Question 1 (In class)

a. ii,iii are true.

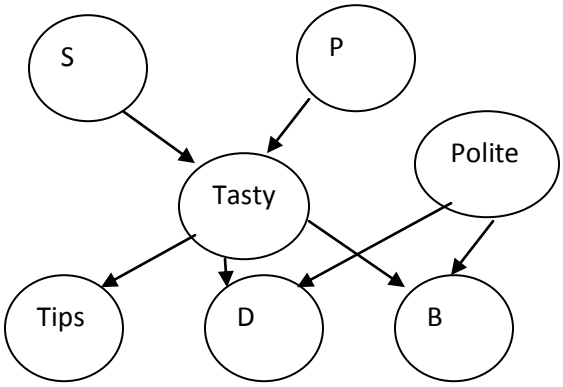
b. **Probability entries in the CPT for this network:**
For each CPT we sum up the number of rows in each table which is $2^1 + 2^1 + 2^3 + 2^4 + 2^3$ =2+2+8+16+8 = 36.

c. **Marginal independence assertions:**
$P \perp S$

d. **Conditional independence assertions:**
$T \perp P|D, B$ and $T \perp S|D, B$

e. $\hat{P}(pepper, salt) = \hat{P}(pepper)\hat{P}(salt)$ are equal when $P \perp S$ (or peper and salt are independent) where $\hat{P}$ represents empirical counts.

f.



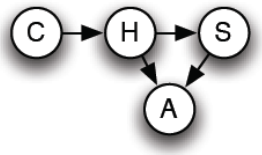
g. Yes, $P(p, \bar{s}|\bar{t}) = \frac{P(p,\bar{s},\bar{t})}{P(\bar{t})} = \frac{P(\bar{t}|p,\bar{s})P(p)P(\bar{s})}{P(\bar{t})} = 0.4 * 0.4 * \frac{0.5}{0.475} = 0.211$

h. Since, $P(Tasty = true|pepper, salt) = \frac{P(Tasty\ )P(pepper\ ,salt\ |Tasty\ )}{P(pepper\ ,salt\ )}$, if we sum over pepper,salt $\sum_{pepper\ ,salt} P(Tasty = true|pepper, salt)\, P(pepper, salt) = 0.25 * (0.8 + 0.6 + 0.6 + 0.1 = .525$, since $Ppepper,salt$ is constant (0.25).

i. Always use either pepper or salt but not the other to conserve them while maximizing taste. This changes Bayes' Net because it is no longer the case that $P \perp S$.

Question 1 (Homework)

1.



2. $$P(A|C=c,S=s) = \alpha \sum_h P(A,h,c,s) = \alpha \sum_h P(c)P(h|c)P(s|h)P(A|h,s);$$

$$\alpha = \frac{1}{\sum_a \sum_h P(c)P(h|c)P(s|h)P(a|h,s)}$$

A	$\frac{1}{\alpha}P(A\|c,s)$	$P(A\|c,s)$
a	$0.3 \cdot 0.6 \cdot 0.9 \cdot 0.01 + 0.3 \cdot 0.1 \cdot 0.7 \cdot 0.5 + 0.3 \cdot 0.3 \cdot 0.3 \cdot 0.2$	0.08
$\neg a$	$0.3 \cdot 0.6 \cdot 0.9 \cdot 0.99 + 0.3 \cdot 0.1 \cdot 0.7 \cdot 0.5 + 0.3 \cdot 0.3 \cdot 0.3 \cdot 0.8$	0.92

3. $P(A|c,s) \approx \frac{1}{5}$ by counting

4.

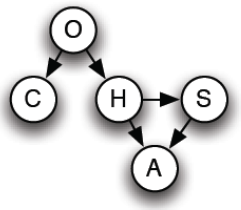
Sample	Weight $= P(c) \cdot P(s\|h)$
$c, f, s, \neg a$	$0.3 \cdot 0.9$
$c, f, s, \neg a$	$0.3 \cdot 0.9$
$c, f, s, \neg a$	$0.3 \cdot 0.9$
c, p, s, a	$0.3 \cdot 0.3$
$c, v, s, \neg a$	$0.3 \cdot 0.7$

, so $P(A|c,s) \approx \frac{0.09}{1.11} = 0.08$

5. The wording of the question was ambiguous, and we apologize for this. What we really wanted to ask is: "Assuming that there had been no rejection for rejection sampling, which sampling process would have been more likely to produce those samples: likelihood weighting or rejection sampling?". The answer is then that likelihood weighting is more likely since it is highly unlikely that rejection sampling wouldn't reject anything (and this is why it is usually more efficient to use likelihood weighting). This is all you were expected to know, and the rest of the solution is only for the curious reader, you won't be responsible for this for the final!

[Optional] The present version could be interpreted as: which sampling scheme is more likely to have produced the given samples: likelihood weighting or rejection sampling (looking at the samples *after* rejection, and so it could be the case that a lot of samples have already been rejected to obtain the one listed). To answer this question, you have to look at what is the probability of seeing a specific sample under each sampling scheme. Under rejection sampling, the probability of a sample is the posterior $P(H, A|C = collar, S = sleepy)$. This is why any table of probabilities that you estimate from these samples will be consistent with the conditional $P(H, A|C = collar, S = sleepy)$. On the other hand, since you don't sample the evidence in likelihood weighting, the probability of a specific sample is $P(H|S) * P(A|H, S)$ (which is quite different from $P(H, A|C = collar, S = sleepy)$ and this is why you need to *weight* those samples to get the right probability estimate). Since each sample is obtained independent from the others, the likelihood of multiple samples will just be the product of each individual probability for the sample (to be totally rigorous, one could worry about having to sum over all possible ordering for the samples, but this only multiply by the same constant for both schemes and is not needed for the comparison). So the probability of the samples under rejection sampling is $\prod_i P(H = h_i, A = a_i | C = collar, S = sleepy)$ where the product is taken over all the samples; whereas the probability of the samples under likelihood weighting is $\prod_i P(H = h_i|S = sleepy)P(A = a_i|H = h_i, S = sleepy)$. If you actually compute both of those quantities, you actually get that likelihood weighting gives a higher probability than rejection sampling, and so that likelihood weighting was more likely to product the sample. But this was not obvious just by looking at the samples.

