USTC

Chapter 14b

Inference in Bayesian Networks

王子磊 (Zilei Wang)

Email: zlwang@ustc.edu.cn

http://vim.ustc.edu.cn

提纲

- 精确推理 (exact inference)
  - 通过枚举 (enumeration)
  - 通过变量消元 (variable elimination)
- 近似推理 (approximate inference)
  - 通过随机模拟 (stochastic simulation)
  - 通过 Markov 链蒙特卡洛 (MCMC)

推理任务

- 简单查询：计算后验边缘概率 $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$

  e.g., $P(NoGas|Gauge=empty, Lights=on, Starts=false)$

- 合取查询：$\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$

- 最优决策：决策网络包括效用信息，概率推理需要计算

  $P(outcome|action, evidence)$

- 信息价值：哪些证据是需要下一步获取的？

- 灵敏度分析：哪些概率值是最重要的？

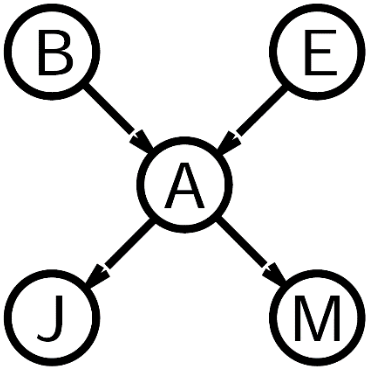
- 解释：我为什么需要一个启动马达？

通过枚举进行推理

- 稍微智能的方法：直接求和联合概率分布中的变量，而不进行实际概率的显式计算

- 在 Burglary 网络上的简单查询

$$
\begin{aligned}
&\mathbf{P}(B|j,m)\\
&= \mathbf{P}(B,j,m)/P(j,m)\\
&= \alpha\mathbf{P}(B,j,m)\\
&= \alpha \textstyle\sum_e \sum_a \mathbf{P}(B,e,a,j,m)
\end{aligned}
$$



- 采用 CPT 项的乘积表达完全联合项

$$
\begin{aligned}
&\mathbf{P}(B|j,m)\\
&= \alpha \textstyle\sum_e \sum_a \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)\\
&= \alpha\mathbf{P}(B) \textstyle\sum_e P(e) \sum_a \mathbf{P}(a|B,e)P(j|a)P(m|a)
\end{aligned}
$$

- 递归深度优先枚举：$O(n)$ 空间复杂度， $O(d^n)$ 时间复杂度

枚举算法

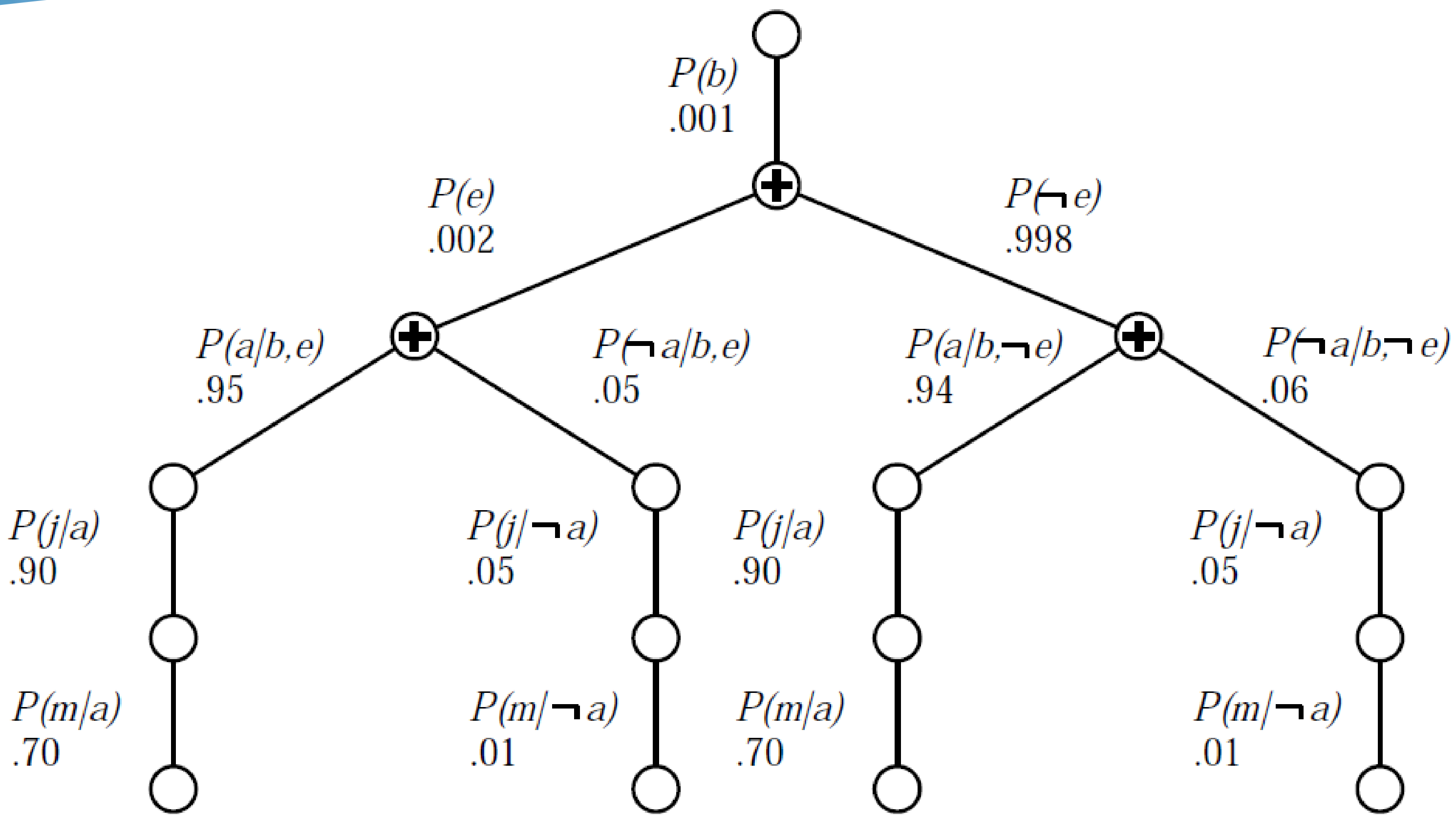
function ENUMERATION-ASK(X, $\mathbf{e}$, bn) **returns** a distribution over X
  inputs: X, the query variable
    $\mathbf{e}$, observed values for variables $\mathbf{E}$
    bn, a Bayesian network with variables $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$ 父节点到子节点的次序

  $\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
  for each value x_i of X **do**
    extend $\mathbf{e}$ with value x_i for X
    $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(VARS[bn], $\mathbf{e}$)
  } 针对查询变量的每个值进行计算（加入到证据变量中）
  return NORMALIZE($\mathbf{Q}(X)$)   归一化

---

function ENUMERATE-ALL($vars$, $\mathbf{e}$) **returns** a real number   递归算法
  if EMPTY?($vars$) **then return** 1.0
  $Y \leftarrow$ FIRST($vars$)
  if Y has value y in $\mathbf{e}$
    then return $P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}$)   观测到
    else return Σ_y $P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}_y$)   未观测到
      where $\mathbf{e}_y$ is $\mathbf{e}$ extended with $Y = y$

