COMP4121 Project Report

How do I influence people on online social network?

Student: YiSong Jiang

ID: z5123920

## Online Social network

A large dimension online social network (e.g. Facebook, Twitter, Instagram, Weibo...) is usually a huge system that millions of users connected with each other. Each of online social network can be treated as a map of social relations (e.g. friends, families, collaborators...). These online social networks are such powerful tools in young people that they have already become main communication style. One estimate says that within a day, Twitter processes over 250 million tweets. Twitter, in particular, provides what we call a direction node link. Because A may subscribe to B tweets but B don't subscribe to A. And indeed, Twitter's success builds on top of micro-blogging and group texting and one-sided or one direction social networking.

Given the popularity of these services, there is no shortage of interesting questions people have raised. One question is, how to measure the influential power of individuals on an online social network. There are variety of possibilities ways. For example, we can measure on Twitter the number of re-tweets of a user's tweets. As we identify through a party of via. We can also look at the number of re-posting of URL's that a user posted. If we have access to the topology of who follows whom, for example, A follow B's tweets, and B follow C's tweets, and so on. Then, this typology can also give us some information about influential power of different nodes which are different Twitter users.

And that relates to the second question, how to leverage the above knowledge of the individual influential power source and actually influence people online. For example, maybe we'll be able to seed some users and ask them to post or tweet certain information about our product or service. And the question is, which nodes to seed? If we know the topology, would it help us to answer this question more precisely? So, the two questions are indeed important.

## Topology Based Influence Models

In order to answer the two questions, we have to, represent our important quantities with some symbols. So, we start with, of course, graphs.

In this project, we mostly focus on undirected graphs. That means all the links are *bidirectional*. Then we can build our *adjacency matrix* **A**. A given graph with N nodes will have an adjacency matrix of dimension N\*N. Aij is 1 if there is a link from node i to j, and 0 otherwise.

#### Measuring node importance

There several approaches to measuring importance of a node.

1. Degree: the number of nodes connected to a node. In a directed graph, it often has two degree --- in-degree and out-degree, represents the number of incoming edges and outgoing edges. Dunbar's number is used by sociologists as an estimate of the largest number of friends in a social network an average person can keep, usually quoted at around 150. But it is obviously not a good one since degree can misleading number to quantify node importance. Similar with page rank algorithm, if we consider each node as a page, then the importance of a node is much more about the number of connected pages. (More discussion about page rank and influence model will be in later part of this report)
2. Eigenvector centrality:  Starting with an initialization vector . We can always write a vector as a linear combination of the eigenvectors  of , arranged in descending order and indexed by i, for some weight constants . That means . So .

As

1. Closeness centrality: closeness of a node is calculated by shortest distance to another node. In this project, the links' weights are treated equally as distance. The distance between node i and j that is the shortest one called as . So, we can tabulate these for all of the node i, j pairs. Pick the largest one. And we call that the *diameter* of this graph. It is the maximum across all pairs. It's the largest to smallest path. Smallest for a given node pair, and largest across all node pairs. We can also look at the average instead of the max of these . So, we can sum up the and then normalize. That says the following, if we pick a node i here fix i and then look at all the other nodes j. And there are N-1 of them. Then look at the distance from node i to all the others. Closeness centrality is the reciprocal of this average:

.

1. Betweenness centrality: If a node is on the (shortest) paths of many other pairs of nodes, then the node is important. Let be the total number of shortest paths between two different nodes, source s and destination t (neither of which is node i itself), and  be the number of such paths that node i sits on. Then betweenness centrality of node i is

(More discussion about will be in the later part of this report)

#### Measuring link importance

What about link's importance? First of all, what is a link? It's not as easy to define. For example, on Facebook. I may have 300 friends, but only have communicated with 100 of them. And maybe mutual wall posting, other than happy birthday, only with 20 of them. So, do I have a link with 20, 100, or 300 people?

