SIEMENS

Learning the Bayesian Network Structure: Dirichlet Prior versus Data

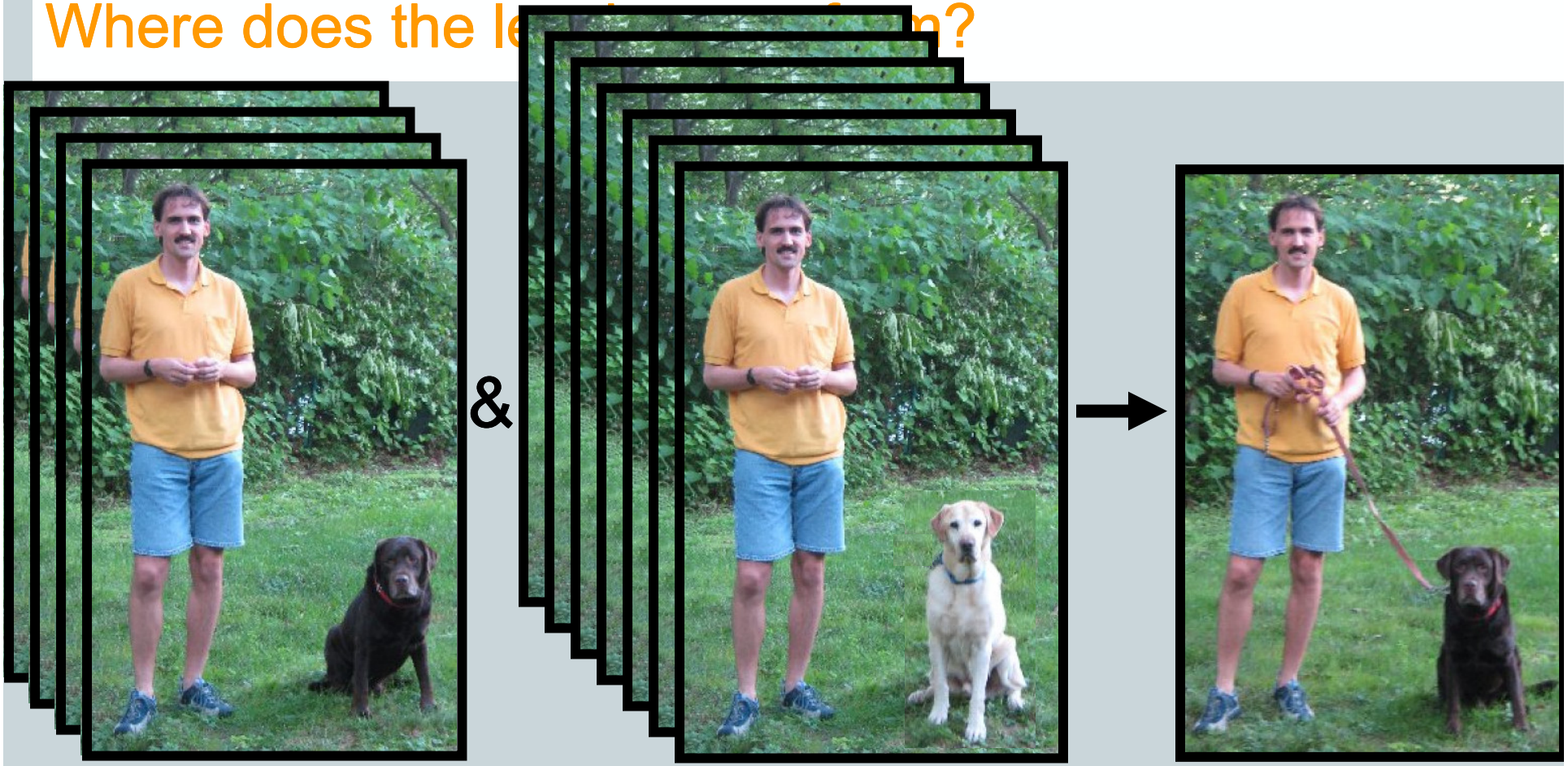
Harald Steck

IKM CKS Siemens Medical Solutions, USA

Harald.Steck@siemens.com

SIEMENS

Where does the le[illegible]m?



