Research & Literature Review

# Overview

An underlying requirement of the project is to incorporate a number of different Computer Science fields in order to gain new information; understanding the background to each of these is the key to being able to represent the real-world phenomena. Specifically, investigating how social networks and personal health interact provides a good starting point for creating an agent-based model. This section will consider relevant and similar works to both validate assumptions and allow the project to expand upon the existing knowledgebase.

# Social Networks

Whilst social networks have always existed they have become a popular area of research in recent years; in particular, by applying an analytical approach, a formal method of representation has been created. At the lowest level, networks are represented using mathematical graph theory [**ref NetMark**]. This means that individual entities within the network are represented as nodes, whilst relations between these entities are edges. Importantly, these edges can be either directed or undirected, which can change their meaning depending on context – for example, directed edges could be used to show how one node likes another [**ref USN 14**]. By extension, a bidirectional edge represents some mutual relationship between the nodes.

Building on this basic set of terms, a number of structural definitions emerge that begin to directly relate the theoretical to the practical. A key concept is that of *triadic closure*, which states that, where two nodes share a common connection, they are more likely to be connected at some point in the future [**ref NetMark pg 44**]. This is particularly influential within social networks, when it comes to working out which pairs of people are going to form connections next. Extending this, the idea of *strong* *&* *weak ties* brings an extra dimension to the edges between nodes. A strong tie may be likened to a friendship, and a weak tie to an acquaintance, for example. Generally, the stronger ties are present in small, connected clusters whereas weaker ties link these clusters together [**ref NetMark 46-48**]. Again relating this to real-world examples, this is similar to groups of close friends being connected to others by acquaintances in other groups. Quantifying a graph is done through a number of measures*.* One of these is *betweenness*, which is calculated by calculating a 'flow' through the graph – edges of high flow are important since they carry the most traffic and, as such, have a high *betweenness* value. This can indicate the strength of a tie; a weak tie is likely to have a high rate of flow, for example, since it is between two highly connected groups and is of high importance in joining these clusters [**ref NetMark 66-67**].

With these terms in mind, further concepts that more directly relate to social networks and human behaviours can be introduced. A key example of this is *homophily*, defined as groups of friends which are similar, where this similarity may manifest itself in beliefs, interests, jobs, or other factors [**USN-18**]. Although the term provides an overview, there are a number of mechanisms that underpin *homophily*. When fixed characteristics such as ethnicity are considered *selection* plays a role, which is the idea that people interact and form relationships with those who they share the most in common. In contrast to this, characteristics which are variable, such as interests or behaviours, show how *socialisation* and *social influence* affect the person. The former is the process of individuals striving to bring themselves closer to others with similar characteristics, whereas the latter is when existing connections to others cause changes to the behaviours or interests, which is effectively the antithesis of *selection* [**ref NetMark 81-2**]. It should be noted that *selection* and *social influence* have an amount of interaction that can result in it being difficult to determine which aspect of *homophily* has contributed towards a connection.

Expanding upon *influence* within a graphical representation of a social network, there are a number of approaches to emulating real-world influence between networked people. Although there are many specialist models that attempt to recreate this, two basic approaches are:

* *Linear Threshold,* which is defined as \***INSERT EQUATION HERE**\*. This is a basic representation of influence that can effectively be summarised as a node taking on behaviours that its neighbouring nodes also exhibit, depending on some predefined boundary. An example of this in a real social network could be that if more than half of someone's friends play football, they will also begin to play football.
* *Independent Cascade*, defined as a series of time-steps during which any 'active' nodes attempt to activate any 'inactive' neighbours with a certain probability. Should a node become 'active', it then tries to activate its neighbours, and so on. Once nodes have attempted neighbour activation, they cannot reattempt and as such, the process ends when no more activations are available. Relating this to human behaviour, it may be likened to someone trying to convince friends about an idea; they will not carry on attempting to convince if they fail, but if someone does adopt the idea, they themselves may spread it further.

