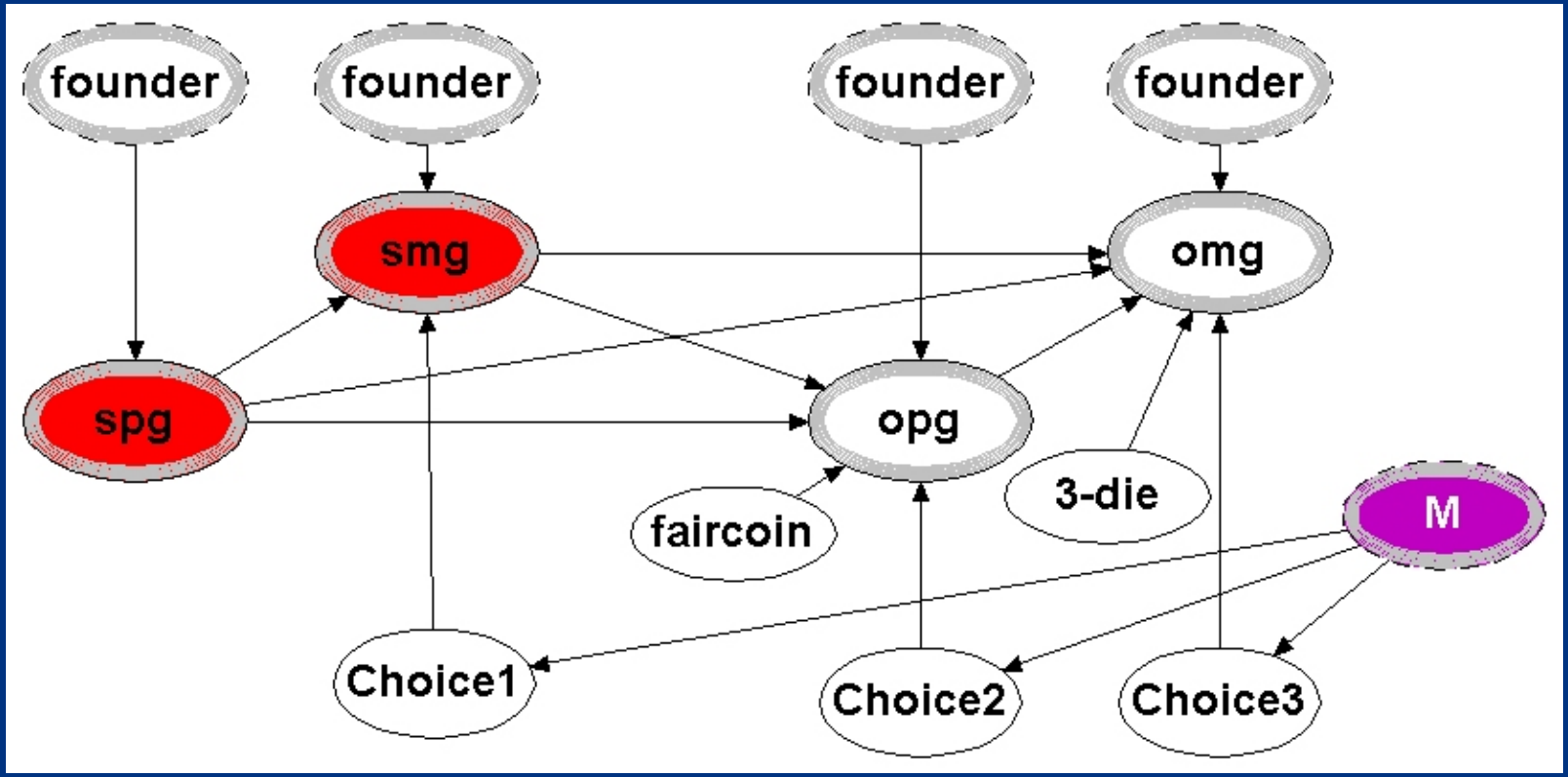
This corresponds to the standard set-up for a Dirichlet process model and *can be represented in a BN using the Pòlya urn scheme*

UGF


D13f
D13f
D13fou
D13founder
UGF
M
s
o
trace
S guilty?