DATA: & DIRICHLET PRIOR: -> RESULT:

Independence & Independence -> Dependence

SIEMENS

Outline

- Formal Problem Statement and Notation
- Explanation, Part I
- Explanation, Part II

SIEMENS

Review: Dirchlet Prior and BDeu score

- random variables with multinomial distribution
- Bayesian network model:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} \theta_{X_i|\Pi_i}$$

- Dirichlet prior over model parameters:
  - … is conjugate prior
  - … ensures likelihood equivalence [Heckerman et al., '95]
  - $$p(\theta_{X_i|\pi_i}) \propto \prod_{x_i} \theta_{x_i|\pi_i}^{\alpha_{x_i,\pi_i}-1}$$

- where:

$\alpha_{x_i,\pi_i} = \alpha \cdot q_{x_i,\pi_i}$

$\alpha > 0$ … equivalent sample size (ESS)

q … prior distribution

SIEMENS

Review: Dirchlet Prior and BDeu score (cont'd)

Scoring function for graph G: marginal likelihood

$$p(D|\alpha, G) = \prod_{i=1}^{n} \prod_{\pi_i} \frac{\Gamma(\alpha_{\pi_i})}{\Gamma(N_{\pi_i} + \alpha_{\pi_i})} \prod_{x_i} \frac{\Gamma(N_{x_i,\pi_i} + \alpha_{x_i,\pi_i})}{\Gamma(\alpha_{x_i,\pi_i})}$$

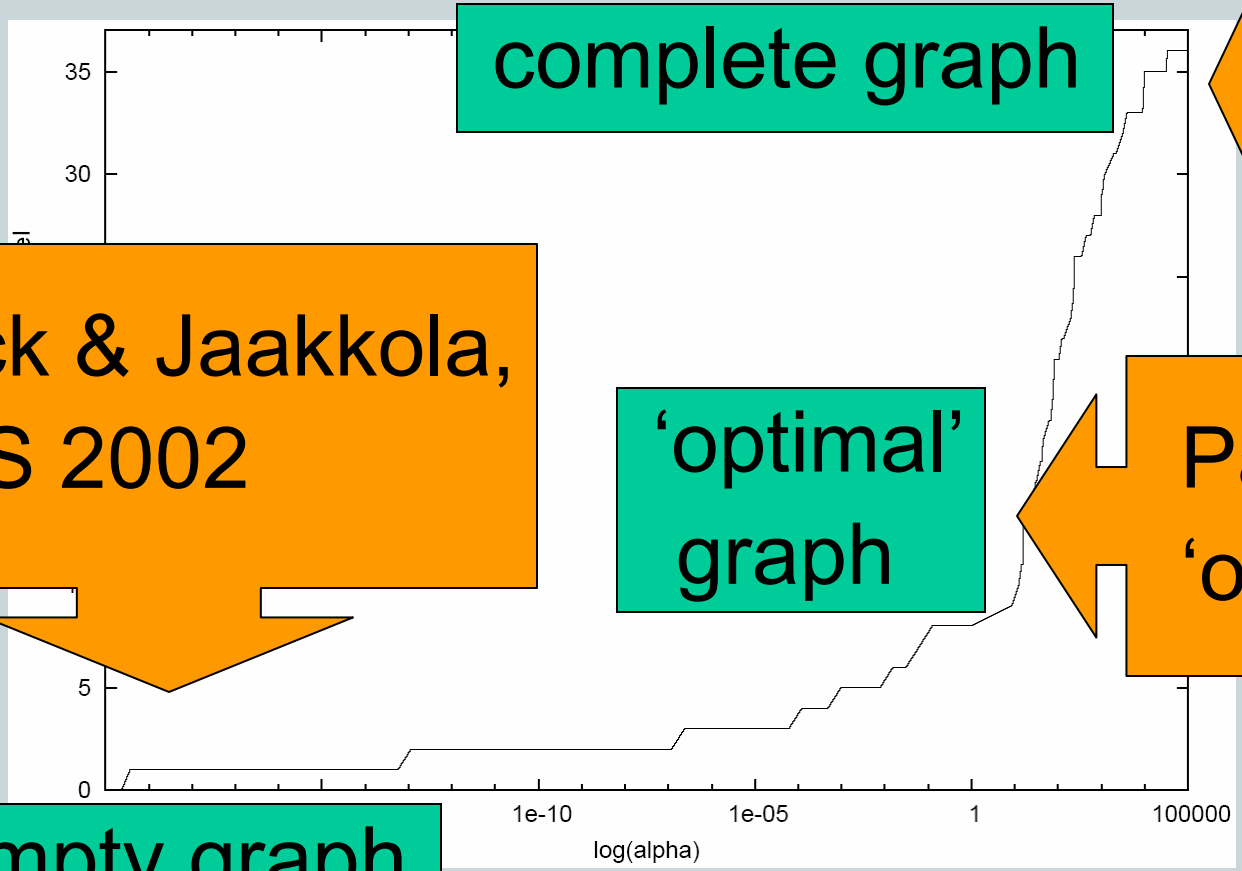
BDeu Score [Buntine, 1991]:

