# Analysis of Social Network Simulation

Reece Jones

## Abstract

This report presents a simulation model for social media networks such as Facebook, and analyses the behaviour and computational complexity of the model. The analysis examines the effect of simulation parameters and network structures, based on profiling of simulation code, and supported with mathematical interpretations and explanations where possible.

The behaviour of the simulation model is shown to be strongly dependent on the initial structure of the modelled network. The computational complexity is shown to be roughly proportional to the connectivity and connection rate of the network (equation 3). Due to the specificity of behaviour, stronger claims are difficult to make in the generalised case; however, a mathematical approximation is derived for randomised networks.

## Notation

| | |
|---|---|
| $N$ | number of people in a network. |
| $P(L)$ | probability of a person liking an interactable post. |
| $P(F)$ | probability of a person following the creator of a liked post. |
| $P$ | set of all people in a network. |
| $P_i$ | person $i$, where $i \in \mathbb{Z}, 1 \leq i \leq N$. |
| ${}^{k}_{i}P^{*}$ | set of creators of posts ${}^{k}_{i}L^{*}$. |
| $Q$ | set of all posts in a network. |
| ${}_{i}Q$ | set of posts made by person $i$. |
| ${}^{k}_{i}F$ | set of people person $i$ is following. |
| ${}^{k}_{i}F^{*}$ | set of people selected by simulation to be followed by person $i$. |
| ${}^{k}_{i}L$ | set of posts liked by person $i$. |
| ${}^{k}_{i}L^{*}$ | set of posts selected by simulation to be liked by person $i$. |
| ${}^{k}_{i}I$ | set of interactable posts for person $i$. |
| ${}^{k}X$ | network parameter $X$ after timestep $k$. $k \in \mathbb{Z}, k \geq 0$.<br>$k = 0$ represents the state before simulation begins.<br>$k = \infty$ represents the state after the simulation has completed. |
| $\mathrm{SRS}(S, p)$ | random subset of $S$, where each element has probability $p$ 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 (${}^{k}_{i}I$) 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 ($P(L)$), and if liked, a chance for the person to follow the creator of the post ($P(F)$). (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:

$$ {}^{k}_{i}I = \left( \bigcup_{p \in {}^{k}_{i}F} {}_{p}Q \cup {}^{k}_{p}L \right) - {}_{i}Q \tag{1a} $$

$$ {}^{k}_{i}L^{*} = \mathrm{RS}\left( {}^{k}_{i}I, P(L) \right) \tag{1b} $$

$$ {}^{k}_{i}F^{*} = \mathrm{RS}\left( {}^{k}_{i}P^{*}, P(F) \right) \tag{1c} $$

$$ {}^{k+1}_{i}L = {}^{k}_{i}L \cup {}^{k}_{i}L^{*} \tag{1d} $$

$$ {}^{k+1}_{i}F = {}^{k}_{i}F \cup {}^{k}_{i}F^{*} \tag{1e} $$

The network is typically seeded with set values of ${}^{0}_{i}F$, ${}^{0}_{i}L$, and ${}_{i}Q$, with further changes to the network 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.) This state may be represented mathematically as:

$$ \sum_{i} \left| {}^{k}_{i}L \right| = \sum_{i} |{}^{\infty}_{i}L| \tag{2a} $$

$$ \sum_{i} \left| {}^{k}_{i}F \right| = \sum_{i} |{}^{\infty}_{i}F| \tag{2b} $$

## 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 $P$: $O(N)$
- Iterating $Q$: $O(|Q|)$
- Iterating ${}_{i}F$: $O(\left| {}_{i}F \right|)$

- Iterating ${}_iQ$: $O(|\ {}_iQ|)$
- Iterating ${}_iL$: $O(|\ {}_iL|)$
- For a person $x$, checking $x \in {}_iF$: $O(1)$
- For a post $x$, checking $x \in {}_iL$: $O(1)$
- Adding a person to ${}_iF$: $O(1)$
- Adding a post to ${}_iL$: $O(1)$
- Finding the creator of a post: $O(1)$

Then the time complexity for evaluating timestep $k$ is:

$$O\left(\sum_i \left(\left|{}^{k-1}_iI\right| + \left|{}^{k-1}_iL^*\right| + \left|{}^{k-1}_iF^*\right| + \sum_{j \in {}^{k-1}_iF} \left(\left|{}^{k-1}_jL\right| + |\ {}_jQ|\right)\right) + |Q|\right) \tag{3}$$

The first three terms in the outer summation account for the post interactions, while the inner summation accounts for the generation of ${}^{k-1}_iI$. The final term accounts for additional processing for each post. (A complete explanation of the code is available in the documentation, from which the time complexities may be derived.)