Whilst these influence models are basic, they are useful when it comes to adapting them for a social network. For the purposes of this project, they are not directly applicable, as different aspects of smoking and smoking behaviours may interact when it comes to influence spread, but they serve sufficiently as a basis.

**Reference to influence paper needs to be here somewhere – include on both**

A final, useful aspect of social network research is that of generating or classifying the type of a network. The most basic type is a randomised network, such as that generated by the Erdős–Rényi model [ref **ERDOS**], where edges between any given pair of nodes have an equal probability of existing. Naturally, this leads to most nodes having similar degree, i.e. the number of edges connected to a node. When compared to real-life networks, this lacks *hubs*, which are nodes that have a higher degree than the network average [ref **ScienAm**]. As such, random networks appear to be too far removed from what one might observe in nature and two other methods emerge with potential uses: *small-world* and *scale-free* networks.

*Small-world* networks are based on the concept of the *small-world phenomenon*, which is where human society exhibits a structure where the number of social connections between two people is, on average, quite low, indicating a high level of connectivity [**ref Milgram**]. More formally, *small-worlds* generally have high clustering and a low average path length, so aim to represent the effect observed by Milgram in a mathematical way – the Watts-Strogatz method provides an approach to generate these networks [**ref SW paper**]. At a high level, given a starting set of nodes where each is connected to its neighbours, the algorithm considers rewiring edges based on a predefined probability. This allows for the formation of more realistic structures such as hubs within the network and could be used to investigate how small social groups are affected by smoking cessation attempts.

On the other hand, *scale-free* networks rely on the previously mentioned concept of *hubs,* who have a higher degree than the average node and are observed in a wide variety of situations, from computer networks, to business alliances and Hollywood actors [**ref SciAm**]. A key principle in building these networks is that of *preferential attachment*, where nodes are more likely to connect to popular, rather than unpopular nodes. Once more, this is seen in a number of situations such as web pages having a higher chance of linking to popular web pages than more obscure sites [**ref SciEm bara**]. Using the Barabási–Albert approach to generate this type of network, the basic idea is that of starting with a simple, connected base and adding nodes incrementally, considering each other node as a connection candidate [**ref BAStat**]. The chance of each connection being successful is relative to the degree of the current node, where a higher degree increases the connection chance. The prevalence of hubs in this type of network is useful in judging the effect of influential members of a community.

# Smoking Cessation & Health

To make an accurate attempt at mapping smoking behaviours to a social network simulation, the basic factors that affect how humans engage in smoking must be understood. There are two aspects to this – both an analysis of the current smoking situation as well as how people go about trying to give up smoking. Furthermore, the impact of socialisation on the health of a person is another important factor in producing this kind of model.

By looking at NHS smoking statistics for 2012 [**ref NHS pg 13**], a number of important pieces of information relative to simulations required for this project. In Britain in 2010, 20% of the adult population were recorded as being actively smoking, with the average number of cigarettes a day being 12.7 [**ref NHS pg 13**]. Using the definition of 'heavy smoker' as somebody who smokes more than 20 cigarettes a day [**ref NHS 14**] (and from this a ‘light smoker’ being somebody who smokes fewer than 20 and one or more cigarettes a day), it can be seen that the average smoker is not a 'heavy smoker'. Adding to this, a study in 2009 displayed that around 67% of smokers wanted to give up, and, of those questioned, people who had attempted to give up smoking in the last five years were more likely to want to repeat this effort [**ref SmokOmn**]. When it comes to commencing smoking, those who begin to smoke after quitting state a number of reasons for their relapse, including stress and their friends being smokers [**ref NHS pg43**], indicating a strongly social aspect to smoking actions. It is also more likely that those who are giving up are more likely to relapse than a non-smoker is to begin smoking [**ref NHS pg43**].

