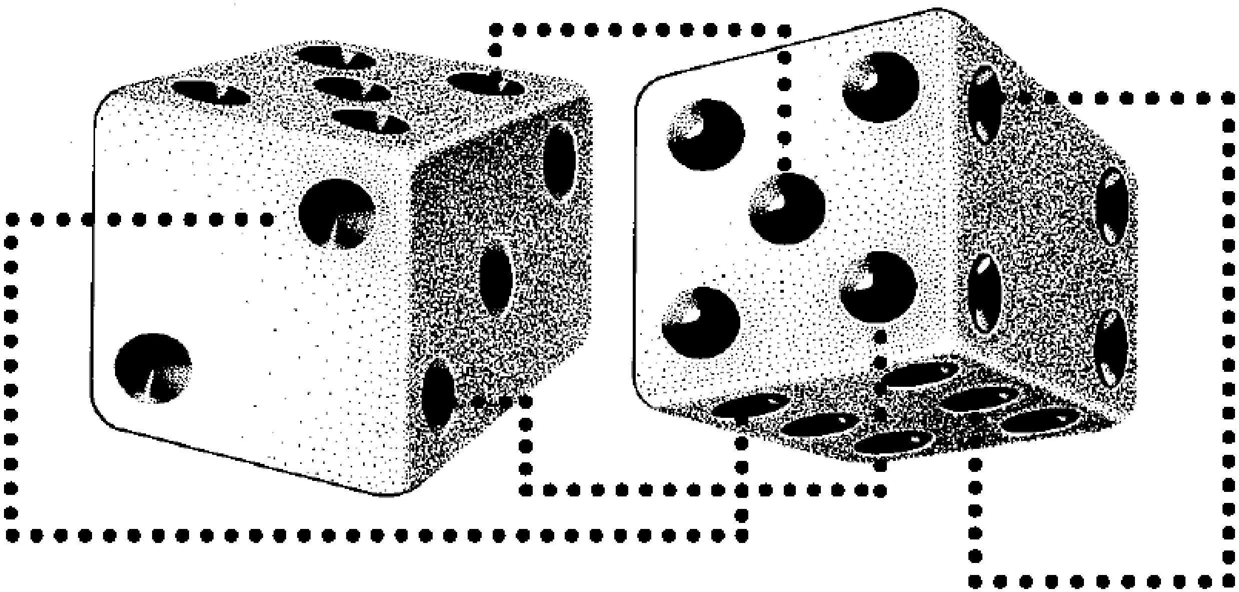
Bayesian Networks

Part 2



QB course

Véronique Lefebvre

v.lefebvre@imperial.ac.uk

Imperial College London

Constructing a Bayesian Network - Guidelines

- **Step 1**: converting influence diagram to BN
  - Literature review and expert help
  - influence diagram : Who’s the parent? Who’s the child?
  - Given the graph and some expert knowledge fill in the CPT for each node

- **Step 2**: Peer review

- **Step 3**: Testing and Learning from observations
  - Data validation
  - Updating structure and probabilities to match data

Bruce Marcot paper

Constructing a Bayesian Network - Guidelines

- Step 1: converting influence diagram to BN
  - Influence diagram:
    - To illustrate the “ecological causal web” of environmental factors affecting species or outcome of interest (e.g.)
    - Simple figure with boxes and arrows showing relevance and influence amongst variables
    - Several diagrams may be constructed at various spatial scales
    - Needs: literature review and expert knowledge (i.e. list environment factors influencing specific species)
  - BN
    - Give states to variables (T/F, low medium high, range..)
    - Give conditional probabilities for each state of child node for all combinations of their parent states.
    - CP can be specified by experts, or functions found in the literature, or be given a uniform distribution if unknown
  - Goal of step 1: get the model to tell you what you think it should tell you
    - i.e. represent expert judgment and any initial empirical data (or equations) on how the system works

Constructing a Bayesian Network - Guidelines

- Tips:
  - Keep number of parents below 3 where possible, and number of states low (but enough to ensure precision) (eg by splitting nodes->adding nodes)
    The number of probabilities in CPTs = nbstate_child x nbstate_p1 X nbstate_p2 x nbstate_p3..
  - Parentless nodes (input) – typically represents predictor habitat / environment factors – should be evaluated from data (e.g. satellite images)
  - Use intermediate nodes to summarise the inputs into major themes
  - BNs can be broken up, with the output of one being the input of another
  - all nodes should be observable and quantifiable or testable entities
  - Keep depth of the model low if possible (e.g. no more than 4 steps between input and output) bigger networks may propagate unnecessary uncertainty. (exception when modelling complex process with measurable intermediate steps)

Constructing a Bayesian Network - Guidelines

- Tips:
  - The model, including the rationale for each node and each linkage, should be fully documented
  - Link input nodes if they are likely to be correlated (an assumption of BNs is that prior probabilities associated with unlinked input nodes are uncorrelated.
  - Use proxy nodes for inputs (e.g. habitat size for habitat suitability) and express degree of confidence in proxy
  - Test the model by inputting values

Constructing a Bayesian Network - Guidelines

- Step 2
  - Consult other expert not involved in modelling
  - Reconcile the 2 models if changes

Constructing a Bayesian Network - Guidelines

- Step 3: Testing, calibrating, validating, updating

  (critical so model does not only reflect existing theory but is grounded from field data and can help develop new theories)

  - Testing BN with case data
    - Confusion matrix: compares predicted outcome with actual outcomes (with baselines..)
    - Include the probability of the outcome state, not just the state (reset baseline, e.g. very rare events)
    - When testing BN with data, no need to have data for the whole BN, just some nodes
  - Updating BN
    - Typical Bayesian thinking, revise prior given evidence (using both model and validation data)
      - 1) calibrate model states to align outcome to data (using test results to determine thresholds in discretizing),

        i.e. adjust baseline (e.g. P(o)>30 may indicate a positive outcome)

        determine more appropriate probability cut-off values in interpreting model predictions
      - 2) adjust CPT to get better data fit, using EM for instance (several learning algorithms here)

- The updated model has a known success rates and is calibrated to provide the best interpretation of results.