A link could also bidirectional or unidirectional. For example, on Twitter, as already mentioned, a link certain to be unidirectional. Some of these links are strong. Some of these links are weak. As we mentioned that if you define links as weak links if the connect nodes that are not otherwise very well connected. Actually, they may be strong in the sense of disseminating information.

And we say if a link is between two nodes that does not have many short path among them other than this direct link. Now we say this is a locally important link. And if this link is connecting two important nodes in a local important way, we can call that a global important link. So, in fact, we can also quantify similar metrics of link's importance as node importance. In particular, we can look at link betweenness. So instead of counting how many shortest path of node pair (s, t) does node i reside on. We can look at on how many shortest path between node pair (s, t) does a particular link (i, j) sits on top. Then we can take the sum or average of that proportion and call that the betweenness important of link (i, j).

(Examples of above content are in the python notebook)

## Contagion Models

Equipped with a basic discussion of importance scores, we can move on to two more influence model with network topology as part of the picture. However, we have to assumed the graph is given and static. In practice, it's often difficult to get an accurate topology. And topology is often time varying, too. So, we're skipping those important practical issues. A contagion is very similar to the tipping idea. In this 2-state model, the initialization has a subset of the nodes adopting one state, for example, the state of 1, while the rest of the nodes adopt the other state, the state of 0.  We can consider the state-1 nodes as the early adopters of a new product, service, or trend, so it is likely they would be a small minority in the network and not necessarily aggregated together.

We basically said that the adoption percentage p next to discreet time slot is a function of the current one. And this function captures the macroscopic effect of each individual's decision by looking at a global adoption percentage.

Instead of global percentage let's look at local percentages. A node will only flip for example from an iPad has user to an Apple watch user. If the known p of is neighbors flipped. So, rather than looking at a percentage of the entire network. Let us just look at the neighbors those that one node have a direct connection. Let's assume these are bidirectional links. And there are 6 of them. If 3 or more of them have adopted Apple watch then the node will also adopt Apple watch.

This is a local density driven decision making. Now each node may have a different threshold of the percentage of local neighbor it needs to see adopting before he decides to adopt. So, this factor p here, may differ from each node to each node. Since it's not about a whole network, global adoption percentage, we have to ask, will the whole network actually flip? To answer this, the first one to talk about is a notion in a graph called a cluster. A cluster is really a sub set of nodes with a certain density p. A class of density p is a set of nodes such that each of these nodes has a least a fraction of p of his neighbors also in the set. So, for example, in the graph below, the set of nodes forming the lower left triangle is a cluster with density 0.5, whereas the set of nodes forming the square is a cluster with density 2/3.

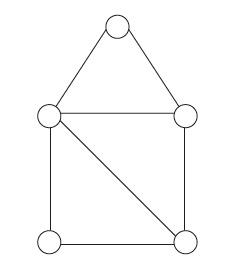


Figure 1

If the entire network flips, then that means there cannot be a cluster of density 1-p or higher among those who have not yet flipped.

Prove: If we have a cluster of density 1-p or higher, all in the state of not flipped. Then this cluster with such a density would become a blocking cluster to prevent the flipping to completely happen inside the network. Because they have enough inertia among themselves, this cluster, this subset, no matter what the rest of network looked like. Their density is 1- p or higher, so, 1- p or more of them will always point to themselves, which means we only have less than p of their neighbors being outside of them. Even if the rest of network all flips, they will still say, look at each other and say well, we do not have yet a high enough density of flipped nodes as our neighbors. So, we will still not flip. At least the only part of this theorem is easy to see.

