The "Viral Marketing" Effect

Andrew Resnikoff

Introduction

Background

In the 1950's, the antibiotic tetracycline started being used by group of doctors in Illinois. Through this analysis, we will explore the "viral marketing" effect by examining how this antibiotic managed to spread among this group of doctors. In this investigation, we hope to show how exactly the drug spread and whether it was related to personal contact between these doctors. We hypothesize that the social network between doctors will have a positive affect on the rate at which tetracycline is adopted.

Data

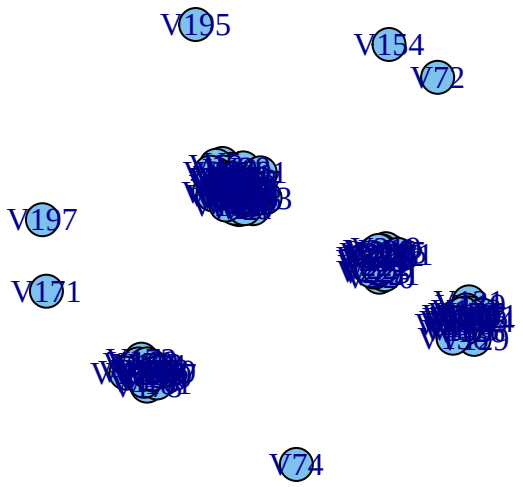
We are given two data sets, the first being a data frame we will call "nodes" and the other being a matrix we will call "network." The nodes data frame initially contains 246 observations of the following variables:

```
##  [1] "city"                          "adoption_date"
##  [3] "medical_school"                "attend_meetings"
##  [5] "medical_journals"              "free_time_with"
##  [7] "discuss_medicine_socially"     "club_with_drs"
##  [9] "drs_among_three_best_friends" "practicing_here"
## [11] "office_visits_per_week"        "proximity_to_other_drs"
## [13] "specialty"
```

The entries in the 246×246 binary matrix "network" represent whether doctor i and j know each other (where i and j are the row and column respectively.)

Since we can not analyze the viral marketing affect on doctors who we do not know when or if they began using it, we exclude these observations from both the "nodes" data frame and the "network" matrix. This leaves us with 125 usable observations.

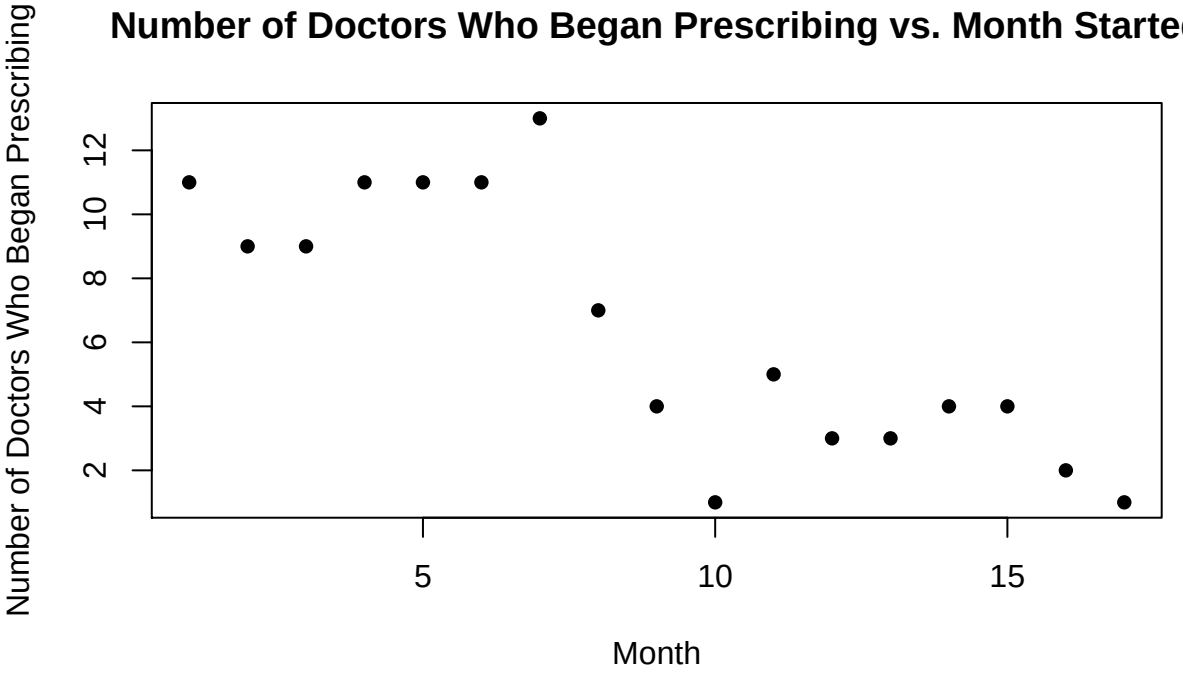
We can take a preliminary look at what the doctors network looks like in the figure below. We see that there are 4 separate networks and then a few unique nodes that don't belong to any networks. This graph tells us that most people are connected to a social network in some way, so we do expect to see the "viral marketing" effect for most of the observations.



