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 at timestep .

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 these relations into exact formulas for and is non-trivial, and is left as an exercise for the reader.  
However, basic observations may still be drawn by evaluating the relations and comparing them with real simulation data. The following data was gathered for and .

Figures 1a and 1b plot and over each timestep, for the mathematical prediction and real simulation, respectively. These values exhibit exponential growth resisted by a limiting factor, which aligns with an intuitive expectation of the simulation’s behaviour: the number of new connections per timestep is proportional to the number of existing connections, however is limited by the size of the network. Equations 9a and 9b have resemblance to the logistic equation, a model of population growth (**citation needed**). Note that typically lags by an amount scaling with , as a post must be liked first before the associated follow can be made.

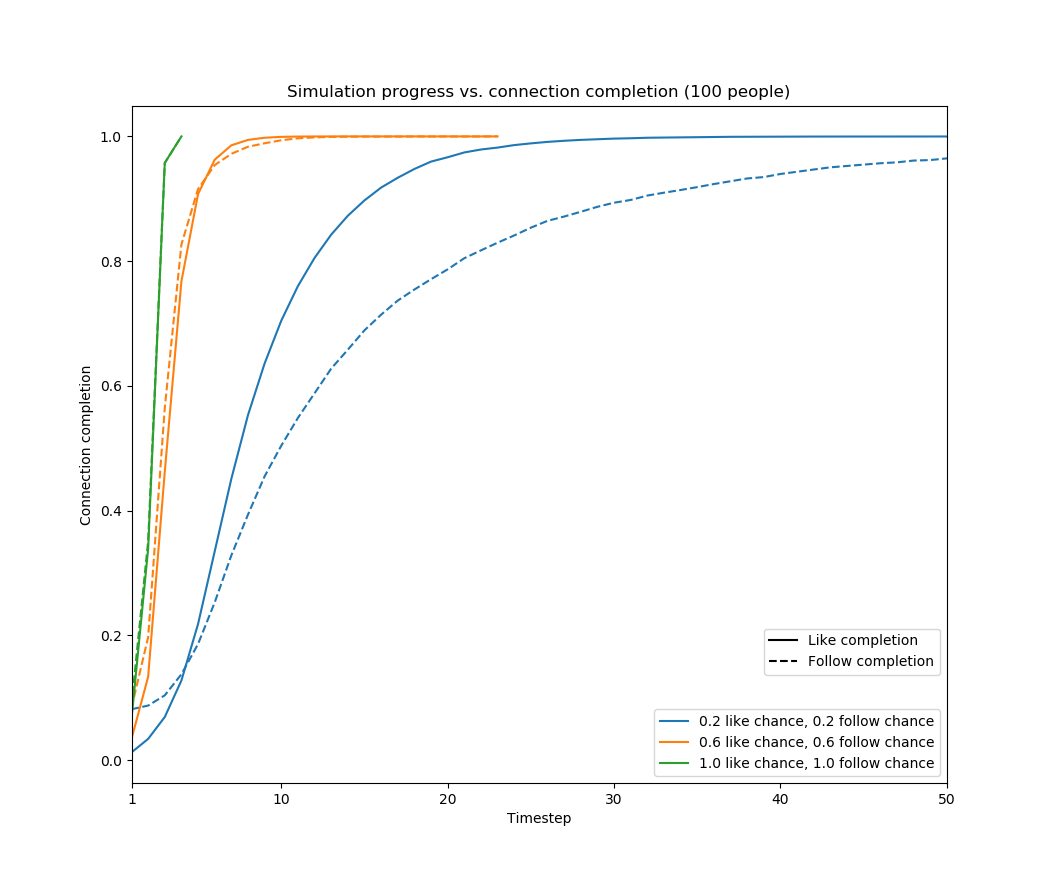


Figure b: simulation progress vs. connection completion.