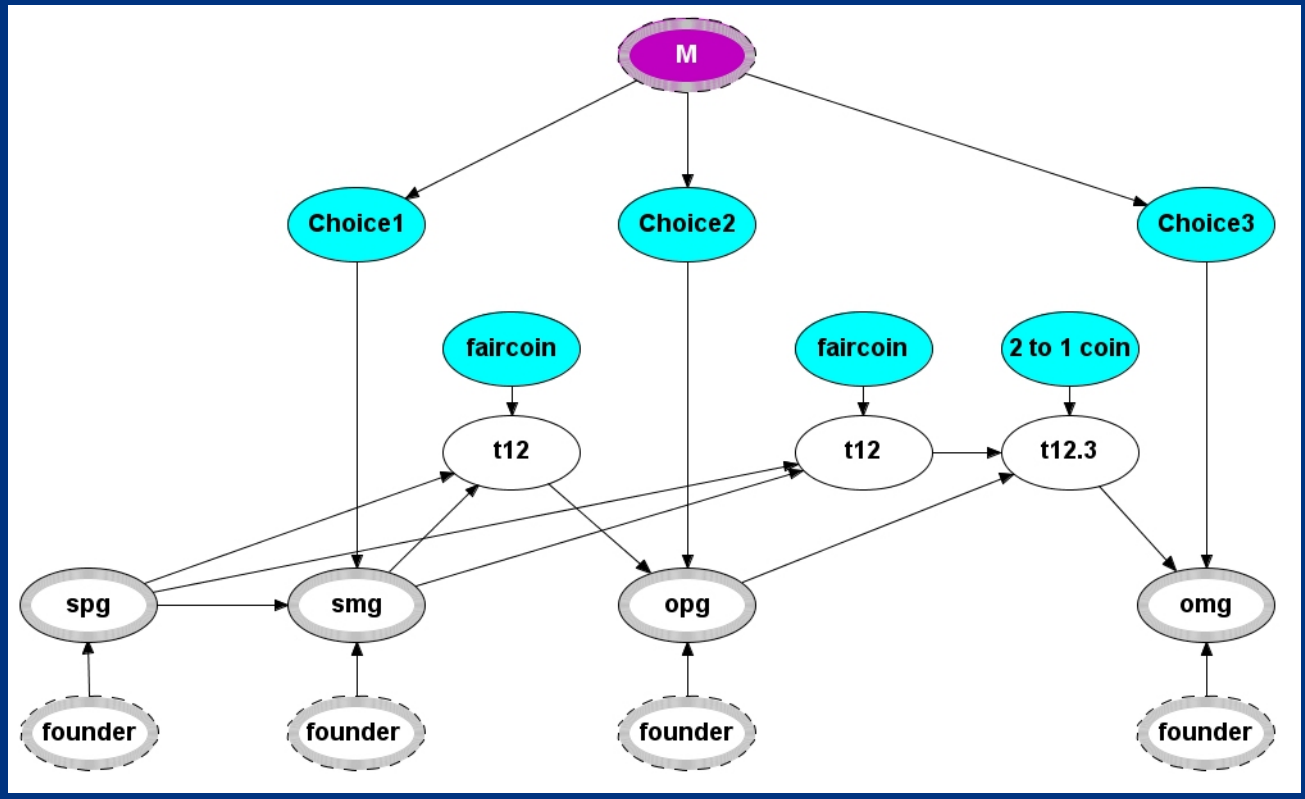

D13foun...
D13foun...
D13foun...
D13founder
founder
founder
founder
founder
M
M
UGF
spg
smg
opg
omg
s
o
trace
S guilty?