## Investigation methodology

In general, it is not expected that there exists a generalised mathematical solution to the simulation model that does not involve simply evaluate the entire simulation. As such, particular network structures will be examined in order to inspect their effects on the behaviour of the simulation. For each network type, the behaviour and computational time complexity will be analysed and visualised with respect to various parameters of the network and evaluated against mathematical predictions and/or explanations.

The simulation code is believed to be optimal or close to optimal in terms of time complexity, as all network operations are $O(1)$ where theoretically possible. It is not, nor is designed to be, optimal in terms of memory usage, as explained in the code documentation. Additionally, as it is implemented in Python, it is difficult to track exact memory efficiency. For these reasons, memory usage is not considered in this investigation.

Simulations were run with the `SocialSim.py` application in simulation mode, with `STATS_ENABLED` set to true in `simulation_mode.py` (refer to the code documentation for further explanation). Input files were generated using `network_generator.py`, which is documented in-code. Multiple executions of a simulation were made where possible, however this was not always feasible for cases that resulted in very long simulations (e.g. hours).
As data collection was automated and consisted of over 3500 invocations of the simulation, the input files and raw output data are not provided.

The Python implementation used was CPython 3.7.5, running on Windows 10 64-bit, with an Intel i7-6700K CPU at 4GHz base clock. Python's `time.perf_counter()` is used to time the evaluation of a timestep, theoretically yielding high accuracy and precision. It is important to keep in mind that such profiling is almost certain to include some error due to process scheduling, CPU clock frequency scaling, and other small interruptions present on modern computers and operating systems.

All figures in this report were produced using the Matplotlib Python library, version 3.0.2. Larger versions are available in the `plots` directory of the project, including additional plots not included in this report.

## 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 $\mu > 0$. People to be followed are chosen by simple random sample (without replacement).

Firstly, we estimate the parameters for each person at some timestep $k$ as uniformly distributed:

$$\left|{}_i^k F\right| \approx {}^k f \tag{4a}$$

$$\left|{}_i^k L\right| \approx {}^k l \tag{4b}$$

$$\left|{}_i^k I\right| \approx {}^k j \tag{4c}$$

$$\left|{}_i^k L^*\right| \approx {}^k l^* \tag{4d}$$

$$\left|{}_i^k F^*\right| \approx {}^k f^* \tag{4e}$$

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

$$\left|\ {}_i Q\right| \approx q \tag{4f}$$

In each timestep, ${}_i^k L^*$ and ${}_i^k L$ may not be disjoint, nor ${}_i^k F^*$ and ${}_i^k F$; that is, follows and likes acquired through simulation may already be present. Note that ${}_i^k L \subseteq {}_i^{k-1} I$, therefore we expect ${}_i^k L^*$ to contain $P(L) \cdot \left|{}_i^k L\right|$ posts already in ${}_i^k L$. For ${}_i^k F^*$, we estimate that the proportion $\frac{\left|{}_i^k F\right|}{\left|{}_i^\infty F\right|}$ are already in ${}_i^k F$ (a better approximation likely exists, but I was unable to derive it).

Then the increase in follows and likes from timestep $k$ to $k+1$ is approximated by:

$${}^{k+1} l \approx {}^k l + {}^k l^* - P(L) \cdot {}^k l \tag{5a}$$

$${}^{k+1} f \approx {}^k f + {}^k f^* \cdot \left(1 - \frac{{}^k f}{{}^\infty f}\right) \tag{5b}$$