- choose q uniform
- ESS is the only free parameter in scoring function:

$$\alpha_{x_i,\pi_i} = \alpha \cdot q_{x_i,\pi_i} = \frac{\alpha}{|X_i| \cdot |\Pi_i|}$$

Result: How many Edges ?

Experiments [Silander et al., UAI 2007]:



complete graph

Part I: large ESS

Steck & Jaakkola, NIPS 2002

'optimal' graph

Part II: 'optimal' ESS

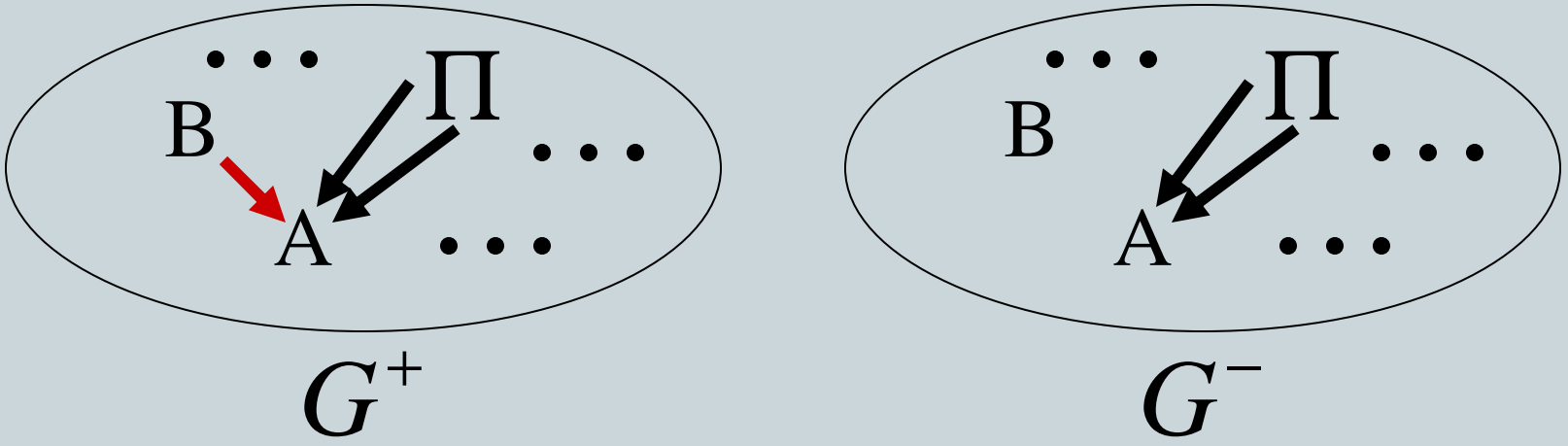
empty graph

SIEMENS

Outline

- Formal Problem Statement and Notation
- Part I: Large Equivalent Sample Size
- Part II: ‘Optimal’ Equivalent Sample Size

