Social Networking Sites

Using Epidemiological Models to analyse changes in active users over time

Eduard A Bruchner sm20698@bristol.ac.uk
Mario Alvarez Junquera zk20505@bristol.ac.uk
Joseph Ishak qb20991@bristol.ac.uk
Abdul Tarafder hg20977@bristol.ac.uk
Ian Vazquez Perez dl20038@bristol.ac.uk
Paola Zabala Arequipa et20798@bristol.ac.uk

Group 3
May 27, 2021

Abstract

This paper proposes the use of a simple epidemiological model composed of mutually exclusive compartments SEIR (susceptible-exposed-infected-recovered) as an analogous way of describing user adoption and abandonment rates of Online Social Networks (OSNs). Adoption corresponds to being infected, whereas abandonment corresponds to being recovered and having gained immunity. The limitations of this model are discussed and an alternative 'infectious recovery' method is presented as a more realistic solution, accounting for the subtleties involved in the process of abandoning the network. Each individual compartment is described by an ordinary differential equation in Python; the whole system is solved numerically under various initial conditions using LSODA from the Fortranlibrary Odepack. A third method involving object oriented programming and web scraping is suggested to combat all the flaws of the first two approaches, by dividing the population into smaller groups based on age and on ease of connectivity to other individuals to be exposed to the network

1 Introduction

Social networking sites form a vital part of today's daily life, with 3.78 Billion active users worldwide - roughly 48% of the population - as of 2021 [1] . This should not be a surprising figure, as Online Social Networks (OSNs) have become pivotal platforms, serving people not only as a means of meeting other individuals and of staying connected, but also as a means of building an audience

with whom they can share entertaining and educational content.

Their widespread presence across the Internet goes hand in hand with the rise of multibillion-dollar industries facilitating such services. As of February 2021, Facebook's market capitalization was 733.6 billion US dollars, making it the third most valuable U.S.-based internet company by market capitalization and making it the leading social network worldwide [2].

The success of such companies is directly proportional to the number of active users in which they can engage. As such, they take measures to create a world in which social media is enmeshed with most aspects of people's lives, by augmenting OSNs to provide a vast array of functionalities and services (e.g. Facebook's Marketplace). This points towards the idea that the modern-day consumer is living in an increasingly "omni-social" world [3].

In this paper, we analyse the patterns in the adoption and abandonment of such social networks by drawing inspiration from rules that govern infectious diseases. This is a sensible analogy to draw, as people tend to join a network if their acquaintances are already active users, but they eventually lose interest in it. There have been previous studies that applied epidemiological models to study OSNs [4] since the most common way of growing a social network is by word of mouth which, from the modelling perspective, is directly comparable to infecting people through contact. Another example of how an epidemiological model was used to study online social networks can be seen in Cannarella and Spechler [5] who studied information diffusion on Twitter. This is a less tangible variable to monitor, but the analogy is still valid as ideas are spread between different people who share them. Eventually, the people who propagate them lose interest and this phenomenon can be thought of as becoming 'recovered' or gaining 'immunity'.

The results from the epidemiological equations presented in this study will be used to interpret the behaviour of any general social network. However, true data about OSN usership is difficult to obtain; as such, data will be used from Google's "Google Trends" service [6] as a proxy for our actual quantity of interest, to compare the validity of our results. The network on which the data is based is MySpace, a platform that has completed its cycle and has now decayed to a level of insignificance.

2 Methods and Results

The core idea of the epidemiological model is to divide the population into mini compartments; the individuals within each compartment are identical in their status with respect to the infection at hand. The SEIR model is an appropriate way to describe the behaviour of Online Social Networks, as the adoption and abandonment of the network are analogous to people being infected and recovered from the virus. The full components are as follows:

- Susceptible(S): the group of people susceptible to the infection (or to join the network). Groups of people that have lost immunity are not included in this model: Once a person abandons the network, it is assumed that they cannot move into the susceptible class again.
- Exposed (E): The group of people that have been exposed to the virus but are not yet infected. They have been in contact with a person that is on the network, but they haven't joined it yet.
- Infected (I): In this group, it is possible to transmit the infection to other susceptible individuals through contact. The people in this class are on social network and can influence others to join.

- Recovered (R): Includes the whole population that has been infected and now has recovered. It is assumed that they gain lifelong immunity, i.e. once an individual leaves the social network, they have no interest in joining it again.