Since ${}_i^k L^*$ is sampled uniformly from ${}_i^k I$:

$${}^k l^* \approx {}^k j \cdot P(L) \tag{6a}$$

Calculating the value of ${}^k f^*$ is more difficult, since $\left|{}_i^k L^*\right|$ and $\left|{}_i^k P^*\right|$ may not be equal if there exists $\left|\ {}_i Q\right| > 1$. The creators of all posts in the network may be considered a multiset, with multiplicity of each element approximately $q$. 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 $s$ from a sequence of $n$ distinct items with replacement, where each distinct item has an equal probability of being selected, the expected number of unique values is given by $n \cdot (1 - \left(1 - \frac{1}{n}\right)^s)$ (user940 2015). Thus:

$${}^k f^* \approx N \cdot \left(1 - \left(1 - \frac{1}{N}\right)^{{}^k l^* \cdot P(F)}\right) \tag{6b}$$

The approximation of ${}^k j$ similarly requires adjustment for repetition of posts:

$${}^k j \approx |Q| \cdot \left(1 - \left(1 - \frac{1}{|Q|}\right)^{{}^k f \cdot \left({}^k l + q\right)}\right) \tag{7}$$

The simulation will be complete when ${}_i^k F = {}_i^\infty F$ and ${}_i^k L = {}_i^\infty L$ for all $i$. For large $N$, for which the network is likely to be a connected graph, we may approximate:

$$ {}_i^\infty F \approx P \tag{8a} $$

$$ {}_i^\infty L \approx Q \tag{8b} $$

Substituting equations 4-8 into equations 5a and 5b yields the recurrence relations

$$ {}^{k+1}l \approx {}^k l + P(L) \cdot \left( |Q| \cdot \left( 1 - \left( 1 - \frac{1}{|Q|} \right)^{{}^k f \cdot \left( {}^k l + q \right)} \right) - {}^k l \right) \tag{9a} $$

$$ {}^{k+1}f \approx {}^k f + N \cdot \left( 1 - \left( 1 - \frac{1}{N} \right)^{|Q| \cdot \left( 1 - \left( 1 - \frac{1}{|Q|} \right)^{{}^k f \cdot \left( {}^k l + q \right)} \right) \cdot P(L) \cdot P(F)} \right) \cdot \left( 1 - \frac{{}^k f}{N} \right) \tag{9b} $$

with initial conditions ${}^0 l$ and ${}^0 f$. Unfortunately converting these relations into exact formulas for ${}^k l$ and ${}^k f$ 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 data for figures 1-3 was gathered for $\left| {}_i^0 F \right| \sim N(5, 4^2)$ and $\left| {}_i Q \right| \sim N(2, 2.5^2)$. Simulation completion for the mathematical prediction was approximated by $\frac{{}^k l}{{}^\infty l} = 0.99$ and $\frac{{}^k f}{{}^\infty f} = 0.99$.

Figures 1a and 1b plot $\frac{\sum_i \left| {}_i^k L \right|}{\sum_i \left| {}_i^\infty L \right|}$ and $\frac{\sum_i \left| {}_i^k F \right|}{\sum_i \left| {}_i^\infty F \right|}$ 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 (Weisstein, n.d.). Note that $\frac{\sum_i \left| {}_i^k F \right|}{\sum_i \left| {}_i^\infty F \right|}$ typically lags $\frac{\sum_i \left| {}_i^k L \right|}{\sum_i \left| {}_i^\infty L \right|}$ by an amount scaling with $P(F)$, as a post must be liked first before the associated follow can be made.

Figures 2a and 2b show the predicted and actual computational time required for evaluating a timestep over the course of a simulation. The time complexity predicted by the approximation for a timestep $k$ is given by:

$$ O\left( N \cdot \left( {}^{k-1} f \cdot \left( {}^{k-1} l + q \right) + {}^{k-1} j + {}^{k-1} l^* + {}^{k-1} f^* \right) + |Q| \right) \tag{10} $$

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 such as ratios between the terms dependent on $P(L)$ or $P(F)$ and those which do not depend on them. In figure 2a, the cost of operations that scale with $P(L)$ and $P(F)$, while still $O(1)$ complexity, is evidently not weighted heavily enough.

Figures 3a and 3b show $N$ against the number of timesteps to achieve ${}_i^k L = {}_i^\infty L$ and ${}_i^k F = {}_i^\infty F$. The mathematical approximation predicts what seems to be logarithmic growth, while the simulation shows more linear growth for low $P(L)$ and $P(F)$. More data is needed to make an exact conclusion, however growth less than linear intuitively seems likely for a random network, as the length of the shortest path between any two people is likely be to much less than $N$.

