

Homework 3

A Probabilistic Model of Information Flow

Hypothesis:

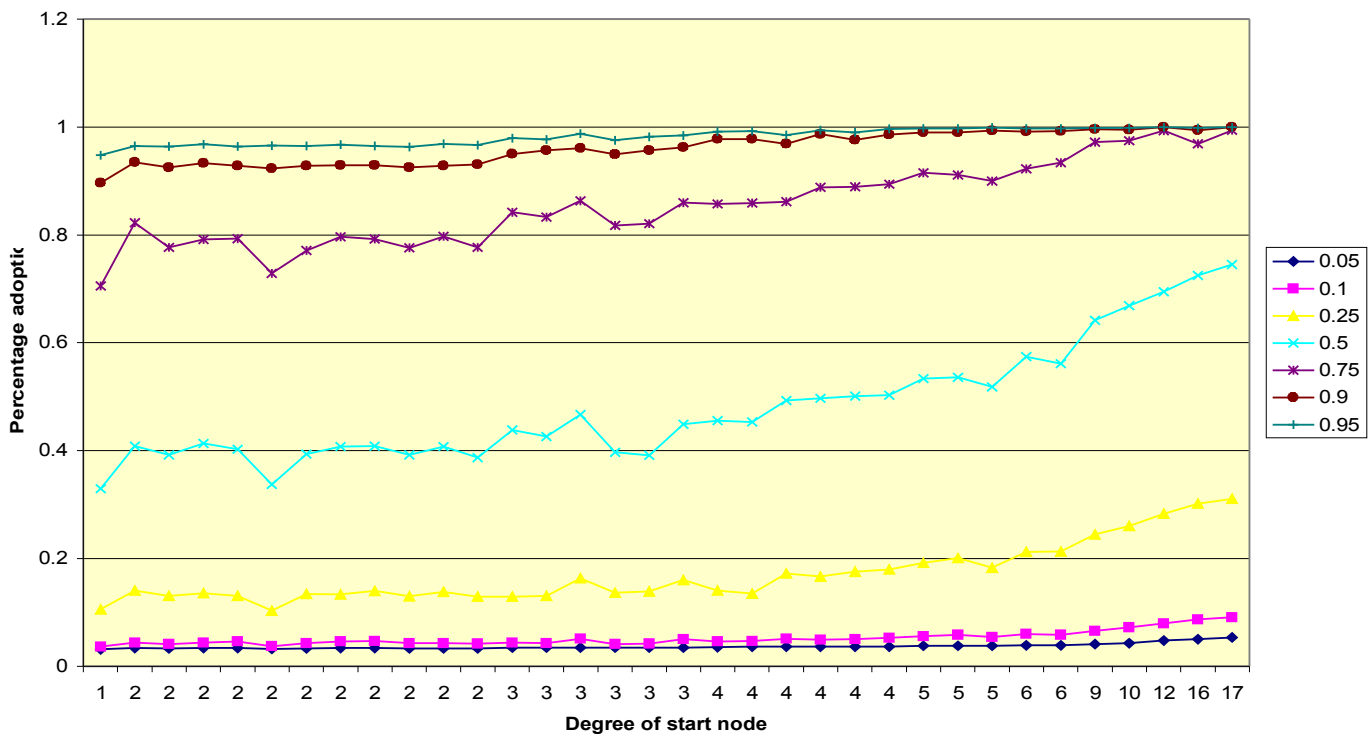
1. For a given user, the percentage of people who will adopt the product is directly correlated to the degree of the user.
2. For a given user, the expected reach is three.

All graphs shown are taken from the Karate Network.

To investigate how the start user's degree might affect the adoption percentage, I found the degree of each node from the input data. I plotted the probability against the percent adoption, creating a separate color line for each possible start degree. This graph will show how changing the degree will affect the percent adoption over different probabilities.

The possible probabilities are $\{0.05, 0.10, 0.25, 0.50, 0.75, 0.9, 0.95\}$. The average number of nodes was taken from 20,000 trials for each pair of (probability, start node).