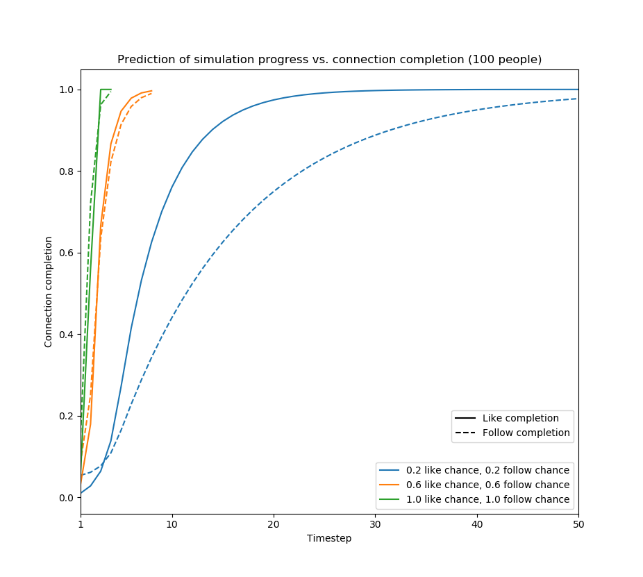


Figure 1a: simulation progress vs. connection completion, using the mathematical prediction. The prediction was run until the connection completion reached 99%.

Figures 2a and 2b show the predicted and actual computational time required for evaluating a timestep over the course of a simulation. As expected, the computational time is seen to be approximately proportional to the connection completion, which is approximately proportional to the total number of interactable posts. Note that the maximum actual computation time per timestep is not the same for different like and follow chances, whereas in the prediction they are equal. This is due to the use of Big-O analysis for the prediction, which ignores constant factors. (In the simulation, lower and equate to fewer operations overall, resulting in a difference by a constant factor).

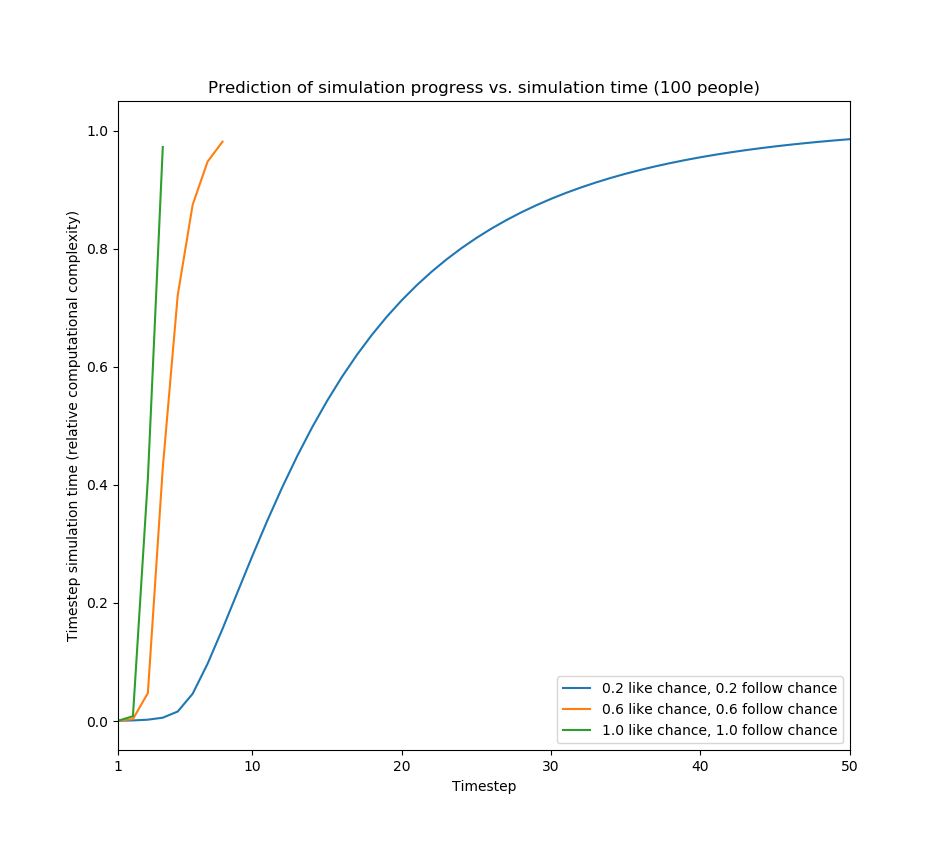


Figure 2a: simulation progress vs. relative timestep time complexity, using the mathematical prediction. The prediction was run until the connection completion reached 99%.