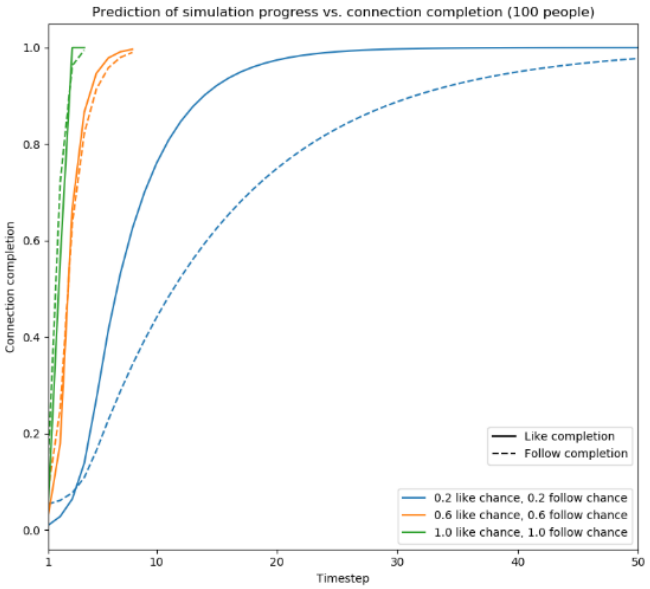


*Figure 1a: simulation progress vs. connection completion, using the mathematical prediction for a random network with N=100.*

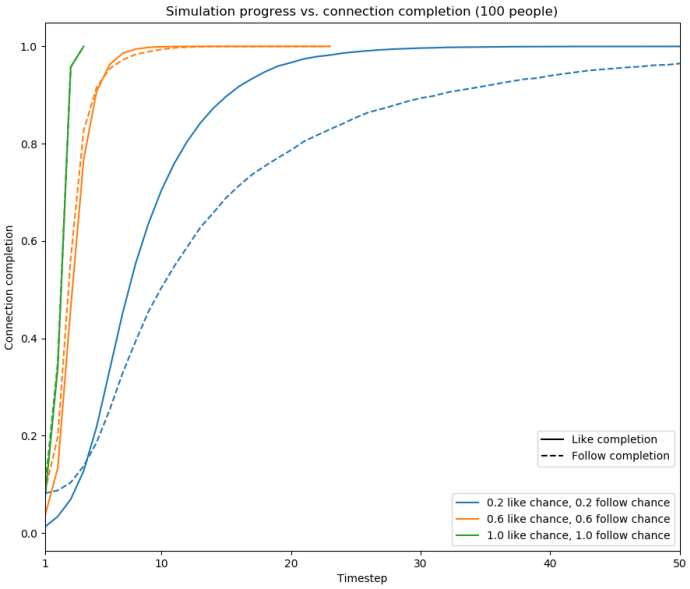


*Figure 1b: simulation progress vs. connection completion for a random network with N=100.*

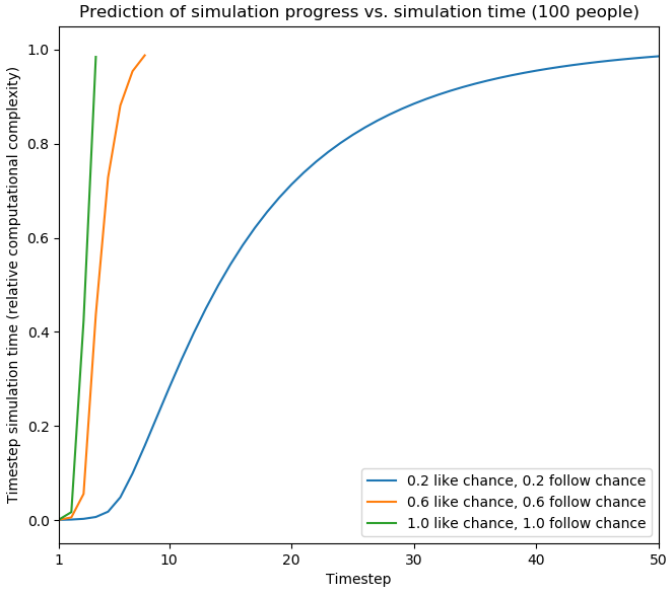


*Figure 2a: simulation progress vs. relative timestep time complexity using the mathematical prediction of a random network with N=100.*