Node UGF: Pólya urn scheme



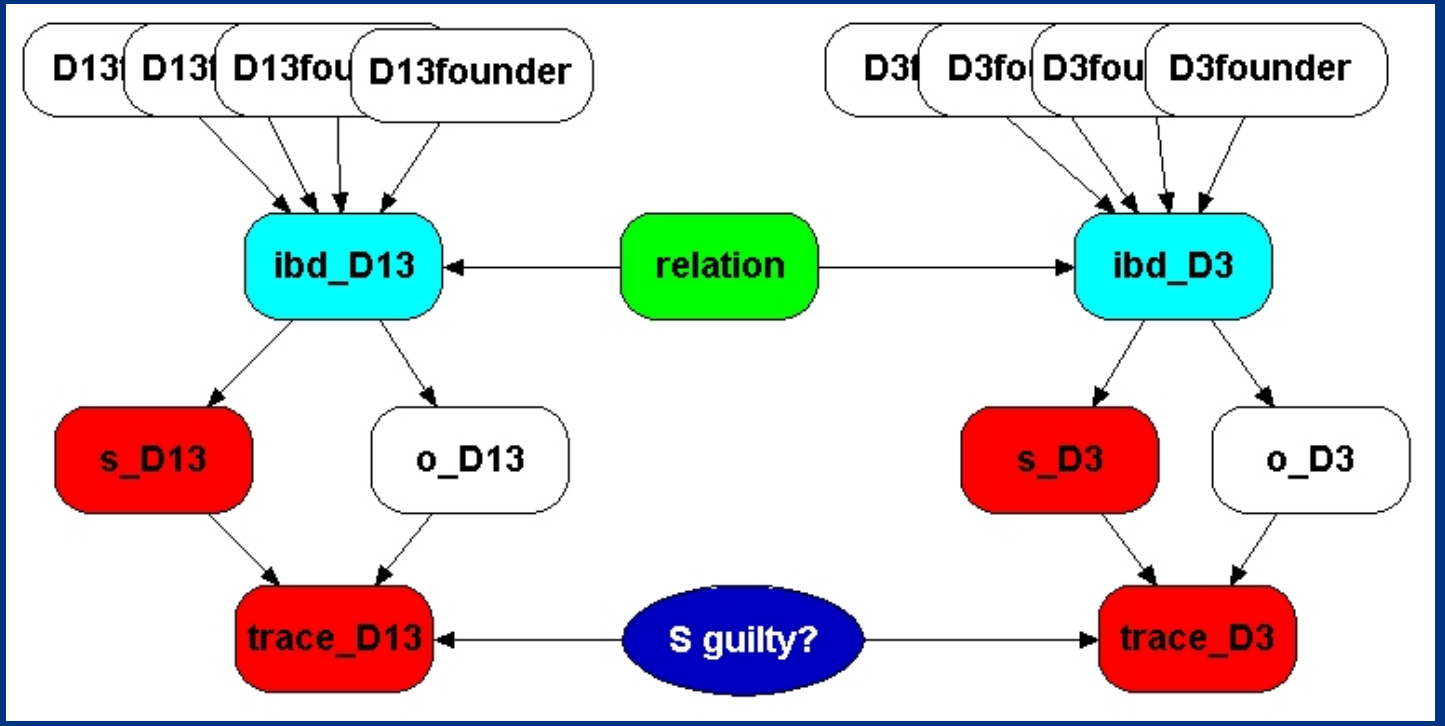
where $\text{Choice}_i \sim Bin(1, i/(M+i))$.