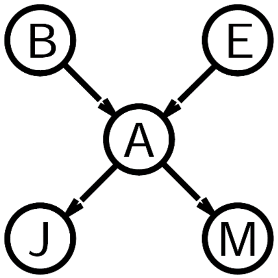
估算树



- 枚举是低效的：存在重复计算
  - e.g., 对 e 的每个值，都计算了 $P(j|a)P(m|a)$

通过变量消元进行推理

- 变量消元 (variable elimination)： 从右向左进行求和操作
  - 存储中间结果 (factors) 以避免重复计算



$$
\begin{aligned}
\mathbf{P}(B|j,m) & \\
&= \alpha \underbrace{\mathbf{P}(B)}_{B} \textstyle\sum_e \underbrace{P(e)}_{E} \textstyle\sum_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&= \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \textstyle\sum_a \mathbf{P}(a|B,e) P(j|a) \underline{f_M(a)} && \text{消除 } M \\
&= \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \textstyle\sum_a \mathbf{P}(a|B,e) \underline{f_J(a)} f_M(a) && \text{消除 } J \\
&= \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \textstyle\sum_a \underline{f_A(a,b,e)} f_J(a) f_M(a) \\
&= \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \underline{f_{\bar{A}JM}(b,e)} \ (\text{sum out } A) && \text{消除 } A \\
&= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \ (\text{sum out } E) && \text{消除 } E \\
&= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b) && \text{消除 } B
\end{aligned}
$$

变量消元：基本操作

- 求和：针对单个变量，累加乘法因子
  - 将所有约束因子移到求和项外部
  - 将其余因子逐点相乘的子矩阵进行累加

$$\sum_x f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \sum_x f_{i+1} \times \cdots \times f_k = f_1 \times \cdots \times f_i \times f_{\bar{X}}$$

这里，假定 $f_1, f_2, \ldots, f_i$ 不依赖于 X

- 逐点相乘：考虑 f_1 和 f_2

$$\begin{aligned} & f_1(x_1, \ldots, x_j, y_1, \ldots, y_k) \times f_2(y_1, \ldots, y_k, z_1, \ldots, z_l) \\ & \quad = f(x_1, \ldots, x_j, y_1, \ldots, y_k, z_1, \ldots, z_l) \end{aligned}$$

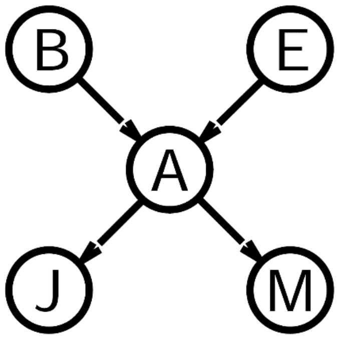
E.g., $f_1(a,b) \times f_2(b,c) = f(a,b,c)$

变量消元算法

function ELIMINATION-ASK(X, $\mathbf{e}$, bn) **returns** a distribution over X
  inputs: X, the query variable
    $\mathbf{e}$, evidence specified as an event
    bn, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

  $factors \leftarrow [\,]$; $vars \leftarrow$ REVERSE(VARS[bn]) 子节点到父节点的次序
  for each var **in** $vars$ **do**
    $factors \leftarrow$ [MAKE-FACTOR(var, $\mathbf{e}$)|$factors$]
    if var is a hidden variable **then** $factors \leftarrow$ SUM-OUT(var, $factors$) 求和消除
  return NORMALIZE(POINTWISE-PRODUCT($factors$)) 显变量点乘

无关变量 (Irrelevant variables)



- 考虑查询：$P(JohnCalls|Burglary=true)$

  $$P(J|b) = \alpha P(b)\sum_e P(e)\sum_a P(a|b,e)P(J|a)\sum_m P(m|a)$$

  - 在 m 上求和为 1； M 是与查询无关的

- 定理1：Y 是查询无关的， 除非 $Y \in Ancestors(\{X\} \cup \mathbf{E})$

- 在上例中，$X=JohnCalls$, $\mathbf{E}=\{Burglary\}$

  $$Ancestors(\{X\} \cup \mathbf{E}) = \{Alarm, Earthquake\}$$

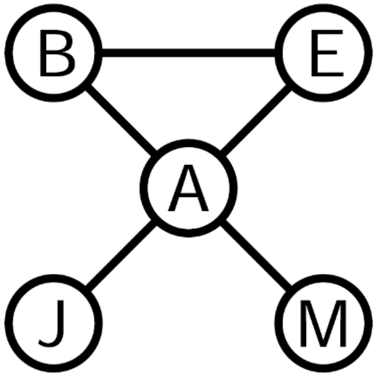
  因此 *MaryCalls* 是无关的

类比于 Horn 子句 KB 的后向链接算法

无关变量 (Irrelevant variables)

- 定义（贝叶斯网络的 moral 图）：连接父节点并消除箭头
- 定义（m-分离, m-separated）：A 通过 C 对 B 是 m-分离的，当且仅当 它们在 moral 图中通过 C 是可分离的
- 定理2：Y 是无关的，如果它通过 E 对 X 是 m-分离的

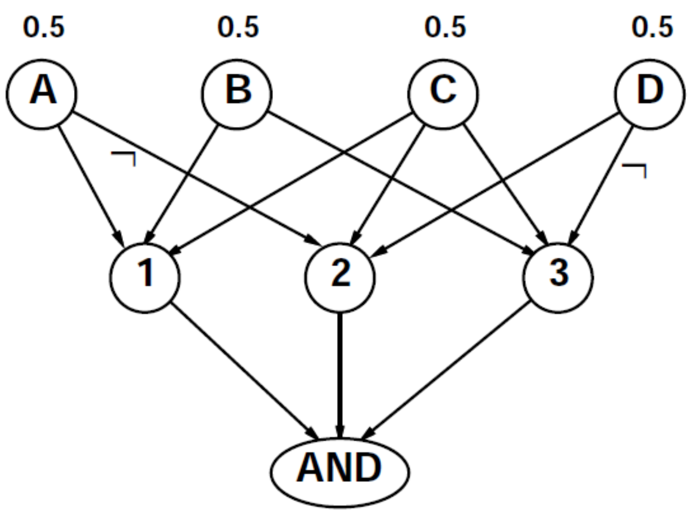
For $P(JohnCalls|Alarm = true)$, both $Burglary$ and $Earthquake$ are irrelevant



确定推理的复杂度 *

- ❖ 单连通网络 (多形树，polytrees)：
  - 任何两个节点之间最多只有一条 (无向) 路径
  - 变量消元的时间和空间代价是 $O(d^k n)$
- ❖ 多连通网络：
  - 能够转换为 3SAT 进行精确推理 → NP-hard
  - 等价于对 3SAT 模型的计数 → #P-complete

1. A v B v C
2. C v D v ¬A
3. B v C v ¬D



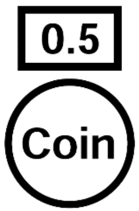
通过随机模拟进行推理

❖ 基本思想：

a) 从样本分布 S 中采用 N 个样本

b) 计算一个近似的后验概率 $\hat{P}$

c) 证明该概率收敛于真实概率 P



❖ 贝叶斯网络采样方法

- 从一个空白网络中进行直接采样
- 拒绝采样：与证据不一致的拒绝样本
- 似然加权：根据证据对采样样本进行加权
- 马尔科夫链蒙特卡洛 (Markov chain Monte Carlo, MCMC)：根据一个随机过程进行采样，它的稳态分布就是真实的后验概率