- Calibrating and updating Bayesian models – structure and/or CPT - can be an on-going, iterative process as new case data are gathered.

BN - Tools

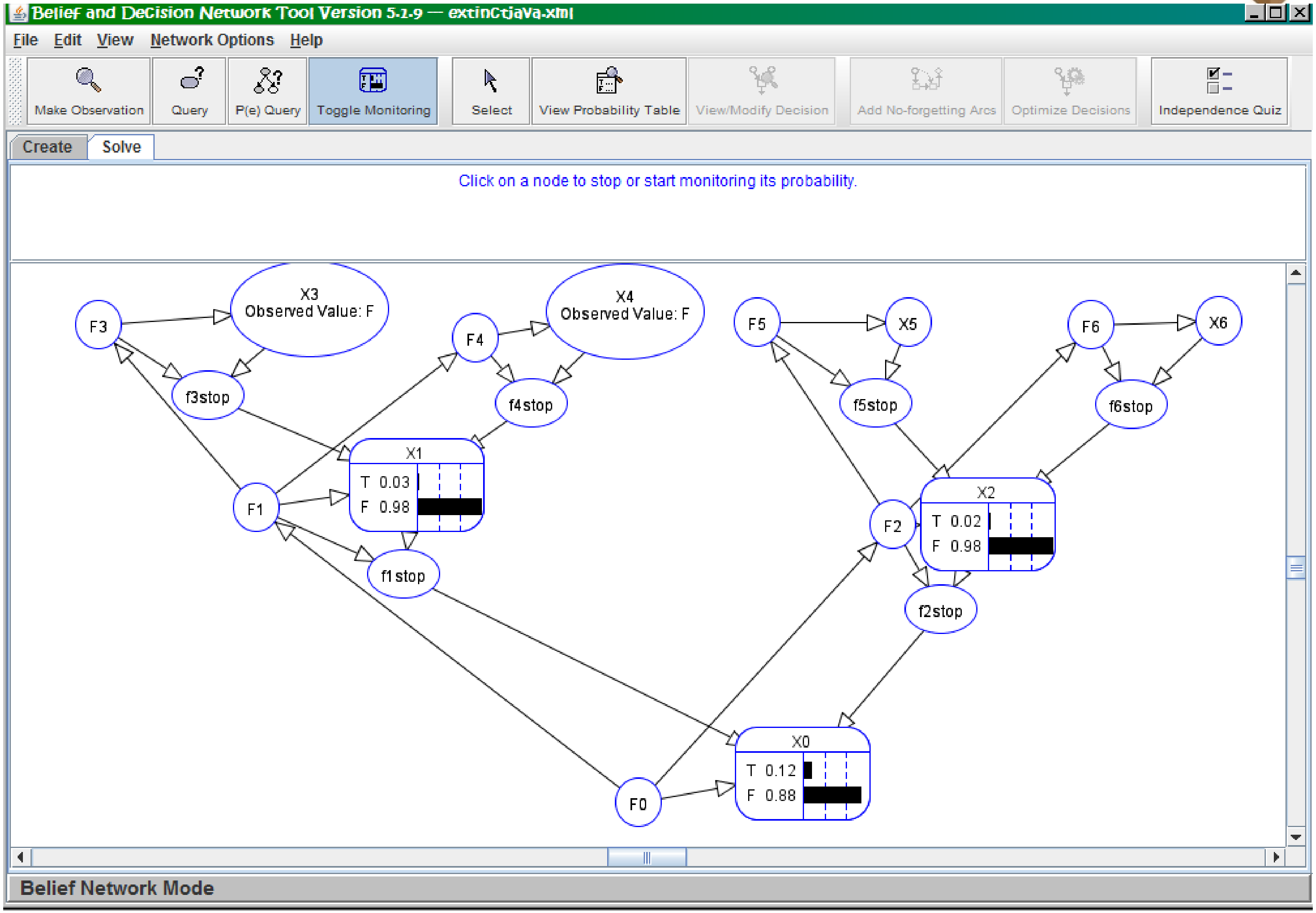
- R packages
- http://www.r-bayesian-networks.org/cookbook
- http://www.bnlearn.com/
- http://cran.r-project.org/web/views/Bayesian.html
- http://www.r-project.org/conferences/DSC-2003/Proceedings/BottcherDethlefsen.pdf

- Python: http://www.bayespy.org/intro.html

- Free Java software with GUI:
  - Belief and Decision Networks http://aispace.org/bayes/

- Netica (free version available)
  - http://www.norsys.com/netica.html
  - The free version cannot save large models

Belief and Decision Networks http://aispace.org/bayes/



Click on a node to stop or start monitoring its probability.