Divorcing

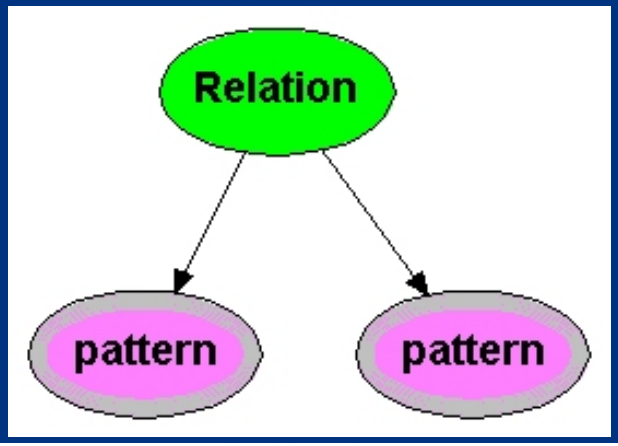


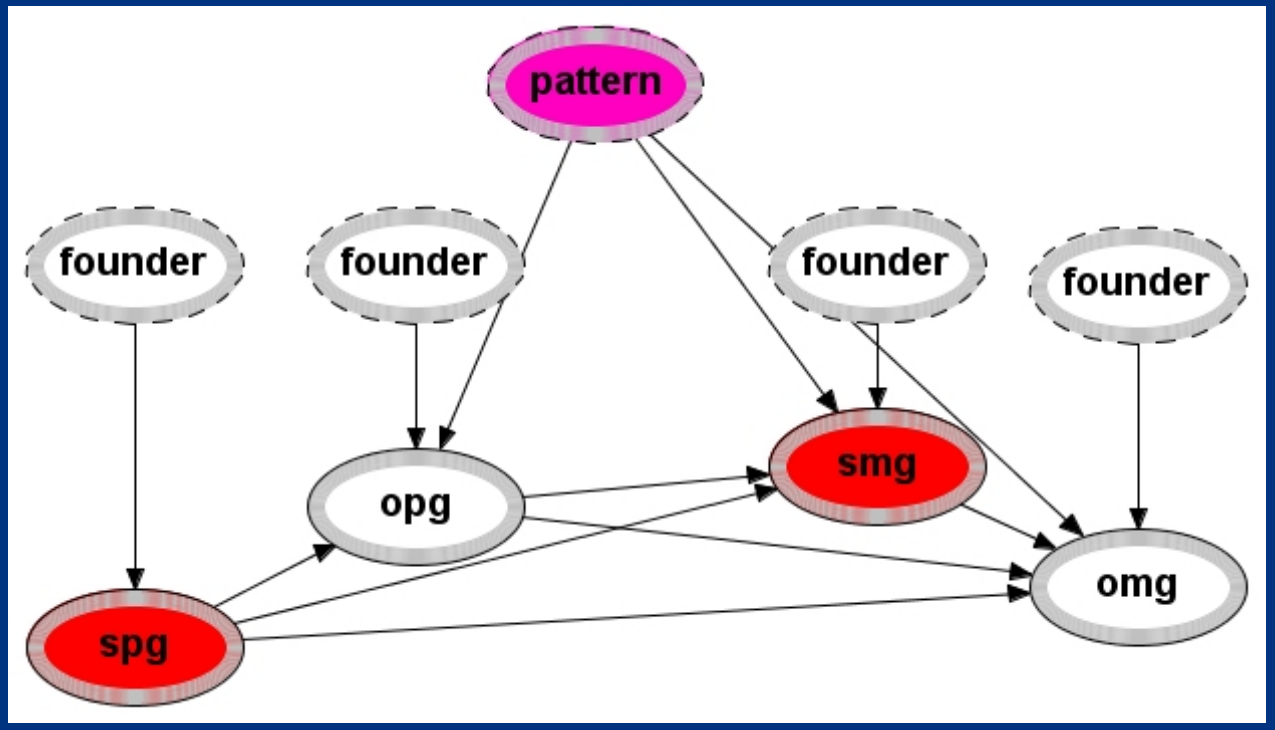
where all choices are now binary, thus reducing the clique table sizes.