从一个空白网络中直接采样

- 无证据变量，直接进行事件采样

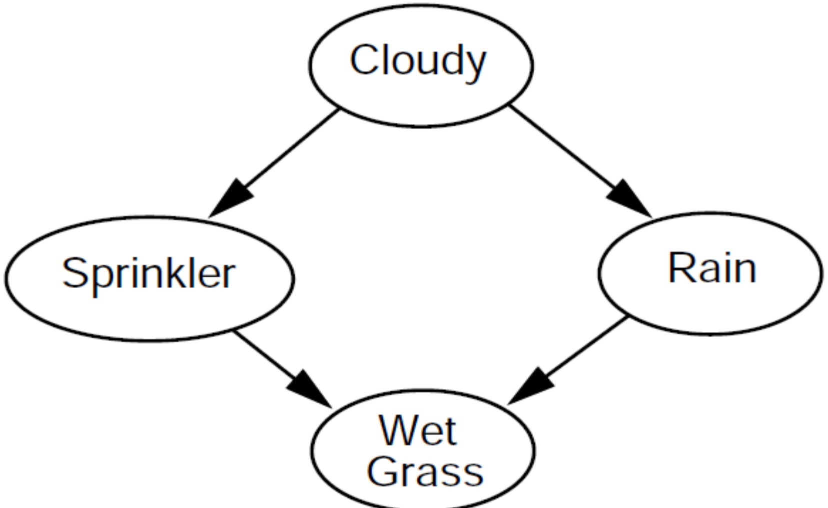
function PRIOR-SAMPLE(bn) **returns** an event sampled from bn
  inputs: bn, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

  $\mathbf{x} \leftarrow$ an event with n elements
  for $i = 1$ **to** n **do**
    $x_i \leftarrow$ a random sample from $\mathbf{P}(X_i \mid parents(X_i))$
      given the values of $Parents(X_i)$ in $\mathbf{x}$
  return $\mathbf{x}$

依据条件概率进行采样

示例

P(C)
.50



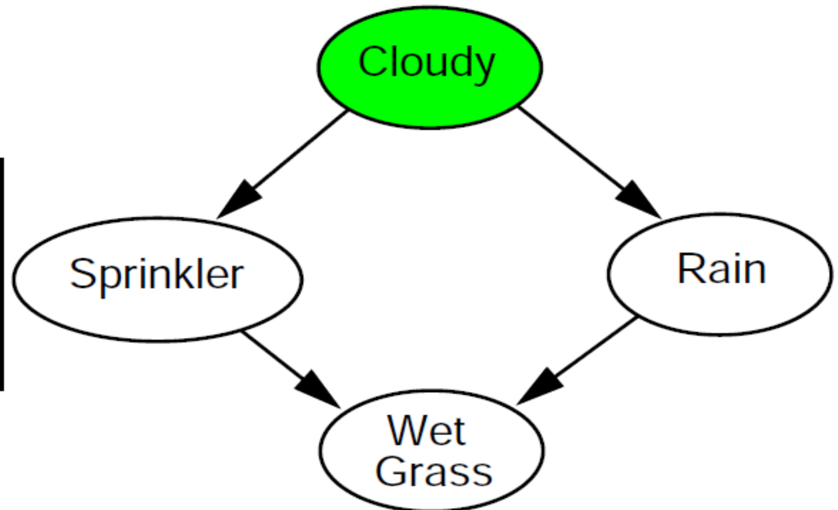
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

