COS 116
The Computational Universe
Laboratory 11: Virus and Worm Propagation in Networks

You learnt in lecture about computer viruses and worms. In this lab you will study virus propagation at a quantitative level. You will use a simple simulation model to gather data and then use Microsoft Excel to understand the data.

The model simulates a network of interconnected computers reminiscent of the internet but with fewer computers (up to 100,000). When you set all relevant parameters and click "Run Simulation," it simulates the spread of the virus/worm in the network and outputs some numbers for you to analyze. Obviously, there are close analogies to spread of diseases in humans, so this lab gives you some insight into that as well.

Lab submission: Submit by Tuesday, May 1 in lecture. ***Turn in all and only the Excel charts that are followed by a star (*) in the lab.*** **Make sure each chart is labeled with the Experiment number and Question number, as well as any relevant additional information (e.g. particular options). You will not receive credit for charts that are unlabeled.** ***Do not turn in the charts without a star; they are just for you to visualize the simulation.*** **Also, turn in answers to all the questions posed in the body of the lab and in the "Additional Questions" section.**

Introduction: Using the Simulator

1. **Download this file to your Desktop:**
http://www.cs.princeton.edu/courses/archive/spring07/cos116/lab11_files/vsim.exe

2. **Double-click the file to run the simulator. Ignore any security warnings.** The simulator has a variety of options to adjust. These options are explained below. Note that there is an element of randomness in the simulator, especially when using the social network settings, so your results may not look exactly like your neighbor's.

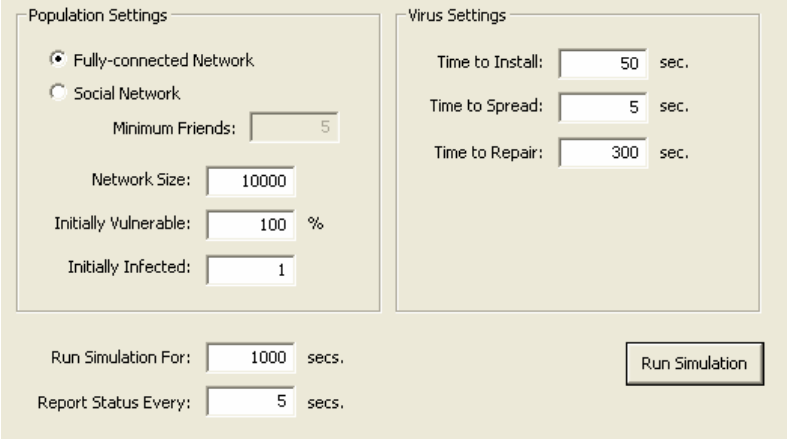

Figure 1: Simulator interface

Here is a brief explanation of all adjustable parameters.

- **Network size**: The number of nodes in the network. This can be interpreted as the number of computers in a computer network or the number of people in a social network.
- **Network type**: In a fully-connected network, every node is directly connected to every other node. In a social network the connections are generated according to a randomized scheme so that the resultant network resembles actual social networks that have been studied. Though the details of the social network are interesting (feel free to ask us questions) they are not relevant for this lab. You just need to know that each node is directly connected to at least the number specified in “Minimum Friends” though some nodes may be connected to much more than this minimum. (In social networks, these are the popular people.)
- **Initially vulnerable**: The percentage of computers that is initially vulnerable to infection. This models the possibility that some computers may already be patched against a weakness, or that some people are immune to a disease.
- **Initially infected population**: The number of computers that are initially infected.
- **Time to install**: The time it takes for an infected computer to begin spreading the virus to other computers.
- **Time to spread**: The time between infection attempts by an infected computer that is spreading the virus.
- **Time to repair**: Time from infection until the virus is removed and the computer is immunized.

How infection works:

An infected computer spreads the virus by randomly choosing a computer to which it is directly connected every "Time to Spread" seconds. If that computer is vulnerable, then it will begin spreading the virus "Time to Install" seconds later. However, the good guys may rush software patches to remove the virus/worm. The "Time to Repair" is the time

after which the computer is considered repaired -- no longer spreading the virus/worm, nor susceptible to infection again.