Degree vs. adoption



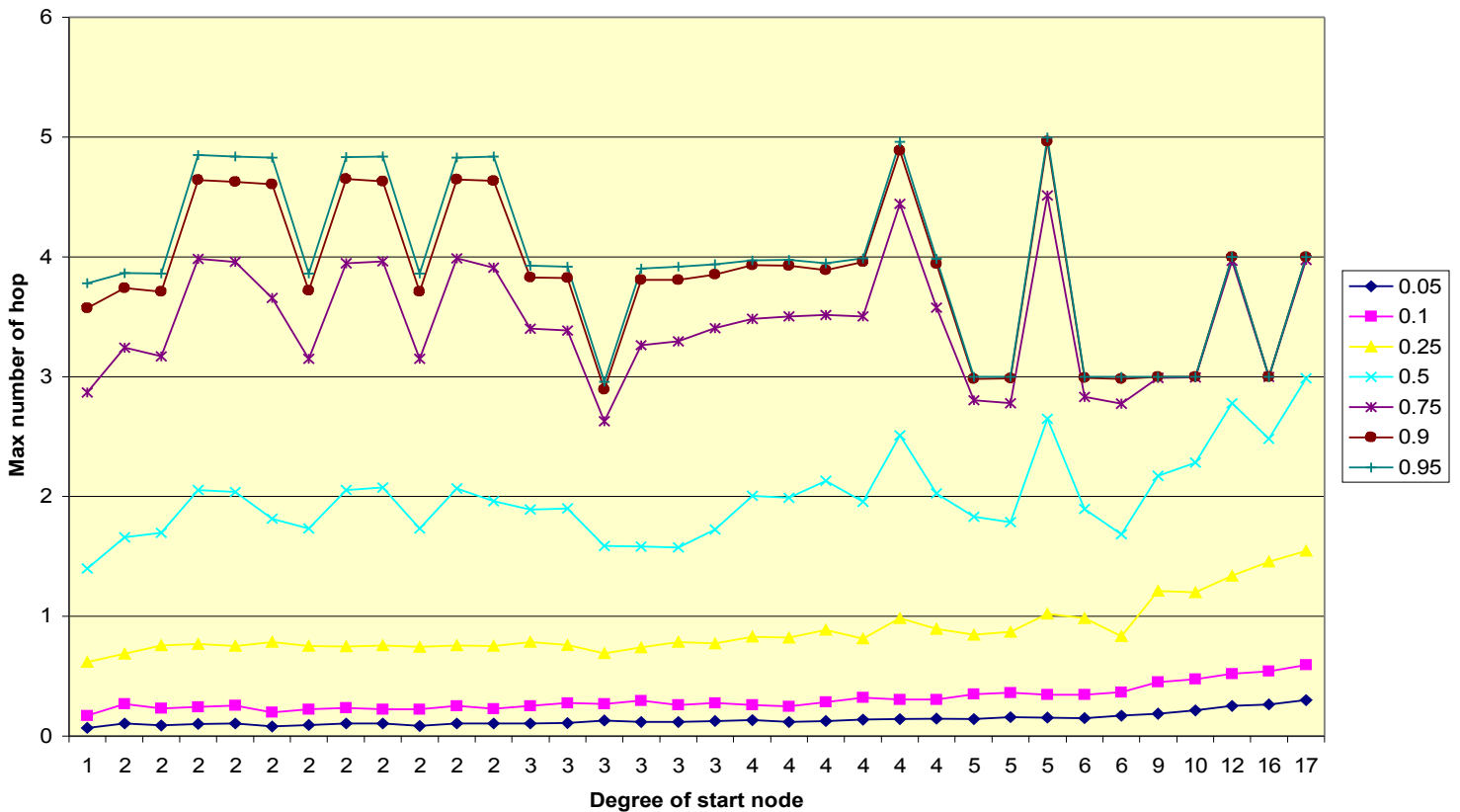
One possible problem with this particular graph is that there are different values along the x-axis for the same start degree. However, this is because they represent different start nodes. The fact that these data points are not identical indicates that start degree is not the only important factor; it also matters where that node is in relation to the rest of the graph.

The lines virtually never overlap in these examples; it is possible that they never overlap. From this chart, it seems that the probability of each node's adoption is much more important than the start node's degree. Even nodes with high degree do not typically spread a new idea very far if the probability of adoption is as low as 0.1. Similarly, even nodes that have degree 1 or 2 will typically spread an idea to most of the rest of the network as long as the probability of adoption is high like 0.9 or 0.95.

Each line trends upwards, but not by very much. This graph supports the hypothesis that higher degree nodes will spread the idea to a greater percentage of the network. However, it is not very strong support, since it seems like the effect of the degree can be dwarfed by other factors.

Examining the lines plotted on a graph is an imperfect technique. In future experiments, it would be better to plot many more probabilities. It would also be helpful to run comparisons for every probability, to figure out whether there are any instances where the lines cross.

Start degree vs. hops away



In retrospect, the hypothesis about the average number of hops away an idea would travel was ill-thought out. A better hypothesis might be the average distance between any two nodes in a graph; or one or two hops less than the maximum number of hops. In this graph, though, an idea does often travel three hops away because of the size of the social network.

The same caveats apply from the previous analysis; the fact that these nodes are identified only by their degree, and not by other statistics about them, makes this a flawed way to examine the data.

In the previous graph, differences between nodes with the same degree were minimal. However, this graph has obvious spikes in some nodes with degree 2, degree 3, or degree 5. These spikes indicate that there is another factor that has much more influence than the degree of the start node; note that some start nodes with degree 2 can spread an idea further, on average, than some start nodes with degree 4.

Like the previous graph, the lines never seem to intersect. This shows that while all other factors about a node are held equal, the probability of adoption can have a strong effect. However, with extreme lows like that in a particular node of degree 3, whatever factors make that node a poor choice for starting a trend are in effect regardless of the probability.

A Model of Peer-Pressure

It didn't seem meaningful to plot the same types of graphs for this model as for the probabilistic model, since there wouldn't be an average to calculate.

For the karate linear network, the optimal initial user is 0, the optimal set of two initial users is $\{0, 32\}$, and the optimal set of three initial users is $\{0, 1, 32\}$.

The optimal set of n nodes seems to include the optimal set of $n - 1$ nodes. However, this may depend on the network configuration. Since I wasn't sure if I could prove this to be true, my code tries all $n!$ combinations to find the optimal start set.

Given a network, I would define $b_{i,j}$ as a ratio of j 's degree to i 's degree: $\frac{d_i}{d_i + d_j}$. The intuition is that someone will often trust the opinion of a friend that he sees as being well-connected. The same idea could be extended to a more complicated definition, weighing the number of friends that j has, and their degrees. However, I think a simple definition that only considers information about i and j is best; it will be the most consistent when applied to graphs of different sizes.

Program usage notes

ProbModelDriver.java must be compiled with ProbGraph.java.
DetermModelDriver.java must be compiled with DetGraph.java.