示例

P(C)
.50

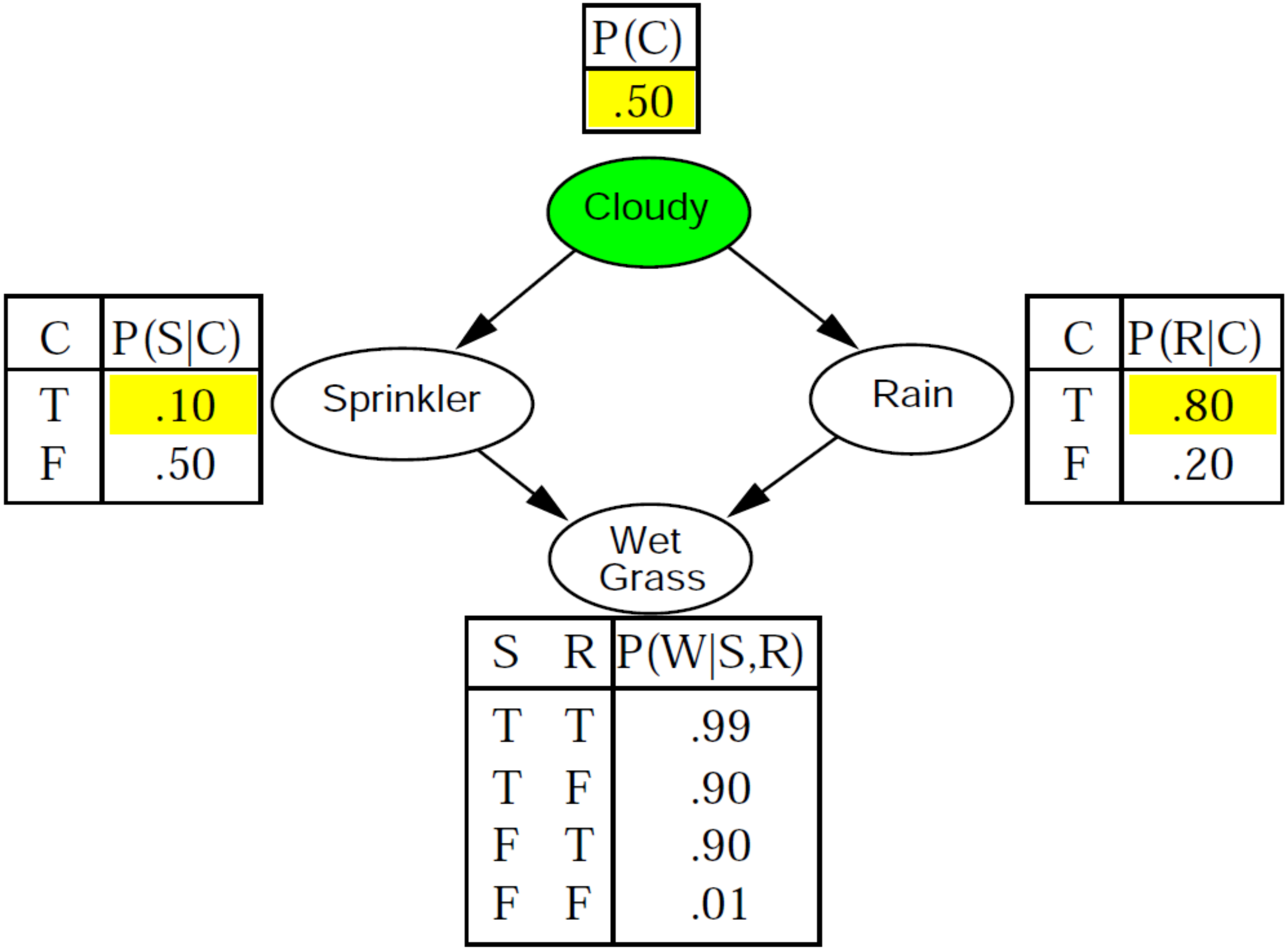


C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

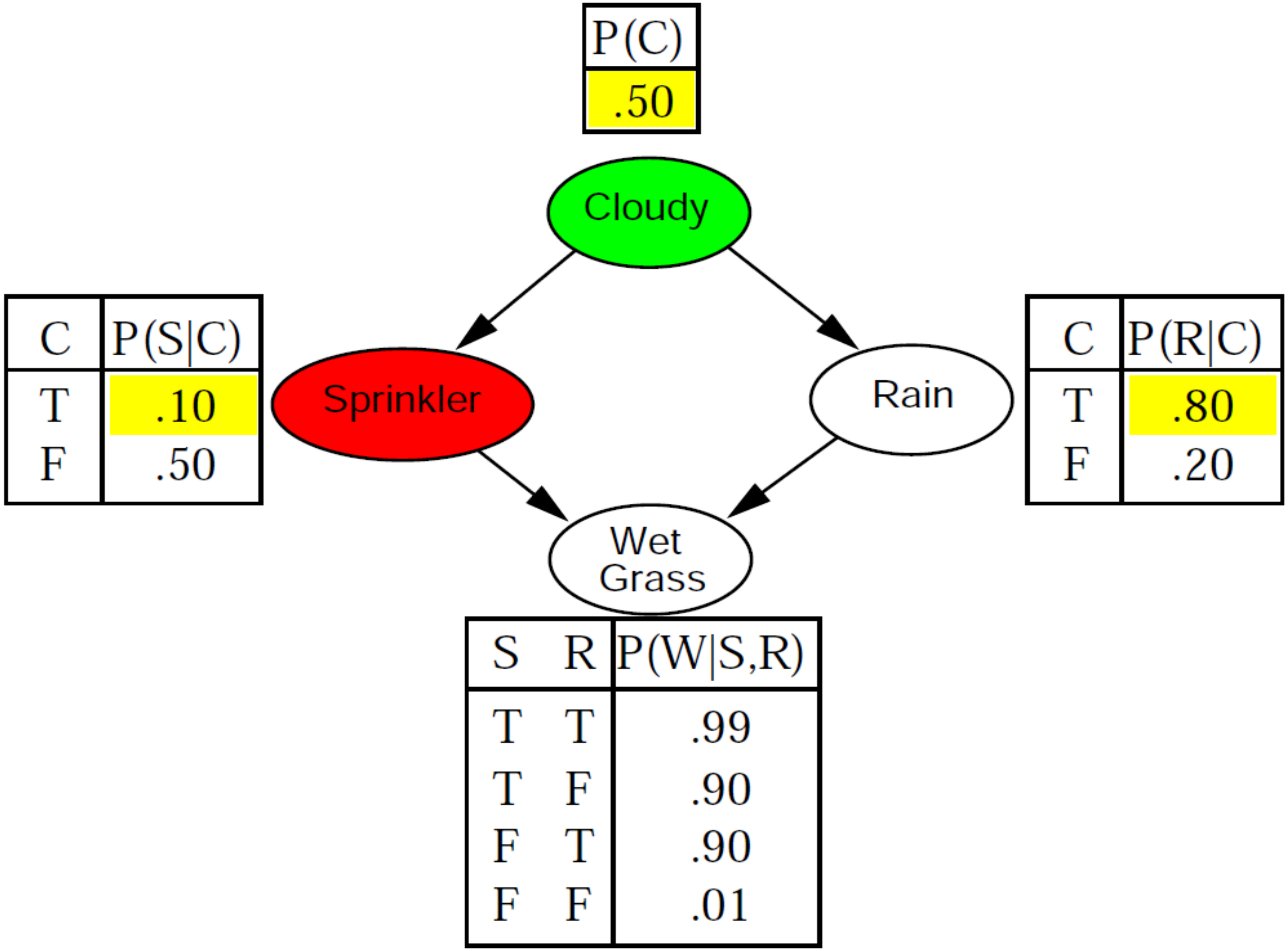
S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

