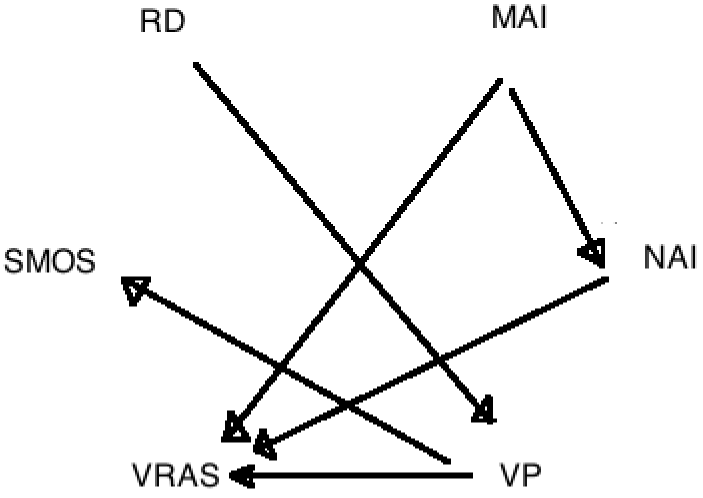
# CS181 Assignment 4

Ashok Cutkosky and Tony Feng

April 9, 2013

## Problem 1.

(a) According to the graph depicted below, we need to describe the following intermediate probabilities: $P(RD), P(MAI), P(NAI \mid MAI), P(VP \mid RD), P(SMOS \mid VP)$, and $P(VRAS \mid VP)$.



- First, we assume that $P(RD) = 0.2, P(MAI) = 0.5$.
- Next, we assume that $P(VP \mid RD)$ is given by the following table

| | RD = 1 | RD = 0 |
|---|---|---|
| P(VP = 1) | 0.9 | 0.0 |

and $P(SMOS \mid VP)$ is given by the following table

| | VP = 1 | VP = 0 |
|---|---|---|
| P(SMOS = 1) | 0.9 | 0.0 |

and $P(NAI \mid MAI)$ is given by the following table

| | MAI = 1 | MAI = 0 |
|---|---|---|
| P(NAI = 1) | 0.1 | 0.9 |

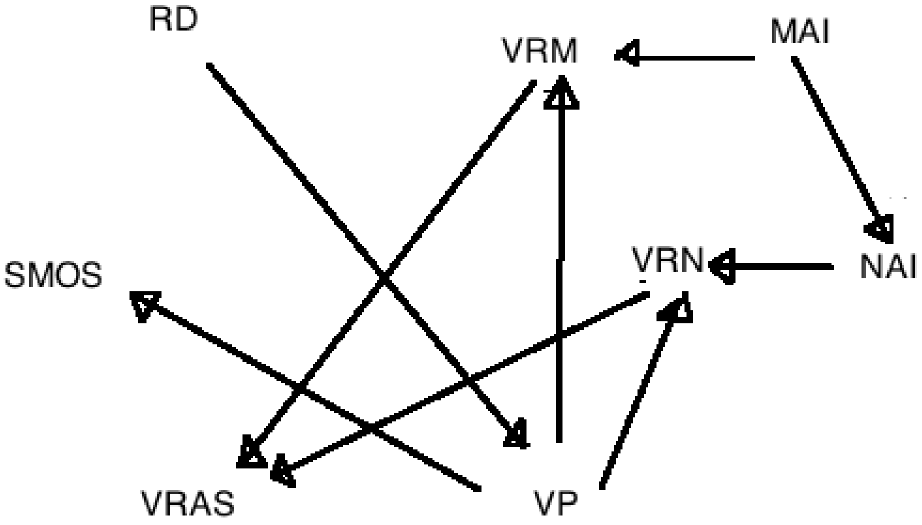
- Finally, we assume that $P(VRAS)$ is given by the following table:

| (VP, MAI, NP) | (0,*,*) | (*, 0, 0) | (1,1,0) or (1,0,1) | (1,1,1) |
|---|---|---|---|---|
| P(VRAS = 1) | 0 | 0 | 0.7 | 0.9 |

(b) Above, we assumed that SMOS (silly messages on screen) was independent of RD (recent download) conditioned on VP (virus present). However, it could be argued that one could have SMOS, e.g. from the ads that tend to inhabit untrusted sites, without actually having a virus. In this case, we should add an edge from RD to SMOS, since we can imagine that RD indicates browsing of the kind of sites that would pop up such silly messages.

We chose not to include this edge for the sake of simplicity, since we expect the majority of silly messages to come from viruses. The problem at hand seems to be that of modeling the virus process, and we can adjust our definition of "silly messages" to be pretty sure that they correspond to viruses.

(c) Here is our new graph.



The things to change are $P(VRM \mid MAI, VP), P(VRN \mid NAI, VP)$, and $P(VRAS \mid ...)$. We are modifying our parametrization slightly so that $P(VRM \mid MAI, VP) = 0.7$ and $P(VRN \mid NAI, VP) = 0.75$, and $VRAS = 1$ if either $VRM$ or $VRN = 1$.

(d) Well, introducing these nodes allows greater flexibility of expression in the modeling process because it allows us to separate the dependencies on the two different antivirus programs. For instance, the graph lets us think about the probability that McAfee detects a virus and the probability that Norton detects a virus separately.