The epidemic disease model shows the changes occurring in the population over time (in weeks) as members transition through the above categories. The population size is assumed to remain constant - demographic factors are excluded, such that births and natural deaths are not taken into account, resulting in the following relation:

$$N = S + E + I + R \tag{1}$$

where N is the total population size.

Let us then consider the following differential equations that are underpinning the epidemiological model:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\beta SI}{N} \tag{2}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta SI - \sigma E \tag{3}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I \tag{4}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \tag{5}$$

These are all subject to the initial conditions S(0) $>$ 0, I(0) $\geq$ 0 and R(0) $\geq$ 0. The constants are defined as follows:

- β is the exposure rate (also known as the transmission rate). In other words, the population within category S decreases with respect to time t at a rate $\frac{\beta I}{N} \times S$. These individuals then are added to category E at that same rate.
- σ is the incubation rate, characteristic to the exposed category, depicting how fast people move from the exposed category to the infected one. This is an addition to the SIR model, accounting for what is known as a latency or incubation period. It allows for the model to be more realistic as most people do not join a network immediately after having been exposed to it. This is analogous to infections showing symptoms after several days following contact.
- γ is the rate at which infected individuals move to the recovered state; that is to say, how fast people are leaving the network and getting removed from the system. Numerically, it is defined as $\frac{t}{t\_infective}$, where t_infective in this model is the time it takes for the infection to run its course; in our context, it is the length of time that people stay on the network, from the time of joining.

Initial value issue:
Having defined the set of ordinary differential equations(ODEs), a set of initial conditions are required to create an Initial Value Problem (IVP), which will be solved numerically using the Odeint facility from Python's Scipy's integrate module.

Let the following initial conditions supplement the equations:

1. N = 100,000
2. t_infective = 50 (weeks)

3. $\gamma = \frac{1}{30}$
4. $\beta = \frac{1}{6}$
5. S, E, I, R = n - 200, 100, 100, 0

To solve this system of ordinary differential equations, LSODA(Livermore Solver for Ordinary Differential Equations) is used from the FORTRAN library ODEPACK, by Alan Hindenmarsh. LSODA is one of nine solvers; it is advantageous over other integrators because it switches automatically between stiff and non-stiff methods. Starting with a non-stiff method, it monitors data dynamically to choose which method to use. As such, the user does not need to determine whether or not the problem is stiff.

When these initial conditions are inputted, the following solutions are plotted:

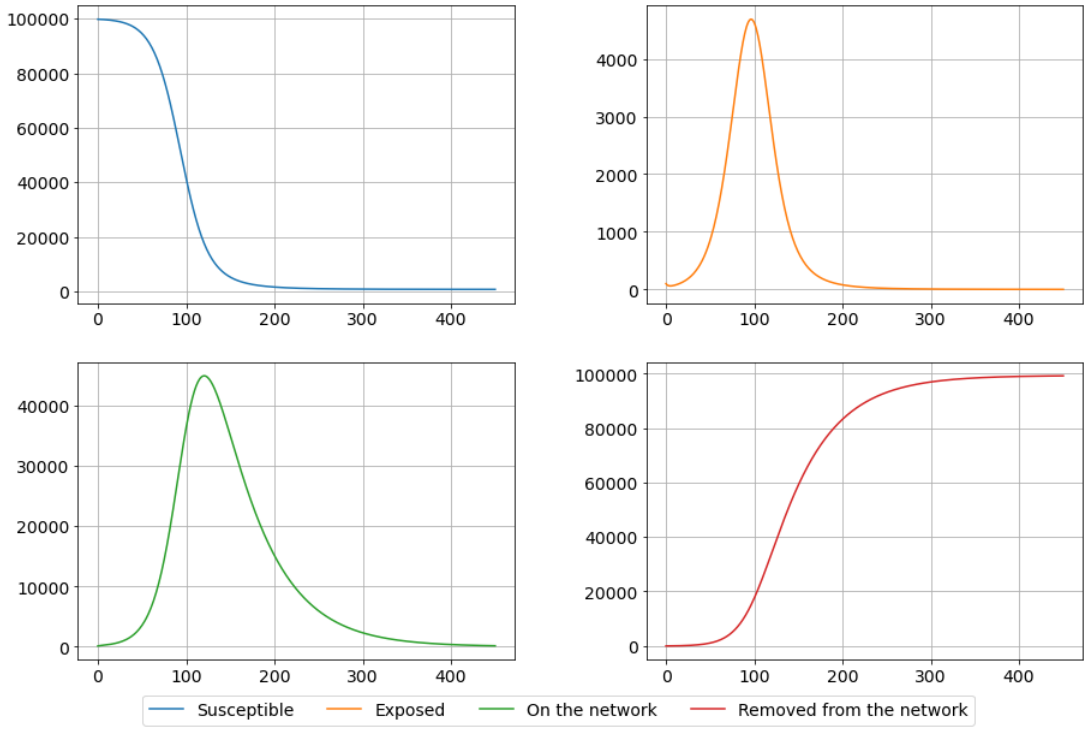

Figure 1: *Initial conditions*

As the third graph shows, over 40% of the population is active on the network by the end of the first 2 years, but the network eventually dies down by the end of the 8th year. A very important variable dictating the highest possible number of 'infected' people is called the threshold basic reproductive number, defined as the average number of secondary infections that occurs when one infective is introduced into a completely susceptible population. As such, the first infective individual can be expected to infect $R_0 = \frac{\beta}{\gamma}$ other people. In the first plot, R_0 was set to 5, assuming that one person would recommend the network to their immediate family or friends. If this value is then increased to 10, the number of people on the network increases to over 60%.

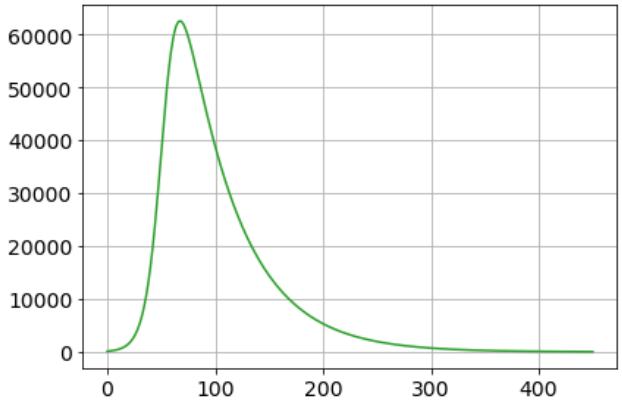

Figure 2: *Active users when R0 is increased to 10*

However, this peak does not last for a long period of time and it starts decreasing rapidly, soon after hitting this all-time high number of active users. Yet this is not what is observed in real-life scenarios. The higher the peak, the longer it should take for the numbers to decrease. In other words, the more friends somebody has on a network, the higher the incentive to stay active to socialise with them. There is generally no reason for a person to abandon the network unless a number of their connections start leaving first (ignoring the smaller percentage of reasons which include, amongst other things, stalking, harassment and online bullying that can potentially put people off).

The set of differential equations can be easily modified to reflect this nuance. The first thing that should be noticed is that the variable t_infective does not apply universally to every single member of the I class to dictate when their transition into the R class happens. The transition depends on whether or not somebody from the I class came in contact with somebody from the R class. It is translated as follows into the differential equations:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\beta SI}{N} \tag{6}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta SI - \sigma E \tag{7}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \frac{\gamma IR}{N} \tag{8}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{\gamma IR}{N} \tag{9}$$

This is the new set of initial conditions to be used:

1. N = 100,000
2. t_infective = 30
3. $\gamma = \frac{1}{t\_infective}$
4. $R_0 = 3.0$
5. $\beta = R_0 \times \gamma$

6. S_0, E_0, I_0, R_0 = N - 1100, 100, 0, 1000

The graph is as follows:

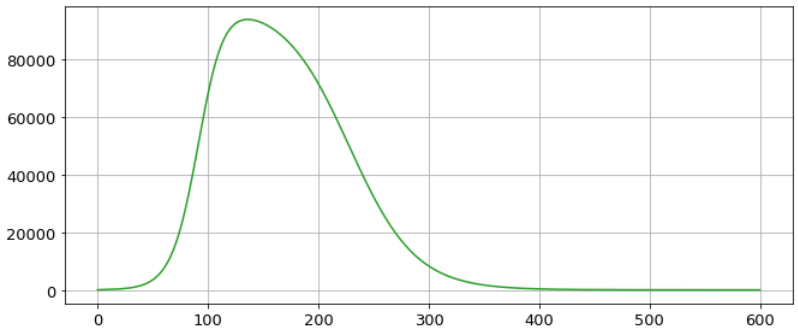

Figure 3: *Active users on the network as a result of the new equations*

There are now two major differences in the graph of active users: firstly, the peak is much higher, at almost 90% of the population. Secondly, once the peak is reached, the numbers do not decrease so rapidly; this fits real-world data: when looking at a network such as MySpace that has transitioned through all the stages, we see the same behaviour. Analysing the data from GoogleTrends (Figure 4), an exponential increase is seen in active users during the first three years, followed by a more gradual (yet still steep) decrease in users over the next four years, just like our model predicted.