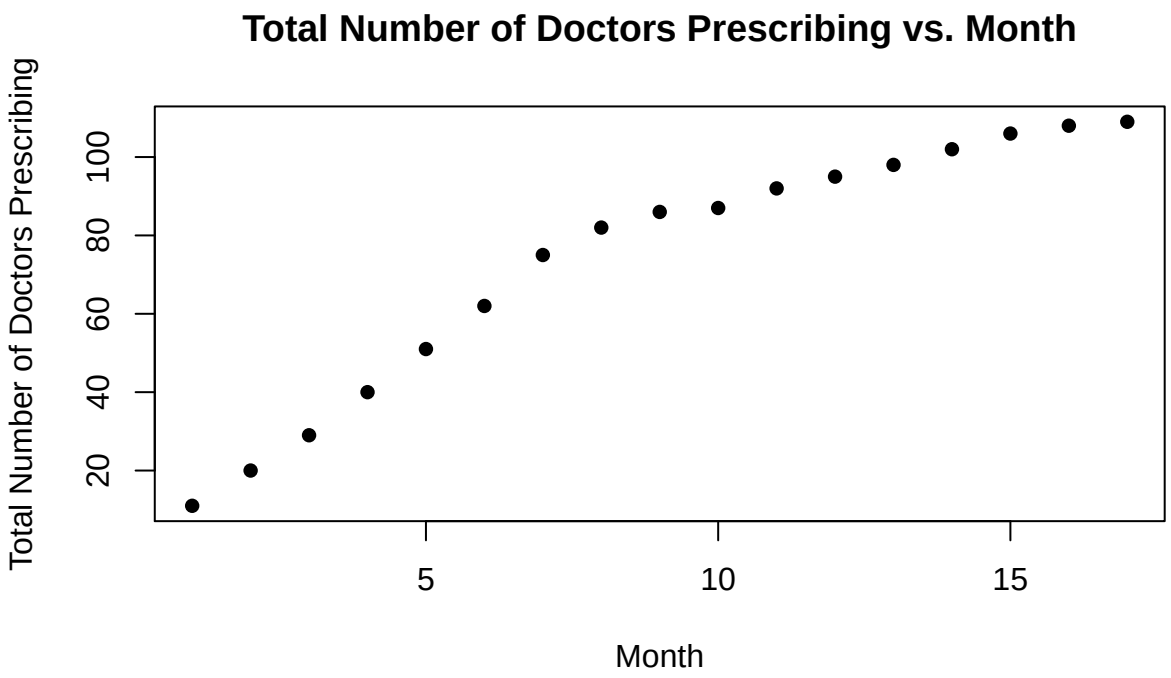
Data Analysis

Plot Tetracycline Adoption Over Time

We begin the analysis by looking at which months doctors began prescribing tetracycline and how many total doctors were prescribing tetracycline after each month.



This first plot shows the number of doctors who began prescribing tetracycline in each month. Notice that in months 1 to 7 many more doctors (around 10) begin prescribing tetracycline than in the later months (around 3).