Whether or not this is beneficial to the modeling process depends on our specific goals and situation. It is conceivable that this extra complexity could lead us towards overfitting. Note also that it could already be captured by the parameters of the simpler model, so one could argue that there is not, strictly speaking, more information in this second model.

## Problem 2

(a) For graph (a), there are no such variables. For graph (b), the independent variables are $F, C$.

(b) For graph (a), $P(A, B, C, D, E, F, G, H, I)$ is

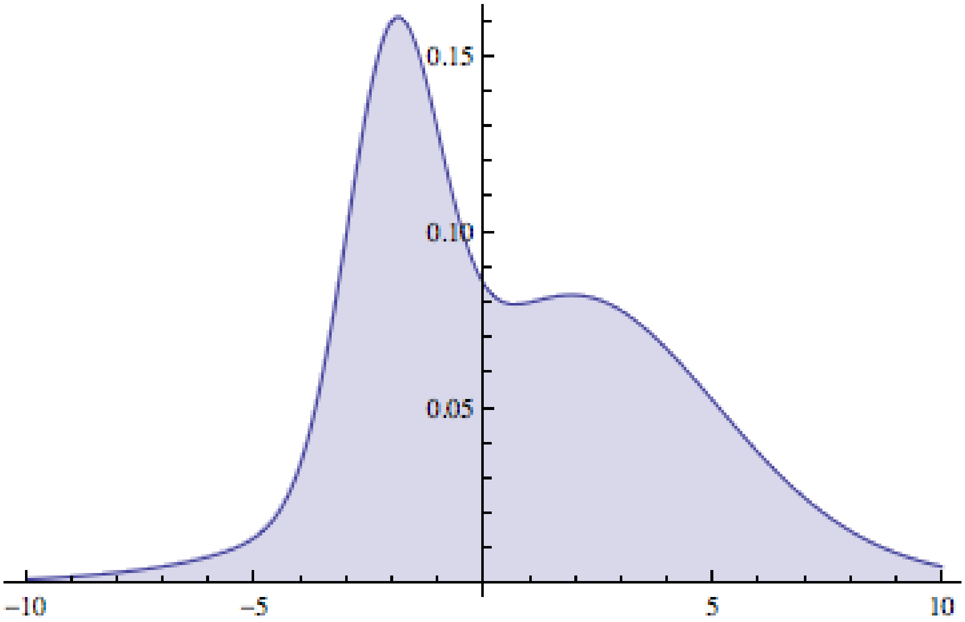
$$P(G)P(H)P(I \mid G, H)P(D \mid G)P(E \mid G)P(F \mid H)P(B \mid D)P(C \mid E, F)P(A \mid B, C).$$

For graph (b), $P(A, B, C, D, E, F, G, H, I)$ is

$$P(B)P(C)P(A)P(D \mid B)P(E \mid B)P(F \mid C)P(G \mid A, D)P(H \mid D, E)P(I \mid E, F)P(J \mid G).$$

## Problem 3

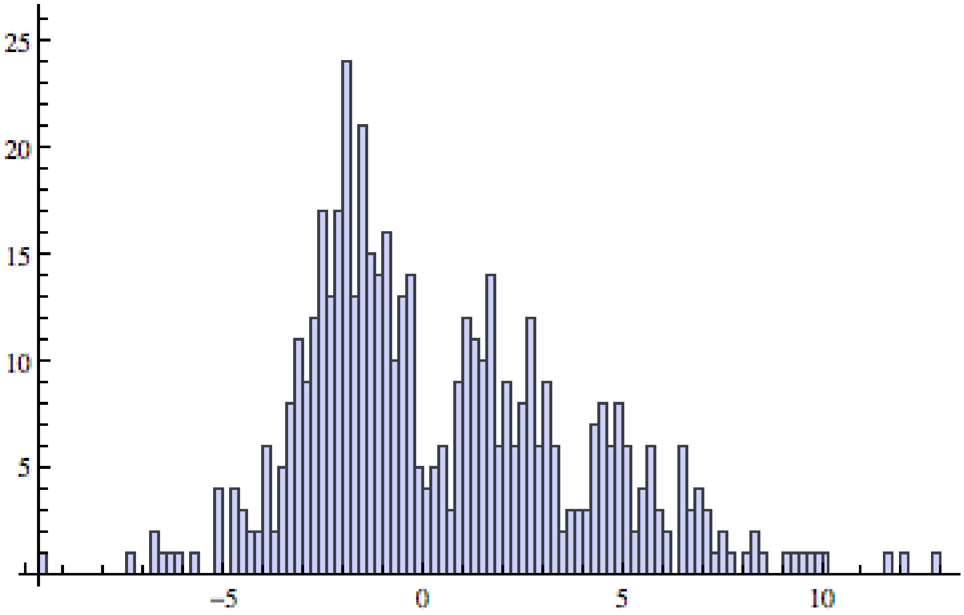
(a)



(b) We draw in two steps.

1. First draw randomly from the discrete distribution to determine from which Gaussian to draw: with probability 0.2 we draw from the first, with probability 0.3 we draw from the second, and with probability 0.5 we draw from the third.
2. Second draw from the Gaussian distribution determined in Step 1.

Using this algorithm, we produced the following histogram.


25
20
15
10
5
−5
0
5
10