CS599: Structure and Dynamics of Networked Information (Spring 2005)
4/4/2005: Epidemic Phenomena
Scribes: Shishir Bharathi, Ashish Vaswani

1 Epidemic Phenomena

Epidemic processes occur very frequently around us in all types of technological and other scenarios. Consider the following examples:

Biological : Diseases, but also mutations. Dominant mutations result in more and more affected entities from generation to generation, until almost all of the species is affected.

Technological : Viruses and worms, but also power failures or car traffic. Power failures in a power grid increase the load on other generating stations causing them to also break down, etc. Congestion on freeways leads to similar behavior.

Social Networks : Rumors or group behavior such as riots. A rumor spreads from person to person until a large subset of people are aware of it. Similarly, peer pressure or "safety in numbers" entices people to participate in behavior shared by others, such as rioting.

Strategies : as a special case of social networks, some trends, such as cell phone usage or car sizes, spreads as a result of "strategy optimization". To not be at a disadvantage on the road with bigger cars like SUVs around, people themselves need bigger cars.

In order to analyze the type of effects that will take hold in a network as a result of epidemic phenomena, we need to first define a model for the effect of the network on individuals' behavior. Here, we can draw on several decades worth of work in sociology, biology, and economics.

2 Schelling's model for segregation

One of the first models studied explicitly in this context was Schelling's model for segregation [3]. Schelling was motivated by the question: why is it that most neighborhoods are very uniform (racially, and in other respects), even though most people profess that they would prefer to live in a diverse community?

Schelling proposed the following model: Assume that roughly $\frac{n^2}{2}$ individuals live on an $n \times n$ grid. Each node wants to make sure to not be completely isolated (the odd person in a community): formally, if less than an ε fraction of v's neighbors (in a small ball around v) are of the same color, v is unhappy and moves to a spot where it is not unhappy (say, the closest, or a random, such spot).

What Schelling observed was that even with a relatively small value of $\varepsilon \approx \frac{1}{3}$, neighborhoods end up mostly segregated: when the process quiesces, about $\frac{4}{5}$ of each node's neighbors are of the same color as the node itself. While this result has been verified experimentally, it appears that no result formally proves this observation as of yet. It is also not known how the result relates with the topology of the neighborhood graph.

3 Granovetter's threshold model

Motivated by the study of outbreaks of riots, Granovetter [1] proposes a threshold model. He notices that cities or neighborhoods with very similar statistical demographics often tend to exhibit very different overall behavior of an epidemic. Thus, he concludes that an accurate description of group behavior is impossible in terms of merely high-level statistical data, and needs to take individuals' tendencies into account.

The threshold model he proposes, in the simplest version, can be described as follows: we assume that each individual (node) has a threshold t_v. If t_v other nodes are already active (have adopted the behavior, such as starting to riot), then v joins (becomes active).

As an example, consider the case of 100 people present in a city square in a tense environment, where the threshold of node v_i is i ($t_{v_0} = 0, t_{v_1} = 1, \ldots, t_{v_{100}} = 100$). In this case, v_0 becomes active first, followed by v_1, etc., until the entire population ends up active. On the other hand, if we make a small change and set $t_{v_1} = 2$, only node 0, and no other node, becomes active, even though 100 of the 101 nodes have the same thresholds as before. While this example may be a bit contrived, it clearly shows that the outcome of an epidemic process can change dramatically even when the experiment itself is almost identical.

One of the first questions we may wish to answer about this model is: given the values of the nodes' thresholds, can we predict how many nodes will be active in the end. We can define $F(x) = |\{v|t_v < x\}|$. Then, the number of active nodes is the smallest x with $F(x) \leq x$. Figure 1 shows the plot of $F(x)$ vs. x. The vertical lines denote the number of nodes active after $0, 1, \ldots$ time steps. The number of nodes that will be active in the next time step is then $F(F(x))$, which we can determine pictorially by moving horizontally to the diagonal, and then vertically to the next intersection with $F(x)$. The process then quiesces at the first point for which the function F crosses the diagonal, as that is a fixed point.