示例



P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

示例



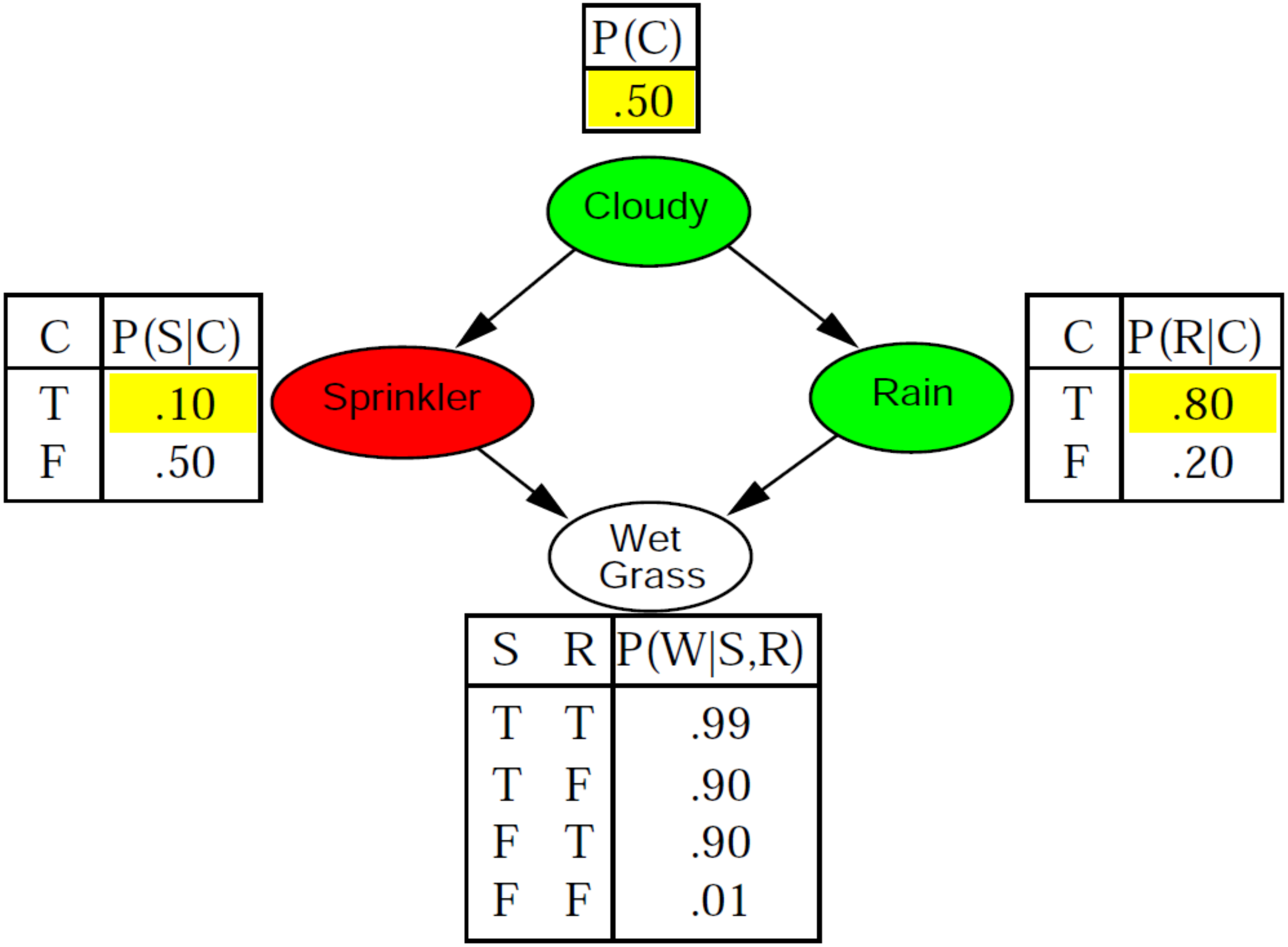
P(C)
.50

C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

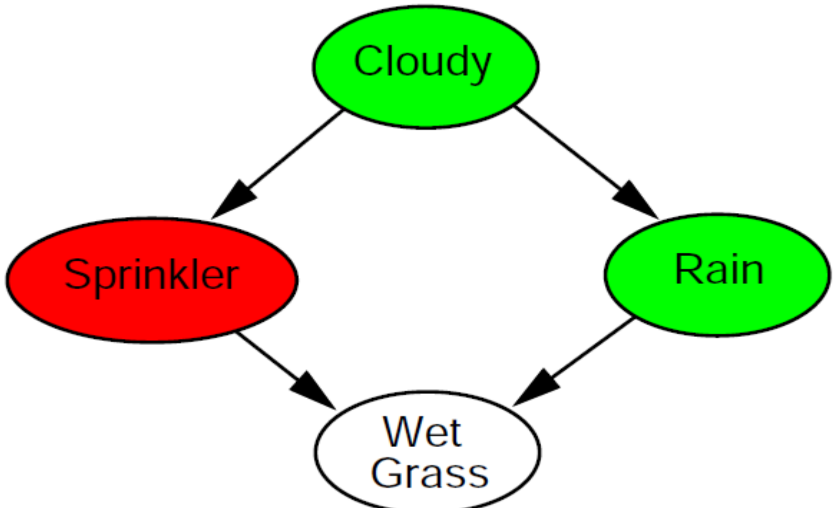
示例



P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

示例

P(C)
.50

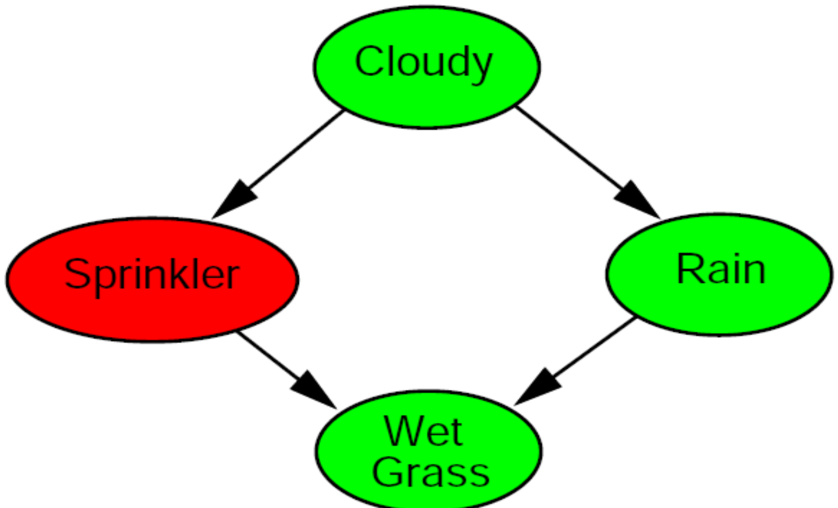


C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

示例



P(C)
.50

C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

从一个空白网络中直接采样

- 直接采样 (PriorSample) 产生一个特定事件的概率为：

$$S_{PS}(x_1 \ldots x_n) = \Pi_{i=1}^{n} P(x_i|parents(X_i)) = P(x_1 \ldots x_n)$$

i.e., 真实的先验概率

E.g., $S_{PS}(t,f,t,t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t,f,t,t)$

- $N_{PS}(x_1 \ldots x_n)$ 表示事件 $x_1, \ldots, x_n$ 的采样样本数，则有：

$$\begin{aligned}\lim_{N\to\infty} \hat{P}(x_1,\ldots,x_n) &= \lim_{N\to\infty} N_{PS}(x_1,\ldots,x_n)/N \\ &= S_{PS}(x_1,\ldots,x_n) \\ &= P(x_1 \ldots x_n)\end{aligned}$$

即：直接采样的概率估计是一致的

$$\hat{P}(x_1,\ldots,x_n) \approx P(x_1 \ldots x_n)$$

拒绝采样 (Rejection sampling)

- 考虑证据变量 e，估计与其一致的概率，即条件概率 $\hat{\mathbf{P}}(X|\mathbf{e})$

```
function REJECTION-SAMPLING(X, e, bn, N) returns an estimate of P(X|e)
   local variables: N, a vector of counts over X, initially zero

   for j = 1 to N do
       x ← PRIOR-SAMPLE(bn)
       if x is consistent with e then
           N[x] ← N[x]+1 where x is the value of X in x
   return NORMALIZE(N[X])
```

正常采样，但只记录一致的样本

- 例如：用 100 个样本估计 $\mathbf{P}(Rain|Sprinkler = true)$
  - 27 个样本满足 $Sprinkler = true$
  - 其中，有 8 个 $Rain = true$，19 个是 $Rain = false$

$$\hat{\mathbf{P}}(Rain|Sprinkler = true) = \text{NORMALIZE}(\langle 8, 19 \rangle) = \langle 0.296, 0.704 \rangle$$

拒绝采样分析

$$
\begin{aligned}
\hat{\mathbf{P}}(X|\mathbf{e}) &= \alpha \mathbf{N}_{PS}(X,\mathbf{e}) && \text{(algorithm defn.)} \\
&= \mathbf{N}_{PS}(X,\mathbf{e})/N_{PS}(\mathbf{e}) && \text{(normalized by } N_{PS}(\mathbf{e})\text{)} \\
&\approx \mathbf{P}(X,\mathbf{e})/P(\mathbf{e}) && \text{(property of PriorSample)} \\
&= \mathbf{P}(X|\mathbf{e}) && \text{(defn. of conditional probability)}
\end{aligned}
$$

- 拒绝采样计算结果是一致的后验概率估计
- 问题：如果 $P(e)$ 比较小，则计算代价非常高
  - $P(e)$ 随着证据变量数目的增加呈指数级下降

似然加权 (likelihood weighting，LW)

- 基本思想：固定证据变量，只采样非证据变量，同时对采样样本根据与证据的似然度进行加权

function LIKELIHOOD-WEIGHTING(X, $\mathbf{e}$, bn, N) **returns** an estimate of $P(X|\mathbf{e})$
  local variables: $\mathbf{W}$, a vector of weighted counts over X, initially zero

  for $j = 1$ to N **do**
    $\mathbf{x}, w \leftarrow$ WEIGHTED-SAMPLE(bn)
    $\mathbf{W}[x] \leftarrow \mathbf{W}[x] + w$ where x is the value of X in $\mathbf{x}$
  return NORMALIZE($\mathbf{W}[X]$)

---

function WEIGHTED-SAMPLE(bn, $\mathbf{e}$) **returns** an event and a weight

返回采样样本及其权重

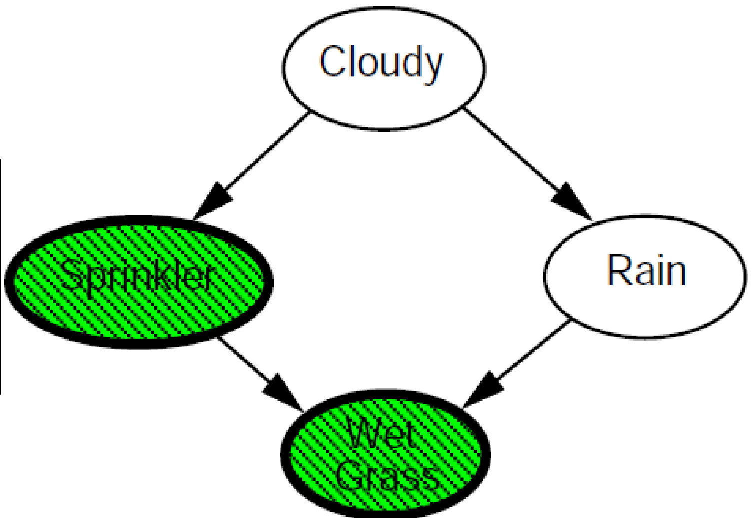
  $\mathbf{x} \leftarrow$ an event with n elements; $w \leftarrow 1$
  for $i = 1$ **to** n **do**
    if X_i has a value x_i in $\mathbf{e}$
      then $w \leftarrow w \times P(X_i = x_i \mid parents(X_i))$