Although often specific examples, factors contributing to both the commencement and cessation of smoking have been monitored. In developing countries such as Malaysia, smoking (and in this case tobacco chewing) is on the rise with around 61% of men being classed as smokers [**ref Malay**]. When analysed with respect to how these people began smoking, a variety of factors, such as gender, ethnicity and alcohol consumption, were measured, however only ethnicity was observed to be an influence in cessation. Whilst this is a single case and may be influenced by a number of socioeconomic factors, an important point to extract is that it is not necessarily a question of thresholds defining when a person starts and stops smoking. Instead, different factors seem to have varying strength in each instance.

More generally, research into cessation factors and their effect on relapse chances shows a multitude of factors that appear to contribute to failed attempts such as previous quitting attempts, presence of other smokers and behaviour/mood changes [**ref UCL cess]**. On the other hand, the work indicated a lack of effect by bodyweight/weight concerns and amount of cigarettes smoked. Although this was an internet based survey study, the observed aspects of the smoking behaviours are interesting as many different factors are involved but only some of these actually effect giving up, whist others only affect relapsing.

In terms of the social aspect of health, the Framingham Heart Study [ref **Fram**] investigates long term health concerns of a large social network. Further work has been conducted using the data from the previously mentioned study, particularly with respect to the spread of obesity [**ref ObPap**]. It was found that there are indications of social interaction playing a role in the presence of obesity within a network. This is crucial, as it is an indication of health being affected by those with whom an individual interacts. It should be noted that the type of tie between persons was significant in its effect; for example, geographically close neighbours had little impact whilst a mutual friendship greatly increases the chance of those involved becoming obese.

Closer to the combination of social networks and smoking cessation, research into the concept of quitting in groups was carried out over 30 years within the Framingham study [**ref droves**] revealed a number of interesting phenomena. Firstly, by the end of the study smoking prevalence was much lower and for those left, there was a higher chance of smokers being connected to other smokers, as well as on the periphery of the non-smoking networks. Adding to this, cessation predictors that occurred within the network were contact with other people who were quitting, type of relationship (e.g. co-worker compared to spouse) and educational status. This reinforces the concept that factors in quitting smoking come from many aspects of life, specifically relationships with others. Finally, due to these facts, group quitting appears to be a more natural approach since it utilises the peer-pressure and the avoidance of having to move to the edge of a social circle, all whilst reducing the number of smoking ties.

# Agent-Based Modelling

Central to the project is the use of an agent-based modelling (ABM) approach to simulation. Fundamentally, it defines a series of agents with attributes, who have a set of distinct behaviours through which they may interact with one another, the aim being that information not initially provided to the system may emerge. Furthermore, agents should display *autonomy* - that is, that they should function without input from outside the model - and that they are *social*, allowing others the ability to influence their behaviour [**ref REPAST PAPER**]. With this in mind, it can be seen how the agents can represent humans with their behaviours and interactions being mapped to smoking-related actions. Furthermore, this means that the model can not only have a certain level of autonomy but also represent an abstract form of socialisation.

On the whole, this technique brings about a number of advantages over other modelling techniques – it allows for emergent phenomena, a more natural way of modelling systems and flexibility [**ref ABM methTech]**. The first is of particular importance since other methods, for example mathematical models, may be bound by strict limits which can in turn limit their possible results to those expected. Autonomy and simple interactions of agents means that beyond the starting state, any number of an extremely large set of end-states can be reached. As such, with elements of randomness involved, unexpected situations can arise in ABMs.

On top of this, the descriptiveness of an ABM is important. As mentioned above, the agents are defined in terms of basic behaviours and interactions which are easily relatable to real-world actions. It is arguably easier to break complex systems into small sub-behaviours than attempt to model the entire environment for not only behaviours, but for all participants. The flexibility of this approach adds to this since once this description is decided upon and implemented, it is easy to add more agents, modify the behaviours and so on without having to redefine the whole model.