This second plot, which shows the total number of doctors who were prescribing tetracycline by each month, confirms what we saw in the first graph. There is a steady increase in the total number of doctors prescribing tetracycline until about month 7. At this point, the total number of doctors prescribing tetracycline begins to level off and increases at a much lower rate.

Estimate Adoption Probabilities

The next step in our analysis involves finding the estimated probabilities that a doctor who has not yet adopted the drug will begin to do so in the next month. We can estimate these probabilities in two different ways. The first way involves estimating based on the total number of people who have adopted tetracycline before this month. The second way involves estimating based on the number of contacts a doctor has who have already adopted the drug.

Method 1

We begin the first method by looking at the number of people who adopted the drug in a given month, the total number of doctors who adopted the drug before that month and the total number of doctors.

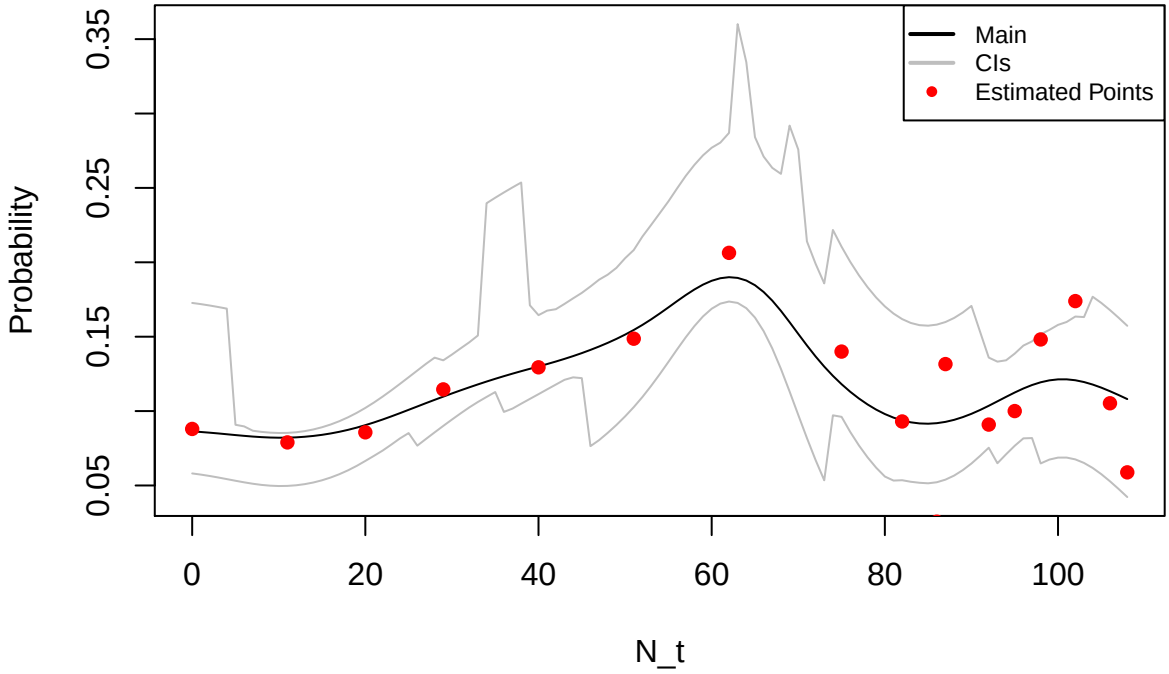
```
# frequency table as vector shows how many doctors adopt each month
this.month<-as.vector(table(nodes$adoption_date)[1:17])
# cumsum ads frequency table to get total who adopted by each month (0 adopted before month 1)
N_t<-c(0,cumsum(this.month))[1:17]
```

Now we need to estimate the probability a doctor will adopt the drug this month, given the number of doctors who have already adopted the drug, N_t. For each N_t, we have the number of doctors who adopted tetracycline that month. Since we are given the total number of doctors, 125, we can estimate the probability a doctor will adopt the drug this month as follows.

$$Pr(\text{Doctor will adopt the drug } | N_t) = \frac{\text{The number of doctors who did adopt the drug } | N_t}{125 - N_t}$$

```
# calculate probabilities
prob<-(this.month/(rep(125,17) - N_t))
```

From this vector, we can estimate the probability density using a kernel regression. The figure below shows our estimated density function, along with the 95% confidence bounds and the estimated probabilities the curve was evaluated from. This curve can be evaluated from 0 to 108, because these are the minimum and maximum values of N_t respectively. We cannot really know for sure what happens beyond these endpoints.