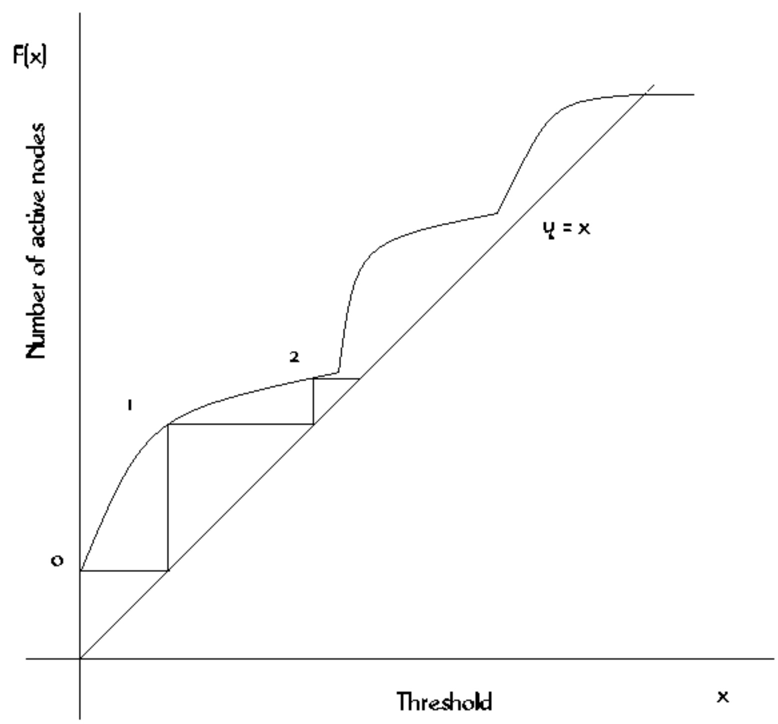

Figure 1: Behavior of F(x)

The threshold values we used in our introductory example were clearly somewhat contrived. Instead, we may wonder what happens when the values are drawn from a more “reasonable” distribution. For instance, we may assume that the thresholds are distributed as Gaussians, with mean, say, $\frac{n}{4}$ and variance σ^2. We could then investigate what is the result of varying σ^2, i.e., making the thresholds more or less sharply concentrated.

For small σ, almost no nodes end up active, because almost no thresholds will be 0 or close to 0. As σ is increased, there is a phase transition from almost no one active to almost everyone active. A further increase of σ leads to a slight drop, steadying around $\frac{n}{2}$.

Granovetter’s threshold model is clearly very simple. We could make it more interesting and realistic in several ways:

Weights : In our model, all nodes have equal influence on all other nodes. Instead, each edge in the graph could have a weight associated with it, representing how much influence one endpoint has over the other. (Edges may be directed, and have different weights in both directions.)

Graphs : We assumed that all nodes affect all other nodes. In a large social network, this is certainly not true. We may consider a graph in which nodes can only affect their neighbors. (Notice that this falls under the case of weights if we assume that the weights of non-existent edges are 0.)

Probabilistic activations : In real world scenarios, some nodes may fail in activating their neighbors. This can be modeled using probabilities on activations. The resulting model may be somewhat different from Granovetter’s.

Non-monotonic behavior : In many scenarios, a large rate of “participation” may discourage nodes from being active. For instance, a new fashion trend, initially spread by copying behavior, may become uninteresting to the trend-setters if many others are following it. As a result, in addition to the activation threshold $t_{1,v}$, we would have a deactivation threshold $t_{2,v} > t_{1,v}$: if more than $t_{v,2}$ nodes are active, then v becomes inactive again. In its most general version, this model subsumes cellular automata and the game of life (and hence, can simulate computation).

Deriving thresholds : So far, we assumed that the thresholds are known. Actually obtaining or estimating this data is an interesting question, and may be partially amenable to data mining techniques. (In the absence of more information, we may assume that thresholds are random, or all identical, but this is only a very crude approximation.) One interesting way to derive thresholds in the strategic settings described earlier is to cast them in terms of the underlying game played by the nodes, e.g., weighing the costs and benefits of a larger car depending on the number of other people having large vs. small cars.

Starting or preventing cascades : Understanding a model is usually only the first step towards acting on it. In the above scenarios, we may be interested in containing the spread of an epidemic, computer virus, or power failure, or promote the adoption of a product.

4 The Contagion model

In investigating the problem of starting or preventing cascades, a first question we may want to answer is how many nodes it takes to infect the entire graph. This leads us to a question considered by Morris [2]. His model is as follows: the graph G is assumed to be infinite, but each vertex has finite degree. All vertices have the same threshold $p \in [0, 1]$: here, this means that they become active if at least a p-fraction of their neighbors is active. (Notice that the number of neighbors may be different for different nodes, but the fraction is the same).

A set $X \subset V$ is called *contagious*, if starting from X, all nodes are activated eventually. The *contagion threshold* $t(G)$ is the supremum of all p such that there exists a finite contagious set for threshold p.

In looking for obstacles to a set X being contagious, we notice that each node that is not infected must have at least a $1 - p$ fraction of its edges to other uninfected nodes. Hence, if X^* denotes the set of all nodes reached from X, then $V \setminus X^*$ is a $(1 - p)$-community in the sense studied in earlier lectures.

We define the community threshold $c(G)$ to be the infimum over all α such that every cofinite[1] set of nodes contains an infinite $(1-\alpha)$ community. In the next lecture, we will prove the following theorem, verifying that communities are indeed the only obstacles to epidemics.

Theorem 1 $c(G) = t(G)$.

References

[1] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83:1420–1443, 1978.

[2] S. Morris. Contagion. *Review of Economic Studies*, 67:57–78, 2000.

[3] T. Schelling. *Micromotives and Macrobehavior*. Norton, 1978.

[1] A cofinite set is a set whose complement is finite.