6.



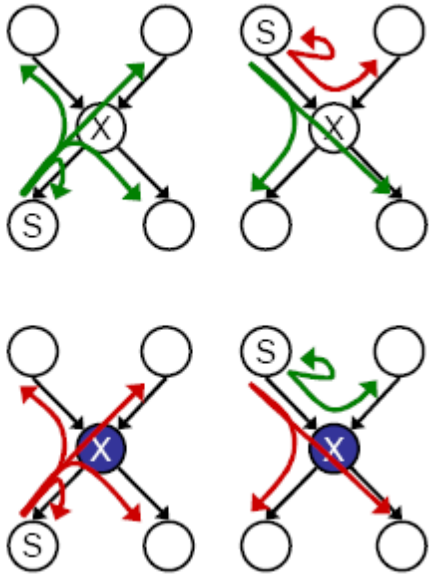
, where O indicates whether the cat has an *owner*.

Question 2 (Homework)

a. A, C, E, F, G
b. A, C, D, E, F
c. A,B,E
d. A,D

There were several errors on this question. Please review the base cases below for determining conditional independence in a Bayes' net.

- Correct algorithm:
  - Shade in evidence
  - Start at source node
  - Try to reach target by search
  - States: pair of (node X, previous state S)
  - Successor function:
    - X unobserved:
      - To any child
      - To any parent if coming from a child
    - X observed:
      - From parent to parent
  - If you can't reach a node, it's conditionally independent of the start node given evidence



Question 3 (Homework)

a. Under G, $P(x|y) = \alpha \sum_z P(x,y,z) = \alpha \sum_z P(x)P(y|x)P(z|y) = \alpha P(x)P(y|x) \sum_z P(z|y)$. Since $P(Z|y)$ is a distirbution over Z, $\sum_z P(z|y) = 1$. So, $P(x|y) = \alpha P(x)P(y|x)$.

Under G', $P(x|y) = \alpha P(x,y) = \alpha P(x)P(y|x)$, which is the same as under G.

In both calculations, α is the normalizing factor $\frac{1}{\sum_x P(x)P(y|x)}$.

b. $$P(q_1, ..., q_k|e_1, ..., e_m) = \frac{\sum_{h_1} ... \sum_{h_p} P(q_1, ..., q_k, e_1, ..., e_m, h_1, ..., h_p)}{\sum_{h_1} ... \sum_{h_p} \sum_{q_1} ... \sum_{q_k} P(q_1, ..., q_k, e_1, ..., e_m, h_1, ..., h_p)}$$

c. Let $\pi(n)$ denote the parents of node n in the Bayes net. Then, we can rewrite

$$P(q_1, ..., q_k, e_1, ..., e_m, h_1, ..., h_p) = \prod_{i=1}^{k} P(q_i|\pi(q_i)) \prod_{j=1}^{m} P(e_j|\pi(e_i)) \prod_{l=1}^{k} P(h_l|\pi(h_l))$$

which just follows from the semantics of Bayes nets. Notice that since h_1 is a leaf, none of the factors above contains h_1 except $P(h_1|\pi(h_1))$. So, we know that

$$\begin{aligned}
P(q_1, ..., q_k|e_1, ..., e_m) &= \alpha \sum_{h_1} ... \sum_{h_p} P(q_1, ..., q_k, e_1, ..., e_m, h_1, ..., h_p) \\
&= \alpha \sum_{h_1} ... \sum_{h_p} \prod_{i=1}^{k} P(q_i|\pi(q_i)) \prod_{j=1}^{m} P(e_j|\pi(e_i)) \prod_{l=1}^{k} P(h_l|\pi(h_l)) \\
&= \alpha \sum_{h_2} ... \sum_{h_p} \prod_{i=1}^{k} P(q_i|\pi(q_i)) \prod_{j=1}^{m} P(e_j|\pi(e_i)) \prod_{l=2}^{k} P(h_l|\pi(h_l)) \sum_{h_1} P(h_1|\pi(h_1)) \\
&= \alpha \sum_{h_1} ... \sum_{h_p} \prod_{i=1}^{k} P(q_i|\pi(q_i)) \prod_{j=1}^{m} P(e_j|\pi(e_i)) \prod_{l=2}^{k} P(h_l|\pi(h_l))
\end{aligned}$$

Where the last step follows from the fact that $\sum_{h_1} P(h_1|\pi(h_1)) = 1$, much like in part (1). Note that the first line is just a restatement of part (2), where α is the normalizing factor (denominator).

d. If we successively remove hidden leaf nodes from the graph (which must be dangling), we will leave $P(q_1, ..., q_k|e_1, ..., e_m)$ unchanged by part (3). Let d be a dangling node. Then,n will eventually be pruned because all its descendants are also dangling. Before n is pruned, there will always be at least one dangling leaf because the graph is acyclic.