The cross-validated MSE of this kernel regression is 0.00169. This low MSE suggests we have a good fit for our data.

From this estimation, we can calculate the average change in probability as N_t increases.

For a one-unit change in N_t, the average predicted change in probability per doctor per month is 1.02×10^{-5}. The standard error of this change is 1.39×10^{-6}.

Method 2

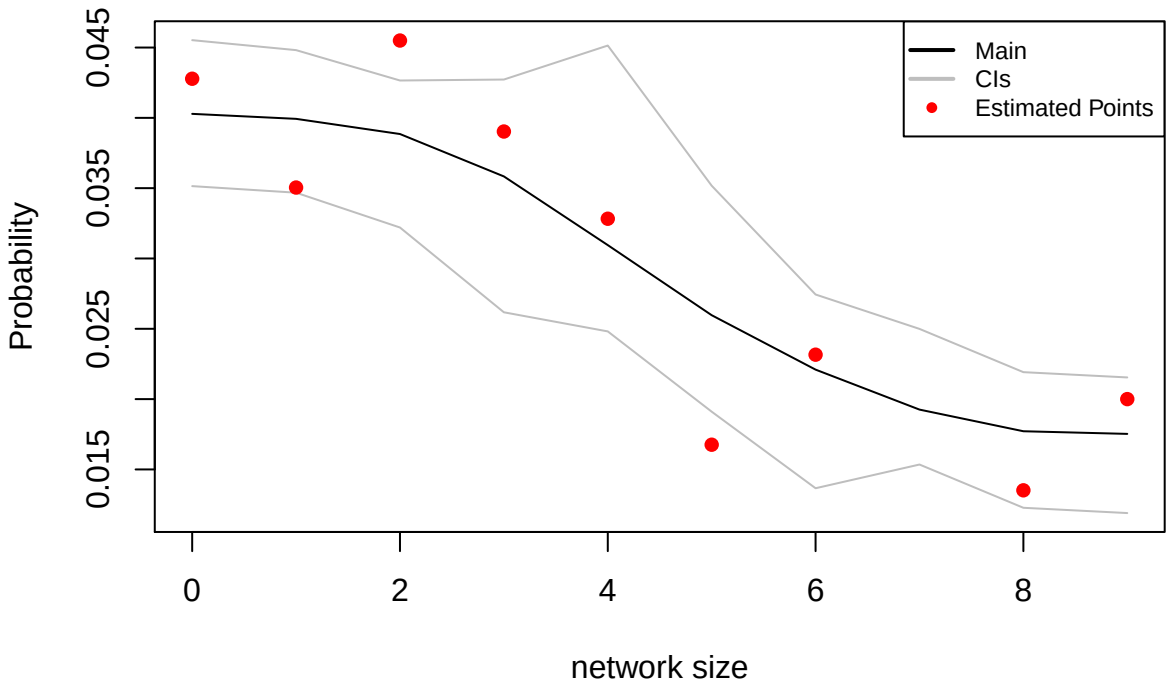
In the second method, we estimate the probability a doctor will adopt the drug this month based on the number of contacts a doctor has who have already adopted the drug. To start, we must first create a data frame recording, for every combination of doctor and month, whether that doctor began prescribing tetracycline that month (`began`), whether that doctor has begun prescribing tetracycline earlier than that month (`before`), and the number of their contacts who began prescribing before that month (`contacts.before`).

To estimate the probability, we need a relationship between the size of a network (of people who are connected to someone who has adopted tetracycline) each month and the number of people who adopted tetracycline each month. We can get the former from our data frame, by going through each doctor and finding the number of contacts they had who had adopted tetracycline before they did. Using this information,

along with the adoption date for each doctor, we can create a matrix (with 17 rows and some amount of columns), where each row represents a month of adoption and each column represents the number of contacts a doctor had who was already prescribing tetracycline. We then add one to each entry $M_{i,j}$ for each doctor who began prescribing in month i and had j contacts who were already prescribing the drug.