Interpreting the results using Excel:

The simulator outputs periodic counts of vulnerable, infected, and repaired computers. You can save this output to a file of type CSV ("Comma Separated Values"), which is a standard format for data files. To graph the data from a CSV file in Excel:

1. Double-click the CSV file to open it in Excel.
2. Use the mouse to select all the columns and rows that have data (or, to do this with the keyboard, press Ctrl + A).
3. From the Insert menu, click Chart.
4. Select the "XY (Scatter)" chart type, and click Finish.
5. Select File … Save As, and save the file as a *.xls file (not as a CSV file).

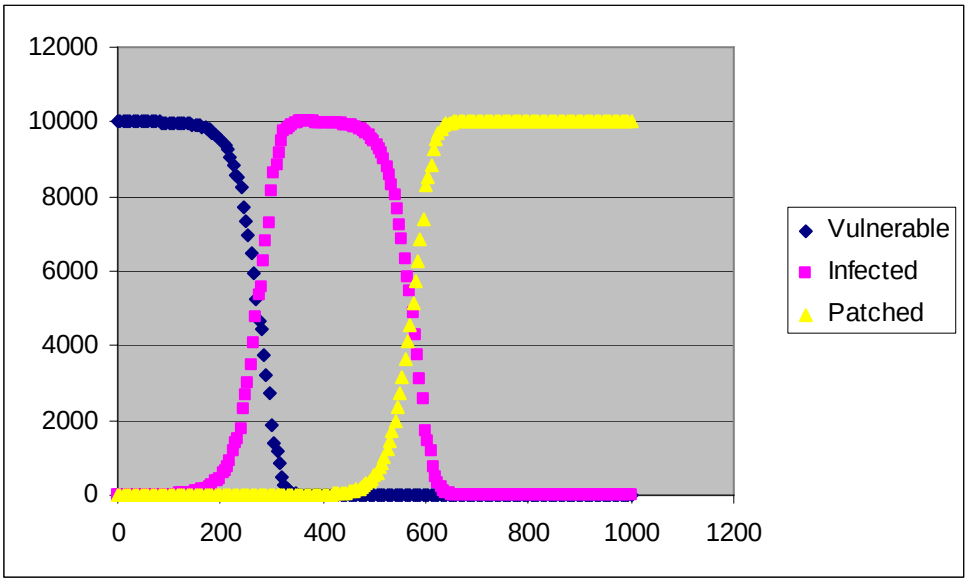

Figure 2: Example chart

Part 1: Worms vs. Email Viruses

You may wish to read the lecture notes to refresh your memory about the difference between viruses and worms. (Your TA is happy to answer questions.)

1. **Run the simulator with the settings depicted in Figure 1 above (which are the default settings), except set "Time to Install" to 10 seconds. Plot the result in an Excel chart.**

2. **Now select "Social Network" and set "Minimum Friends" to 20. Also, change "Time to Install" to 150 seconds. Run the simulator, and plot the result in an Excel chart.**

3. **Which of the previous simulations is a better model for the spread of worms, and which is a better model for the spread of email viruses? Explain your answer in terms of what you know about how worms/email viruses are propagated.**

   The fully-connected network was a better model for worm propagation, since worms spread randomly from computer to computer. The social network is a better model for email viruses, since they spread from person to person by emailing themselves to everyone in an infected computer's email address book.

4. **Report the total number of computers that were infected during each simulation.** To do this, look at the last line of output for the simulation, and note the number of infected and repaired computers. The total number of computers infected during the simulation is the sum of these two numbers.

5. **The two totals you report in (4) should be similar to each other (if not, see the TA). Still, worms are considered more damaging than email viruses. Explain why, based on the charts of the simulations.**

   Because they require a human to open an infected email in order to spread, email viruses spread much more slowly than worms. Worms, conversely, spread without human intervention.

Part 2: Vulnerability of the Network

1. **Change back to the default settings (see Figure 1 above), and plot the result in an Excel chart*.**

2. **Now change "Initially vulnerable" to 50%, and plot the result in an Excel chart. Explain any difference between the two charts.**

   Only half the computers were ever infected in the second experiment. Also, the rate of infection was slightly slower.

Part 3: Infection Rate

1. **Change back to the default settings, except set "Time to Spread" to 2 seconds. Plot the result in an Excel chart*.**

2. **Repeat Step 1 for "Time to Spread" values of 4, 6, 8 and 10 seconds. Plot the results in separate Excel charts*.** Hint: It will probably be faster to generate all the data first, and then create the Excel charts. When saving the output to CSV files, use descriptive names so that you don't forget what the output represents.