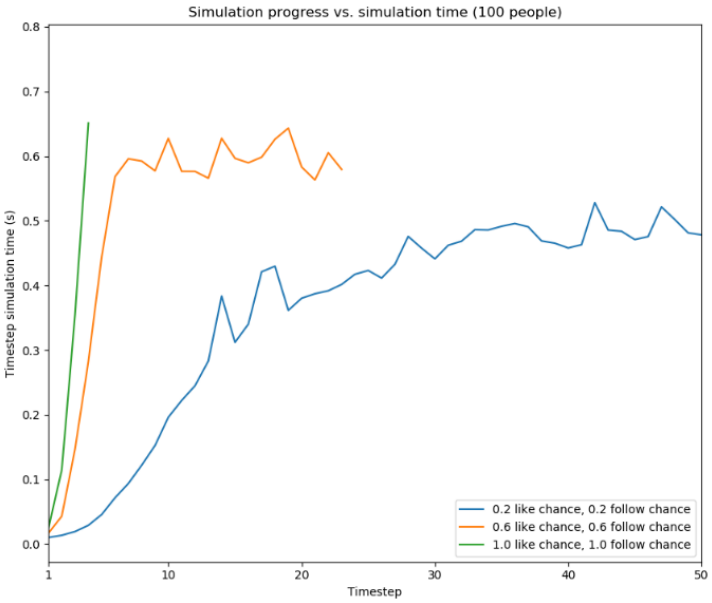


*Figure 2b: simulation progress vs. timestep evaluation time for a random network with N=100.*

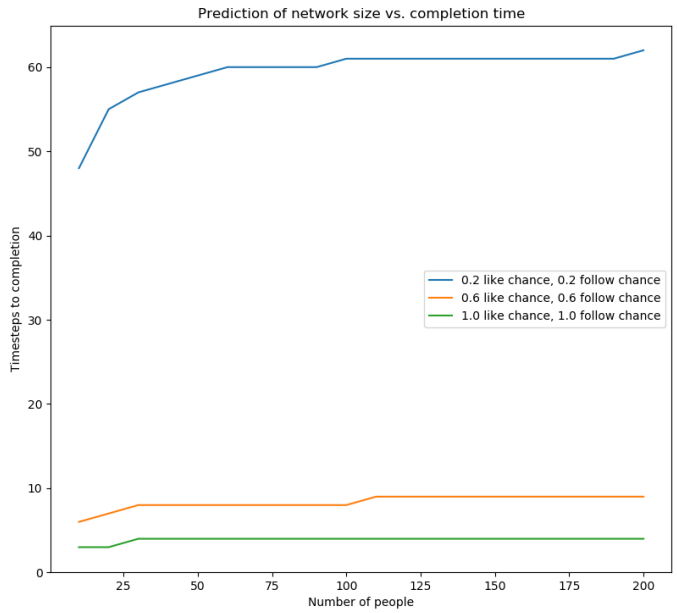


*Figure 1a: network size vs. timesteps to completion, using the mathematical prediction of a random network.*