Contacts Already Prescribing	Estimated Probability
0	0.043
1	0.035
2	0.045
3	0.039
4	0.033
5	0.017
6	0.023
8	0.014
9	0.020

From these estimated probabilities, we can fit a kernel regression. The figure below shows our estimated density function, along with the 95% confidence bounds and the estimated probabilities the curve was evaluated from. This curve can be evaluated from 0 to 9, because these are the minimum and maximum observed values of the size of a doctor's network already prescribing respectively (when a doctor is not already prescribing). Thus, our data does not give us enough information to accurately predict what happens outside of those endpoints.



The cross-validated MSE of this kernel regression is 4.94×10^{-5}. This low MSE suggests we have a good fit for our data.

Looking at this plot, it may seem like the relationship between network size and probability is a negative one, contrary to what we would expect, but actually, this is not the case. First, it is important to note that mean and median number of contacts a person has are 3.84 and 3 respectively. These probabilities do

not take into account how many contacts a person actually had in total in the estimation. Notice that the estimated probabilities for having 2 or 3 contacts already prescribing the drug are the two highest (other than for having 0, but for this networking effect to exist some people will have to initially try it without it being recommended to them by a doctor in their network.) This suggests that if most of the doctors in a doctors' network are using the drug, the doctor is more likely to start prescribing it as well. Also, the drop off in probability as the number of contacts using prescribing tetracycline also makes sense. If the first 2 or 3 doctors could not convince this doctor to start using tetracycline, this doctor may only start prescribing once almost everyone in their network is using it. And some will never start prescribing it at all, despite their contacts who have started prescribing it.

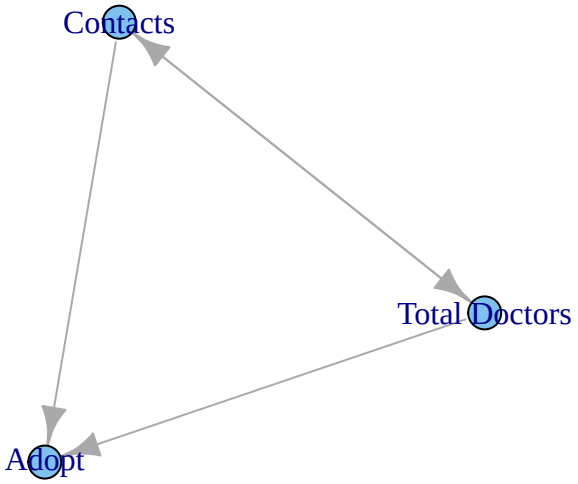
Ultimately, this plot does show support the hypothesized network effect. A curve that does not support the network effect hypothesis might show a constant probability for all network sizes prescribing tetracycline. Since this plot shows that once 2 or 3 of a doctors' friends begin using tetracycline, the doctor becomes more likely to start using it, it seems that these other doctors do have an effect on that doctor.

For a one-unit change in network size prescribing tetracycline, the average predicted change in probability per doctor per month is -1.07×10^{-5}. The standard error of this change is 8.65×10^{-7}.

Estimation Comparison

Looking at our two methods of probability estimation, there is a good amount of consistency between them. The first method shows that over time, there is a gradual buildup to the number of people who begin prescribing. At first, only a few people are willing to try the new drug. Eventually, word of this drug begins spreading through these social networks. Once there are is good portion of doctors using it (suggested to be around 40 in the first method plot), most doctors will have a few friends (between 2 and 4) in their network using it and thus there is a spike in the probability that a doctor will begin using tetracycline. After this spike has passed, most of the people left are doctors who will either never start prescribing tetracycline. There will still be a few doctors who will wait for even more of their friends to start using it, which explains why both plots show the probability decreases after the spike.

Since we see that these probability estimations seem to be deeply related, this implies that these estimation methods are confounding and that the average change we estimated for each method is not quite the causal effect on adoption. For the average change in adoption rate for a one unit change total number of doctors prescribing tetracycline (or the total number of contacts prescribing tetracycline) to be a fair estimate, it would be necessary that the number of contacts prescribing tetracycline be independent of the total number of doctors prescribing tetracycline. This is an unreasonable assumption.