X3
Observed Value: F

X4
Observed Value: F

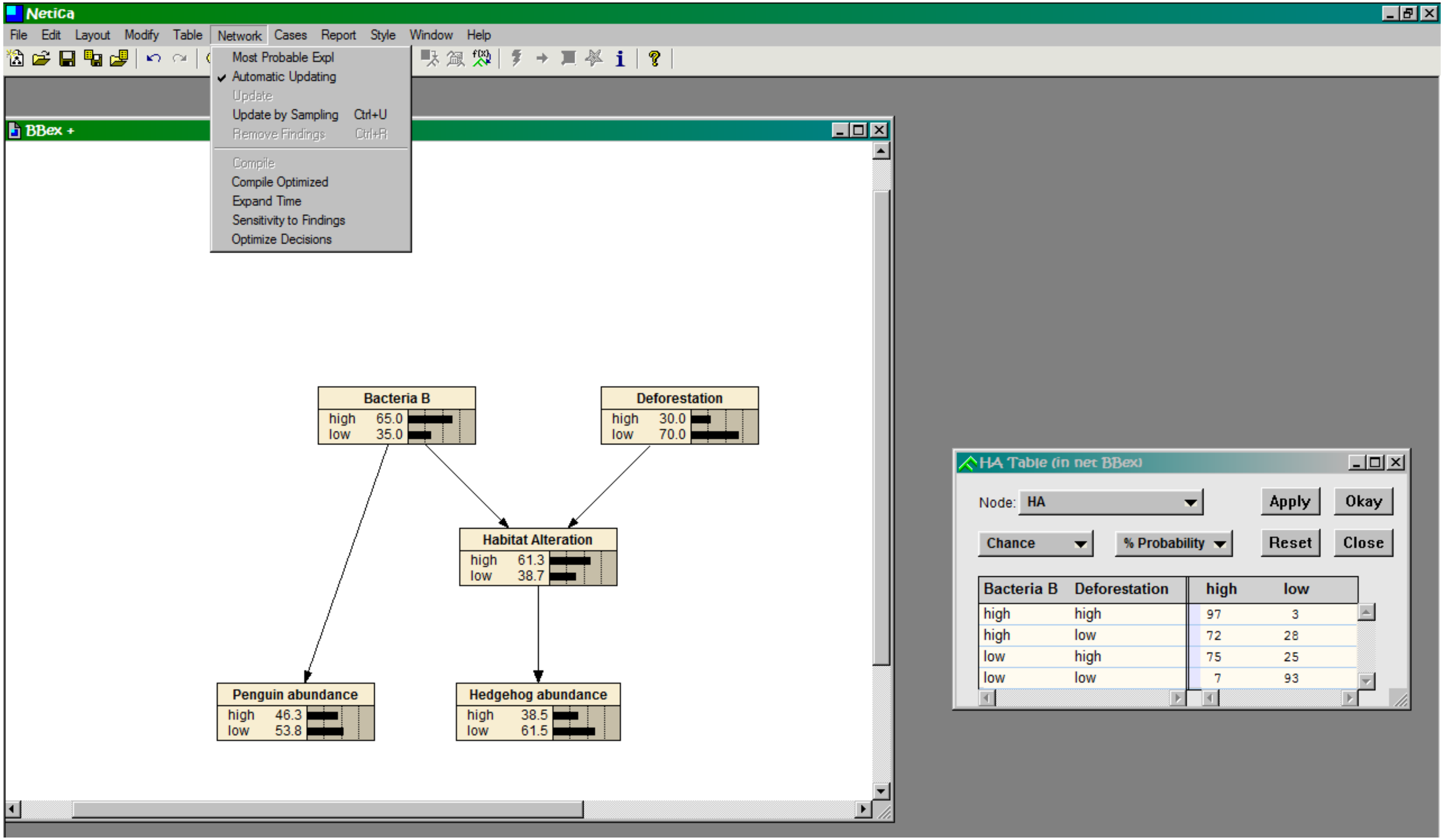
X1
T 0.03
F 0.98

X2
T 0.02
F 0.98

X0
T 0.12
F 0.88

Belief Network Mode

Netica (free version) http://www.norsys.com/netica.html



Node: HA

Chance

% Probability

Bacteria B	Deforestation	high	low
high	high	97	3
high	low	72	28
low	high	75	25
low	low	7	93

Advantages and limitations of BN

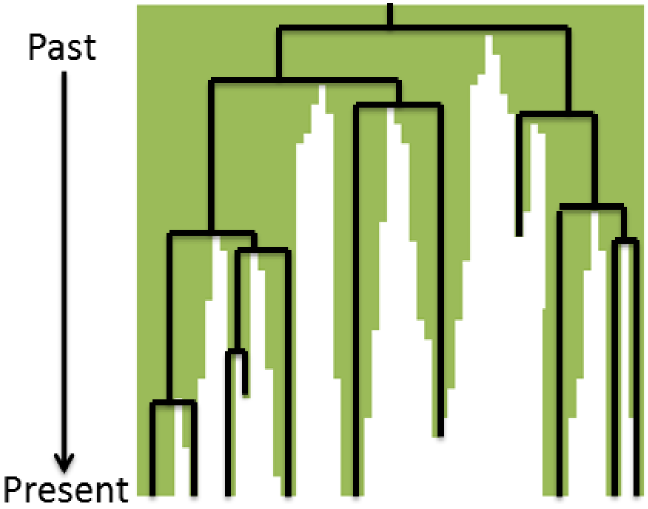
- The advantages of Bayesian Networks :
  - It is easy to recognize the dependence and independence between various nodes.
  - Bayesian networks can handle situations where the data set is incomplete since the model accounts for dependencies between all variables.
  - Bayesian networks can map scenarios where it is not feasible/practical to measure all variables due to system constraints (costs, not enough sensors, etc.)
  - Can be used for any system model - from parameters all known to none