Figure 4: *Google Trends*

This edited model of the SEIR is called irSEIR, standing for 'infectious recovery SEIR' - a term firstly coined by Cannarella and Spechler [7] serves to contrast a core assumption of the standard SEIR model, namely that there is a universal 'recovery' time. Contrary to characteristic recovery times for diseases, no users join a network and expect to automatically leave at some pre-established time in the future. The main reason for which they decide to abandon the network is when they see their friends and family members dropping out as well. When this happens, the recovery spreads infectiously. The consequence for the mathematical model is that if there is no initial recovered population, equation 9 will be 0 and I would grow until the whole population is infected. This is the consequence of needing a small recovered population for the model to work. Whilst there are evident benefits of choosing this over the traditional model as highlighted above, there are still some important limitations that are worth mentioning.

3 Limitations

The SEIR model is more suited (as it was originally intended) to model infectious diseases and global epidemics as opposed to OSNs. This is due to some variables (such as incubation time, infection rate, recovery time) being assumed to be universal and immutable. Social networks have a slightly different dynamic, as already shown via the infectious recovery model, where active user dropout rate is determined by the contacts between an infectious and a recovered person. Moreover, our model has a major limitation as it doesn't take into account demography. Having a constant population, with the natural birth rates balancing out the natural death rates and as a result, the change in population being zero at all times, i.e.

$$\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} = 0 \tag{10}$$

does not account for natural changes that occur in the population. For example, if birth rates were included, another factor could have been added whereby parents recommend the network to their children, or perhaps where parents who have abandoned the network ensure that their children stay away from the network. This in turn feeds into the assumption that once somebody leaves the OSN, they will never rejoin. This becomes a limitation as it is not rare for someone to leave a network and rejoin a few months or years later, with the purpose of perhaps tracing back old friends.

Another factor which this report does not take into account is the idea of having a conditional probability for each individual to join the network, based on age demographic. This would make for a more realistic model, as younger people are more likely to keep up with the flow of information and to join the network at an earlier stage than elderly people.

The whole scheme would vary greatly depending on the country in question. Countries with high populations yet which are in developing stages may not observe high percentages of active users as a direct consequence to the difficulty in accessing technological devices. Importing tables via Web Scraping in Python would overcome this and help in changing the R_0 value based on how easy it is to interact with other people in order to move them from the Susceptible to the Exposed Category.

4 Discussion and Conclusion

In this paper, we have used epidemiological models to analyse the trends in adoption and abandonment within Social Networks, reaching the conclusion that traditional models such as the SEIR do a poor job at depicting real world data. A modified, 'infectious recovery' model is more effective at depicting the behaviours of people as it moves past the reductionist assumption that people leave networks after a predetermined time. In order to take this further, we recommend a modification entailing web scraping to provide live data from various countries, as well as a modification to the source code entailing a more object oriented programming approach with different classes of people. This would make possible the predictions of how current OSNs such as Instagram and TikTok will develop in the near future.

5 Appendix

Python files can be accessed via the GitHub link below:
`https://github.com/ianvazperez/mdm1-rep3`

References

[1] H. Tankovska. How many people use social media in 2021?, 2021. https://www.oberlo.co.uk/statistics/how-many-people-use-social-media.

[2] Joseph Johnson. Market capitalization of the largest u.s. internet companies, 2021. https://www.statista.com/statistics/209331/largest-us-internet-companies-by-market-cap/ date accessed 13/05/2021.

[3] Rhonda Hadi Andrew T. Stephen Gil Appel, Lauren Grewal. The future of social media in marketing, 2021. https://link.springer.com/article/10.1007/s11747-019-00695-1#Sec17 date accessed 13/05/2021.

[4] H.S. Rodrigues. Application of sir epidemiological model: new trends, 2016. Business School, Viana do Castelo Polytechnic Institute, and CIDMA - Center for Research and Development in Mathematics and Applications, university of Aveiro, Portugal.

[5] Spechler J.A. Cannarella, J. Epidemiological modeling of online social network dynamics, 2014. https://arxiv.org/abs/1402.1225v2.

[6] Google trends, 2021. http://trends.google.com date accessed 13/05/2021.

[7] John Cannarella and Joshua A. Spechler. Epidemiological modeling of online social network dynamics, 2021. https://www.researchgate.net/publication/259783329_Epidemiological_modeling_of_online_social_network_dynamics date accessed 13/05/2021.