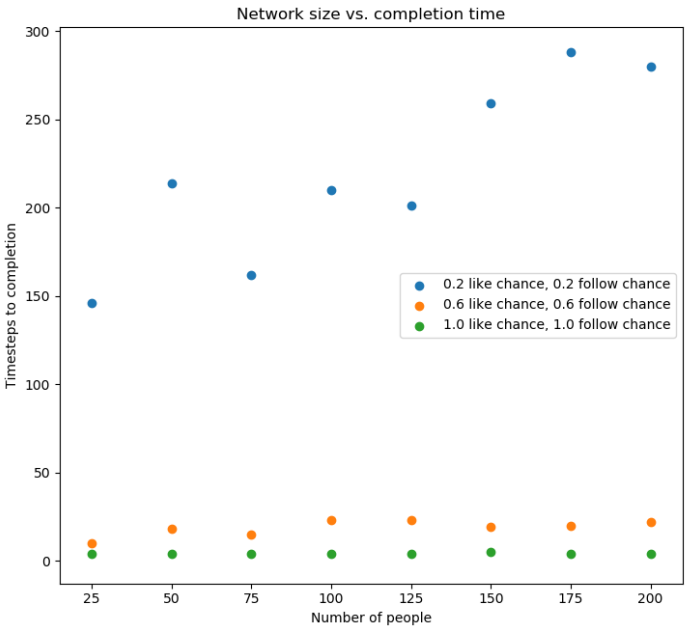


*Figure 3b: network size vs. timesteps to completion for a random network.*

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 $\frac{\left|{}_i^k L\right|}{\left|{}_i^\infty L\right|}$ nor $\frac{\left|{}_i^k F\right|}{\left|{}_i^\infty F\right|}$ 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 $P(L)$ and $P(F)$ are low. Finally, the predicted follows appear to lag likes slightly too much, as visible in figure 1. Improvement of the prediction and elimination of these flaws is likely possible with a more advanced and rigorous analysis of the random structure of the network.

## Linear network structure

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

$$ {}_i^0 F = \{P_{i+1}\} \text{ for } 1 < i \leq N $$

$$ {}_1^0 F = \emptyset $$

$$ {}_i Q = \emptyset \text{ for } 1 < i \leq N $$

$$ {}_1 Q = \{q\} $$

$$ {}_i^0 L = \emptyset $$

Such an example is unlikely to occur in a real social network, yet I wish to examine it in contrast to the random network to demonstrate the effect of network structure. I have not derived a mathematical solution to this network type as it proved to be too difficult; therefore, this section will give only a general, surficial analysis of the simulation. The data shown in figures 4-7 was collected with one simulation for $P(L) = 1$, $P(F) = 1$, and two simulations each for $P(L) = 0.6$, $P(F) = 0.6$ and $P(L) = 0.2$, $P(F) = 0.2$.

The most notable feature of this network structure, unsurprisingly, is its linearity. With only one initial follow per person, ${}_i^k I$ "depends" directly on ${}_{i-1}^{k-1} I$, and $q \in {}_i^k L \rightarrow q \in {}_{i-1}^{k-1} L$. Therefore, it will take at minimum $N - 1$ timesteps before $q \in {}_N^k L$ and $P_1 \in {}_N^k F$ and the simulation is complete; i.e. proportional to $N$. Figure 4 illustrates this relationship. In comparison with the random network, a significantly larger amount of timesteps required to complete a simulation.

Since $P\left(q \in {}_i^k L^*\right)$ is proportional to $P(L)$, and $P(P_1 \in {}_N^k F^*)$ to $P(F)$, the number of timesteps to completion is also inversely proportional to $P(L)$ and $P(F)$. Figures 5 shows this relationship. Note that $P(F)$ does not seem to have a particularly large effect on the completion time, except perhaps for very low values. A plausible explanation for this is that most of the time is spent "waiting" for $q \in {}_i^k I$, which takes time proportional to $i$ and $P(L)$, after which $P_1 \in {}_i^k F$ takes time proportional only to $P(L) \cdot P(F)$.

Despite many timesteps being required for completion of a simulation, each timestep is evaluated quickly, as shown in figure 6. Since $\left|{}_i^k L\right|$, $\left|{}_i^k I\right|$, $\left|{}_i^k L^*\right|$, $\left|{}_i^k F^*\right|$ and $\left|\ {}_i Q\right|$ are either 0 or 1 and $\left|{}_i^k F\right|$ is either 0, 1, or 2, evaluating a timestep has an average case time complexity of $O(N)$. Even so, evaluation time still scales with $k$, as these values approach their maximums as $k$ approaches ∞.
Note that the measured times are likely subject to relatively high error due to the small values, as can be seen with the outliers for $P(L) = 0.6$ and $P(F) = 0.6$ around $k = 150$.