- The limitations of Bayesian Networks:
  - The quality of the results of the network depends on the quality of the prior beliefs or model. A variable is only a part of a Bayesian network if you believe that the system depends on it (although this is true for many models)
  - Calculation of the network is NP-hard (nondeterministic polynomial-time hard), so it is very difficult and possibly costly. (but there are solutions)
  - Calculations and probabilities using Bayes' rule and marginalization can become complex and are often **characterized by subtle wording**, and care must be taken to calculate them properly.
  - Subtle wording is essential, words description of some “tool” nodes used can be very tricky!
  - Discrete states
  - No loop

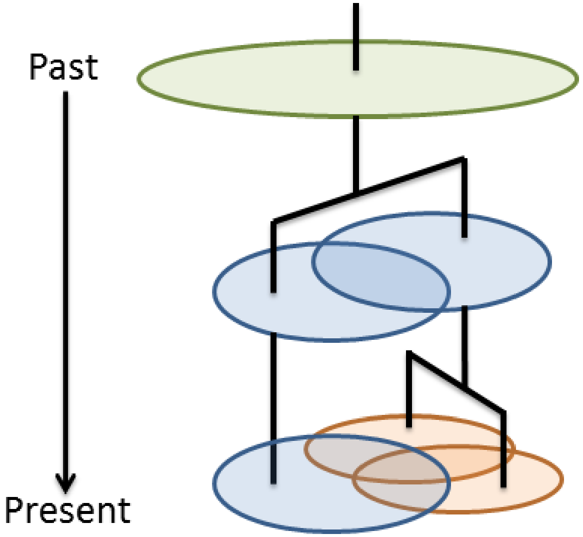
Current example: The Terragenesis model

Using landscape history to predict biodiversity patterns in fragmented landscapes

- Theory:
  - Through a succession of deforestation events a forest becomes split into fragments
  - The family tree of such fragments is named the terrageny
  - Reduction in size of local habitat results in species extinction
  - Fragments that are recently separated may have more species in common
  - The Terragenesis model can help predict biodiversity in human-modified landscapes



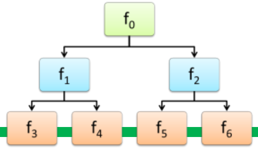
Sets of fragments



Sets of species

- Hypothesises:
  - The terrageny and fragment sizes can be learned from satellite images
  - The expected number N of species in a fragment is a function of fragment area A: $N = A^z$ (Species Area Relationship)
  - Persistence of a species in a fragment does not affect other species and they all have equal chances
  - The species composition of a fragment is determined by the species in its ancestors, but not in its siblings (no dispersal)

The Terragenesis BN



- The BN represents the presence/absence of a single species i in all the fragments of the terrageny
- The nodes of the terrageny are the fragments and the links between them represent their transformation into smaller fragments. The graph of the BN is the terrageny.

- Each node (corresponding to a fragment k) is the binary random variable “presence of species i in the fragment k”

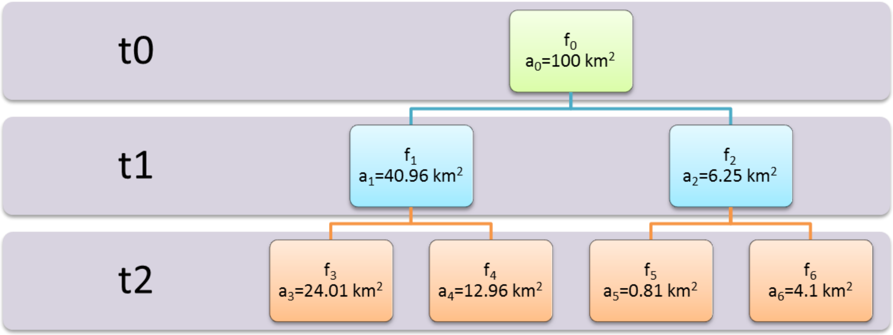
  F_{ki} : binary variable describing the event “species *i* is present in f_k “

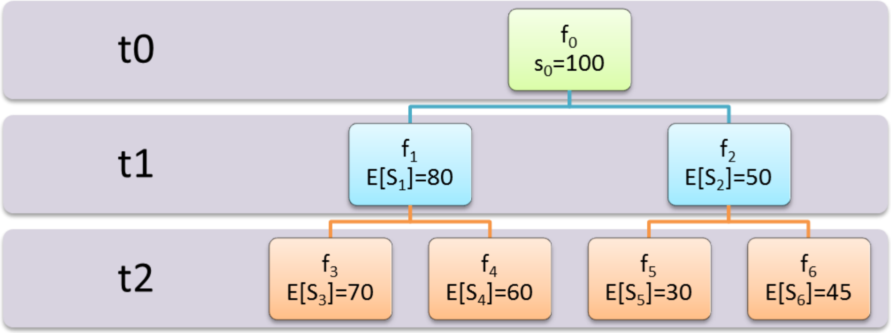
- Species i is one of the s_0 present in the original fragment f_0

- The probability of species i to persist after a splitting event is the proportion of “species places” (species capacity) in the new fragment out of the number of species places in the current fragment (they have equal chances)

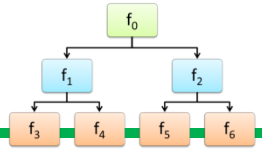
- The number of species in a fragment randomly varies and is not necessarily equal to the species capacity (except for the original landscape which is fixed).
- The species capacity is the expected value of the number of species given by the species area relationship.