The graph above shows how the total number of doctors prescribing and the total number of contacts prescribing might affect the probability that a doctor might adopt tetracycline. Notice that for us to accurate measure the causal effect for either the total number of doctors prescribing tetracycline or the total number of contacts prescribing tetracycline, we would need these values to be independent of one another, which is most likely not the case. To get an accurate estimation for one of the methods, we would need to control for the other method.

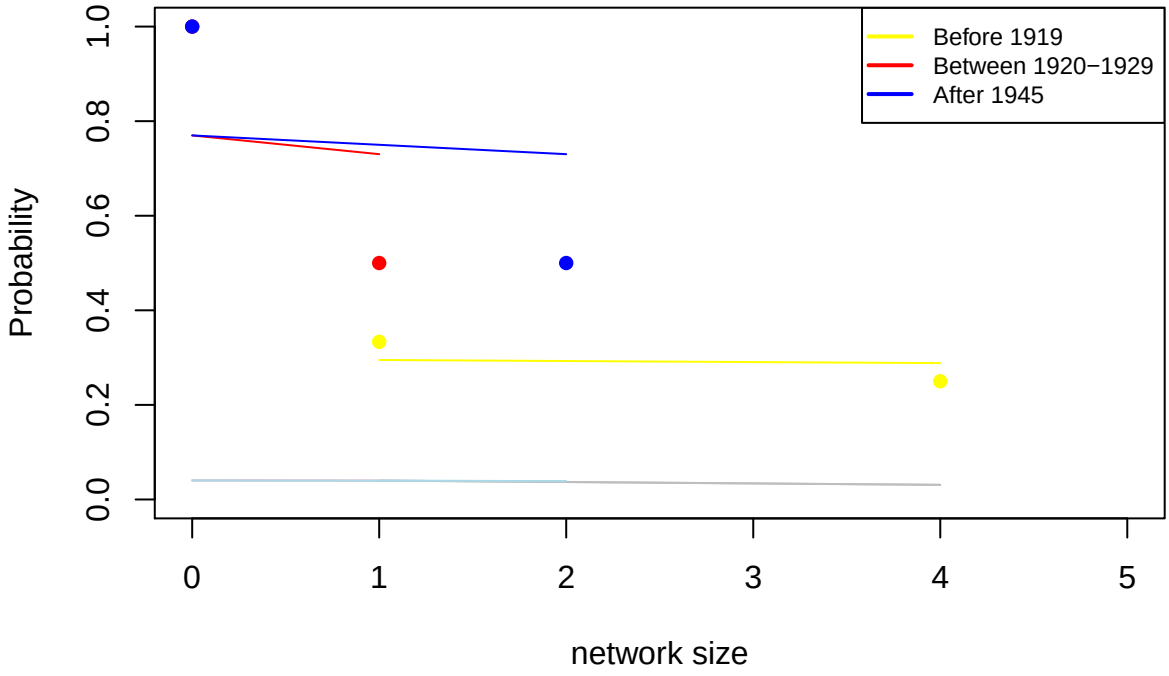
Additional Analysis

We can also explore the effects of some other factors on the probability that a doctor will adopt tetracycline. We can estimate the probability a doctor will adopt tetracycline based on the number of contacts they have who have already begun prescribing the drug, when the doctor graduated medical school, whether the doctor attends medical-society meetings and how many medical journals they read.

We first must isolate the observations of interest. Since the median number of medical journals read is 5, we will say a doctor reads the minimum number of journals if they read less than 5. We can get the observations that read less than 5 medical journals and who don't attend medical-society meetings. From that group, when we can then split the data by the years the doctors graduated medical school.

We now have 3 frames of 4 observations, 2 observations and 2 observations for the graduation periods before 1919, the 1920's and after 1945 respectively. Using, method 2, we can create estimate the probabilities of a doctor adopting tetracycline for each of these frames.

Since we do not have as many observations, our estimated models will not be as complete as they were previously. In each case, we end up fitting a line on 2 points. These lines can be seen, along with 95% confidence bands and the actual estimated probabilities, in the plot below.