对证据变量不采样

      else $x_i \leftarrow$ a random sample from $\mathbf{P}(X_i \mid parents(X_i))$
  return $\mathbf{x}$, w

似然加权示例

P(C)
.50



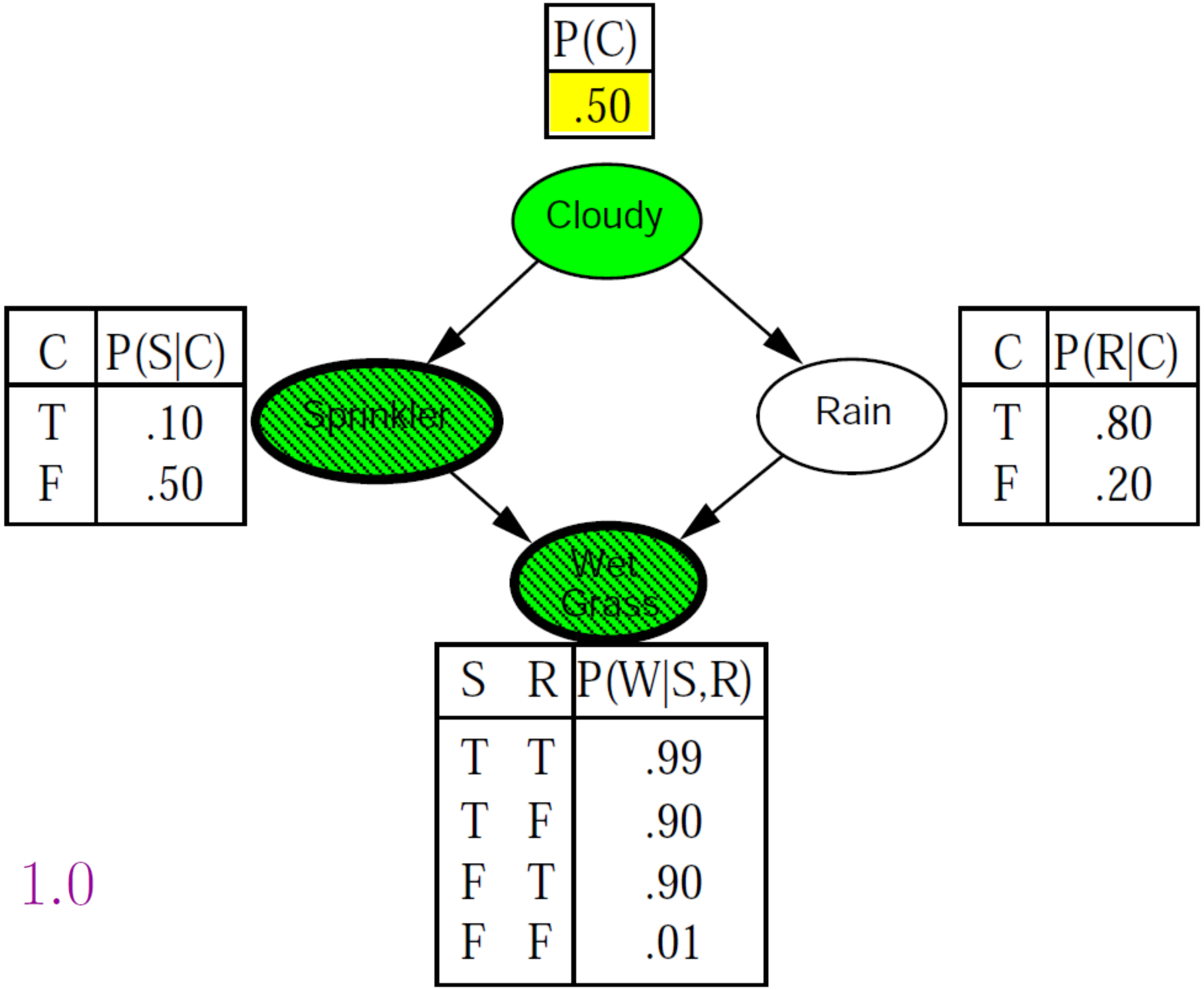
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0$$

似然加权示例



P(C)
.50

C	P(S|C)
T	.10
F	.50

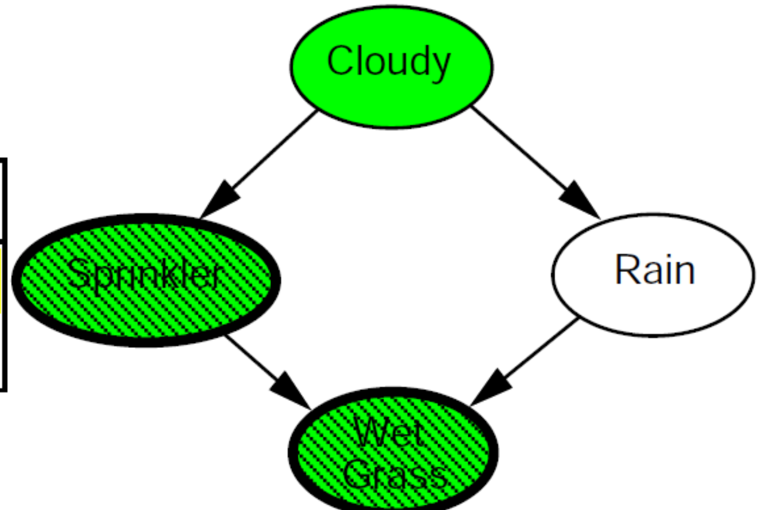
C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

似然加权示例

P(C)
.50



C	P(S|C)
T	.10
F	.50

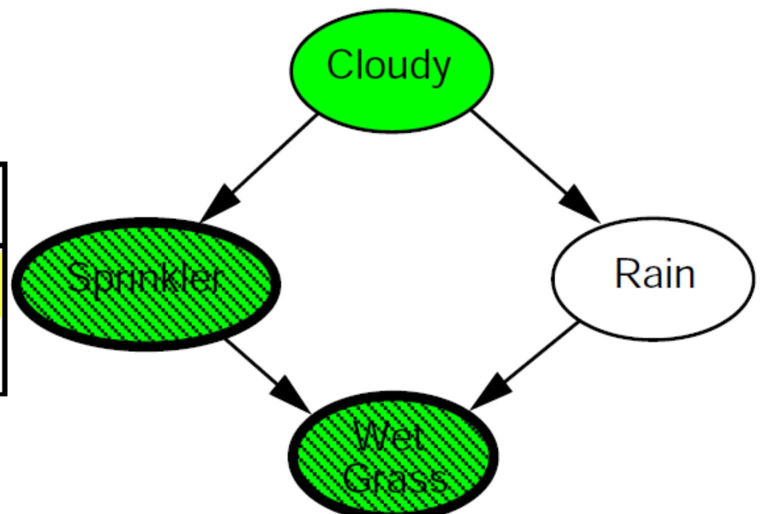
C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0$$