Absolute Score -> Relative Score

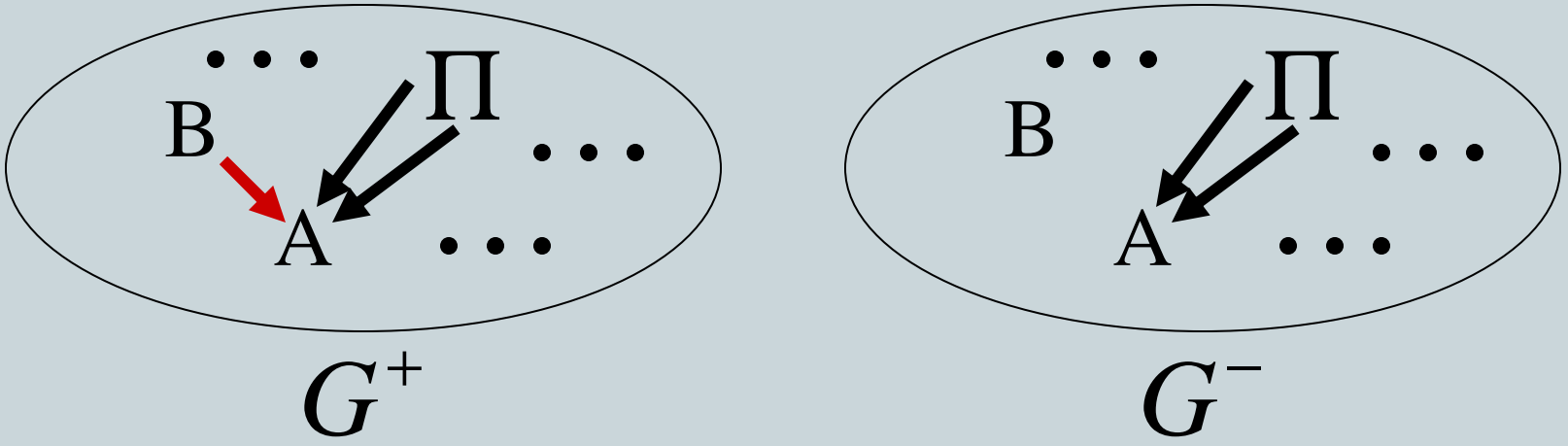


log Bayes Factor (log BF):

$$\log \frac{p(D|\alpha, G^+)}{p(D|\alpha, G^-)} = \sum_{a,b,\pi} \log \frac{\Gamma(N_{a,b,\pi} + \alpha_{a,b,\pi})}{\Gamma(\alpha_{a,b,\pi})} + \sum_{\pi} \log \frac{\Gamma(N_\pi + \alpha_\pi)}{\Gamma(\alpha_\pi)}$$
$$- \sum_{a,\pi} \log \frac{\Gamma(N_{a,\pi} + \alpha_{a,\pi})}{\Gamma(\alpha_{a,\pi})} - \sum_{b,\pi} \log \frac{\Gamma(N_{b,\pi} + \alpha_{b,\pi})}{\Gamma(\alpha_{b,\pi})}$$

SIEMENS

Absolute Score -> Relative Score



G^+ G^-

log Bayes Factor (log BF):

log BF> 0 $\Leftrightarrow$ edge present $\Leftrightarrow$ A,B dependent cond. on Π

log BF< 0 $\Leftrightarrow$ edge absent $\Leftrightarrow$ A,B independent cond. on Π