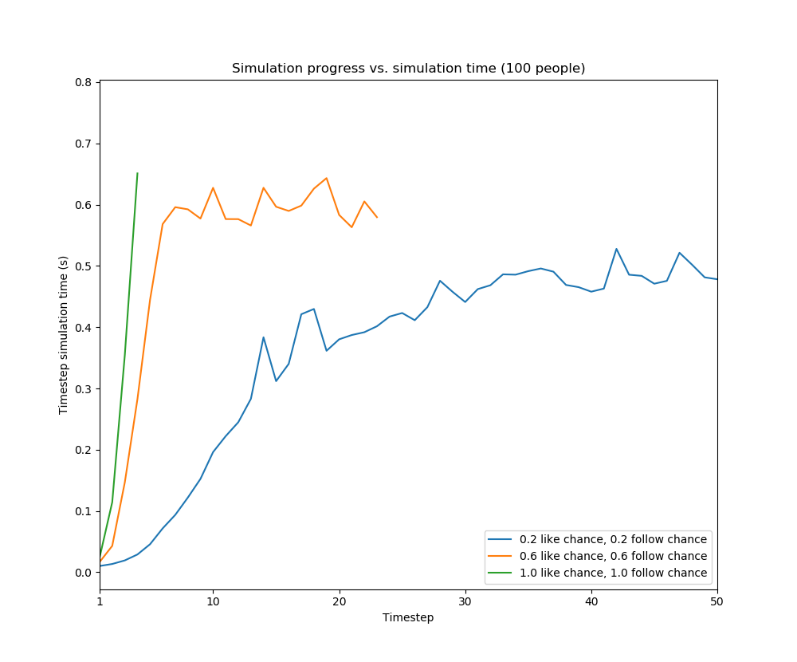


Figure b: simulation progress vs. timestep evaluation time.

Figures 3a and 3b show against the number of timesteps to achieve and . This is approximated for the mathematical prediction by and . The mathematical approximation predicts what seems to be logarithmic growth, while the simulation shows more linear growth for low and . More data is needed to make an exact conclusion, however growth less than linear intuitively seems likely for a random network, as the shortest path between any two people, and thus the maximum number of timesteps for a post to become directly connected, is likely be to much less than .