Finally, figure 7 plots $\frac{\sum_i \left|{}_i^k L\right|}{\sum_i \left|{}_i^\infty L\right|}$ and $\frac{\sum_i \left|{}_i^k F\right|}{\sum_i \left|{}_i^\infty F\right|}$ across timesteps. As I have not derived a solution for the linear network, I can provide only speculation on the observed behaviour. Interestingly, the relationship appears

to be linear, despite the probabilistic nature of the simulation. Examining the behaviour for $P_2$, we find that $P\left(q \in {}_2^k L\right) = 1 - (1 - P(L))^k$, which is already a non-linear function of $k$; presumably $P\left(q \in {}_i^k L\right)$ for $i > 2$ are also non-linear. However, since the post may only extend its reach by one person per timestep, $P\left(q \in {}_i^k L\right) = 0$ for $i > k + 1$. I suspect that this contributes to the observed linearity; further investigation is required.

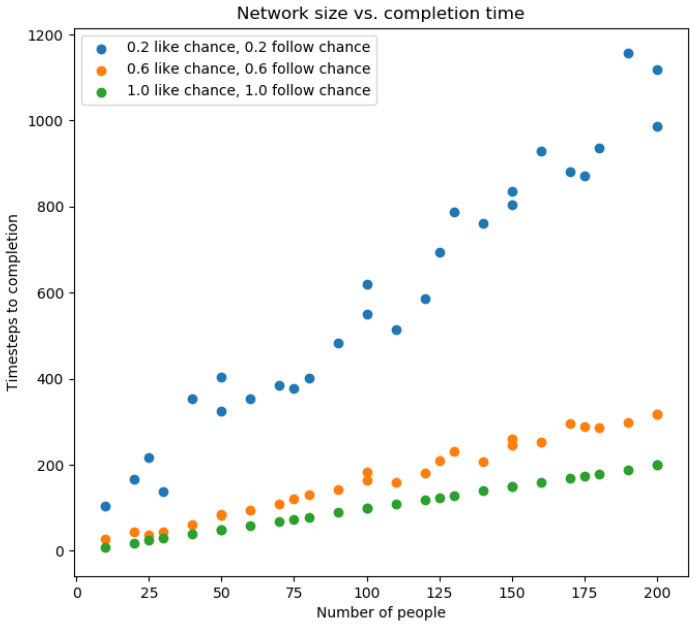


*Figure 4: network size vs. timesteps to completion for a linear network.*

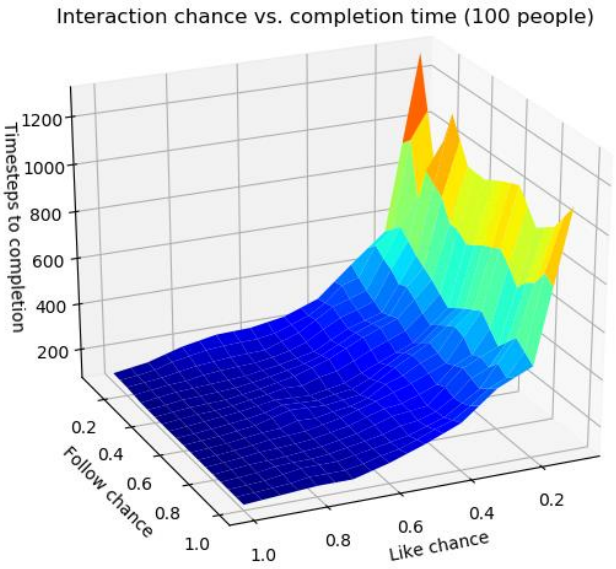


*Figure 5: P(L) and P(F) vs. timesteps to completion for a linear network with N=100 (linearly interpolated surface).*