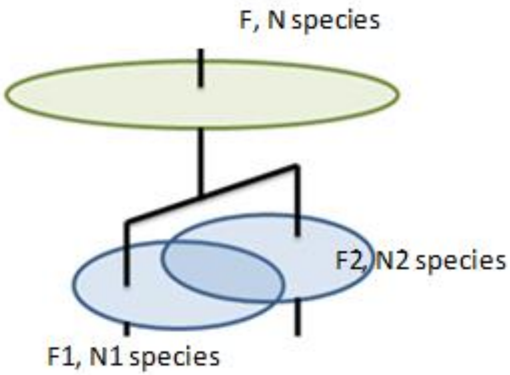
- **Conditional independence**: the probability of species i to be in a fragment given it is in its parent is independent to all other fragments.





The Terragenesis BN





The Terragenesis BN gives us

- the distribution of species i in the terrageny,
- the probability for all fragments to contain species i, P(F_{ki})

All s_0 species are have equal chances to persist and are mutually independent

→ For all fragments:
The events F_{k1}, F_{k2}, … F_{ks0} for all species are mutually independent and identically distributed.
Their sum for a fragment f_k corresponds to the number S_k of species in f_k:

$$S_k = \sum_{i=1}^{s0} F_{k_i}$$

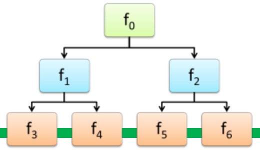
E.g.:

in f_0: F_{01}=1, F_{02}=1, F_{03}=1, F_{04}=1 → s_0 = 4
in f_1: F_{11}=1, F_{12}=0, F_{13}=1, F_{14}=1 → S_1= 3

Bernoulli and Binomial

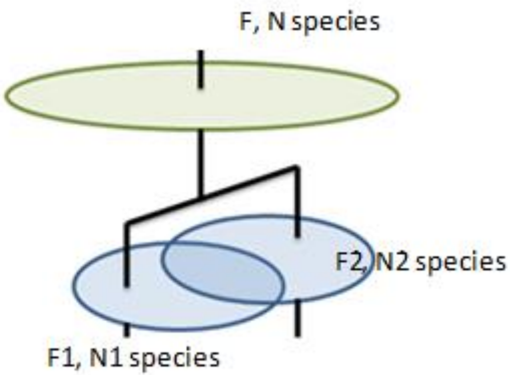
- The Fk_i are Bernoulli variables
  - $P(Fk_i = 1) = p$
  - $P(Fk_i = 0) = 1 - p$
  - Their expected value is their probability to be true.
  - $E[Fk_i] = 0 \cdot (1 - p) + 1 \cdot p = p$
  - Variance
  - $Var(Fk_i) = E\left(Fk_i^2\right) - E(Fk_i)^2 = E(Fk_i)\left(1 - E\left(Fk_i\ \right)\right) = p(1 - p)$
- S_k the sum of the Fk_i
  - The expected value of their sum, for i=1→N:
  - $E[S_k] = E[\sum_{i=1}^{N} Fk_i] = \sum_{i=1}^{N} E(Fk_i) = N\ \times p$
  - The Fk_i independent for all i (cov = 0) so the variance of their sum:
  - $Var(Fk_i) = N \cdot Var(Fk_i) = N \times p(1 - p)$
  - S_k is a binomial variable of parameters N and p
  - A sum of N i.i.d. (independent and identically distributed) Bernoulli variables with probability p, is a binomial variable of parameters N and p

Fragments similarity



- From the Terragenesis BN we have the probability of species i to be in each fragment
- And from the Binomial distribution we have the expected values of the number of species in each fragment, and the probability than the number of species is x.

- We are interested in the similarity between 2 fragments:
- The amount of species shared between 2 fragments (intersection) over the total number of species in the 2 fragments(union).
- To compute this for all fragments we can first compute the probability for a single species i to be in the 2 fragments (intersection), and in one or the other or both (union)