OOBN network for criminal identification with IBD for 2 Markers

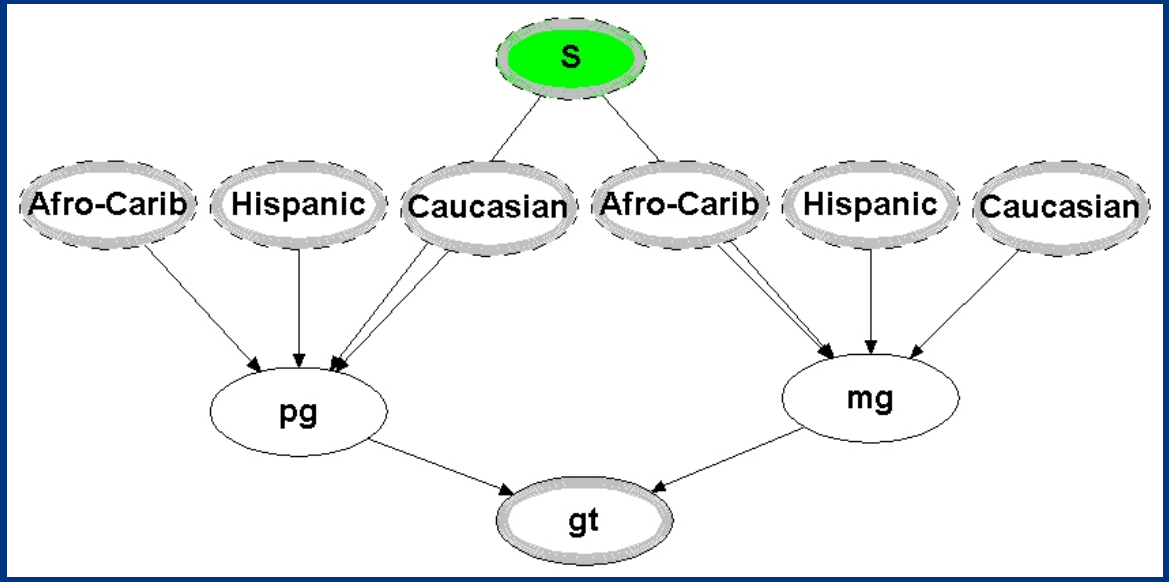


Networks representing relation R and IBD





Network for genotype when uncertainty in subpopulation



This induces dependence between markers, m. S is same for all m so mixing across supopulations is not the same as using mixture of allele frequencies.

Computing across-marker inferences using within-marker BNs

Let R be a latent variable (codes for relationship among individuals), then since $T \perp\!\!\!\perp R$ *a priori*:

$$p(E|T) = p(T)^{-\#(M)} \sum_R p(R) \prod_m p(E_m, T|R)$$

Now $p(E_m, T|R)$=$p(E_m|R)p(T|E_m, R)$ can be obtained from a BN (directly in GRAPPA). The per-marker LRs

$$p(E_m|T) = p(T)^{-1} \sum_R p(R)p(E_m, T|R)$$

and the PR does not hold.

Within-marker latent variables

Let $\pi = \{\pi_m, m = 1, 2, \ldots, M\}$ be within-marker latent variables (for IBD these code the pattern of identity among genes). Assume
$p(T, R, \pi, E) = p(T)p(R) \prod_{m=1}^{M} \{p(\pi_m|R)p(E_m|T, \pi_m)\}$
then

$$p(E|T) = \frac{1}{p(T)^{\#(M)}} \sum_R p(R) \prod_m \left\{ \sum_{\pi_m} p(\pi_m|R)p(E_m, T|\pi_m) \right\}$$

Can get the combined inference from within-marker BN (for each m and π_m). The BN is simpler, since R not needed. Computational cost of each depends on the numbers of values in R and $\{\pi_m\}$.

Likelihood ratios LRs

	Standard	UGF	IBD	Subpop
D13	138.9	106.6	88.7	126.7
D3	1162.8	194.6	111.9	3488.4
D5	27.7	23.6	20.5	35.6
D7	16.9	14.6	13.7	11.8
	Overall $Log_{10}LR$ for 8 markers			
exact	13.38	12.10	7.71	13.85
product rule	13.38	12.10	11.54	13.57

Overall LR for UGF is about 20 times smaller than baseline, whereas true IBD it is roughly 460×10^3 smaller than baseline and 7×10^3 smaller than product rule.