Another most useful question is about seeding a contagion, seeding a flipping. Let's say that we have fixed a budget of $10,000 and want to stage a viral campaign to sell certain new product by word of mouth online. We want to seed or buy off very important knows, could be people whose tweets are watched and re-tweeted, by many other people. Always, they command a certain prize to be seeded to help us. Let's assume these prices are also increasing functions, of the influential power as perceived by or themselves. In order to fitting the budget, we cannot buy off everyone. Buy off a small number of them, and the question is which ones to buy off? Even without considering the price to pay them, just consider each is one unit of a dollar. It's a graph theoretic and combinatorial optimization problem

There is a trade-off between seeding nodes that they are concentrated enough. And they can have a chance of flipping at least some cluster. Versus seeding nodes so that they are far apart and therefore they may be able to flip entire network. We want to spread them out so that they can reap different corners, and yet we don't want them spread out so much that they lose the chance of flipping an even small corner of the network.

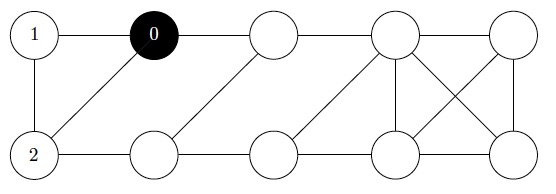


Figure 2

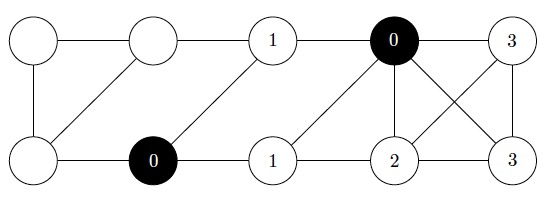


Figure 3

Another point is that seeding is not necessarily successfully refineable. There is a small example, it's supposed to want to seed a contagion in this graph here. And we've got ten nodes, with the links drawn in Figure2. Then which one shall we seed?

If we can only seed one node: At times zero shade it dark meaning that it's flipped from, not buying Apple watch to buying one for example. Flipped from one state of to the other state of the mind. Assume every other node has a flipping threshold of 0.49. If 49% or higher of node neighbors flipped, then the node will flip next as well. Then for this node. It's got two neighbors and one of the neighbors flipped. After two iterations, two of its neighbors flipped and none of the other five nodes would flip. This is the best we can do. If we can only seed one node, seed this node and we can flip two more as the equilibrium are flipping in the contagion.

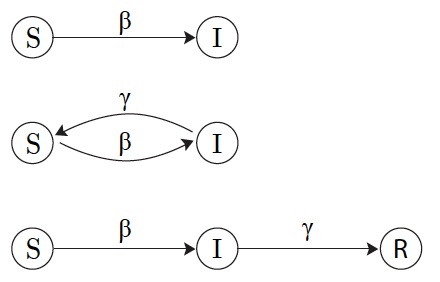
But suppose we can actually seed two nodes as we got a slightly bigger budget. It turns out the optimal way to seed is the choice in Figure3. You seed these two nodes. Even though they are not directly connected, and this is the best that we can do, as it turns out.

The interesting point is that the optimal seeding strategy not successfully refineable. We cannot say, if allow me seed of seed one node, I will seed a particular node. And then when allow me to seed one more node, I could just add one more node.

## Infection Models

Unlike the other models, in the infection models, we'll see state transition across more than two states. It's not just flipping from one to the other states. We're going got see three states, for example. And they are, variants with even more states. Also, instead of looking at discrete, we look at continuous time through the more convenient mathematical representation of differential equation.

We'll talk about three variants of the infection model. S stands for susceptible, I stands for infected and R stands for recovered population. We'll be looking at also the percentage of the overall population rather than the absolute values. And in the Figure4 we see that if you are susceptible to an infection disease, this is where that infection model started, originally, thus these particular choices of that node have a chance of being infected. And the rate of that happening is represented by .



Figure

And if a node is infected, then we have a choice of deciding what to say in the model. In SI model, the simpler ones, says that then the node just stay infected forever. In another SIS model, the node may actually be able to recover. Or at least go back to susceptible with a certain rate . And in the SIR models, the node actually truly recover and become immunized from this infectious disease with a certain rate .