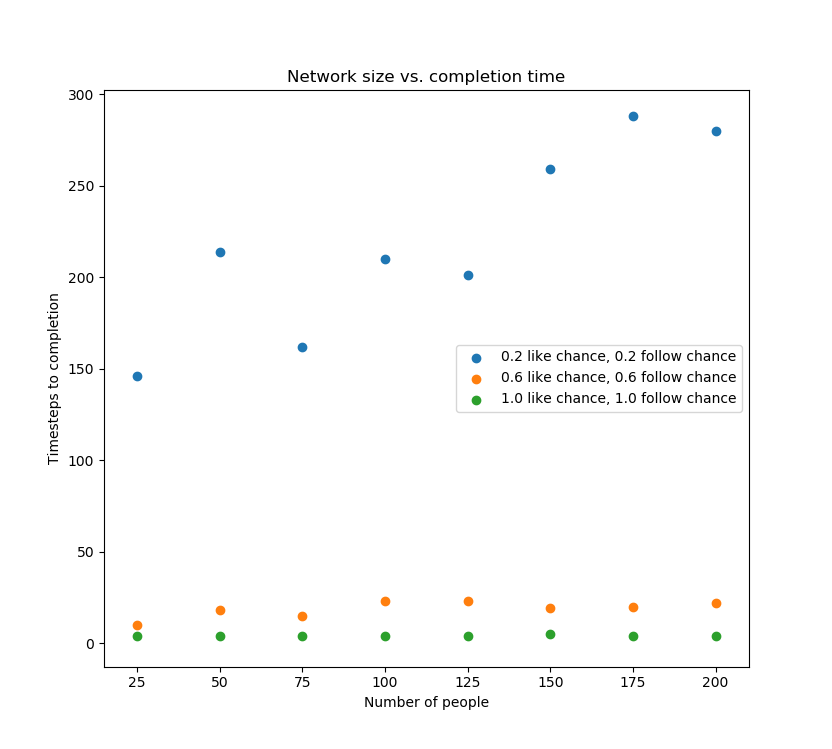


Figure 3b: network size vs. timesteps to completion.

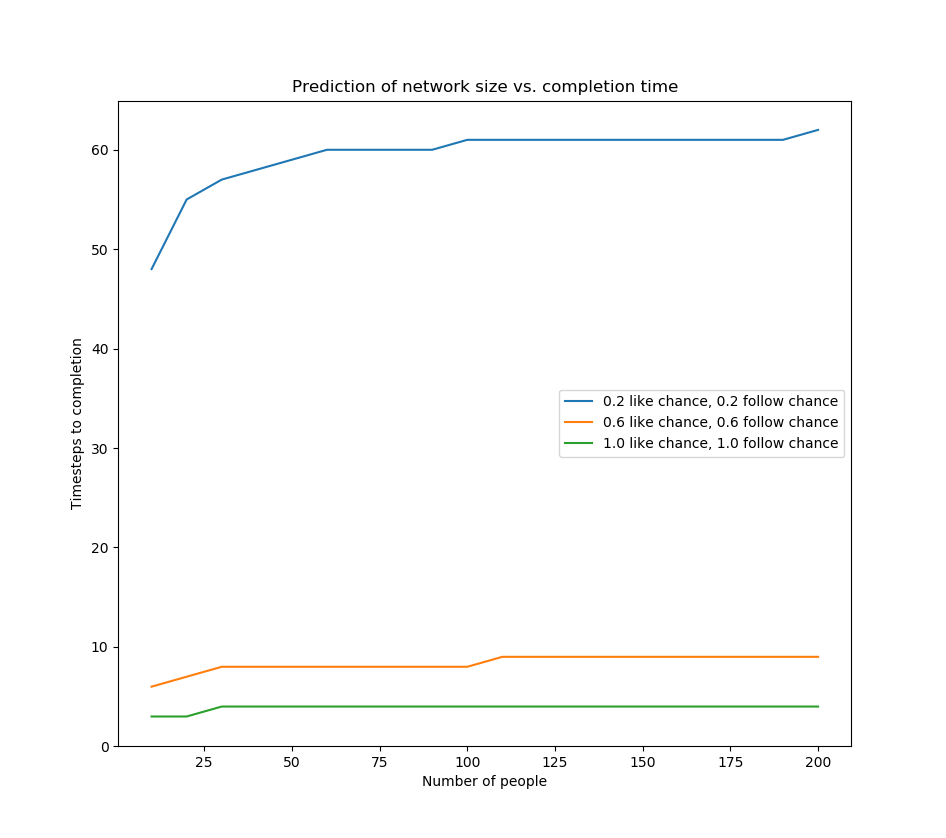


Figure a: network size vs. timesteps to completion, using the mathematical prediction.

Overall, the mathematical approximation analyses the measured results decently, but with definite flaws. Most notably, it cannot be used to accurately predict the timestep at which the simulation is completed; due to its unquantized, asymptotic nature, neither nor reach 1, whereas such a state does occur in the real simulation. Additionally, its behaviour in the first few timesteps is inaccurate, as can be seen in figures 1 and 2, particularly when and are low. Improvement of the prediction and elimination of these flaws is likely possible with a more advanced and rigorous analysis of the structure of the random network.

Linear network structure

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

for

for

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

For timestep 1, for , and , therefore:

And in general:

Note that at timestep , the maximum number of connections post may have propagated through is ; hence .

Since the network is linear, , and hence:

for

References

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

<http://mathworld.wolfram.com/LogisticEquation.html>