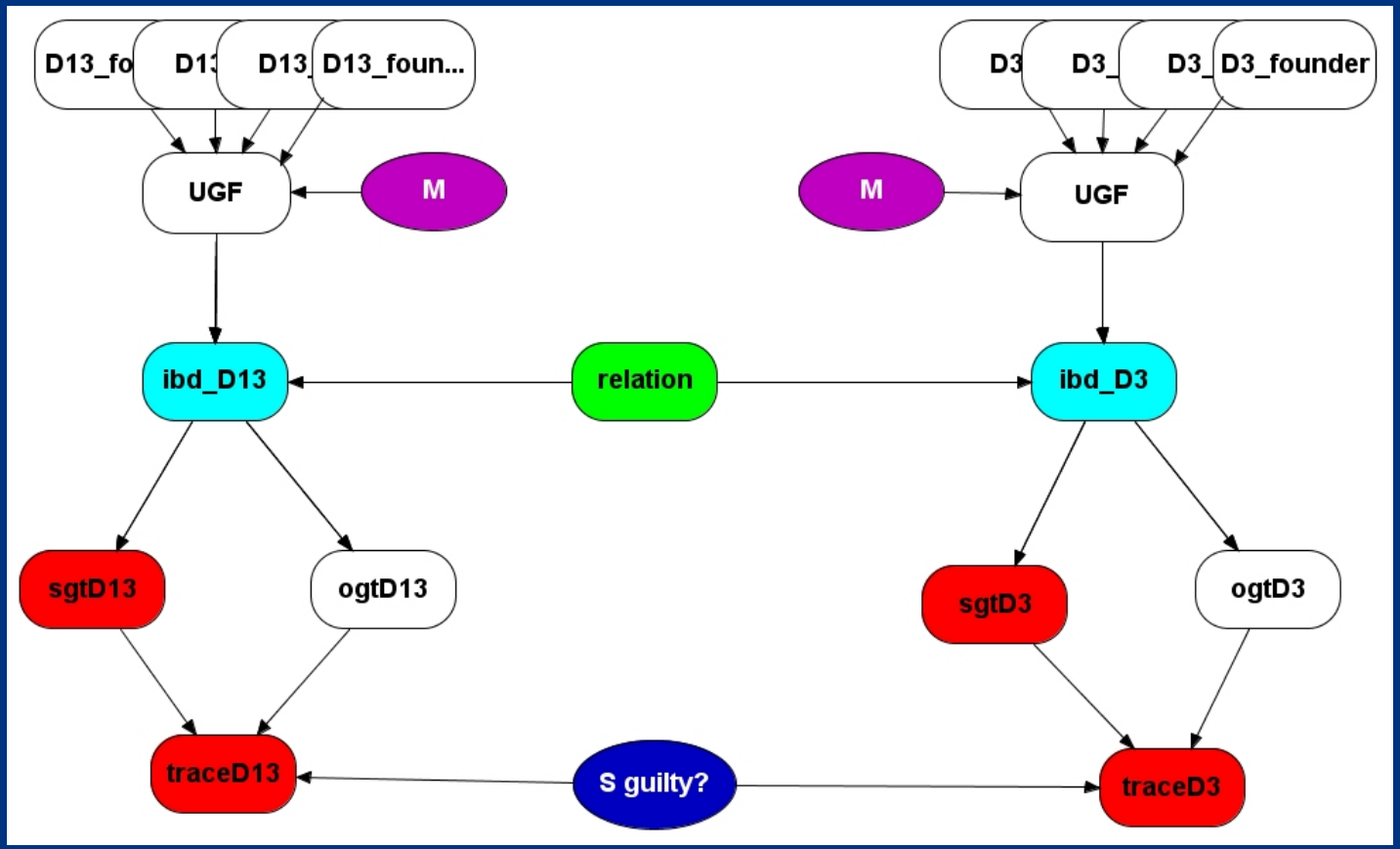
LRs for Subpopulation

suspect	mixed population			
other	mixed	Cauc	Afro-Car	Hisp
D13	126.70	138.89	432.90	70.58
D3	3488.37	1162.79	∞	∞
D5	35.56	27.70	55.02	33.22
	Overall $Log_{10}LR$ for 8 markers			
true	13.85	13.38	∞	∞
product rule	13.57	13.38	∞	∞

The LR when suspect and alternative are both from a heterogeneous mixed SUBPOP is twice as large than for product rule.