It seems inappropriate to try to build and analyze a legitimate predictive model off of so few data points. We can still try to calculate the average change in probability of adoption for a one unit change in the network size.

Before 1919 For a one-unit change in network size prescribing tetracycline, the average predicted change in probability per doctor per month is -3.11×10^{-6}. The standard error of this change is 1.02×10^{-6}.

1920-1929 For a one-unit change in network size prescribing tetracycline, the average predicted change in probability per doctor per month is -1.86×10^{-5}. The standard error of this change is 1.24×10^{-7}.

After 1945 For a one-unit change in network size prescribing tetracycline, the average predicted change in probability per doctor per month is -1.86×10^{-5}. The standard error of this change is 1.24×10^{-7}.

These changes are valid estimates assuming that there are no other confounding factors that would also influence the probability for adoption that are influenced by the same factors that influence how the number of contacts prescribing the drug a doctor has affects the probability of adoption. In other words, assuming that controlling for medical school graduation date, medical journals read and medical-society meeting attendance satisfied the back-door criterion, these would be valid estimates.

Conclusions

Based on our probability estimations, we can see strong evidence for the idea that tetracycline spread through Illinois at least partially due to the social network that existed through doctors. Our first estimation of the probability that a doctor would adopt (based on the total number of doctors who have begun prescribing tetracycline) demonstrated that once about a third of the doctors began prescribing tetracycline, the probability that a doctor would begin prescribing tetracycline had a sharp increase. Then, the probability fell back to close to its initial levels.

This pattern indicates that the spread of the drug influenced other doctors to begin prescribing it. After most the doctors who allowed themselves to be influenced by what the other doctors were doing had all begun prescribing tetracycline, the doctors who held out began adopting with the same probability as if they found it on their own (as in they didn't listen to other doctors and did their own research to conclude that they should begin using the drug) or as if so many other doctors were using it they realized that it was a good idea for them to use it as well.

Moving on to our second estimation, we see similar results. Doctors begin prescribing tetracycline while having no doctors in their network prescribing it with relatively high probability. This makes sense as some people would need to begin the process of "viral marketing." We see that once 2 or 3 of the doctors in a doctor's network have begun prescribing tetracycline, the doctor becomes way more likely to use it. Since the mean and median number of doctors in another doctors network are both around 3, this means that once the doctors in another doctor's circle have adopted tetracycline, that doctor is with relatively high probably going to adopt it too. We also see that the probability decreases as the number of contacts goes up. Doctors who have enough contacts to be able to have more than 5 or 6 doctors in their network and did not begin using it after 2 or 3 of those doctors in their network had were not convinced by the first few doctors in their network to begin using tetracycline. These doctors are less likely to begin prescribing at all. If they do start prescribing, it is probably for a reason unrelated to the social network effect. Or, if basically everyone in their network was using it, perhaps it was more of a peer pressure type situation. Regardless, we see that for most doctors, the social network does play a significant role.

In calculating the estimated effect of a one unit increase in the total number of doctors prescribing tetracycline and the size of a doctor's network prescribing tetracycline, we obtained the following effects averaged over months and doctors.

For a one-unit change in the total number of doctors prescribing tetracycline, the average predicted change in probability per doctor per month is 1.02×10^{-5}. The standard error of this change is 1.39×10^{-6}.

For a one-unit change in network size prescribing tetracycline, the average predicted change in probability per doctor per month is -1.07×10^{-5}. The standard error of this change is 8.65×10^{-7}.

It is interesting to note that these effects are basically of equal magnitude and opposite sign. Even though the negative sign from the effect of the second method may seem counter-intuitive, remember that the probability peaks quickly for small network sizes and then decreases quickly from there. While these estimated values may be confounded by other influential variables, they are the values that we have found in this analysis and can at least serve as a benchmark until the effects have been fully investigated.

In a future analysis, we could potentially improve this investigation by exploring the 4 disjoint networks we saw in the graph of the networks. There is potentially a different effect for each of these networks. Additionally, with some more data, we could explore the effects of more of the predictor variables on the average change.