Obviously, this does not come without disadvantages. One of the key issues with all modelling approaches is that a 'general model' cannot be constructed so the model is only useful for its original purpose [**ref UCL paper**]. By extension, this means that the model has to focus on one area of behaviour (which could be very wide itself), thus removing those which are considered external. This can limit the results of the model since many 'external' aspects may have some effect on those modelled. In addition to this, mapping some actions to an ABM is difficult, particularly those in humans such as subjectiveness **[ref ABM methTech**]. When understanding the results of simulations, this kind of omission from the model must be considered since these behaviours can have a major impact on the course of events; for example, irrational choices by one agent might cause a 'butterfly effect' over the course of the rest of the run which would change the results dramatically.

Although it is a relatively new approach to modelling, there are a number of examples which demonstrate that it is widely applicable. In a similar capacity to this project, work has been done to model how viruses spread through humans [**ref IEEE paper**]. Specifically, the inclusion of real-life data allows the simulation to be built and set up using a realistic base, with the aim of understanding how governmental decisions affected the H1N1 epidemic in Mexico. A particular finding of this study was that a lot of agent-based models use survey data as a basis, resulting in a lack of representation of the way in which humans move over time; this is because of the difficulty in tracking and gaining information from specific people. To avoid this, data sources that allows the tracing of individuals, such as phone records, were used to build in this travelling behaviours.

Furthermore, ABM has also been used to investigate how emergency response can work optimally; from terrorist attacks to floods, the technique can be used to understand how both current emergency processes can be improved and new response actions can be added [**ref emerPap**]. These two methods, optimising existing behaviours and adding new actions, can apply to ABMs which aim to provide understanding into effecting human behaviour. It is noted that for systems which propose such changes where human life is at stake, an amount of verification and validation of the model must be carried out. This emphasises the fact that, for the data to be relevant to real-world situations, the simulation must display an acceptable degree of similarity to said real-world situations.

# Similar Work

To conclude this section, it is worth considering other similar pieces of work, as these can be useful in informing the direction of the project. It does appear that the particular combination of areas that this project is using has not been widely explored, but there are a number of examples that display some common features.

The first is an analysis of how epidemiology, the study of how diseases spread, can be analysed by using a social network and agent-based modelling approach [**ref epid**]. By combining these approaches, the researchers found that it allowed a much more complete view of the situation than by studying individual effects alone. This is due to the agent-based approach that provides the opportunity to define interactions and behaviours, many of which can be handled at once. Furthermore, the inherent ability of ABMs to model social interaction means that the social aspects of epidemiology can be investigated more thoroughly. Although they were found to be useful, it was also the case that the researchers emphasised the need for validation, as mentioned above, and that the scope of the model needs to be restricted to avoid the high level of complexity associated with modelling humans.

Closer to the aims of this project, there are a number of pieces of research in regards to the relationship between social networks and smoking [**ref combo paper**]. Specifically, a three-part approach is taken; the first focuses on modelling addiction and cessation as functions, the second considers influence, and the third deals with generating realistic networks. An extremely detailed solution is proposed in general, where a probability based approach was used within the model to represent different aspects of human behaviours and character [**ref pap1**]. This means that mathematical functions can be defined to produce these probabilistic representations. In terms of influence research, it was found that targeting audiences and indirect influence applied to individuals can be very effective when attempting to change behaviours whilst trying to ‘force the issue’, paint the behaviour as bad or being overly explicit in the message caused a lack of receptiveness[**ref pap2**]. In general, the model uses a combination of peer pressure, implemented in various ways, and health concern when it comes to influencing the decision of whether to cease smoking. [**ref 3**]

## Summary

In general, whilst there are few similar projects, if the fields are separated into social networks, agent based modelling and smoking cessation, a solid foundation can be constructed, upon which this research may be built. Although the approach will be detailed in full in the next section, this research indicates that a graph based representation of a social network, using humans as nodes within an agent based simulation, offers a common and fruitful way to address this problem.