3. **Find the time point in each graph when the population becomes completely infected. How does this time point change as "Time to Spread" is increased? Why do you think this happens?**

   As "Time to Spread" increased, the time of peak infection increased. This was because each computer took longer to infect other computers.

Part 4: Social Networks

1. **Run the simulator with the default settings, except select "Social Network", and set "Minimum Friends" to 2. Plot the result in an Excel chart*.**

2. **Compare the rate of infection in this chart (indicated by the steepness of the infection curve) with the rate of infection in Part 2, Step 1. Explain any differences in terms of network structure.**

   The curve was less steep for the social network than for the fully-connected network. This is because, in the social network, each infected computer could potentially only infect two other computers, and not all other computers.

3. **Now set "Minimum Friends" to 1, run the simulator again, and plot the result in an Excel chart*. Explain why this change from "2" to "1" resulted in such a dramatic difference. (Hint: 'exponential growth'.)**

   If at every time step, each newly-infected computer infects two other computers, then the total number of infected computers will grow like 1, 2, 4, 8, 16, etc. On the other hand, if at each time step, each newly-infected computer infects one other computer, then the total number of infected computers will grow like 1, 2, 3, 4, 5, etc.

Part 5: Disease Modeling

The simulator can also be used to model virus spreading in humans. When used this way, a "repaired" node represents a person who has either recovered from the illness or has been killed by it, so that he is no longer infecting others.

1. **Change back to the default settings, except select "Social Network", and set "Time to Install" to 10 seconds and "Time to Spread" to 10 seconds. Run the**

simulator, and plot the result in an Excel chart. This model will represent the spread of the flu.

2. **Change "Time to Repair" to 10 seconds. Re-run the simulation, and plot the result in an Excel chart. This model will represent the spread of a more deadly virus, such as Ebola. (Note that in case of Ebola, people don't get "repaired"; they die. But the net effect is still that they stop spreading disease any further.)**

3. **Now reduce "Time to Spread" for the deadly virus model and re-run the simulation until the total number of people infected is the same for both models (this may not possible; just get as close as you can). What does "Time to Spread" represent in our models? Why do you suppose that outbreaks of deadly viruses tend to be quite localized, even if they are highly contagious?**

   Since viruses like Ebola kill so quickly, their hosts don't have a chance to spread them to very many other people.

Part 6: Additional Questions

1. **Which model is suited for studying the spread of the flu: fully connected network or social network? Write a ballpark estimate for "Minimum Friends" in this setting and briefly explain how you arrived at it. Just guess -- no need for extensive Googling. (Hint: Whom does the person infect? Everybody he/she meets? Shakes hands with? Sneezes at? Some small subset of the above?)**

   Social network. While a person is contagious with the flu, she might meet 50 people.

2. **Explain briefly but clearly why the following measures are recommended during the cold season. Use terminology from the lab, such as "Minimum Friends", and "Time to Spread" (i) Governments and employers should get a large fraction of the population vaccinated against the flu. (ii) Cover your mouth while sneezing and frequently wash hands. (iii) If you have a bad cold or flu, stay home and rest. (Note: Of course this is a good idea for the infected person, but why is it also a good public health measure?)**

   (i) Decreases "% Vulnerable".
   (ii) Increases "Time to Spread"
   (iii) Increases "Time to Spread", decreases "Time to Repair". Also perhaps decreases "Minimum Friends".

3. **The phrase "tipping point" has recently become popular in the press and several academic fields. (See http://www.gladwell.com/tippingpoint/index.html.) It refers to the fact that phenomena like crime, disease, etc. can respond in an extremely nonlinear fashion to small changes in tactics. For instance, increasing**

the police force by 50% may cause crime rates drop precipitously. Did you observe a "tipping point" phenomenon in any of the experiments we did today?

The jump in infection rate when changing "Minimum Friends" from 1 to 2.

4. **By examining your data, write a couple of lines on why an attacker might prefer to write computer worms instead of viruses.**

   Worms spread much faster than viruses.

5. **However, computer viruses are more common than worms. Why? (Hint: The answer uses things you learnt in class, such as social engineering, and the comparative difficulty of writing a virus versus writing a worm.)**

   A worm writer must exploit a vulnerability in a computer's software. A virus writer often only has to exploit a vulnerability in human psychology, and this is typically easier, or at least requires less specialized knowledge.