SIEMENS

Bayes Factor for large & finite ESS

Leading-order Taylor-expansion for $\alpha >> N$:

$$\log \frac{p(D|\alpha, G^+)}{p(D|\alpha, G^-)} = \frac{N}{2\alpha}\left\{N \cdot U(\hat{p}(A, B|\Pi)) - d_F\right\} + \mathcal{O}\left(\frac{N^2}{\alpha^2}\right)$$

where $\hat{p}(a, b|\pi) = N_{a,b,\pi}/N_\pi$

$$d_F = |\Pi|(|A| - 1)(|B| - 1)$$

- $N \cdot U(\hat{p}(A, B|\Pi)) - d_F > 0$
  $\Leftrightarrow$ edge $A \leftarrow B$ present for suff. large α

SIEMENS

new Uniformity Measure *U*

$$U(p(A,B|\Pi)) = \sum_{a,b,\pi} p(a,b,\pi)\Big(|A,B,\Pi|\cdot p(a,b,\pi) - |A,\Pi|\cdot p(a,\pi) - |B,\Pi|\cdot p(b,\pi) + |\Pi|\cdot p(\pi)\Big)$$

- **symmetry**: $U(p(A,B|\Pi)) = U(p(B,A|\Pi))$
- **non-negativity**: $U(p(A,B|\Pi)) \geq 0$
- **minimality**: $U(p(A,B|\Pi)) = 0$ if and only if
  - (conditional) independence:
    $p(A,B|\Pi) = p(A|\Pi)p(B|\Pi)$
  - and, for each state π with $p(\pi) > 0$, at least one of the marginal distributions is uniform:
    $p(A|\pi) = 1/|A|$ or $p(B|\pi) = 1/|B|$.

SIEMENS

Presence of Edge A <- B

$$N \cdot U(\hat{p}(A, B|\Pi)) - d_F > 0$$

$$\Leftrightarrow \text{ edge } A \leftarrow B \text{ present for suff. large } \alpha$$

- cf. AIC $= N \cdot I(\hat{p}) - d_F > 0$:

  $I(\hat{p})$ notably larger than *0*

- In words: For sufficiently large ESS,

  A <- B present if $\hat{p}$ implies

  - notable dependence between A and B given Π
  - OR notable skewness of both distributions

    $\hat{p}(A \mid \Pi)$ and $\hat{p}(B \mid \Pi)$

SIEMENS

Skewness --- An Intuitive Explanation

Average as regularized parameter estimate:

$$\bar{\theta}_{x_i|\pi_i} \equiv E_{p(\theta_{x_i|\pi_i}|D,G)}[\theta_{x_i|\pi_i}] = \frac{N_{x_i,\pi_i} + \alpha_{x_i,\pi_i}}{N_{\pi_i} + \alpha_{\pi_i}}$$

… weighted sum of uniform prior *q* and (skewed) empirical distribution

$$N \cdot \hat{p}(A)\hat{p}(B) + \alpha \cdot q(A)q(B) \neq c \cdot p(A)p(B)$$

=> In general: dependence

Outline

- Formal Problem Statement and Notation
- Part I: Large Equivalent Sample Size
- Part II: ‘Optimal’ Equivalent Sample Size:

  What properties of the data determine the value of the optimal ESS?

SIEMENS

‘Optimal’ ESS

- ESS treated as an additional parameter to be learned
- Objective [Silander et al., UAI 2007]:

$$(\alpha^*, G^*) = \arg\max_{(\alpha,G)} p(D|\alpha, G)$$

- Coordinate-wise ascent:

1. optimize the graph for a fixed ESS-value:
   $G_k^* = \arg\max_G p(D|\alpha_{k-1}^*, G)$,
2. optimize the ESS-value for a fixed graph:
   $\alpha_k^* = \arg\max_\alpha p(D|\alpha, G_k^*)$.