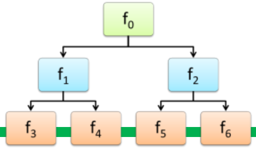


Independence $\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B)$

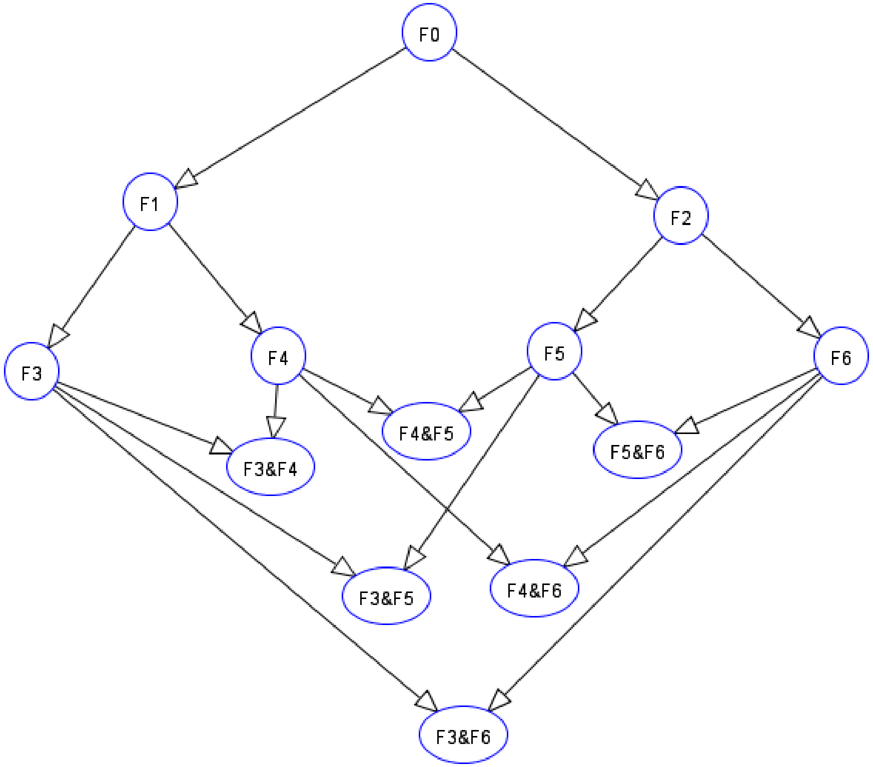
Inclusion-exclusion principle $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$

Disjoint events $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$

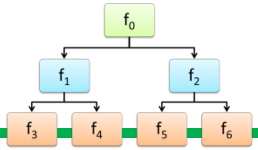
Fragments similarity



- We can add functional ‘AND’ nodes in the graph, to represent the variable species i is present in both fragments, (and then sum for all species) to obtain the expected number of shared species.
- Species i presence in a fragment is conditionally independent to its presence in another given it is present in their most recent common ancestor
- Expected number of shared species will be a question in the practical



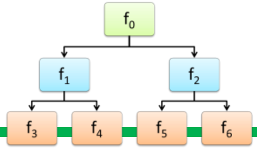
Extinct species (Not part of practical)



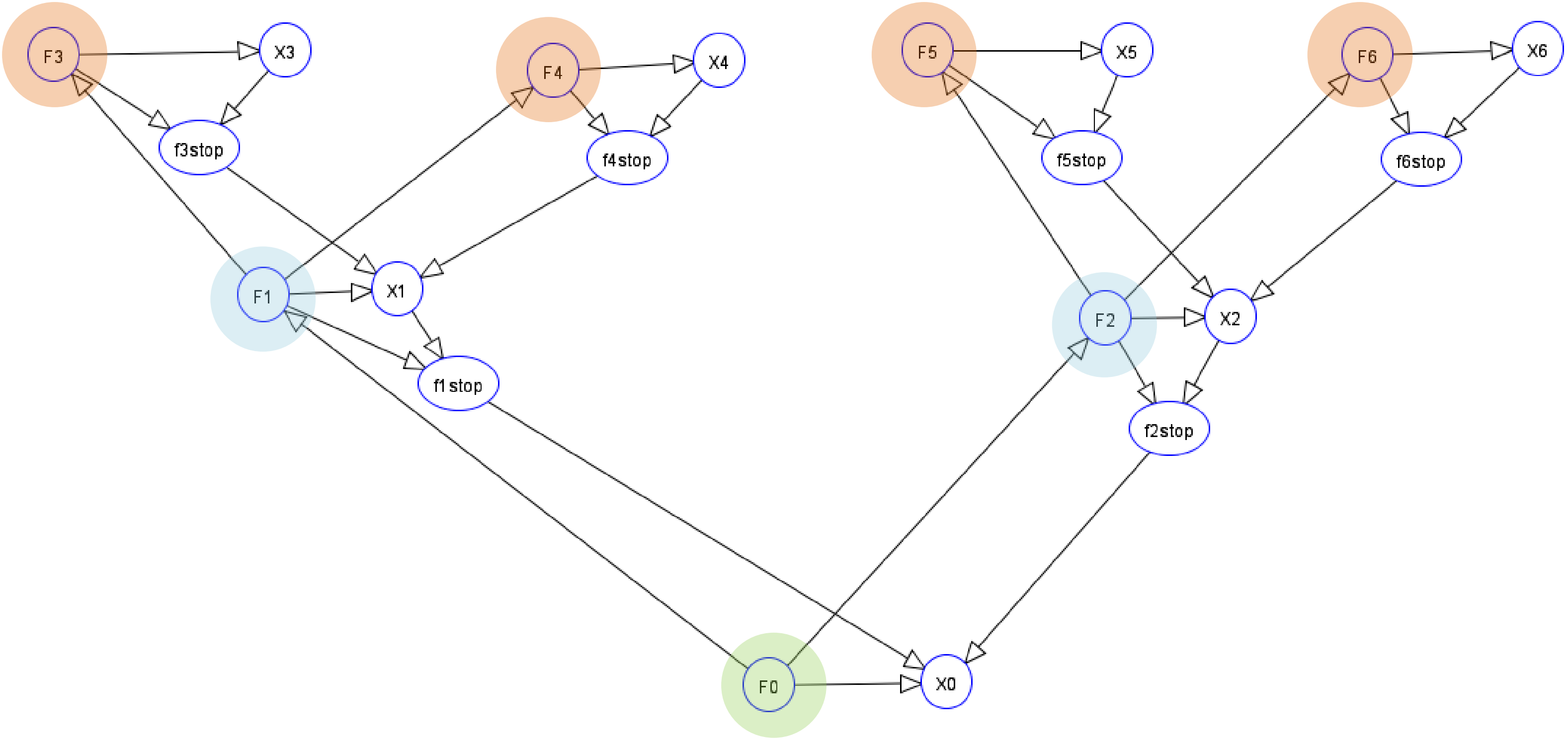
- Species can become extinct after splitting events
- A species in f1 is extinct from f1 if it did not persist in f3 and it did not persist in f4 (descendants)
- A species in f1 is extinct from f2 if it did not persist in f5 and it did not persist in f6
- A species in f0 is extinct from f0 if:
  - It did not persist in f1 OR it did AND is extinct from f1
  - And it did not persist in f2 OR it did AND is extinct from f2