似然加权示例

P(C)
.50



C	P(S|C)
T	.10
F	.50

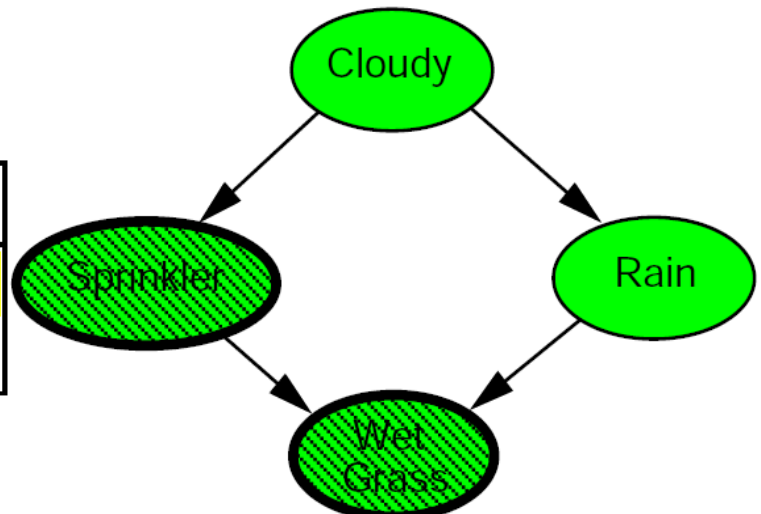
C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0 \times 0.1$$

似然加权示例

P(C)
.50



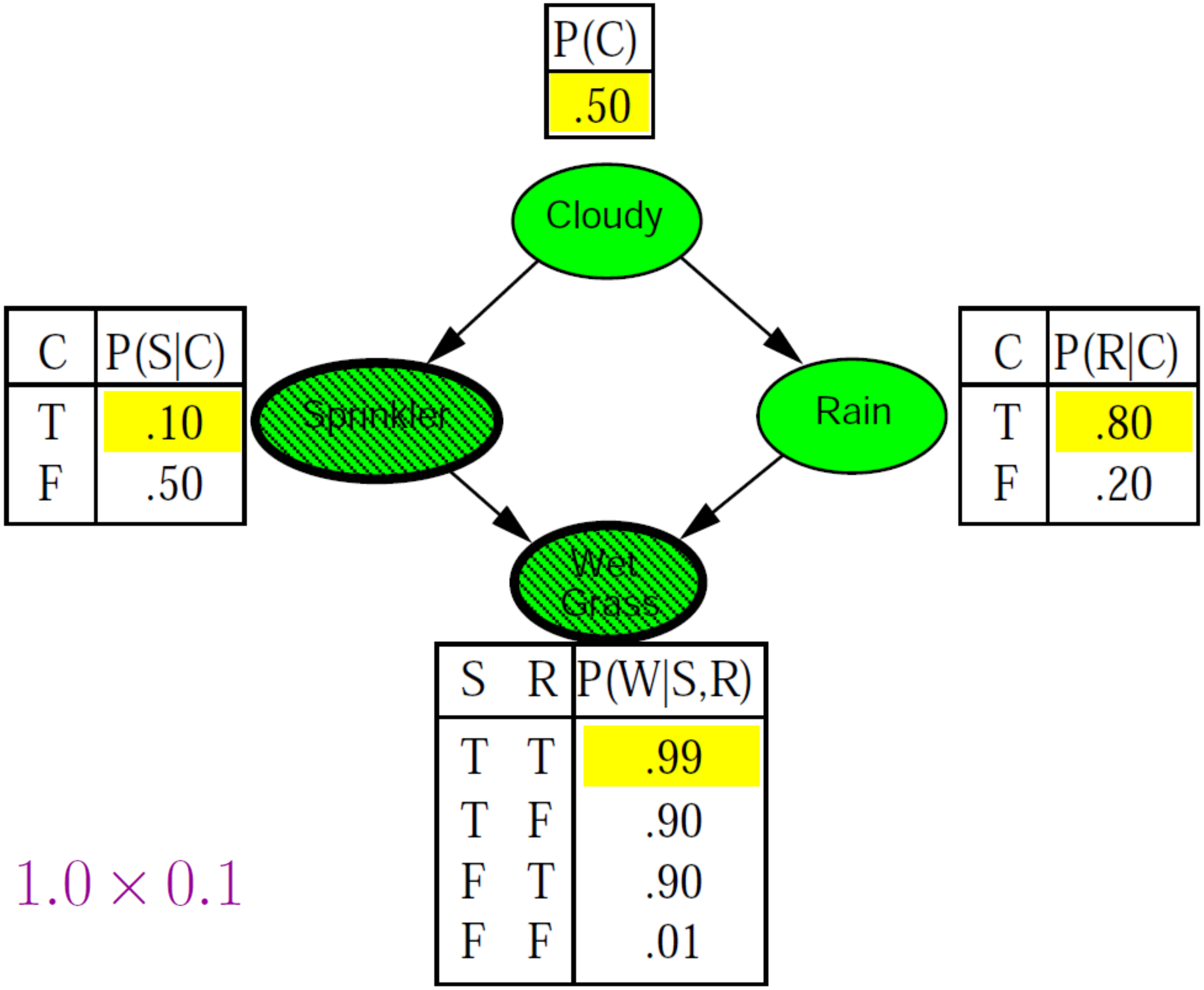
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0 \times 0.1$$

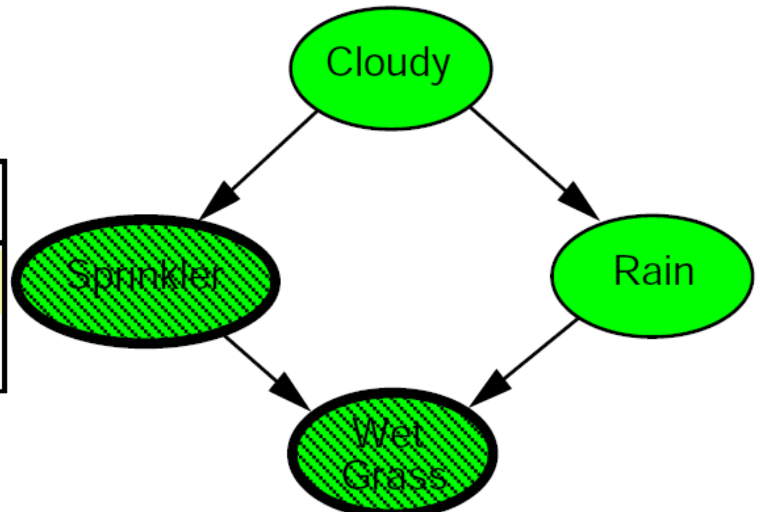
似然加权示例



$$w = 1.0 \times 0.1$$

似然加权示例

P(C)
.50



C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0 \times 0.1 \times 0.99$$
$$= 0.099$$

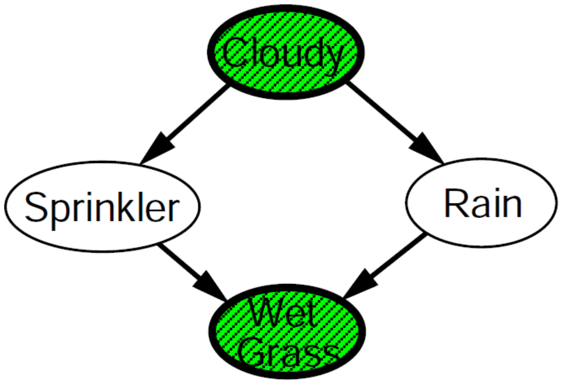
似然加权分析

- WeightedSample 的采样概率为

$$S_{WS}(\mathbf{z},\mathbf{e}) = \prod_{i=1}^{l} P(z_i|parents(Z_i))$$

注意：这里只关注了 ancestors 中的证据变量

➔ 存在先验和后验分布之间的情况



- 给定样本 **z, e** 的权重为：

$$w(\mathbf{z},\mathbf{e}) = \prod_{i=1}^{m} P(e_i|parents(E_i))$$

- 加权的采样概率是

$$\begin{aligned} & S_{WS}(\mathbf{z},\mathbf{e})w(\mathbf{z},\mathbf{e}) \\ &= \prod_{i=1}^{l} P(z_i|parents(Z_i)) \ \prod_{i=1}^{m} P(e_i|parents(E_i)) \\ &= P(\mathbf{z},\mathbf{e}) \text{ (by standard global semantics of network)} \end{aligned}$$