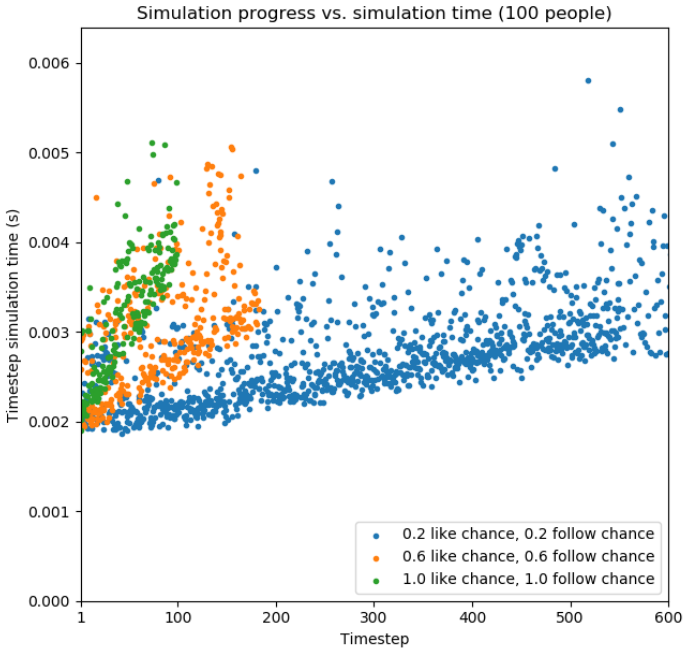


*Figure 6: simulation progress vs. timestep evaluation time, for a linear network with N=100.*

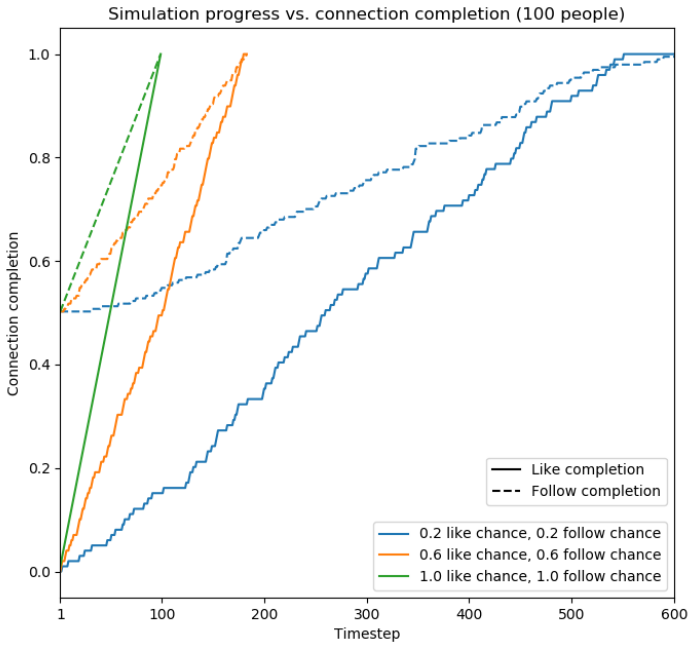


*Figure 7: simulation progress vs. connection completion, for a linear network with N=100.*

# Conclusion

The behaviour of the network simulation is difficult to analyse exactly in the general case, as it is largely dependent on the initial network structure. However, through comparisons of different types of network structures, it is possible to derive some useful understanding. As demonstrated by the randomised network, a more densely connected network evolves similarly to that predicted by models of population growth. In addition, the relatively high amount of connections between entities causes a high computational complexity involved in simulating the model. At the other extreme, a linear network structure demonstrates bottlenecking of connection spread, resulting in significantly lower simulation computational complexity. In general, the time complexity of the simulation is shown to be proportional to the connectivity and connection rate.

Overall, specific behaviours were decently predicted or explained theoretically. The mathematical model for the random network predicts the actual behaviour surprisingly well given the amount of approximations made. While no such approximation or solution was found for the linear network, none of the results defied explanation, with the exception of connection completion as a function of simulation progress.

There is still a significant amount of research into this simulation model that may be done. The approximation of the random network may be refined, including solving recurrence relations 9a and 9b, and better utilising the properties of its randomness. The linear network remains largely unsolved mathematically, primarily requiring a model of its connection completion over time. Finally, there is potential in investigating other network structures, perhaps hybrids of those presented in this report, in order to gain more understanding of real-world social networks that do not have such homogeneous structures.

# References

Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9 (3): 90-95. https://doi.org/10.1109/MCSE.2007.55.

user940. 2015. “Expected amount of repeats in a random sequence of integers.” Mathematics Stack Exchange. https://math.stackexchange.com/q/1386590.

Weisstein, Eric W. n.d. “Logistic Equation.” Wolfram MathWorld. http://mathworld.wolfram.com/LogisticEquation.html.