- A species in f3,f4,f5 or f6 cannot be extinct

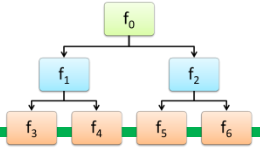
Extinct species



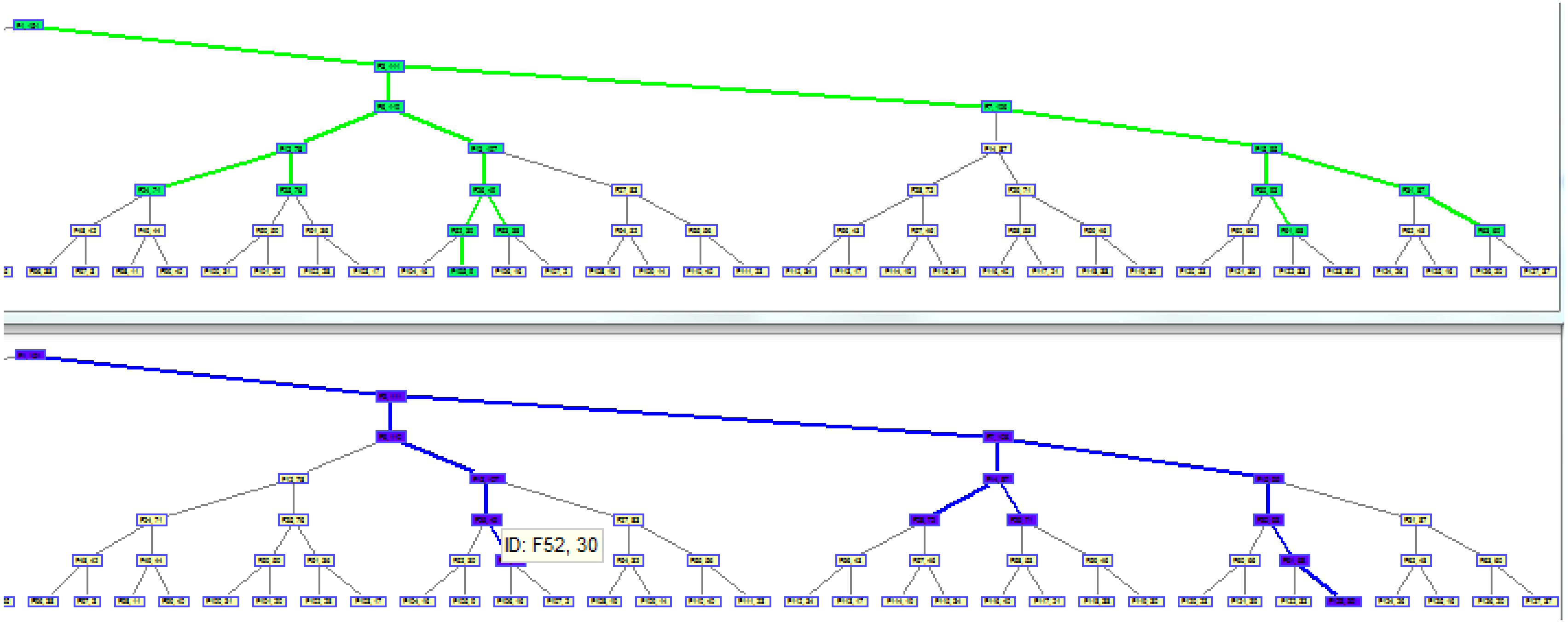
- We can add in the BN the variables “Xk: species i is extinct from fk” and logical nodes to relate them to the Fki following the “extinction” rules.
- Then we can predict the expected number of extinct species in a landscape, given a terrageny.



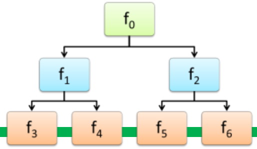
Endemic species (Not part of practical)



- A species is endemic to the landscape if it is present in only 1 present day fragment
- It is linked to extinct species and also require a recursive definition
- We define a species to be “dominated” by a fragment fk if it is in fk and the only present day fragment that contains the species (if any) is a direct descendent of fk (subtle wording !)
- If fk is a present-day fragment then dominated is the same as endemic



Endemic species



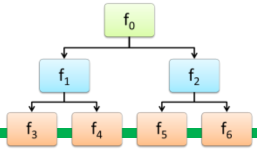
- A species is dominated by fk (given that it is dominated by its parent) if:
  - It is in fk
  - AND
  - It is not present in his siblings or is but got extinct from it

(similar to extinct but it starts from the top)

