Analysis of Social Network Simulation

Reece Jones

Notation

number of people in a network.

probability of a person liking an interactable post.

probability of a person following the creator of a liked post.

set of all people in a network.

person , where .

set of creators of posts .

set of all posts in a network.

set of posts made by person .

set of people person is following.

set of people selected by simulation timestep to be followed by person .

set of posts liked by person .

set of posts selected by simulation timestep to be liked by person .

set of interactable posts for person .

network parameter after timestep . .  
 represents the state before simulation begins.  
 represents the state after the simulation has completed.

random subset of , where each element has probability of being included.

Simulation model

The simulation is based on a social media network consisting of people and posts made by those people. Each person may *follow* zero or more other people and may *like* zero or more posts, forming connections within the network. The simulation models the evolution of these connections through time. The flow of time in the network model is quantised into *timesteps*. At each timestep, the evolution algorithm is applied, producing a (typically) new network structure.

The algorithm iterates the people in the network and possibly causes them to form new connections to other people and posts. For each person, a set of “interactable” posts () is generated, consisting of posts made and liked by people that person is following. This emulates the behaviour of a real social media application, in which one typically is shown a list of posts/activities from connected people.  
For each interactable post, there is a chance for the person to like the post (), and if liked, a chance for the person to follow the creator of the post (). (An additional like chance scaling factor may be present for a specific post, however this will be ignored in this analysis for simplicity.) These probabilities are constant parameters of the simulation.  
The simulation can be represented mathematically as:

The network is typically seeded with set values of , , and , with further changes to the network structure made only by the simulation algorithm.

The simulation is considered complete when each person likes every post they possibly can and follows every person they possibly can via simulation alone. Note that this is dependent on the initial network structure and is not the same as a fully connected graph; see the code and documentation for details.

Computational complexity of simulation

The simulation code has been developed and optimised for performance. The time complexity of basic network operations are as follows:

* Iterating :
* Iterating :
* Iterating :
* Iterating :
* For a person , checking :
* For a post , checking :
* Adding a person to :
* Adding a post to :
* Finding the creator of a post:

(A complete explanation of the code is available in the documentation, from which the time complexities may be derived.)

Then the time complexity for evaluating timestep is:

That is, proportional to the number of total number of interactable posts across all people. However, the values of and vary throughout simulation and are heavily dependent on the network structure. The following sections will derive solutions for certain structures.

Randomised network structure

In this network, each person follows a random number of randomly selected people and has a random number of posts. The number of follows and posts per person are normally distributed with . People to be followed are chosen by simple random sample (without replacement).

To calculate the time complexity of simulating one timestep, we approximate the number of follows and post likes for any person at some timestep as uniformly distributed:

And approximate the number of posts for any person as uniformly distributed:

Then the time complexity of timestep is approximately:

In order to calculate the time complexity for the entire simulation, we must evaluate and at each timestep, as well as calculate the total number of timesteps before completion.

In each timestep, and may not be disjoint, nor and ; that is, follows and likes acquired through simulation may already be present. Note that , therefore we expect to contain posts already in . For , we approximate that the proportion are already in (a better approximation likely exists, but I was unable to derive it).

Then the increase in follows and likes from timestep to is approximated by:

Since is sampled uniformly from :

Calculating the value of is more difficult, since and may not be equal if there exists . The creators of all posts in the network may be considered a multiset, with multiplicity of each element approximately . Notably, if selecting a post creator at random, the probabilities of selecting each are approximately equal.  
In general, when selecting a random sample of size from a sequence of distinct items with replacement, where each distinct item has an equal probability of being selected, the expected number of unique values is given by (**citation needed**). Thus:

The approximation of similarly requires adjustment for repetition of posts:

The simulation will be complete when and for all . For large , for which the network is likely to be a connected graph, we may approximate:

Substituting equations 3, 6, 7, and 8 into 5 yields the recurrence relation

with initial conditions and . Unfortunately converting this relation into an exact formula for and is non-trivial, and is left as an exercise for the reader.

However, observations basic may still be drawn and compared with real simulation data.

Linear network structure

In this network, the set of follows resembles a linear graph, with the last person having one post. More precisely:

for

Such an example is unlikely to occur in reality; however, it is somewhat simple to analyse.

References

https://math.stackexchange.com/questions/1386527/expected-amount-of-repeats-in-a-random-sequence-of-integers