- 似然加权采样获得了一致的概率估计，但随着证据变量的增多其性能仍会急剧下降
  - 只有很少的样本具有近似完整的权重

通过 MCMC 进行近似推理

- 网络状态 = 对所有变量的当前赋值
  - 给定 Markov 覆盖，通过采样一个变量来产生下一个状态——Gibbs采样
  - 证据变量保持不变，轮流采样其他的每个变量

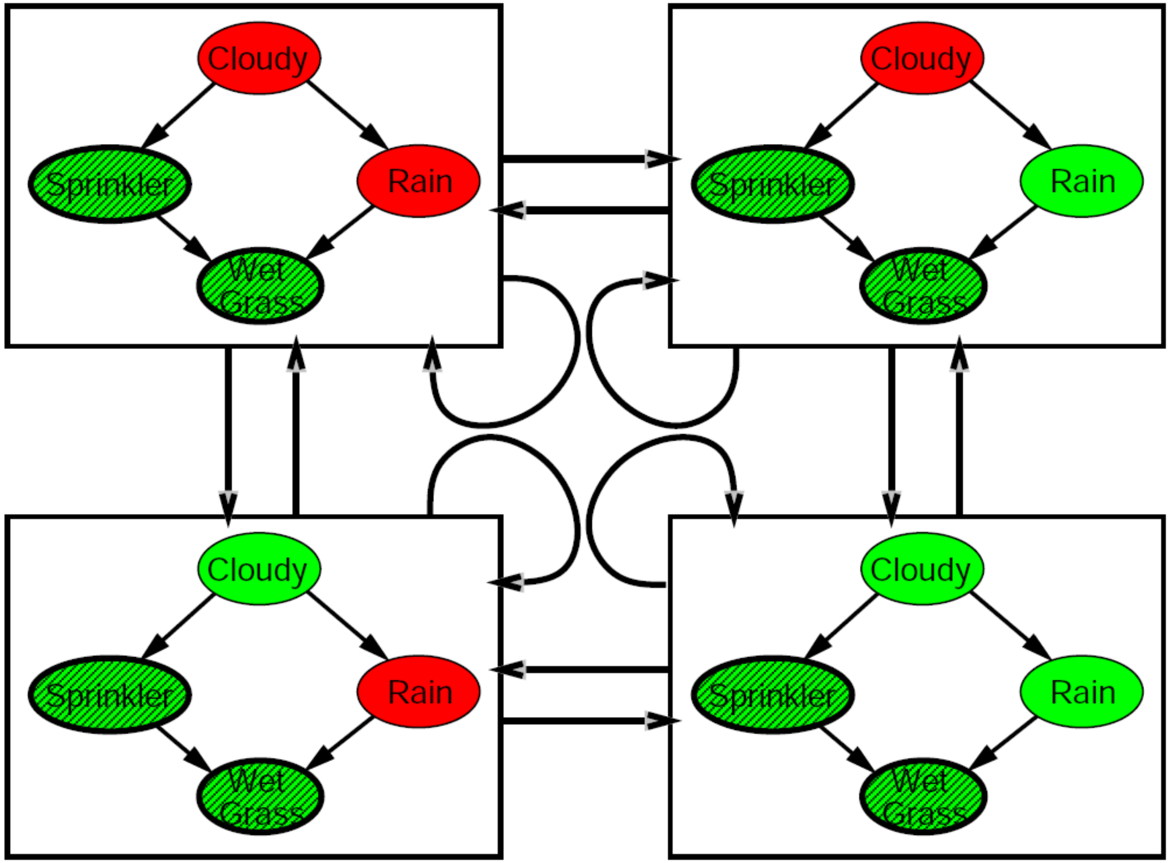
```
function MCMC-ASK(X, e, bn, N) returns an estimate of P(X|e)
    local variables: N[X], a vector of counts over X, initially zero
                     Z, the nonevidence variables in bn
                     x, the current state of the network, initially copied from e

    initialize x with random values for the variables in Y
    for j = 1 to N do
        for each Z_i in Z do
            sample the value of Z_i in x from P(Z_i|mb(Z_i))
                given the values of MB(Z_i) in x
            N[x] ← N[x] + 1 where x is the value of X in x
    return NORMALIZE(N[X])
```

使用 markov 覆盖
计算采样概率

- 也可以每次随机选择一个变量进行采样 （均匀分布）

Markov 链

❖ 针对 $Sprinkler = true, WetGrass = true$，有四种状态



MCMC 示例

- 估计 $\mathbf{P}(Rain|Sprinkler=true, WetGrass=true)$
  - 给定 *Cloudy* 和 *Rain* 的 Markov覆盖，重复采样
  - 统计 *Rain* 为 *true* 和 *false* 的次数

- E.g., 总共访问了 100 个状态
  - 31 个是 *Rain* = *true* 的，69 个是 *Rain* = *false* 的

$$\hat{\mathbf{P}}(Rain|Sprinkler=true, WetGrass=true) = \text{NORMALIZE}(\langle 31, 69 \rangle) = \langle 0.31, 0.69 \rangle$$

- 定理（链稳态分布）：长期来看，在每个状态上所消耗时间与其后验概率成正比

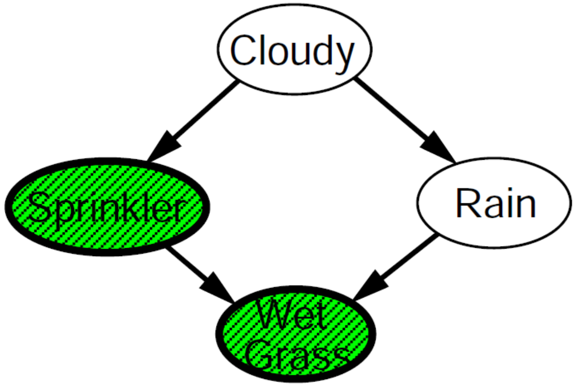
Markov 覆盖采样

- *Cloudy* 的 Markov 覆盖是

  Spinkler 和 *Rain*

- *Rain* 的 Markov 覆盖是

  Cloudy, *Spinkler* 和 *WetGrass*



- 给定 Markov 覆盖下，概率计算公式为：

  $$P(x_i'|mb(X_i)) = P(x_i'|parents(X_i))\prod_{Z_j\in Children(X_i)} P(z_j|parents(Z_j))$$

  - 在消息并行传递系统中是容易实现的

- 主要计算问题：
  - 判断是否已经收敛是困难的
  - 如果 Markov 覆盖较大，计算可能是浪费的，因为 $P(X_i|mb(X_i))$ 不会变化太大

总结

- 通过变量消元进行精确推理
  - 在多形树上是多项式时间的，在一般图上是 NP-hard
  - 空间复杂度与时间复杂度相当，对拓扑很敏感
- 通过 LW, MCMC 进行近似推理
  - 当存在大量下游证据变量时，LW 是低效的
  - LW, MCMC 通常对拓扑是不敏感的
  - 当概率接近于 1 或 0 时，收敛速度将很慢
  - 能够处理离散变量和连续变量的任意组合

谢谢聆听！