All species are dominated by f0

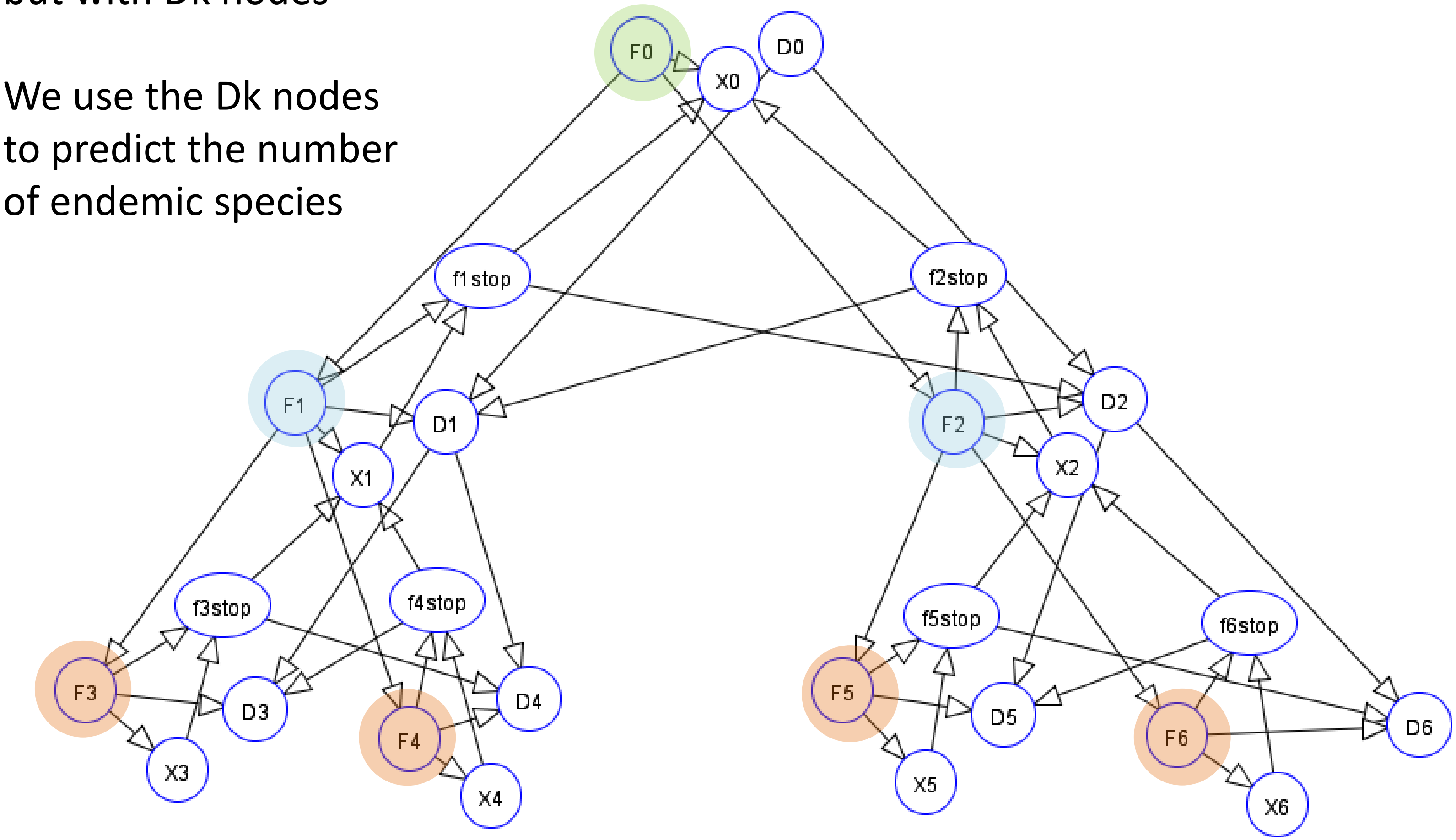
We can add another layer of nodes to represent it:

Endemic species



Same as extinct BN, but with Dk nodes

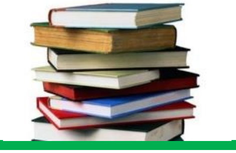
We use the Dk nodes to predict the number of endemic species



Practical :

- Finding equations of terrageny model
  - As a review of random variables and BN
  - All rules and theorems will be given
  - Needs pen and paper
- Solving simple example BNs manually
- Implementing the Terragenesis model or other examples with a BN GUI software

Bibliography / suggested readings



Articles:

- Charniak, Eugene. "Bayesian networks without tears." AI magazine 12.4 (1991): 50.
- Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation – Marcot et al., 2006 Canadian Journal of Forest Research
- Koller, Daphne, et al. "2 Graphical Models in a Nutshell." Statistical relational learning (2007): 13.
- Nicholson, O. Woodberry and C. Twardy (2010). The "Native Fish" Bayesian networks. http://bayesian-intelligence.com/bwb/2012-03/how-to-model-with-bayesian-networks/

Web resources:

- An Intuitive Explanation of Bayes' Theorem - by Eliezer Yudkowsky
  http://yudkowsky.net/rational/bayes
- Fallacies about probabilities - by Norman Fenton
  http://www.agenarisk.com/resources/probability_puzzles/Making_sense_of_probability.html
- A Brief Introduction to Graphical Models and Bayesian Networks - by Kevin Murphy
  http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html
- Machine Learning course by Doina Precup
  http://www.cs.mcgill.ca/~dprecup/courses/ML/lectures.html
- Virtual Laboratories in Probability and Statistics
  http://www.math.uah.edu/stat/
- Bayes nets library:
  http://www.norsys.com/netlibrary/index.htm