SIEMENS

Approximation of optimal ESS

Marginal likelihood : predictive accuracy in prequential sense [Dawid,'84] :

$$p(D|\alpha, G) = \prod_{j=1}^{N} p(x^{(j)}|D^{(j-1)}, \alpha, G)$$

(frequentist) test error

true distribution has form:

$$\tilde{p}_{x_i,\pi_i} = \frac{N_{x_i,\pi_i} + \alpha^* \cdot q_{x_i,\pi_i}}{N + \alpha^*}$$

AIC = training error + penalty

Approximation of optimal ESS

Approximation for optimal ESS given graph G:

$$\alpha^* \approx \frac{d_G^{\text{eff}}}{E_{\hat{p}(X)}[\log \hat{p}(X|G)] - E_{q(X)}[\log \hat{p}(X|G)]} + \mathcal{O}\left(\frac{\alpha^{*2}}{N}\right)$$

d_G^{eff} ... effective number of parameters in BN

Denominator is positive for uniform q:

$$E_{\hat{p}(X)}[\log \hat{p}(X|G)] - E_{q(X)}[\log \hat{p}(X|G)]$$
$$= H(q(X|G)) - H(\hat{p}(X|G)) + \mathrm{KL}(q(X|G)||\hat{p}(X|G))$$

Properties of optimal ESS

Approximation for optimal ESS given graph G:

$$\alpha^* \approx \frac{d_G^{\text{eff}}}{E_{\widehat{p}(X)}[\log \widehat{p}(X|G)] - E_{q(X)}[\log \widehat{p}(X|G)]} + \mathcal{O}\left(\frac{\alpha^{*2}}{N}\right)$$

d_G^{eff} ... effective number of parameters in BN

Large skewness or dependences in data:

$\Rightarrow$ large ML

$\Rightarrow$ optimal ESS is small

Properties of optimal ESS

Approximation for optimal ESS given graph G:

$$\alpha^* \approx \frac{d_G^{\text{eff}}}{E_{\hat{p}(X)}[\log \hat{p}(X|G)] - E_{q(X)}[\log \hat{p}(X|G)]} + \mathcal{O}\left(\frac{\alpha^{*2}}{N}\right)$$

d_G^{eff} ... effective number of parameters in BN

Sample size N: has no explicit impact

$\Rightarrow$ for large N, we have $\alpha^* \approx const$

Properties of optimal ESS

Approximation for optimal ESS given graph G:

$$\alpha^* \approx \frac{d_G^{\text{eff}}}{E_{\hat{p}(X)}[\log \hat{p}(X|G)] - E_{q(X)}[\log \hat{p}(X|G)]} + \mathcal{O}\left(\frac{\alpha^{*2}}{N}\right)$$

d_G^{eff} ... effective number of parameters in BN

Number n of nodes in graph:

enumerator and denominator are additive in the number of nodes

$\Rightarrow$ optimal ESS approx. unaffected by n

SIEMENS

Experimental Validation of Approximations

Comparison to exact results from [Silander et al., UAI 2007] :

=>

α^* has correct order of magnitude

Data	N	n	α^I	α^M	α^*	k
balance	625	5	1...100	48	44	1
iris	150	5	1...3	2	2	2
thyroid	215	6	2...2	2	3	5
liver	345	7	3...6	4	3	3
ecoli	336	8	7...10	8	8	3
abalone	4177	9	6...6	6	7	3
diabetes	768	9	3...5	4	3	3
post op	90	9	3...5	3	3	3
yeast	1484	9	1...6	6	6	2
cancer	286	10	6...10	8	7	2
shuttle	58000	10	1...3	3	3	2
tictac	958	10	51...62	51	60	2
bc wisc	699	11	7...15	8	5	3
glass	214	11	5...6	6	6	4
page	5473	11	3...3	3	3	2
heart cl	303	14	13...16	13	9	3
heart hu	294	14	5...6	5	5	3
heart st	270	14	7...10	10	10	4
wine	178	14	8...8	8	7	3
adult	32561	15	48...58	50	49	3

SIEMENS

Conclusions

- large ESS:

  even if data and prior imply independence, the **complete** graph may have the highest marginal likelihood.

- ‘optimal’ ESS:
  - approx. independent of sample size and number of nodes
  - … is small if strong skewness or dependencies in data
  - … is large if small skewness and weak dependencies in data