Combination of Scenarios

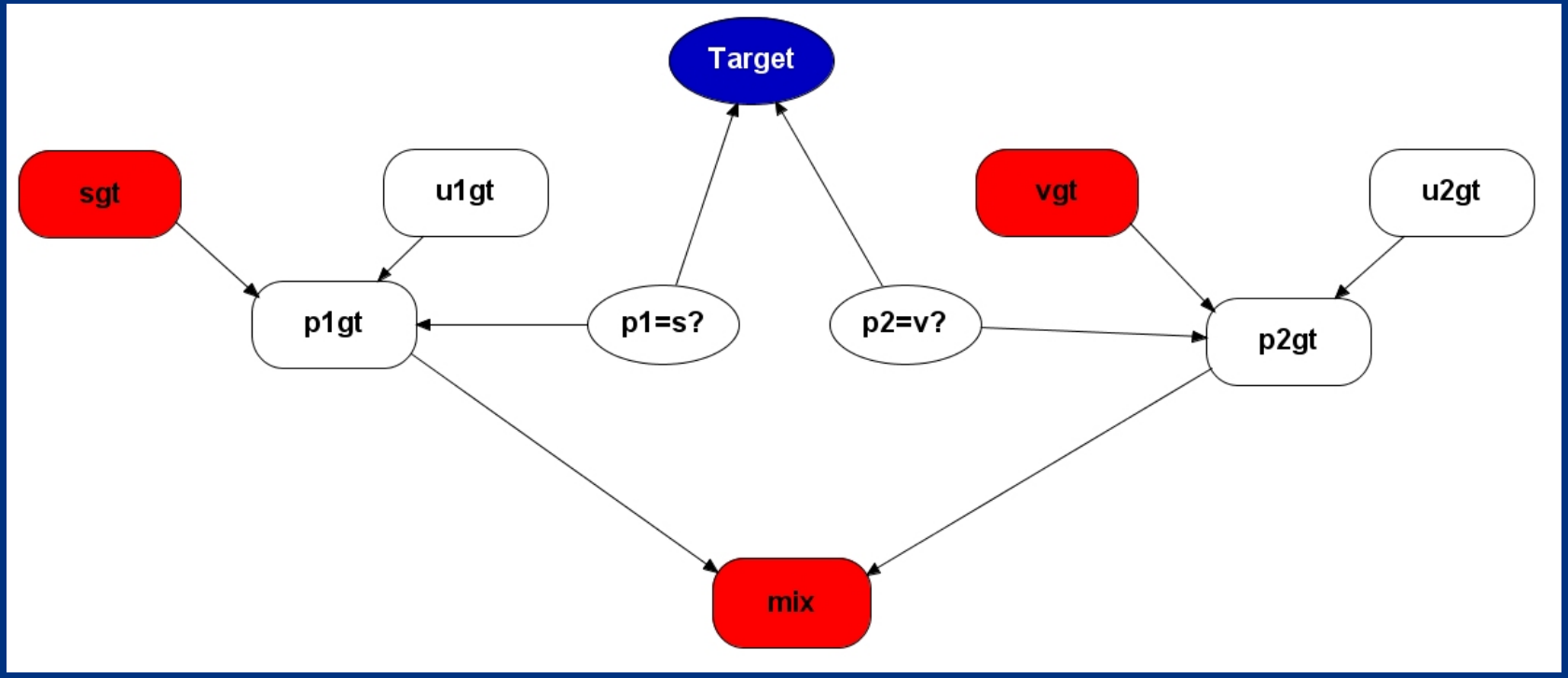
Thanks to the modularity of BN we can combine UGF+IBD and UGF+HET



Results: Overall $\log_{10}$LR

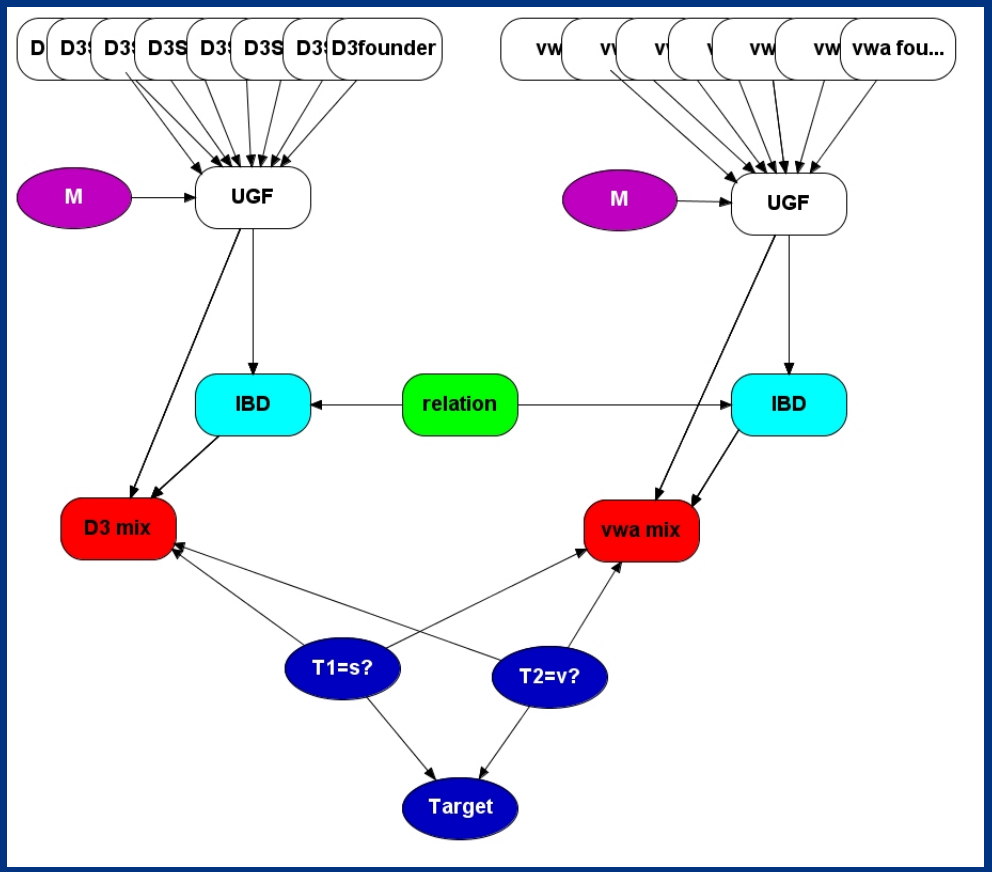
	Base	UGF	IBD	HET	UGF+ IBD	UGF+ HET
D13	138.9	106.6	88.7	126.7	71.7	113.9
D3	1162.8	194.6	111.9	3488.4	74.3	583.7
D5	27.7	23.6	20.5	35.6	18.2	33.4
Overall $Log_{10}LR$ for 8 markers						
exact	13.38	12.10	7.71	13.85	7.49	12.57
PR	13.38	12.10	11.54	13.57	10.95	12.96

OOBN for DNA Mixture



Note: $4 \times 2 = 8$ founding genes in this case.

UGF plus IBD for a DNA Mixture



LR for UGF plus IBD

Target: $H_0 : s\&v$ **vs.** $H_1 : v\&u$

UGF with $M = 99$ **(**$\theta = 0.01$ **Balding correction)**

	D3	VWA	FGA
unrelated	**50.90**	**11.52**	**14.61**
parent-child	**7.12**	**2.94**	**2.94**
half-sibs	**12.49**	**4.69**	**4.89**
mix over R	**34.84**	**9.45**	**11.25**

Suspect and U1 (alternative suspect) possibly related

Conclusions

- Freeware software GRAPPA in R by Peter Green (http://www.stats.bris.ac.uk/$\sim$peter/Grappa) for construction of and inference in discrete BNs.
- We have a range of different methods. Possibly some of these could be applicable to other areas. UGF $\rightarrow$ Pólya urn could be useful for other BN with uncertainty on founders?
- Other examples: simple and complex paternity testing have been analysed.
- Can infer the posterior probability of a specific relationship R among actors conditional on their DNA profiles. Useful in immigration cases.

- IBD and HET induce dependence among markers which can be handled it in one big net or using smaller nets and looping over latent variables.
- IBD more subtle than the standard θ (FST) approach.
- Results show that effects of IBD, UGF and HET can be quite dramatic.
- **Constrained Steepest descent: CSD**
  Aim: bound differences $|h(f) - h(f_0)|$ in terms of $||f - f_0||$ subject to constraints, e.g. $f_i \geq 0$, $\sum f_i = 1$ and for fixed marginals at each f.
- **Linear Fractional Programming: LFP**

Aim: Find min and max of $h(f)$, subject to linear constraints and linear bounds, e.g.
$\max_{\boldsymbol{i}} |(f - f_0)_{\boldsymbol{i}}| \leq \varepsilon$.